On the evaluation of clustering results: measures, ensembles, and gene expression data analysis

Pablo Andretta Jaskowiak

Pablo Andretta Jaskowiak

On the evaluation of clustering results: measures, ensembles, and gene expression data analysis

Doctoral dissertation submitted to the Instituto de Ciências Matemáticas e de Computação - ICMC-USP, in partial fulfillment of the requirements for the degree of the Doctorate Program in Computer Science and Computational Mathematics. *FINAL VERSION*

Concentration Area: Computer Science and Computational Mathematics

Advisor: Prof. Dr. Ricardo José Gabrielli Barreto Campello
Coadvisor: Prof. Dr. Ivan Gesteira Costa

USP – São Carlos
January 2016

Ficha catalográfica elaborada pela Biblioteca Prof. Achille Bassi
e Seção Técnica de Informática, ICMC/USP,
com os dados fornecidos pelo(a) autor(a)

A555o

Andretta Jaskowiak, Pablo
On the evaluation of clustering results: measures, ensembles, and gene expression data analysis / Pablo Andretta Jaskowiak; orientador Ricardo José Gabrielli Barreto Campello; co-orientador Ivan Gesteira Costa. -- São Carlos, 2015.
152 p.

Tese (Doutorado - Programa de Pós-Graduação em Ciências de Computação e Matemática Computacional) -- Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo, 2015.

1. cluster analysis. 2. clustering. 3. clustering validation. 4. clustering evaluation. I. José Gabrielli Barreto Campello, Ricardo, orient. II. Gesteira Costa, Ivan, co-orient. III. Título.

Pablo Andretta Jaskowiak

Sobre a avaliação de resultados de agrupamento: medidas, comitês e análise de dados de expressão gênica

Tese apresentada ao Instituto de Ciências Matemáticas e de Computação - ICMC-USP, como parte dos requisitos para obtenção do título de Doutor em Ciências - Ciências de Computação e Matemática Computacional. *VERSÃO REVISADA*

Área de Concentração: Ciências de Computação e Matemática Computacional

Orientador: Prof. Dr. Ricardo José Gabrielli Barreto Campello

Coorientador: Prof. Dr. Ivan Gesteira Costa

USP – São Carlos
Janeiro de 2016

To my mother. Without your love, patience, and help, this thesis wouldn't exist.

A minha mãe. Sem seu amor, paciência e ajuda, esta tese não existiria.

Life is one huge lottery where only the winning tickets are visible.

Jostein Gaarder

The Solitaire Mystery

Acknowledgements

A minha querida e amada mãe, Lidete Maria Andretta, que sempre me apoiou incondicionalmente durante toda esta jornada. Sem seu amor, carinho e suporte (inclusive financeiro) eu não teria chegado até aqui. Obrigado por se fazer presente em toda minha vida.

Ao meu orientador, Prof. Ricardo J. G. B. Campello, que teve papel fundamental no meu crescimento pessoal e profissional, desde o meu mestrado. Sua dedicação e constante busca pela excelência são inspirações que levo comigo para toda a vida, não só a acadêmica.

Ao meu co-orientador, Prof. Ivan G. Costa, excelente profissional pelo qual tenho grande admiração, que mesmo distante fisicamente, esteve sempre presente ao longo deste trabalho enriquecendo todas nossas discussões. Meu obrigado também pela sua recepção na Universidade Federal de Pernambuco (UFPE) durante os dois mesês em que estive visitando seu laboratório.

Ao Prof. Jörg Sander, por me receber sob sua supervisão durante um ano na Universidade de Alberta (U of A), em Edmonton, AB, Canadá. Meu obrigado também aos amigos Arthur Zimek e Davoud Moulavi, com quem tive o privilégio de colaborar neste período e dividir alguns pistaches.

A todos amigos e amigas do laboratório Biocom, pelas calorosas discussões científicas (ou não) e por toda a convivência extra laboratório. Extendo os agradecimentos a todos os amigos e amigas que fiz em São Carlos em todos esses anos. Os palquinhos e churrascos trarão saudades. A Amanara Potykytã, por todo seu amor, calma e compreensão, você tornou tudo mais fácil.

Sou grato também aos membros do comitê avaliador, Prof. Wagner Meira Júnior, Prof. Renato Tinós, Prof. Ana Carolina Lorena e Prof. Alexandre Cláudio Botazzo Delbem, por seu tempo e suas considerações e contribuições ao trabalho. Elas estão refletidas nesta versão do texto.

A Fundação de Amparo à Pesquisa do Estado de São Paulo (FAPESP), pelo suporte financeiro ao projeto, tanto de bolsa regular de doutorado (Processo FAPESP #2011/04247-5) quanto de estágio de pesquisa no exterior (Processo FAPESP #2012/15751-9) e a Coordenaçao de Aperfeiçoamento de Pessoal de Nível Superior (CAPES) pelo auxílio financeiro inicial ao projeto.

Abstract

Clustering plays an important role in the exploratory analysis of data. Its goal is to organize objects into a finite set of categories, *i.e.*, clusters, in the hope that meaningful and previously unknown relationships will emerge from the process. Not every clustering result is meaningful, though. In fact, virtually all clustering algorithms will yield a result, even if the data under analysis has no "true" clusters. If clusters do exist, one still has to determine the best configuration of parameters for the clustering algorithm in hand, in order to avoid poor outcomes. This selection is usually performed with the aid of clustering validity criteria, which evaluate clustering results in a quantitative fashion. In this thesis we study the evaluation/validation of clustering results, proposing, in a broad context, measures and relative validity criteria ensembles. Regarding measures, we propose the use of the Area Under the Curve (AUC) of the Receiver Operating Characteristics (ROC) curve as a relative validity criterion for clustering. Besides providing an empirical evaluation of AUC, we theoretically explore some of its properties and its relation to another measure, known as Gamma. A relative criterion for the validation of density based clustering results, proposed with the participation of the author of this thesis, is also reviewed. In the case of ensembles, we propose their use as means to avoid the evaluation of clustering results based on a single, ad-hoc selected, measure. In this particular scope, we: (i) show that ensembles built on the basis of arbitrarily selected members have limited practical applicability; and (ii) devise a simple, yet effective heuristic approach to select ensemble members, based on their effectiveness and complementarity. Finally, we consider clustering evaluation in the specific context of gene expression data. In this particular case we evaluate the use of external information from the Geno Ontology for the evaluation of distance measures and clustering results.

keywords: clustering, clustering validation

Resumo

Técnicas de agrupamento desempenham um papel fundamental na análise exploratória de dados. Seu objetivo é a organização de objetos em um conjunto finito de categorias, *i.e.*, grupos (*clusters*), na expectativa de que relações significativas entre objetos resultem do processo. Nem todos resultados de agrupamento são relevantes, entretanto. De fato, a vasta maioria dos algoritmos de agrupamento existentes produzirá um resultado (partição), mesmo em casos para os quais não existe uma estrutura "real" de grupos nos dados. Se grupos de fato existem, a determinação do melhor conjunto de parâmetros para estes algoritmos ainda é necessária, a fim de evitar a utilização de resultados espúrios. Tal determinação é usualmente feita por meio de critérios de validação, os quais avaliam os resultados de agrupamento de forma quantitativa. A avaliação/validação de resultados de agrupamentos é o foco desta tese. Em um contexto geral, critérios de validação relativos e a combinação dos mesmos (*ensembles*) são propostas. No que tange critérios, propõe-se o uso da área sob a curva (AUC — *Area Under the Curve*) proveniente de avaliações ROC (*Receiver Operating Characteristics*) como um critério de validação relativo no contexto de agrupamento. Além de uma avaliação empírica da AUC, são exploradas algumas de suas propriedades teóricas, bem como a sua relação com outro critério relativo existente, conhecido como Gamma. Ainda com relação à critérios, um índice relativo para a validação de resultados de agrupamentos baseados em densidade, proposto com a participação do autor desta tese, é revisado. No que diz respeito à combinação de critérios, mostra-se que: (i) combinações baseadas em uma seleção arbitrária de índices possuem aplicação prática limitada; e (ii) com o uso de heurísticas para seleção de membros da combinação, melhores resultados podem ser obtidos. Finalmente, considera-se a avaliação/validação no contexto de dados de expressão gênica. Neste caso particular estuda-se o uso de informação da *Gene Ontology*, na forma de similaridades semânticas, na avaliação de medidas de dissimilaridade e resultados de agrupamentos de genes.

palavras chave: agrupamento de dados, validação de agrupamentos

Contents

List of Figures

List of Tables

Notation

This document adopts the following convention:

- Scalars are given in lower case italics;
- Vectors are given in lower case bold;
- Matrices are given in upper case bold;
- Sets are given in upper case italics.

Notation is as follows:

$\|.\|$	The cardinality of a given set
$\mathbf{X}$	A set of objects, *i.e.*, a dataset
n	The number of objects in $\mathbf{X}$, *i.e.*, $\|\mathbf{X}\|$
$\mathbf{x}_i$	An object from $\mathbf{X}$, *i.e.*, a m-dimensional vector $\mathbf{x_i} = (x_1, \ldots, x_m)$
m	The number of dimensions (*i.e.*, features) for a given object or dataset
$\mathcal{C}$	A set of k clusters, *i.e.*, a clustering or partition, with $\mathcal{C} = \{C_i, \ldots, C_k\}$
$\overline{\mathbf{c}}_i$	The centroid from cluster i
$\mathbf{D}$	A dissimilarity matrix
k	Number of clusters
k^*	The *optimal* (desired) number of clusters as defined by a reference partition
k_{max}	The superior limit considered for the cluster number, typically $k_{max} = \lceil \sqrt{n} \rceil$

Lista de Abreviaturas

ALOI	Amsterdam Library of Object Images
AL	Average-Linkage
ARI	Adjusted Rand Index
ASSWC	Alternative Simplified Silhouette Width Criterion
ASWC	Alternative Silhouette Width Criterion
AUC	Area Under the Curve
BP	Base Pairs
cDNA	Complementary Deoxyribonucleic Acid
CL	Complete-Linkage
DBCV	Density Based Clustering Validation
DBSCAN	Density-Based Spatial Clustering of Applications with Noise
DB	Davies-Boudlin (Criterion)
DNA	Deoxyribonucleic Acid
FN	False Negative
FP	False Positive
HCAs	Hierarchical Clustering Algorithms
IBSA	Intrinsic Biological Separation Ability

ISA	Intrinsic Separation Ability
mRNA	Messenger Ribonucleic Acid
PBM	Pakhira, Bandyopadhyay, and Maulik's (Criterion)
PB	Point-Biserial (Criterion)
PCA	Principal Component Analysis
PCR	Polymerase Chain Reaction
RI	Rand Index
RNA	Ribonucleic Acid
ROC	Receiver Operating Characteristics
rRNA	Ribosomal Ribonucleic Acid
SL	Single-Linkage
SSE	Sum of Squared Errors
SSWC	Simplified Silhouette Width Criterion
SWC	Silhouette Width Criterion
TIFF	Tagged Image File Format
TN	True Negative
TP	True Positive
tRNA	Transfer Ribonucleic Acid
VRC	Variance Ratio Criterion

CHAPTER

1

Introduction

We are embedded in a world of data. As our capacity to collect and store data from the most varied sources continues to evolve, so does the need of developing efficient methods to analyze and extract useful information from them. The field of Data Mining embraces methods and algorithms from different areas of research, such as Artificial Intelligence, Machine Learning, and Statistics, that aim to uncover valuable information from data (Fayyad et al., 1996; Tan et al., 2006). Its methods are usually categorized into different tasks, which can be broadly regarded as supervised and unsupervised, taking into account the learning strategy they employ (Tan et al., 2006).

Cluster analysis, or simply clustering, is an unsupervised Data Mining task. Given that no prior knowledge is used during the clustering process, it finds great applicability in the exploratory analysis of data. Its goal is to organize data objects into a finite set of categories, *i.e.*, clusters, by abstracting the underlying structure of the data, in the hope that meaningful and previously unknown relationships among objects will emerge as a result of the process (Hartigan, 1975; Jain and Dubes, 1988). The lack of a globally accepted definition for the term cluster in the literature drove the development of several clustering paradigms and numerous clustering algorithms within each paradigm (Estivill-Castro, 2002; Jain and Dubes, 1988). These have been applied to the most diverse areas of expertise, such as astronomy, economics, psychology, and bioinformatics, just to mention a few (Jain, 2010; Tan et al., 2006; Xu and Wunsch II, 2009; Zhang, 2006).

One of the first steps of the clustering procedure is to choose an adequate clustering algorithm and set its parameters properly for the application in hand. The choice of a particular algorithm and its corresponding parameterization, the so-called model selection

problem, is, however, far from trivial in an unsupervised environment. Fortunately, over the past years a number of mathematical indexes that can be used to guide these choices in a quantitative way have been developed. In the clustering literature, these indexes are usually referred to as clustering validity criteria, as they can also be used to evaluate/validate the relevance of the clustering results from a statistical perspective (Jain and Dubes, 1988; Xu and Wunsch II, 2009). In real world applications, clustering validity criteria known as internal and relative find broad applicability. In brief, internal criteria quantify the quality of a clustering solution using only the data itself, whereas relative criteria are internal measures that can go further and also compare two clustering structures, pointing out which one is better, in relative terms. Although a few exceptions do exist, these criteria are often based on the general idea of measuring, somehow, the balance between within-cluster scattering (compactness) and between-cluster spread (separation), with differences arising mainly from different formulations of these two fundamental concepts.

The validation of clustering results has been historically described as challenging (Milligan and Cooper, 1985). Indeed, in their classical book on clustering, Jain and Dubes (1988) state that:

> "*The validation of clustering structures is the most difficult and frustrating part of cluster analysis. Without a strong effort in this direction, cluster analysis will remain a black art accessible only to those true believers who have experience and great courage.*"

After almost 30 years and despite achievements observed in this particular area, we believe that the above statement remains true. This has motivated us to study, develop, and propose methods to the evaluation of clustering results. In the general[1] context of clustering validation this thesis follows two main lines of investigation. The first one comprises the study and development of new relative validity measures, whereas in the second one we investigate and propose the use of ensembles of relative validity criteria, based on a careful selection of members.

With respect to measures, we propose the application of concepts from the supervised domain of classification to the relative validation of clustering results. More specifically, we study and adapt the Area Under the Curve (AUC) of the Receiver Operating Characteristics (ROC) curve to the validation of clusterings. We then investigate some of its theoretical properties, and provide an empirical evaluation of the measure. This leads to our first research hypothesis, as stated below:

- **Hypothesis 1:** *The Area Under the Curve of the Receiver Operating Characteristics curve, which is commonly adopted in the supervised learning setting (Fawcett, 2006), can be effectively employed in the unsupervised setting, as a relative validity criterion.*

Still regarding measures, a relative criterion for the validation of density based clustering results, proposed with the participation of the author of this thesis (Moulavi et al., 2014), is also reviewed.

Regarding ensembles, our second line of investigation, we start by acknowledging that there is a plethora of relative validity criteria proposed in the literature (Vendramin et al., 2009, 2010). The

[1] By general we mean that these lines of investigation are not tied to any particular application domain.

variety of measures by itself suggests that a single evaluation index cannot capture all the aspects involved in the clustering problem and, therefore, may fail in particular application scenarios. In fact, due to the subjective nature of the problem, it is well-known that no index can systematically outperform all the others in all scenarios. As conjectured by Bezdek and Pal (1998), a possible approach to bypass the selection of a single relative validity criterion is to rely on multiple criteria in order to obtain more robust evaluations. The rationale behind this approach follows essentially the same intuitive idea of combining multiple experts into a committee so as to get more stable and accurate recommendations, which has been well studied in the realm of ensembles for classification (Rokach, 2010), clustering (Ghosh and Acharya, 2011), and outlier detection (Zimek et al., 2013), for instance. The belief that this topic has not received sufficient attention considering the validation of clustering results has motivated the formulation of the following hypotheses:

- **Hypothesis 2:** *Ensembles of relative validity criteria built on the basis of an ad-hoc selection of their constituent members provide very limited (if any) practical benefits.*

- **Hypothesis 3:** *Ensembles built on the basis of a simple, yet principled selection of their constituent members, perform better than those built in an ad-hoc fashion and provide more reliable evaluations than the ones obtained with individual relative validity criteria.*

The previous hypotheses are formulated in a broad sense, that is, with no particular application domain in mind. Some application domains, however, provide peculiar challenges that require custom tailored developments. This is the case of the clustering of gene time-series, coming from gene expression experiments (Zhang, 2006). In this particular domain, one can hardly find externally labeled datasets for the empirical evaluation of clustering algorithms and methods. Moreover, due to their peculiar characteristics, custom distance measures have also been developed in the past years, *e.g.*, (Balasubramaniyan et al., 2005; Möller-Levet et al., 2005), but no systematic evaluation of them has been provided in the literature. This has motivated us to investigate the use of biological information from the Gene Ontology (GO) (Ashburner et al., 2000) to the evaluation of distance measures and clustering results in this particular domain. These two lines of investigation are summarized below by two different hypotheses.

- **Hypothesis 4:** *External information, in the form of semantic similarities extracted from the Gene Ontology (Ashburner et al., 2000), can be employed to evaluate the suitability of distances among pairs of gene time-series for the task of clustering, independently from the bias of a particular clustering algorithm.*

- **Hypothesis 5:** *External information, in the form of semantic similarities extracted from the Gene Ontology (Ashburner et al., 2000), can be employed in the relative evaluation of clustering results, whether alone or combined with statistical similarities from the data.*

The assessment of the aforementioned hypotheses was carried out through empirical analysis, which were performed considering both synthetic and real datasets.

1.1 Contributions

With the previous hypotheses in mind, the contributions of this thesis are summarized below.

1. Empirical evaluation of ad-hoc ensembles of relative validity criteria, showing that, in general, arbitrary ensembles provide poor performance, with very limited practical benefits.

2. Proposal of a heuristic approach for building effective ensembles of relative validity criteria, showing that ensembles derived from our heuristic provide superior results than ad-hoc ones. These also provide more robust evaluations than those obtained with single validity criteria.

3. A review of different approaches that can be employed to combine relative validity criteria outcomes into ensembles. These fall under two categories, namely, value and rank based.

4. The proposition of the Area Under the Curve (AUC) of the Receiver Operating Characteristics (ROC) curve as a relative validity criterion for evaluating clustering results.

5. Formal proofs regarding: (i) the expected value of the AUC; and (ii) the relation between AUC and the Gamma Index (Baker and Hubert, 1975). Regarding (ii), we reduced the computational time required to compute Gamma from $O(n^2m + n^4/k)$ to $O(n^2 \log n)$, where n is the number of objects, m is the number of features, an k is the number of clusters.

6. An empirical evaluation of AUC/Gamma, showing that it ranks among the best alternatives regarding a comprehensive pool of relative validity criteria. Given its reduced computational cost and reasonable overall results, AUC/Gamma arises as an useful and viable alternative to the clustering practitioner, specially in the relational clustering setting.

7. A review of 15 distances for the clustering of gene expression data, including, for the first time, a collection of distances specifically developed for the clustering of gene time-series.

8. The proposal of a methodology that uses biological information to the evaluation of distance measures regarding gene time-series, namely Intrinsic Biological Separation Ability (IBSA).

9. Investigation on the use of biological information, in the form of semantic similarities extracted from the Gene Ontology, in the relative evaluation of gene clustering results.

10. The evaluation of distance measures for the clustering of gene expression data, w.r.t.: (i) different methodologies; (ii) robustness to noise; and (iii) considering different gene expression data obtained with different technologies, namely, microarrays and RNA-Seq.

11. Empirical evaluation of different factors and their respective effects in the clustering of cancer samples from RNA-Seq data, covering: (i) expression estimates, (ii) number of features, (iii) data transformation, (iv) clustering algorithms, and (v) distance measures.

1.2 Outline

This thesis is structured in eight chapters. The remainder of the thesis is organized as follows:

- **Chapter 2 - Cluster Analysis:** The basic concepts of cluster analysis are presented. Classical clustering algorithms from the literature (employed in this thesis) are briefly reviewed. The issue of clustering validation is addressed, with emphasis on relative validity criteria. A review of procedures for the evaluation of relative validity criteria is provided.

- **Chapter 3 - Gene Expression Data:** Key biological concepts to the understanding of gene expression data are reviewed. High throughput technologies employed to measure gene expression are presented, namely microarrays and RNA-Seq. The chapter closes with a discussion on the importance of clustering in the analysis of gene expression data.

- **Chapter 4 - Ensembles for Relative Validity Criteria Evaluation:** Relative validity criteria ensembles are investigated. Ensembles built in an ad-hoc fashion are discussed and evaluated. A principled heuristic strategy for the selection of ensemble members, based on concepts of effectiveness and complementarity, is developed and assessed experimentally. The developments and results from this particular chapter have already been published, please see (Vendramin et al., 2013) and (Jaskowiak et al., 2015).

- **Chapter 5 - Relative Validation of Clustering Results:** Two contributions regarding the relative validation of clustering results are presented. In the first half of the chapter, the Area Under the Curve (AUC) of the Receiver Operating Characteristics (ROC) is introduced as a relative validity criterion. Properties of the AUC of a clustering solution and its relation to the Gamma index (Baker and Hubert, 1975) are theoretically explored. An empirical analysis of AUC is also provided. In the remaining of the chapter, a relative validity criterion (DBCV - Density Based Clustering Validation) proposed to the validation of density-based clustering results is reviewed. The work regarding DBCV was performed in collaboration with Davoud Moulavi (main contributor), during the author's one year internship at the University of Alberta, Edmonton, Alberta, Canada, under the supervision of Prof. Jörg Sander. Results from the second part of this chapter have already been published, see (Moulavi et al., 2014).

- **Chapter 6 - Distances for Clustering Gene Expression Data:** The selection of distance measures in the context of gene expression data clustering is addressed. A total of 15 distances are reviewed. Methodologies for the evaluation of distance measures are discussed. A methodology that employs biological information from the Gene Ontology (Ashburner et al., 2000) to evaluate distances between pairs of genes without the inherent bias of a particular clustering algorithm is introduced. Results regarding the selection of distance measures for the clustering of microarray and RNA-Seq datasets are presented. Results from this chapter have already been published in international peer reviewed journals, which can be found in references (Jaskowiak et al., 2013) and (Jaskowiak et al., 2014).

- **Chapter 7 - Biological Validation of Gene Clustering Results:** The use of biological information from the Gene Ontology in the evaluation/validation of gene time-series clustering results is considered. More specifically, we evaluate the potential of semantic similarities extracted from the Gene Ontology in the relative evaluation of clustering results from gene time-series. Some comments on undesired properties regarding a validity measure with biological bias, which is commonly employed in the literature, are also provided.

- **Chapter 8 - Conclusions:** The main contributions of the work are presented. Limitations and opportunities for future work are discussed. Works published in the form of conference and journal articles throughout the author's PhD are highlighted, and the thesis is concluded.

CHAPTER

2

Cluster Analysis

Cluster analysis, or simply clustering, comprehends the set of methods and algorithms frequently employed during the exploratory analysis of data. The broad use of clustering as an exploratory tool arises from the fact that it requires few assumptions about the data under investigation (Jain and Dubes, 1988), being characterized as an unsupervised task from the viewpoint of Machine Learning (Tan et al., 2006). In this chapter we present a concise review of cluster analysis. We start by introducing its basic concepts in Section 2.1. In Section 2.2 we review some well-known clustering algorithms from different clustering paradigms employed in this thesis. Finally, in Section 2.3, we discuss clustering validation techniques, which aim to assess the quality of the results produced by different clustering algorithms.

2.1 Basic Concepts

Given the unsupervised nature of the clustering process, there is usually no *a priori* information[1] on how the data under analysis is structured. In this context, the aim of clustering algorithms is to organize objects from the data in a natural manner, in the hope that previously unknown relations from the data will emerge as a result of the clustering process. Although there is no single globally accepted definition in the literature for the term cluster (Jain and Dubes,

[1] In the case of semi-supervised clustering (Bilenko, 2004) there is actually some *a priori* knowledge that is provided as input to the clustering algorithm, usually in the form of must-link (ML) and cannot-link (CL) restrictions between pairs of data objects. The study of semi-supervised clustering algorithms is beyond the scope of this thesis.

1988), a plethora of clustering algorithms have been introduced in the past decades and applied to the most diverse areas of expertise, including bioinformatics (Jiang et al., 2004; Zhang, 2006).

In general, clustering algorithms can be divided into two major categories, namely partitional and hierarchical (Kaufman and Rousseeuw, 1990). Given a dataset $\mathbf{X} = \{\mathbf{x_1}, \ldots, \mathbf{x_n}\}$ with n objects embedded in a space with m dimensions (features), where $\mathbf{x_i} = \{x_1, \ldots, x_m\}$, a partitional clustering algorithm divides the data into a finite number of mutually exclusive clusters. Let $\mathcal{C} = \{C_1, \ldots, C_k\}$ denote the result of a partitional clustering algorithm (*i.e.*, a partition) with k clusters, then these clusters respect the following rules:

$$C_1 \cup \cdots \cup C_k = \mathbf{X}$$
$$C_i \neq \emptyset, \forall i$$
$$C_i \cap C_j = \emptyset, \forall i, j \text{ with } i \neq j$$

The definition provided above comprehends partitional clustering algorithms known as *hard* or *crisp*, for which objects in a partition belong to only one cluster. It is worth noticing, however, that there are other subclasses of partitional clustering algorithms in which objects in a partition can belong to more than one cluster at a time. Algorithms that belong to this subclass are usually referred to as *soft*. This is the case of fuzzy clustering algorithms (Bezdek, 1981; Bezdek et al., 1984). In this particular case, each object belongs to all the k clusters with different membership levels, varying from 0 (lowest membership) to 1 (highest membership). Given that the focus of this thesis is on hard clustering algorithms, we refer the reader to the works of Bezdek (1981), Xu and Wunsch II (2009), and Vendramin (2012) for more information on fuzzy clustering methods.

Hierarchical Clustering Algorithms (HCAs) produce as a results not a single partition, but a set of nested partitions, *i.e.*, a hierarchy of partitions. HCAs can be further categorized into two major subclasses, considering how the final hierarchy is obtained, namely: divisive and agglomerative. The general workflow for divisive HCAs is as follows: (i) given n objects, assign all objects to a single cluster; (ii) divide the initial cluster into two clusters according to a given criterion; (iii) recursively apply step (ii) to the two clusters generated by the initial division until each cluster has a single object, that is $k = n$. For agglomerative HCAs the general process is as follows: (i) given n objects, assign each object to a different cluster, that is $k = n$; (ii) merge the two most similar clusters into a new cluster; (iii) apply step (ii) until all objects belong to a unique cluster, that is, $k = 1$. Due to their high computational cost, divisive Hierarchical Clustering Algorithms are rarely employed in the literature (Xu and Wunsch II, 2009), with a few exceptions, *e.g.*, (Campello et al., 2013). The result of a HCA is usually depicted in the form of a dendrogram, as shown in Figure 2.1. In this figure, each leaf node of the dendrogram represents an object. The dissimilarities depicted in the y axis indicate the points at which a cluster is formed (in the case of agglomerative HCAs) or dissolved (in the case of divisive HCAs). It is worth noticing that the figure is presented for illustrative purposes and that other criteria than dissimilarities can be

employed in order to merge or split clusters as is the case in HDBSCAN* (Campello et al., 2013), for instance.

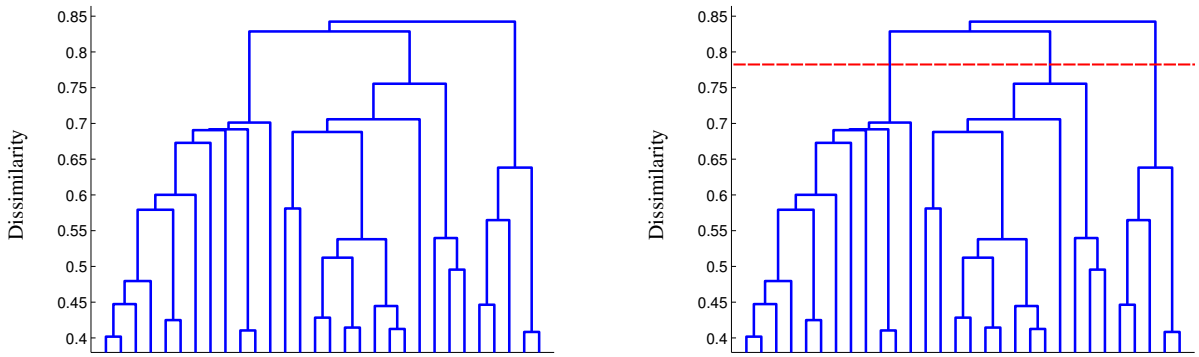


Figure 2.1: Result from a Hierarchical Clustering Algorithm, depicted as a dendrogram (left). A hard partition with k clusters can be derived by cutting the dendrogram at the desired level. The same dendrogram with a cut (red dashed line) that produces a partition with $k = 3$ is shown (right).

Note that the definitions of partitional and hierarchical clustering as provided above are quite general, *i.e.*, they do not define a specific algorithm for clustering, leaving room for the definition of different clustering algorithms. These algorithms emerge from the adoption of different biases during the partitioning of the data and/or different definitions for the term cluster itself. Following closely from Jain and Dubes (1988) and Everitt (1974), for instance, a cluster can be defined as:

1. A set of alike objects, with objects from different clusters being not alike;
2. A region characterized by a high density of objects, separated from other such dense regions (other clusters) by a region with relatively low density of objects.

From the first definition objects belong to the same cluster if they are alike. This notion is usually represented mathematically/algorithmically in terms of proximity measures, either in the form of distances or similarities. Therefore, objects within the same cluster should have small distances among themselves and large distances to objects belonging to a cluster other than their own. Such definition of the term cluster has led, for instance, to the development of partitional algorithms such as the well-known k-means (MacQueen, 1967) and k-medoids (Bishop, 2006), and Hierarchical Clustering Algorithms such as the Average-Linkage (Jain and Dubes, 1988). Note that although these algorithms seek clusters that fall under the same definition, their differences arise from their different biases. The second definition is the basis for clustering algorithms that belong to the density-based clustering paradigm, for which examples are the well-known partitional algorithm called Density-Based Spatial Clustering of Applications with Noise, also known as DBSCAN (Ester et al., 1996) and the more recently hierarchical clustering algorithm called HDBSCAN* (Campello et al., 2013, 2015). Having made such considerations, in the sequel we review specific clustering algorithms from these paradigms that are employed in this thesis.

2.2 Clustering Algorithms

In this section we discuss clustering algorithms that are related to the development of the thesis.

2.2.1 k-means and k-medoids

The k-means clustering algorithm (MacQueen, 1967) is perhaps the most well-known clustering algorithm from the literature (Xu and Wunsch II, 2009). Indeed, it was deemed one of the top 10 most influential algorithms in Data Mining by a representative pool in the the data mining community described by Wu et al. (2008). One of the main characteristics of the k-means clustering algorithm is its simplicity. It is described bellow:

1. Given a dataset $\mathbf{X} = \{\mathbf{x_1}, \ldots, \mathbf{x_n}\}$ and a desired number of clusters k;
2. Select k cluster prototypes (centroids) randomly from the data objects;
3. Assign each object to the cluster with nearest centroid;
4. For each cluster recalculate its centroid as the mean of the objects within the cluster;
5. Repeat Steps 2 and 3 until there is no change in cluster memberships.

The k-means clustering algorithm has a time complexity of $O(nkmi)$, where n is the number of objects, k the desired number of clusters, m the number of dimensions (considering a distance measure with linear time complexity) of the data, and i is the number of iterations. Given that k, m and, i are generally smaller than n, the algorithm is regarded to have linear complexity in n.

It is worth noticing that the k-means clustering algorithm has its convergence properties guaranteed only for the Squared Euclidean distance. If other measures are to be employed, the centroid calculation must be redefined to maintain k-means optimization and convergence properties, as pointed out by Steinley (2006). In order to avoid convergence problems, when different distance measures are employed a counterpart of k-means is usually adopted, namely k-medoids (Bishop, 2006). The k-medoids clustering algorithm is similar to k-means in every aspect, except for the definition of its cluster prototypes. Note that in k-means the prototypes of each cluster are artificial objects (centroids). In k-medoids these are replaced by actual objects (medoids). The medoid of each cluster is then defined as the object that has the minimum distance to all other objects within the same cluster. It is important to note that with such a replacement, the k-medoids clustering algorithm has an $O(n^2)$ time complexity.

Both k-means and k-medoids are not deterministic, *i.e.*, for different initial sets of prototypes (centroids or medoids) the algorithms may produce different result partitions. Moreover, these algorithms do not guarantee convergence to a global optimum solution. For these reasons, it is common practice to run the algorithms several times with different initializations

(*e.g.*, 50 initializations) and, at the end, given a fixed number of clusters (k), to select the partition with minimum Sum of Squared Errors (SSE), given by Equation (2.1), as the best result. In Equation (2.1) $\overline{\mathbf{p_i}}$ represents the prototype of cluster C_i, *i.e.*, its centroid or medoid, whereas $d(,)$ is a distance function.

$$SSE(\mathcal{C}) = \sum_{i=1}^{k} \sum_{\mathbf{x} \in C_i} d(\mathbf{x}, \overline{\mathbf{p_i}})^2 \tag{2.1}$$

Note that the procedure above can be employed solely for cases in which the number of clusters (k) is fixed *a priori*, given that SSE values tend to decrease as the number of clusters increase. In order to compare partitions with different number of clusters and select a single result, one can employ, for instance, relative validity criteria. These are discussed in Section 2.3.2.

2.2.2 Hierarchical Clustering Algorithms

Hierarchical Clustering Algorithms (HCAs) are quite popular in the literature, in part given to their visual appeal. This is specially true in the field of bioinformatics, in which it is common to report a picture of the resulting dendrogram (Jiang et al., 2004; Zhang, 2006). Among the different HCAs available in the literature, three agglomerative methods are commonly employed, *i.e.*, Single-Linkage (SL), Average-Linkage (AL), and Complete-Linkage (CL). These methods follow the standard procedure for generating a nested hierarchy of partitions as bellow:

1. Given a dataset $\mathbf{X} = \{\mathbf{x_1}, \ldots, \mathbf{x_n}\}$, assign each one of its n objects to a different cluster;
2. Identify and merge the two closest clusters, according to a *distance between clusters*;
3. Repeat Step 2 until there are no more clusters to be merged (all objects belong to one cluster).

The difference among these three algorithms lie in how they define the distance (often called linkage) between a pair of clusters. These are provided in Table 2.1 for each one of them:

Table 2.1: Distances between clusters commonly used in HCAs.

Algorithm (Linkage)	Distance between clusters
Single-Linkage (SL)	$\min_{\mathbf{x_o} \in \mathbf{C_i}, \mathbf{x_p} \in \mathbf{C_j}} \{d(\mathbf{x_o}, \mathbf{x_p})\}$
Average-Linkage (AL)	$\frac{1}{\lvert\mathbf{C_i}\rvert \lvert\mathbf{C_j}\rvert} \sum_{\mathbf{x_o} \in \mathbf{C_i}, \mathbf{x_p} \in \mathbf{C_j}} d(\mathbf{x_o}, \mathbf{x_p})$
Complete-Linkage (CL)	$\max_{\mathbf{x_o} \in \mathbf{C_i}, \mathbf{x_p} \in \mathbf{C_j}} \{d(\mathbf{x_o}, \mathbf{x_p})\}$

Agglomerative Hierarchical Clustering Algorithms have an $\Omega(n^2)$ computational complexity. Note that the algorithms presented above produce a hierarchy of nested partitions, with $1 \leq k \leq n$.

If a single partition of k clusters is desired, a horizontal cut in the dendrogram has to be performed, as previously depicted in Figure 2.1.

2.2.3 DBSCAN

Density-Based Spatial Clustering of Applications with Noise, or simply DBSCAN (Ester et al., 1996), is among the most well-known clustering algorithms from the density-based clustering paradigm. The general idea of the algorithm is to find density-connected regions in the data. Each density-connected region is defined as a cluster, whereas points that do not belong to any density-connected region are deemed as noise. In order to find dense regions in the data, the algorithm requires two input parameters from the user, named Epsilon (ϵ) and $MinPts$. Given these two parameters, a dataset $\mathbf{X} = \{\mathbf{x_1}, \ldots, \mathbf{x_n}\}$ and a distance between objects, DBSCAN can be derived from Definitions 1 to 5, which follow from its original publication (Ester et al., 1996).

Definition 1. (core-object) Object $\mathbf{x_i}$ is a core-object if it has at least $MinPts$ in its neighborhood considering ϵ, which is given by: $N_\epsilon(\mathbf{x_i}) = \{\mathbf{x_j} \in \mathbf{X} | d(\mathbf{x_i}, \mathbf{x_j}) \leq \epsilon\}$, *i.e.*, $|N_\epsilon(\mathbf{x_i})| \geq MinPts$.

Definition 2. (directly density-reachable) An object $\mathbf{x_j}$ is directly density-reachable from $\mathbf{x_i}$, with respect to ϵ and $MinPts$, if $\mathbf{x_i}$ is a core-object and $\mathbf{x_j}$ is in its neighborhood, that is $\mathbf{x_j} \in N_\epsilon(\mathbf{x_i})$.

Definition 3. (density-reachable) A given object $\mathbf{x_j}$ is said to be density-reachable from $\mathbf{x_i}$, with respect to ϵ and $MinPts$, if there is a chain of objects connecting $\mathbf{x_i}$ to $\mathbf{x_j}$, such that every object that belongs to the chain is directly density-reachable from its predecessor, with $\mathbf{x_i}$ and $\mathbf{x_j}$ as the first and last objects from the chain, respectively.

Definition 4. (density-connected) An object $\mathbf{x_i}$ is density-connected to an object $\mathbf{x_j}$, with respect to ϵ and $MinPts$, if there is an object $\mathbf{x_k}$ such that both $\mathbf{x_i}$ and $\mathbf{x_j}$ are density-reachable from $\mathbf{x_k}$.

The relation given by Definition 4 is symmetric. DBSCAN clusters are given by Definition 5.

Definition 5. (density-connected set) A density-connected set, *i.e.*, a cluster, with respect to ϵ and $MinPts$ is a non-empty and maximal set, in which all of its objects are density-connected.

Objects that do not belong to any cluster are deemed as noise. DBSCAN generates clusters by examining each object from the dataset. For a given object the algorithm initially establishes if it is a core-point. If that is the case, a cluster is formed and its expansion starts. Expansion occurs by adding to the current cluster all other objects that are density-reachable from objects within the cluster. The cluster is expanded until no further object can be added to it. Objects that are added to a cluster are excluded from further examination. If no indexing structures are used during the examination and expansion processes, DBSCAN has $O(mn^2)$ time complexity (Ester et al., 1996). Regarding cluster shapes DBSCAN is less restrictive than k-means and k-medoids, given that it is capable of uncovering non-convex clusters from data, *i.e.*, clusters of arbitrary shape.

The clusters found by DBSCAN are formed by core-objects and border objects. Border objects are those that are density connected to other objects within the cluster but are not core-objects. A variant of DBSCAN in which border points are deemed as noise, is referred to as DBSCAN* (Campello et al., 2013, 2015). More recently, a hierarchical version of the DBSCAN* algorithm was introduced by Campello et al. (2013, 2015). The algorithm in question, called HDBSCAN*, requires as input only $MinPts$ (Minimum Number of Points) and $MinClSize$ (Minimum Cluster Size, which is an optional parameter) and can derive all possible DBSCAN* partitions (w.r.t. all possble ϵ values) in $O(mn^2)$ time, where m is the number of features and n is the number of objects from the data.

2.3 Clustering Validation

Most of the clustering algorithms from the literature will produce an output (partition or hierarchy) given that their required inputs are provided. In such a scenario, an algorithm may "find" clusters even if the data has none. In an extreme case, an algorithm will find clusters even in uniformly distributed data. The first problem is, one usually does not know beforehand if the data has clusters or not. Even if one assumes that the data has clusters, their number and distribution are usually unknown. In order to avoid the use of what we shall call spurious clustering results, *i.e.*, meaningless or poor results, one can resort to clustering validation techniques. According to Jain and Dubes (1988), clustering validation can be defined as the set of tools and procedures that are used in order to evaluate clustering results in a quantitative and objective manner.

Regarding the first problem listed above, Jain and Dubes (1988) and Gordon (1999) provide a review of statistical procedures that can be employed before the actual clustering of the data, in order to verify if there are clusters in the data under analysis or, put in other words, if it has cluster tendency (Xu and Wunsch II, 2009). Note that this procedure does not determine the actual clusters from the data, but rather provides a quantification of how far the data under analysis is from a data with no cluster structure. The core ideas behind cluster tendency analysis remain roughly the same as when they were presented by Jain and Dubes (1988) and will not be further addressed here.

Even by assuming the existence of clusters in the data one still has to figure out, among other issues, which clustering algorithm to apply and how to select the best configuration of parameters for it. In this particular context, clustering validation techniques can provide an objective and quantitative evaluation of clustering results. The results provided by these techniques can, in turn, help the practitioner to select the "final" clustering result for further and detailed analysis. According to Jain and Dubes (1988) clustering validation techniques can be divided into three major categories, namely: external, internal, and relative. We review these in the sequel.

2.3.1 External Validation

External validity criteria measure the agreement between two different partitions. Usually, but not necessarily, one of the partitions under analysis is the output of a clustering algorithm, whereas the other is the desired solution, *i.e.*, the gold standard partition[2]. Note that in a real clustering application the gold standard partition is not available *a priori*, therefore, the use of external validity criteria is commonly associated with controlled experiments in which, for instance, one wants to determine the best clustering algorithm for a particular application in hand.

The external measure known as Adjusted Rand Index (ARI), from Hubert and Arabie (1985), is one of the most commonly employed in the clustering literature. This measure is based on the Rand Index (RI) (Rand, 1971), which is given by Equation (2.2). In this equation a indicates the number of pairs of objects that are in the same cluster in both partition $\mathcal{C}$ and partition $\mathcal{G}$; b indicates the number of pairs of objects that are in the same cluster in $\mathcal{C}$ and in different clusters in $\mathcal{G}$; c represents the number of pairs of objects that are in different clusters in $\mathcal{C}$ and in the same cluster in $\mathcal{G}$; and d represents the number of pairs of objects that are in different clusters in both $\mathcal{C}$ and $\mathcal{G}$. Note that, if one considers $\mathcal{C}$ as the partition under evaluation and $\mathcal{G}$ as the desired solution (gold standard), then a, b, c, and d correspond to the number of True Positives (TP), False Positives (FP), False Negatives (FN), and True Negatives (TN), respectively. These can be represented in the form of a contingency table (confusion matrix). The Adjusted Rand Index is an extension of the Rand Index which accounts for chance agreements. It assumes that the values previously described follow the generalized hyper-geometric distribution, *i.e.*, both partitions under evaluation are selected randomly, considering their original number of classes and objects. It is given by Equation (2.3).

$$RI(\mathcal{C},\mathcal{G}) = \frac{a+b}{a+b+c+d} \tag{2.2}$$

$$\begin{aligned} ARI(\mathcal{C},\mathcal{G}) &= \frac{RI(\mathcal{C},\mathcal{G}) - RI_{Expected}(\mathcal{C},\mathcal{G})}{RI_{Max}(\mathcal{C},\mathcal{G}) - RI_{Expected}(\mathcal{C},\mathcal{G})} \\ &= \frac{a - \frac{(a+c)(a+b)}{(a+b+c+d)}}{\frac{(a+c)(a+b)}{2} - \frac{(a+c)(a+b)}{(a+b+c+d)}} \end{aligned} \tag{2.3}$$

It is worth noticing that other works, besides that of Hubert and Arabie (1985), also introduced adjusted versions of the Rand Index, as discussed by Milligan and Cooper (1986). Due to problems on their formulation these are usually not employed. For such a reason, any mention to Adjusted Rand Index in this thesis refers only to the version from Hubert and Arabie (1985), which is the external measure employed during our experimental evaluations. We note that there are other external measures proposed in the literature, with new measures being introduced from time to

[2]External validity criteria have also been employed to assess cluster stability through resampling (Dudoit and Fridlyand, 2002) and cluster diversity (Naldi et al., 2013), just to mention a few possible distinct applications.

time, see, for instance de Souto et al. (2012). A number of works have focused on the study and the evaluation of these measures, such as in (Amigó et al., 2009; Meila, 2005; Milligan and Cooper, 1986). We refer to these works for the description and discussion of further external indices.

2.3.2 Internal and Relative Validation

Internal relative validity criteria evaluate clustering results without the use of any external information. The evaluation provided by these measures is based on the data itself and its partitioning, as provided by a clustering algorithm, for example. The Sum of Squared Errors (SSE) from Section 2.2.1, is an example of internal criteria. Note that this particular measure cannot be employed to compare partitions with different numbers of clusters, as already discussed.

Relative validity criteria are defined as internal criteria that are able to compare two partitions and indicate which one is better in relative terms, without being biased by the number of clusters from the partitions under evaluation[3]. There is a number of relative measures in the literature. Even though they have different formulations, the intuition behind most of them is similar, *i.e.*, they favor cluster solutions characterized by a higher "within-cluster-similarity" than "between-cluster-similarity". In the sequel we review a collection of 28 relative validity criteria, that have this preference in common. These measures will be employed later in Chapter 4.

2.3.2.1 Silhouette Width Criterion and Variants

The Silhouette Width Criterion (SWC) was introduced by Rousseeuw (1987). Given an object $\mathbf{x_i}$, its Silhouette is given by Equation (2.4). In this equation a_i is the average distance of $\mathbf{x}_i$ to all the objects within its cluster. In order to define b_i proceed as follows: (i) select a cluster different than that of $\mathbf{x}_i$; (ii) compute the average distance of $\mathbf{x}_i$ to all the objects of that cluster; (iii) repeat the process to all clusters (except the one of $\mathbf{x}_i$); (iv) take the minimum average distance as b_i.

$$s(\mathbf{x}_i) = \frac{b_i - a_i}{\max\{a_i, b_i\}} \tag{2.4}$$

The Silhouette of a clustering solution is then given by Equation (2.5). Its values lie within the $[-1, 1]$ interval, with greater values indicating better partitions. In the case of singletons, *i.e.*, clusters with only one object, $s(\mathbf{x}_i)$ is defined as 0, preventing preference for partitions with $k = n$.

$$SWC(\mathcal{C}) = \frac{1}{n}\sum_{i}^{n} s(\mathbf{x}_i) \tag{2.5}$$

The original SWC has inspired the proposal of three different variants by Hruschka et al. (2004). The first one of these is the so-called Alternative Silhouette Width Criterion (ASWC).

[3] Although most external validity criteria also obey such a definition, the term relative validity criteria is more commonly employed to internal criteria that are also relative, a convention that we adopt hereafter.

Its difference lies in the definition of the Silhouette of an individual object, which is given by Equation (2.6), where ϵ is a small constant employed to avoid division by zero when $a_i = 0$. ASWC has the same rationale behind SWC, with a non-linear component in individual Silhouettes.

$$s(\mathbf{x}_i) = \frac{b_i}{a_i + \epsilon} \tag{2.6}$$

The second variation, called Simplified Silhouette Width Criterion (SSWC) was introduced as a less expensive alternative than the original measure. Its difference lies in the definitions of a_i and b_i for each object. In the SSWC, these are simply the distance of the object to the centroid of its cluster (a_i) and the distance to the closest neighboring centroid (b_i).

The third and final variation, which is called the Alternative Simplified Silhouette Width Criterion (ASSWC), is given by the combination of the two previous variants (ASWC and SSWC).

2.3.2.2 Variance Ratio Criterion

The criterion introduced by Calinski and Harabasz (1974), usually referred to as Variance Ratio Criterion (VRC), is given by Equation (2.7), where n is the number of objects, k is the number of clusters for the partition under evaluation, and $\mathbf{W}$ and $\mathbf{B}$ are the $n \times n$ within-group and between-group dispersion matrices, which are given by Equations (2.8) and (2.9), respectively.

$$VRC(\mathcal{C}) = \left(\frac{\text{Trace}(\mathbf{B})}{\text{Trace}(\mathbf{W})}\right)\left(\frac{n-k}{k-1}\right) \tag{2.7}$$

$$\mathbf{W} = \sum_{i=1}^{k} \sum_{\mathbf{x}_j \in \mathbf{C}_i} (\mathbf{x}_j - \overline{\mathbf{c}}_i)(\mathbf{x}_j - \overline{\mathbf{c}}_i)^T \tag{2.8}$$

$$\mathbf{B} = \sum_{i=1}^{k} n_i (\overline{\mathbf{c}}_i - \overline{\mathbf{c}})(\overline{\mathbf{c}}_i - \overline{\mathbf{c}})^T \tag{2.9}$$

In these definitions, n_i is the number of objects in cluster C_i, $\overline{\mathbf{c}}_i$ is the centroid of that cluster, and $\overline{\mathbf{c}}$ is the mean of all data points, *i.e.*, the data centroid. The second term in Equation (2.7) accounts for increases in the number of clusters. The greater the value, the better is the partition, according to VRC.

2.3.2.3 Dunn and Variants

Dunn's relative index (Dunn, 1974) is given in Equation (2.10)

$$Dunn(\mathcal{C}) = \min_{\substack{C_i, C_j \in \mathcal{C}, \\ C_i \neq C_j}} \left(\frac{\delta_{C_i, C_j}}{\max_{C_l \in \mathcal{C}} \Delta_{C_l}}\right) \tag{2.10}$$

where δ_{C_i,C_j} is the minimum distance between two objects from distinct clusters (the distance between two clusters), given in Equation (2.11), and Δ_{C_l} represents the diameter of a cluster, *i.e.*, the maximum distance among its objects, which is given in Equation (2.12).

$$\delta_{C_i,C_j} = \min_{\substack{\mathbf{x_r} \in C_i \\ \mathbf{x_s} \in C_j}} ||\mathbf{x}_r - \mathbf{x}_s|| \tag{2.11}$$

$$\Delta_{C_l} = \max_{\substack{\mathbf{x}_r, \mathbf{x}_s \in C_l \\ \mathbf{x}_r \neq \mathbf{x}_s}} ||\mathbf{x}_r - \mathbf{x}_s|| \tag{2.12}$$

Bezdek and Pal (1998) introduced different alternative definitions for δ_{C_i,C_j} and Δ_{C_l}, giving rise to a total of 18 variants of the measure (including the original index). The alternative definitions for inter cluster distance are given through Equations (2.13) to (2.17).

$$\delta_{C_i,C_j} = \max_{\substack{\mathbf{x_r} \in C_i \\ \mathbf{x_s} \in C_j}} ||\mathbf{x}_r - \mathbf{x}_s|| \tag{2.13}$$

$$\delta_{C_i,C_j} = \frac{1}{n_i n_j} \sum_{\mathbf{x}_r \in C_i} \sum_{\mathbf{x}_s \in C_j} ||\mathbf{x}_r - \mathbf{x}_s|| \tag{2.14}$$

$$\delta_{C_i,C_j} = ||\overline{\mathbf{c}}_i - \overline{\mathbf{c}}_j|| \tag{2.15}$$

$$\delta_{C_i,C_j} = \frac{1}{n_i + n_j} \left(\sum_{\mathbf{x}_r \in C_i} ||\mathbf{x}_r - \overline{\mathbf{c}}_i|| + \sum_{\mathbf{x}_s \in C_j} ||\mathbf{x}_s - \overline{\mathbf{c}}_j|| \right) \tag{2.16}$$

$$\delta_{C_i,C_j} = \max \left\{ \max_{\mathbf{x}_r \in C_i} \min_{\mathbf{x}_s \in C_j} ||\mathbf{x}_r - \mathbf{x}_s||, \max_{\mathbf{x}_s \in C_j} \min_{\mathbf{x}_r \in C_i} ||\mathbf{x}_r - \mathbf{x}_s|| \right\} \tag{2.17}$$

The alternative definitions for cluster diameter are provided by Equations (2.18) and (2.19).

$$\Delta_{C_l} = \frac{1}{n_l(n_l - 1)} \sum_{\substack{\mathbf{x}_r, \mathbf{x}_s \in C_l, \\ \mathbf{x}_r \neq \mathbf{x}_s}} ||\mathbf{x}_r - \mathbf{x}_s|| \tag{2.18}$$

$$\Delta_{C_l} = \frac{2}{n_l} \sum_{\mathbf{x}_r \in C_l} ||\mathbf{x}_r - \overline{\mathbf{c}}_l|| \tag{2.19}$$

All variants of the Dunn criterion (including the original version) are maximization criteria.

2.3.2.4 Davies-Bouldin

In order to define the criterion introduced by Davies and Bouldin (1979), which receives the names of the authors, let us first define the average distances within a group C_i, denoted here by $\overline{d}_i$. This is given by Equation (2.20), for which $\overline{\mathbf{c}}_i$ is the centroid of the cluster and n_i its number of objects.

$$\overline{d}_i = \frac{1}{n_i} \sum_{\mathbf{x}_j \in \mathbf{C}_i} ||\mathbf{x}_j - \overline{\mathbf{c}}_i|| \tag{2.20}$$

Let us also denote the distance between two clusters C_i and C_j as the difference between their centroids, denoted by $\overline{\mathbf{c}}_i$ and $\overline{\mathbf{c}}_j$, respectively. This is given by Equation (2.21).

$$d_{i,j} = ||\overline{\mathbf{c}}_i - \overline{\mathbf{c}}_j|| \tag{2.21}$$

Based on these two definitions, the Davies-Bouldin criterion can be defined by Equation (2.22).

$$DB(\mathcal{C}) = \frac{1}{k} \sum_{i=1}^{k} \max_{j \neq i} \left(\frac{\overline{d}_i + \overline{d}_j}{d_{i,j}} \right) \tag{2.22}$$

The criterion adds up the worst case for each cluster under evaluation, considering the distance within the cluster and the distance between clusters. DB is a minimization criterion.

2.3.2.5 PBM

The criterion introduced by Pakhira et al. (2004), which is usually referred to by its authors initials, *i.e.*, PBM, is given by Equation (2.23). Terms E_1, E_k, and D_K denote the sum of the differences from each object to the centroid of the whole data (Equation (2.24)), the sum of the distances of each object to the centroid of its cluster (Equation (2.25)), and the maximum distance between cluster centroids (Equation (2.26)), respectively. The first term in Equation (2.23) penalizes for the number of clusters. Good partitions are indicated by high values of PBM.

$$PBM(\mathcal{C}) = \left(\frac{1}{k} \frac{E_1}{E_K} D_K \right)^2 \tag{2.23}$$

$$E_1 = \sum_{i=1}^{n} ||\mathbf{x}_i - \overline{\mathbf{c}}|| \tag{2.24}$$

$$E_K = \sum_{i=1}^{k} \sum_{\mathbf{x}_r \in C_i} ||\mathbf{x}_r - \overline{\mathbf{c}}_i|| \tag{2.25}$$

$$D_K = \max_{\substack{C_i, C_j \in \mathcal{C}, \\ C_i \neq C_j}} ||\overline{\mathbf{c}}_i - \overline{\mathbf{c}}_j|| \tag{2.26}$$

2.3.2.6 C-Index

The criterion known as C-Index (Hubert and Levin, 1976) is given by Equation (2.27). The term S_W is given by Equation (2.28) and accounts for the sum of distances of pairs of objects within the same cluster. The terms S_{Min} and S_{Max} are the minimum and maximum possible values for S_W. Assuming the existence of n_p pairs of objects within the same cluster (no matter which cluster), S_{Min} is the sum of the n_p smallest distances from the dataset, whereas S_{Max} is the sum of the n_p largest ones. C-Index is a minimization criterion.

$$CI(\mathcal{C}) = \frac{S_W - S_{Min}}{S_{Max} - S_{Min}} \tag{2.27}$$

$$S_W = \sum_{C_i \in \mathcal{C}} \sum_{\mathbf{x}_r, \mathbf{x}_s \in C_i} ||\mathbf{x}_r - \mathbf{x}_s|| \tag{2.28}$$

2.3.2.7 Point-Biserial

The Point-Biserial (Milligan, 1981) criterion is inspired by the homonymous correlation, which is defined to work with a continuous variable (distances here) and a binary variable (cluster memberships, in this case). It is given by Equation (2.29), in which t is equal to the number of pairs of objects, that is $n(n-1)/2$. The terms w_d and b_d indicate the total number of pairs of objects within the same cluster and in different clusters, respectively. The average distance of pairs of objects within the same cluster and in different clusters are denoted by d_w and d_b. Term s_d accounts for the standard deviation in the distances from the data. It is a maximization criterion.

$$PB(\mathcal{C}) = \frac{(d_b - d_w)}{s_d} \sqrt{\frac{w_d b_d}{t^2}} \tag{2.29}$$

2.3.2.8 $C/\sqrt{k}$

The $C/\sqrt{k}$ criterion (Hill, 1980; Ratkowsky and Lance, 1978) considers the contribution of each feature from the dataset in the distances within the same cluster and between different clusters individually. The measure is given by Equation (2.30). The first term is a penalization factor accounting for the number of clusters. The second term sums the contributions of each one of the m features separately. The term WDD_i is the difference, for each object $\mathbf{x}_j$ to the centroid of the data $\bar{\mathbf{c}}$, considering only the i^{th} feature. Similarly, WCD_i is the within cluster distance, considering only the i^{th} feature. $C/\sqrt{k}$ is a maximization criterion.

$$C/\sqrt{k} = \frac{1}{n\sqrt{k}} \sum_{i=1}^{m} \sqrt{\frac{WDD_i - WCD_i}{WDD_i}} \tag{2.30}$$

$$WDD_i = \sum_{\mathbf{x}_j \in \mathbf{X}}^{n} ||x_i - \bar{c}_i||^2 \tag{2.31}$$

$$WCD_i = \sum_{l=1}^{k} \sum_{\mathbf{x}_j \in C_l} ||x_i - \bar{c}_i||^2 \tag{2.32}$$

2.4 Relative Validity Criteria Evaluation

In the previous section 28 relative validity criteria from the literature were reviewed. These are likely to have arisen as a consequence of the vague definition surrounding the term cluster. In this sense, each relative validity criterion tries to provide a formulation for a different scenario or cluster definition. The problem is, given the number of formulations available, to determine which ones make the most sense, at least in general. In order to try to answer this question, or to provide some guidelines on the selection of these measures, different works have evaluated them systematically (Milligan and Cooper, 1985; Vendramin et al., 2009, 2010). The main idea behind these works is to use external information during the evaluation of the measures. Note that external information is not available in real applications, but for the sake of evaluation it can be generated (or obtained with a field specialist), so that controlled experiments can be conducted. Considering the nature of the external information employed during the evaluation, two different strategies can be defined. The first one, which we refer to as Traditional Methodology, comes from the work of Milligan and Cooper (1985). The second one, which attempts to overcome some drawbacks from the Traditional Methodology is referred to as Alternative Methodology here, and was introduced by Vendramin et al. (2009, 2010). Both methodologies are reviewed in the sequel. They will be employed during our experimental evaluations, for which results are provided in Chapters 4 and 5.

2.4.1 Traditional Methodology

The methodology introduced by Milligan and Cooper (1985) evaluates relative validity criteria by their ability to identify the correct number of clusters in the data. The correct number of clusters is, of course, defined previously by the external partition. In this sense, given a dataset $\mathbf{X}$ with n objects and an *a priori* defined number of classes (k^*), the evaluation proceeds as follows:

1. Generate a collection of partitions by employing different clustering algorithms and selecting different number of clusters (k), *e.g.*, $2 \leq k \leq \sqrt{n}$ (Vendramin et al., 2009, 2010);

2. Determine the quality of each partition with all the relative validity criteria under evaluation;

3. For each relative validity criterion, verify if the best partition it indicates has the correct number of clusters, as defined by the desired solution. If that is the case, mark it as a hit.

The procedure above is based on a single dataset. In practice, a collection of datasets is usually employed. At the end of the process (repeated for all datasets from the collection), the sum of hits for each relative criterion under evaluation is reported as an indicative of its overall quality.

2.4.2 Alternative Methodology

More recently, Vendramin et al. (2009, 2010) introduced an alternative methodology for the evaluation of relative validity criteria. The authors introduce their methodology in an attempt to overcome the three major deficiencies associated with the traditional methodology from (Milligan and Cooper, 1985). These drawbacks are discussed in detail bellow.

1. The methodology from Milligan and Cooper (1985) implies that the quality of a relative criterion can be quantified accordingly to the number of times that it correctly identifies the number of clusters of a dataset, as defined by the external partition. This premise, however, is not always valid. Consider, for instance, the two clustering solutions obtained with the k-means clustering algorithm depicted in Figure 2.2. From this figure, it becomes clear that a "correct" number of clusters does not necessarily always indicates a good partition.

2. Milligan and Cooper (1985) assume that errors made by a criterion are equally bad. That is, identifying $k = (k^* + \Delta)$ or $k = (k^* - \Delta)$ clusters is equally bad according to the methodology. Note that a partition with a greater number of clusters than that of the gold standard partition usually breaks large clusters into small ones, providing thus a finer granularity. This is not the case of a partition with fewer clusters than those defined by the gold standard partition, in which information is usually lost as a result of the cluster merging.

3. Finally, the evaluation obtained with the traditional methodology is based on the number of clusters of the single best partition indicated by each relative validity criterion. Even though readily available, the performance of each criterion when evaluating other partitions is, therefore, completely ignored. By ignoring the results of these partitions, the traditional methodology does not evaluate the robustness of relative validity criteria, which can be important in some practical application scenarios, as pointed out by Vendramin et al. (2010).

Bearing such drawbacks in mind, the alternative methodology introduced by Vendramin et al. (2009, 2010) takes into account all the partitions generated during the process, alongside with their associated quality. The quality of each partition, as identified by the relative criterion, is then compared to that obtained with the use of an external measure. The procedure is as follows:

1. For a given dataset, generate a collection of partitions by employing different clustering algorithms and selecting different number of clusters (k), which is usually as $2 \leq k \leq \sqrt{n}$;

2. Determine the quality of each partition with all the relative validity criteria under evaluation;

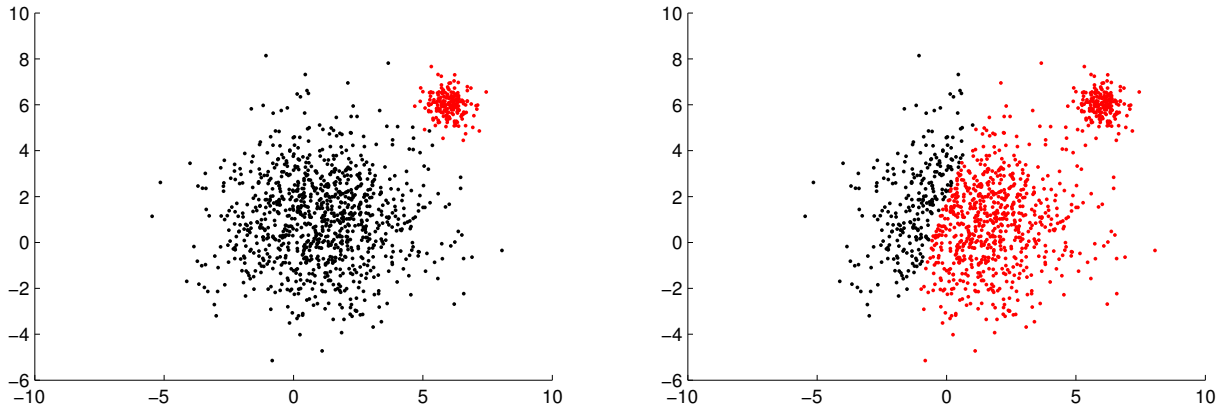


Figure 2.2: Two partitions of the same data, with $k = 2$ clusters, denoted in red and black. It is intuitive that the clustering on the left is more *natural* than the one presented on the right plot.

3. Also determine the quality of each partition with an external validity criterion, *i.e.*, ARI;

4. For each relative criterion, correlate its evaluations with the ones from the external criterion.

Note that Steps 1 and 2 are exactly the same from the traditional methodology from Milligan and Cooper (1985). The difference here lies in the fact that all partitions that were generated in Step 1 integrate the evaluation process. To that end, they are also evaluated with the use of an external criterion. At the end, the correlation between the relative and external criteria indicate the quality of the relative measure, *i.e.*, its agreement with external labeling. Common choices for measuring correlation are the well-known Pearson (Pearson, 1895) and Spearman (Spearman, 1904), although other measures can also be employed.

2.5 Chapter Remarks

In this chapter we provided a concise overview of cluster analysis. We started by presenting the basic concepts of clustering (Section 2.1) and some of the most well-known clustering algorithms from the literature (Section 2.2). These are the clustering algorithms ultimately employed in this thesis to generate different sets of clustering results. We then proceeded to the topic of clustering validation. The Adjusted Rand Index (ARI), which is one most well-known external validity criterion in the literature, was reviewed. This particular external validation measure will be employed in virtually all the controlled experiments presented later on this thesis. Following on the discussion of clustering validation we presented a collection of 28 relative validity criteria. These will be particularly important in the discussions presented in Chapter 4. Finally, we elaborated on the evaluation of relative validity criteria. This particular topic is of fundamental importance to the understanding of Chapters 4 and 5 of the thesis, in which different validity criteria are evaluated.

CHAPTER

3

Gene Expression Data

Living organisms, as we know, are made of cells, the so-called fundamental unities of life (Lodish et al., 2012). Although most of the cells in a particular living organism carry the very same genetic information (Belk and Borden, 2003), *i.e.*, the genetic code responsible for the characteristics of that being, they can become highly specialized in order to participate in the most diverse processes. Highly specialized cells are possible because of different gene expression levels, which ultimately define which proteins and what quantities of them will be produced by each cell. It is, therefore, the intricate process of gene expression that makes life as we know possible. Even though it is endowed with some level of built-in robustness and error correction mechanisms, certain disturbances in the gene expression process cannot be corrected, leading to the appearance of diseases, such as cancer (Alberts et al., 2014; Lodish et al., 2012).

An advanced knowledge about gene expression can help, for instance, the identification of gene functions, leading to a better understanding of how different types of cells and living organisms work. It also has the potential to help unveil the mechanisms behind different types of diseases on a molecular level, so that genes associated with them can be identified and later targeted in personalized treatments. In order to measure gene expression, different technologies have been developed in the last years, allowing researchers to quantify gene expression for the most diverse organisms and conditions of interest. Given the massive amounts of data generated by these technologies, their analyses has been usually described as a great challenge in the literature (Brazma and Vilo, 2000; Fan et al., 2014; Finotello and Di Camillo, 2014; Sherlock, 2001; Zhang, 2006).

In this chapter we provide a concise review of gene expression data, providing the necessary background for the experiments and results that will be discussed on Chapters 6 and 7. We start by briefly presenting and discussing basic biological concepts associated with gene expression in Section 3.1. The two major high throughput technologies employed to quantify gene expression levels are discussed in Section 3.2. Datasets from these two technologies are employed in this thesis. In Section 3.3, we highlight the applications of clustering in the analysis of gene expression data. Finally, in Section 3.4, we provide a brief overview of the Gene Ontology (GO), from which external biological information will be later employed in the evaluation of distance measures and clustering results. We acknowledge that Molecular Biology is a complex and extensive area of research. In fact, as pointed out by Setubal and Meidanis (1997), for each rule in biology there are several exceptions, thus making the subject even more complex. Bearing that in mind, we do not have the ambition to cover the subject in depth and refer the interested reader to the following books for further details (Alberts et al., 2014; Lesk, 2008; Setubal and Meidanis, 1997).

3.1 Biological Background

Before presenting the technologies that provided some of the data employed in this thesis, we review biological concepts related to gene expression, which is measured by those technologies.

3.1.1 Nucleic Acids

Cells are complex entities capable of reproduction, information processing, among other diverse functions (Lodish et al., 2012). They can organize and give rise to multicellular organisms, like us, or present themselves alone, as in many unicellular life forms, such as yeast. Although cells can present themselves in different forms (eukaryotic and prokaryotic) and differentiate in order to perform different tasks, most of them have something in common, that is, they carry the same and whole genetic information that defines the living organism they belong to (Alberts et al., 2014). The molecules responsible for encoding, storing, and transmitting genetic information are called nucleic acids. There are two main forms of nucleic acids, namely DNA (Deoxyribonucleic Acid) and RNA (Ribonucleic Acid). We discuss each one of these molecules in the sequel.

DNA is usually found in a double strand structure (also referred to as double helix structure), as presented in Figure 3.1. It is composed of four base pairs, namely Adenine (A), Thymine (T), Cytosine (C), and Guanine (G). These bind to each other selectively, obeying the following two rules: Adenine (A) bases bind only with Thymine (T); and Cytosine (C) bases bind only with Guanine (G). One key property of DNA, is hybridization, by which two single stranded complementary DNA molecules will bind together to form a double strand (this property is the foundation of technologies such as microarrays). The length of a DNA sequence is usually given in base-pairs (BP), *i.e.*, the number of pairs it contains. In eukaryotic cells (cells that contain a defined nucleus) DNA molecules are further organized in chromosomes, with the whole set of

chromosomes from an organism being called its genome. The human genome, for instance, is composed of 23 chromosome pairs, with a DNA length of about 3.2×10^9 BP (Alberts et al., 2014). In each organism there are several regions of DNA that store code for specific molecules, such as RNA and proteins, these particular regions are known as genes (Alberts et al., 2014).

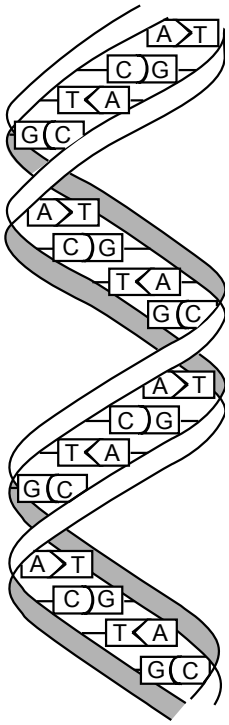


Figure 3.1: Representation of a double stranded DNA molecule, with its base pairs: Adenine (A), Thymine (T), Cytosine (C), and Guanine (G). Adapted from Human Genome Program (1992).

RNA molecules are closely related to those of DNA, with at least two main differences. The first one is that RNA is usually found in the form of a single strand structure. The second difference lies in the replacement of Thymine (T) base by Uracil (U). The other bases remain the same, alongside with their binding properties (with Uracil (U) now binding to Adenine (A) in RNA). There are three main subtypes of RNA, called messenger RNA (mRNA), transfer RNA (tRNA), and ribosomal RNA (rRNA). These different forms of the RNA molecules play key roles in cellular regulation and in the process of protein synthesis, as we discuss in the following.

3.1.2 Gene Expression

Gene expression is the process by which the coding region of a gene, present in the DNA, is used to build molecules known as proteins (Alberts et al., 2014). The process of gene expression, alongside with the process of DNA replication, is part of the central dogma of the molecular biology, which establishes how genetic information is transmitted in the cell (Setubal and Meidanis, 1997). To provide an overview of these processes we make use of Figure 3.2. The process of gene expression is represented through steps 1 to 3 in this figure. In the first step, called transcription, the information stored in DNA, more specifically in a gene, is transcribed into

a pre-mRNA molecule. Non-coding regions of the pre-mRNA are removed, in 2, resulting in a mRNA molecule. The mRNA, which now carries the information necessary to build a protein leaves the cell nucleus and moves to its cytoplasm, more specifically to the ribosome, where translation takes place, in 3. During the process of translation, the mRNA molecule is read by ribosomal subunits. Every three bases from the mRNA sequence, which are also known as a codon, code for a specific amino acid. The amino acid specified by each one of the codons is acquired within the cell's cytoplasm by a tRNA molecule and transported to the ribosome, which assembles the amino acids in a chain, called protein. The process continues until a stop codon is identified by the ribosome. Note that although a three letter code with four possibilities in each position, as specified by the mRNA, can code up to 64 different codons, there are only 20 different types of amino acids, with some of the codons coding for the same amino acid. The process of replication, which is also part of the central dogma, is depicted in 4. By the process of replication a DNA molecule can duplicate itself, resulting in two identical copies of the original DNA molecule.

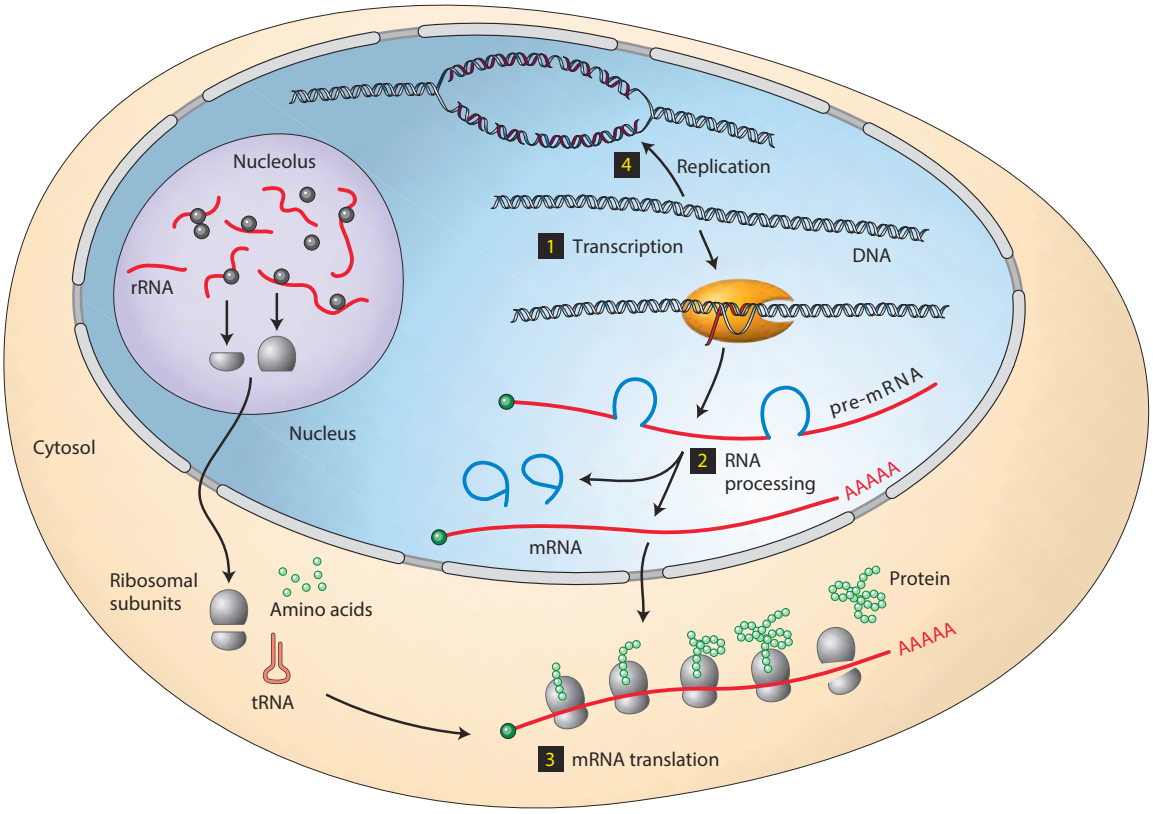


Figure 3.2: The central dogma of molecular biology. Figure adapted from Lodish et al. (2012).

3.2 Measuring Gene Expression

Technologies for measuring gene expression can be broadly divided into two major categories based on the amount of data they produce, namely low and high throughput (Kuo et al., 2004). Low throughput technologies, such as Polymerase Chain Reaction (VanGuilder et al., 2008)

and Northern Blot (Lodish et al., 2012), allow expression level measurement for a handful of genes, usually with a high precision. Their high precision is counterbalanced, however, by the quantity of genes that can be analyzed. These technologies are, therefore, mostly employed to confirm or reject experimentally previously formulated hypothesis (Kuo et al., 2004; VanGuilder et al., 2008). Technologies that fall under the high throughput category, on the other hand, can provide a complete snapshot of the current state of a cell. These technologies are able to produce massive amounts of data that can be used to *formulate* different hypothesis within a global perspective. Microarrays (Harrington et al., 2000) are probably the best known representatives of high throughput technologies, given their widespread use. More recently, the emergence of Next Generation Sequencing (NGS) (Reis-Filho, 2009), has enabled the development of the RNA-Seq technology (Ozsolak and Milos, 2011; Wang et al., 2009), which overcomes several of the drawbacks associated with microarrays. We briefly review these two technologies in the sequel.

3.2.1 Microarrays

Microarray technology is based on the concept of hybridization, by which complementary base pairs tend to spontaneously join together (Zhang, 2006). Although there are different variants of the technology, each one with its subtleties, the main underlying idea behind them is the same (Dalma-Weiszhausz et al., 2006; Duggan et al., 1999; Harrington et al., 2000; Schena et al., 1995; Tarca et al., 2007). Microarrays are usually made of glass or nylon, where regions (also called spots) are defined for different genes under investigation. At each one of these regions, thousands of sequences that belong to a specific gene are synthesized in situ, *i.e.*, printed on the array, or deposited robotically. When an experiment takes place, genetic material from a biological sample is extracted and prepared with substances that allow posterior identification, *i.e.*, markers, such as a fluorescent material. The material is then placed over the array, so that the hybridization process can take place. Afterwards the array is washed, in order to remove genetic material that did not hybridize to any of its predefined regions. At the end of the process, the array is submitted to a light source, so that the markers added to the biological sample can be detected and quantified. The amount of marker substance in a given spot on the array is assumed to be proportional to that present in the cell. Images of the array are obtained and with the use of image processing techniques, the expression of each gene is quantified numerically for posterior analysis.

There are different microarray technologies available. From these, cDNA (Schena et al., 1995) and Affymetrix (Lockhart et al., 1996; Tarca et al., 2007) microarrays are among the ones that first gained popularity. Even though serving the same purpose, they have some differences regarding their experimental process. The manufacture and experimental process of these two technologies are highlighted in Figure 3.3. Here we focus on key differences regarding the data obtained from each technology, setting aside specific issues regarding their preparation. First, it is important to note that Affymetrix microarrays (also referred to as high density oligonucleotide microarrays) measure gene expression levels in absolute terms, whereas cDNA microarrays measure expression

levels in relative terms. To that end, note that in Figure 3.3 (right) two arrays are necessary in order to measure expression levels coming from two different samples, whilst in Figure 3.3 (left) only one cDNA microarray is sufficient. For such a reason, Affymetrix microarrays are usually referred to as single-channel whereas cDNA microarrays are referred to as double-channel. At the end of the process, expression levels obtained from two Affymetrix arrays can be combined into a single value, if necessary. In practice, however, expression values from Affymetrix microarrays are presented in absolute terms. For cDNA microarrays, the final expression of a gene (e_g) is necessarily a ratio, which is usually obtained by the log-ratio between the two fluorescent markers present in the array, that is $e_g = \log(Cy5/Cy3)$, where $Cy5$ and $Cy3$ correspond to the expression intensities for each sample being assessed (as measured by their fluorescence).

Each one of the steps that encompass a microarray experiment, from its manufacture to its analysis, has been subject of research. Describing the whole process in detail is beyond the scope of this thesis. A detailed review of Affymetrix GeneChip microarrays is provided by Dalma-Weiszhausz et al. (2006). Details regarding cDNA microarrays are provided by Duggan et al. (1999). A good review regarding image acquisition from microarrays is provided by Esteves (2007) and references therein. Zhang (2006) covers microarrays from their manufacture to its analysis, offering diverse pointers to more detailed material on each one of the topics.

3.2.2 RNA-Seq

Microarray technology has changed the way in which biological experiments are conducted (Jiang et al., 2004; Zhang, 2006). It has allowed the examination of genome-wide expression levels for different conditions, providing a snapshot of the cell's current state. Notwithstanding, microarrays have several limitations, among which is notable the need to determine *a priori* the sequences that will be placed on the array. Furthermore, background noise and cross-hybridization (sequences that are not completely alike, but hybridize) interference, may result in difficulties to detect genes expressed at very low or high intensities (Zhao et al., 2014).

The development of Next-Generation Sequencing (NGS) (Metzker, 2010), has decreased the cost associated with sequencing and enabled the sequencing of RNA transcripts (RNA-Seq) from a given cell of interest (Ozsolak and Milos, 2011; Wang et al., 2009; Zhang et al., 2015). Unlike microarrays, RNA-Seq is based on the sequencing of the genetic material obtained from a biological sample. Given that no previous definition of sequences is necessary, approaches based on the RNA-Seq technology can identify and quantify expression level for genes that undergo alternate splicing or have variation in their sequences (Metzker, 2010). Another advantage of the technology is a higher resolution, in comparison to that of microarrays. Given that it does not suffer from cross-hybridization problems, genes expressed at low and high levels can be more accurately identified, with no lower or upper limit intrinsic to the technology, as is the case in microarrays.

In a simplified manner, a RNA-Seq experiment starts with the collection of mRNA molecules from a cell of interest. The mRNA molecules are fragmented into shorter molecules, depending

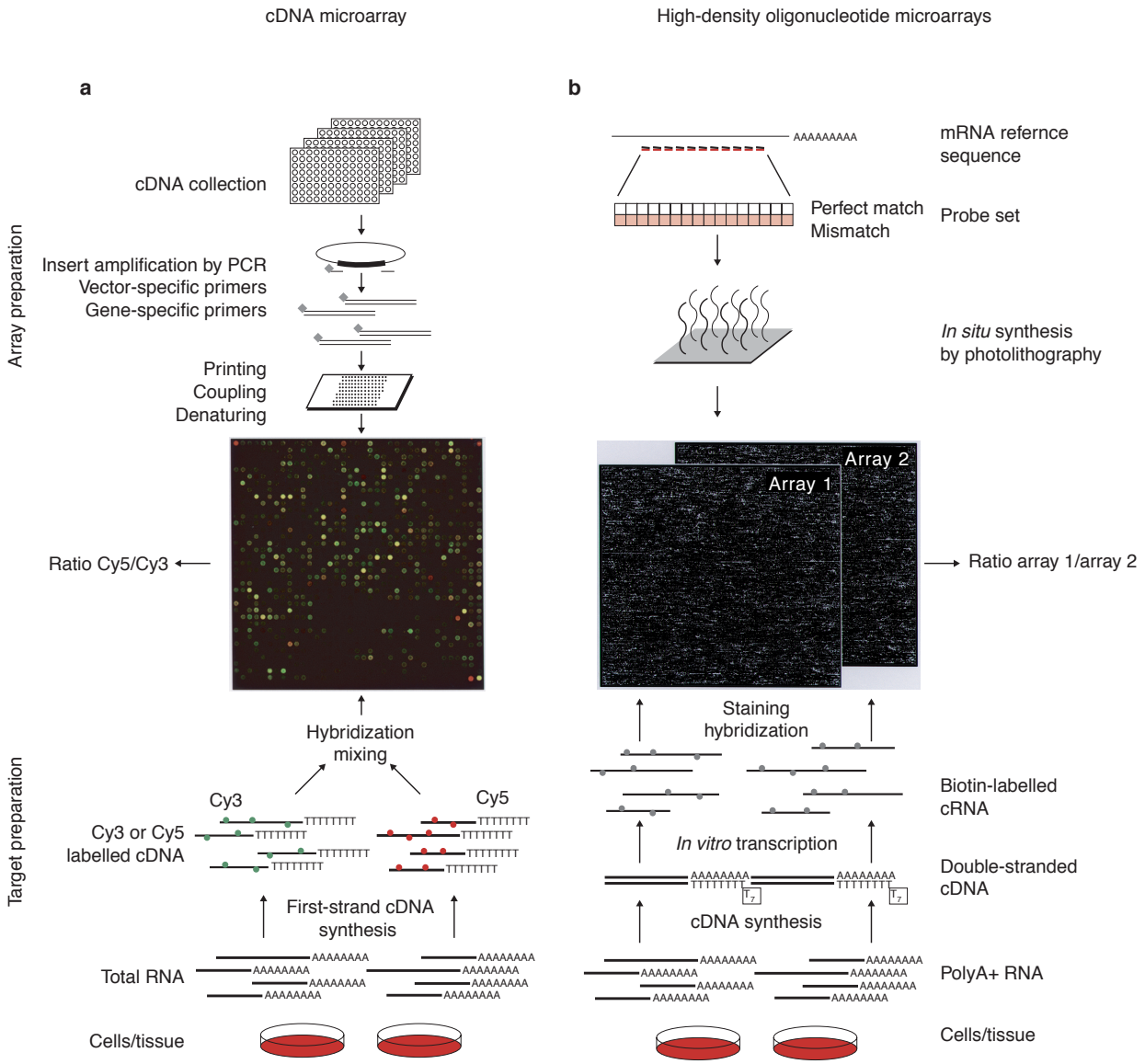


Figure 3.3: Manufacture (top) and experimental (bottom) processes for Affymetrix (right) and cDNA (left) microarrays. Affymetrix microarrays (referred to as high-density oligonucleotide microarrays) assess gene expression levels for one sample at a time (single-channel), whereas a single cDNA microarray requires two samples (double-channel). Note that for a single Affymetrix microarray expression levels are absolute. In the case of cDNA microarrays expression levels are relative, given that one array examines two samples. Figure from Schulze and Downward (2001).

on the sequencing technology that will be employed. In order to sequence the genetic material, sequence adapters (short constant sequences) are added to the fragmented mRNA molecules. These can have different roles in the process, also accordingly to the technology employed. Once the genetic material is ready, it is provided as input to the sequencing technology, which, at the end of the process, provides for each sequence found in the biological sample a discrete expression estimate, *i.e.*, its count. Another advantage of RNA-Seq technology is the possibility to summarize genes on distinct perspectives (exons, isoforms, splice junctions or whole transcript levels), capturing different levels of variability from the biological process. We note that the details related to RNA-Seq technology are far more complex than those related to microarrays. We refer the interested reader to the work of Metzker (2010) for a review of different NGS technologies. A detailed overview of RNA-Seq is provided by Wang et al. (2009).

3.3 Clustering Gene Expression Data

The development of high throughput technologies has made the collection of massive amounts of data possible. Considering the capabilities of these technologies, the bottleneck to understand complex biological phenomena has thus changed from data generation to data analysis (Brazma and Vilo, 2000; Sherlock, 2001). Gene expression data has peculiar characteristics that make its analysis challenging. The number of samples available is usually limited, because of the financial costs associated to experiments[1]. In the case of microarrays, for each biological sample an array is necessary, whereas in the case of RNA-Seq the cost is associated to sequencing the biological material from a given sample. Since each experiment measures expression levels for a large number of genes, the gene expression data matrix is usually composed of a small number of samples and a large number of genes, as depicted in Figure 3.4. In the case of microarrays, due to experimental errors, before data analysis one usually has to deal with the problem of missing values. de Souto et al. (2015) showed recently, however, that advanced missing value imputation methods provide no difference in the clustering and classification of the data, when compared to simple methods, *e.g.*, mean. This work was developed with participation of the thesis's author.

Different types of gene expression experiments can be conducted in order to investigate the most diverse biological phenomena. One can, for instance, investigate the expression of genes during cellular division (or other time related processes). In such experiments, usually referred to as time-course, each sample provides a snapshot of the cell's state in a different instant of time. Because of the time component involved, each gene under analysis can be regarded as a time-series, although traditional time-series techniques usually cannot by employed due to their short length (Son and Baek, 2008). Apart from time dependent experiments one can investigate the expression of genes for samples coming from different patient groups, related to different types or subtypes of diseases.

[1]The term sample designs the expression levels of all genes, as measured by a microarray or RNA-Seq experiment.

	Sample 1	Sample 2	...	Sample s
Gene 1	$e_{1,1}$	$e_{1,2}$	...	$e_{1,s}$
Gene 2	$e_{2,1}$	$e_{2,2}$	...	$e_{2,s}$
⋮	⋮	⋮	⋱	⋮
Gene g	$e_{g,1}$	$e_{g,2}$	...	$e_{g,s}$

Figure 3.4: Depiction of a gene expression data matrix. Each biological sample from the matrix comes from a different experiment. In the case of microarrays, for instance, each sample is associated with a distinct microarray experiment. Gene expression data is characterized by a large number of genes (thousands) and a small number of samples (usually dozens or a few hundreds).

Cluster analysis plays a key role in the analysis of both types of experiments previously discussed. In the case of time-series experiments, clustering is usually performed on genes. The clustering of gene time-series may help, for instance, to identify genes that share the same regulatory mechanisms or functions (D'haeseleer, 2005; Heyer et al., 1999; Kerr et al., 2008). When samples are associated with different patients, clustering is usually performed on samples. In this application scenario the main objective is to detect previously unknown clusters of biological samples, which are usually associated with unknown types of cancer (de Souto et al., 2008). Since the seminal work presented by Golub et al. (1999), the clustering of cancer samples has drawn quite an attention of the research community and has been employed in a number of works, *e.g.*, Alizadeh et al. (2000); Alon et al. (1999); Kao et al. (2009); Lapointe et al. (2004); Ramaswamy et al. (2003). In these works, novel cancer subtypes were unveiled with the application of clustering algorithms. Once cancer signatures are identified for different subtypes on a genomic level, specific drugs can be developed to aim particular variants of the disease, improving treatment efficacy while reducing its side effects.

Taking into account the peculiarities of each one of the aforementioned scenarios, a number of clustering algorithms has been developed specifically to the clustering of cancer samples, *e.g.*, (Ben-Dor et al., 1999; McLachlan et al., 2002; Sharan et al., 2003), and particularly to the clustering of gene time-series, *e.g.*, (Costa et al., 2005; Ernst et al., 2005; Hestilow and Huang, 2009; Heyer et al., 1999; Si et al., 2013). Moreover, classical algorithms from the clustering literature have been borrowed and employed with success to the analysis of gene expression data, including, but not limited to, hierarchical methods (Jain and Dubes, 1988) and k-means (MacQueen, 1967). In this thesis, we focus on two distinct aspects of gene expression clustering, as we discuss in the following.

The first one concerns the proper selection and evaluation of distance measures for gene expression data clustering. In view of gene expression data characteristics, objects (genes or samples) should be regarded as similar if they exhibit trend or shape similarity (Heyer et al., 1999). In a number of clustering algorithms this requires the definition of proper distance measures. There is a variety of measures capable of identifying trend similarity available in the general clustering

literature. Additionally, some distances have been specifically introduced aiming the clustering of gene time-series, *e.g.*, (Balasubramaniyan et al., 2005; Heyer et al., 1999; Möller-Levet et al., 2005; Son and Baek, 2008), taking into account its temporal characteristics. In this context, in Chapter 6 we introduce a methodology for the evaluation of distance measures based on the Gene Ontology (Ashburner et al., 2000). We also evaluate the combined influence of classical clustering algorithms and different distance measures, for microarray and RNA-Seq data analysis, significantly extending the previous investigations conducted by the author during his Master's Degree (Jaskowiak, 2011).

The second aspect concerns the validation of clustering results in the specific context of clustering of genes. Although we already discussed clustering validation and its importance in a broad perspective in Chapter 2, the validation of gene clusters has its own peculiarities. As a real clustering application, there are no comprehensive datasets with external information in the form of a golden standard partition available for the evaluation of methods in this scenario. The current knowledge about different genes is, however, structured in different formats, such as the Gene Ontology (GO) (Ashburner et al., 2000) or the Kyoto Encyclopedia of Genes and Genomes (KEGG) (Kanehisa et al., 2010). In the following we provide a brief overview of the GO, which was the source of biological information employed later in Chapter 6 and Chapter 7.

3.4 Gene Ontology

In a broad context, ontologies provide ways to represent and share knowledge about a particular domain. Ontologies aim to accomplish these goals by creating an unified and standardized notation in which information is represented. According to Uschold and Grüninger (1996) the knowledge regarding a particular domain is usually structured in an ontology through terms, relations between terms, and their definitions. The terms that compose an ontology and their respective meanings are defined though the so-called vocabulary of terms, which is a part of the ontology itself.

The Gene Ontology (GO) (Ashburner et al., 2000) employs the aforementioned concepts aiming to standardize the representation of genes, their products (*i.e.*, any molecule that is derived from a gene), and their interactions across different organisms. In the case of Gene Ontology, each one of its terms is associated with one or more genes. When such an association exists, we say that the GO term *annotates* the respective gene. Given the different natures of the terms that compose the Gene Ontology, these are actually subdivided into three main distinct ontologies, namely:

- **Biological Process (BP)**: Terms in this ontology are associated with different biological processes, such as cell death (`GO:0008219`), for example. The annotation of a gene by a term from the BP ontology implies its participation in that particular process.
- **Cellular Component (CC)**: The terms from the Cellular Component ontology are related to different locations within the cell, such as the nucleus (`GO:0005634`), for example. Annotation here implies that the given gene has activity in that region of the cell.

- **Molecular Function (MF)**: This ontology describes the molecular functions of genes and their products. These may be associated with the transport of different substances within the cell, such as miRNA transporter activity, specified by the term `GO:0061717`, for example.

At this point it is important to mention that the annotation of a given gene by a term can be supported by different types of sources, which are distinguished by the so-called evidence codes. These provide information about how the information regarding annotation was obtained, such as Traceable Author Statement (TAS), Inferred from Experiment (EXP), and Inferred by Curator (IC), to mention a few[2]. Although evidence codes do provide information regarding the origin of the annotation, in general they cannot be used to determine the overall quality of the annotation itself, since different methods employed for annotation within each evidence code can provide different levels of reliability (The Gene Ontology Consortium, 2015).

Regarding structure, each one of the three ontologies within GO are structured in the form of a Directed Acyclic Graph (DAG) (Chartrand, 1985). In these graphs, each one of the vertices corresponds to a particular term, whereas relations between terms are given by edges connecting vertices. Although other types of relations between terms do exist, the *is_a* relationship is among the most used ones. This kind of relation indicates that a specific term is a particularity of a more general one, as exemplified in Figure 3.5. In this example note that *neuron death* is a subtype of *cell death*, that is, *neuron death* is a specific case of *cell death*. Note that this relation is also transitive, that is, *neuron death* is likewise a subtype of *cellular process* and *death*.

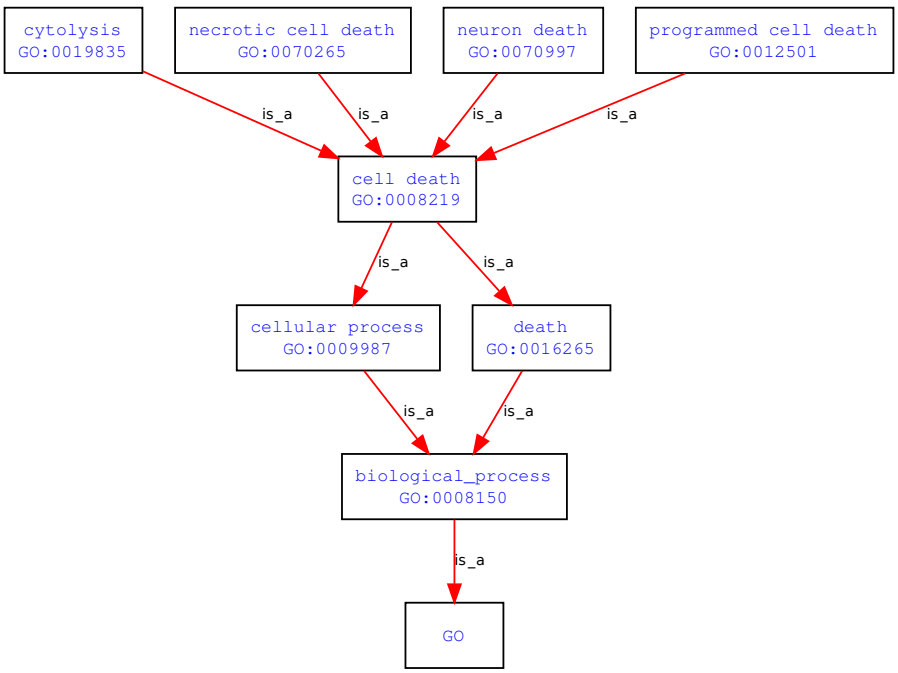


Figure 3.5: Example of relations between terms in the Gene Ontology. In this particular example we consider a part of the Biological Process (BP) ontology. Note that each term has its own unique identifier (which is given by `GO:XXXXXXX`) and an unique name, *e.g.*, *death*, for ease of use.

[2]Complete information regarding evidence codes can be obtained at `http://geneontology.org/`.

The transitivity observed in the Gene Ontology graph also has implications in the annotation process. To that end, when a gene is annotated by a term t it is automatically annotated by all other terms for which t is a subtype. Note that although the gene is annotated by a set of terms, these posses different levels of information, regarding its specificity. Indeed, annotating a gene with the term *biological process* is much more vague than annotating it with *cell death*, for example. Here it is important to note that the height of a given term in the GO graph[3] is not related to the granularity of information it provides (Pesquita et al., 2008, 2009), since two terms at a same height can provide different level of specificity. This has motivated the use and development of semantic similarities between genes (Pesquita et al., 2008, 2009), which try to characterize how similar they are given the set of terms that annotate them. Semantic similarities will be employed in Chapter 6 and Chapter 7 in the evaluation of distance measures and gene clustering solutions. There we provide further details on a particular semantic similarity, which is employed in our experiments.

3.5 Chapter Remarks

This chapter presented a brief review of gene expression data. We started by presenting a concise review of molecular biology concepts necessary to the understanding of gene expression in Section 3.1. Afterwards, in Section 3.2, we briefly discussed high throughput technologies, namely microarrays and RNA-Seq, which made the measurement of expression levels in a genomic scale a reality. In Section 3.3, the role of clustering in the analysis of gene expression data was highlighted. The distinction between the clustering of genes and samples was discussed. Finally, in Section 3.4, we briefly reviewed the Gene Ontology (GO), which will in later chapters provide external biological information to the selection of distances and validation of clustering results. These two specific aspects of gene expression data clustering will be further examined in Chapters 6 and 7, where related work pertinent to each topic will be further explored.

[3]The height of a term in the Gene Ontology is given by the number of edges between the term and the root of the ontology. If more than one path is possible between the root and the term, then the shortest one is considered as height.

CHAPTER

4

Ensembles for Relative Validity Criteria Evaluation

In Chapter 2 we provided an overview on data clustering and the procedures employed in the validation of its results. The evaluation and validation of clustering results has historically been described in the literature as one of the most difficult and frustrating steps of cluster analysis (Jain and Dubes, 1988; Milligan and Cooper, 1985). When it comes to validation, there is a plethora of criteria proposed in the literature (as we reviewed in Section 2.3.2) with new ones still being proposed from time to time — *e.g.*, the density-based relative validity criterion from Moulavi et al. (2014), that was developed with participation of the author and will be discussed in Chapter 5. It becomes clear from the variety of relative validity criteria described in the literature that no single measure can capture all different aspects of the clustering problem and, as such, each of them is prone to fail in particular application scenarios. In fact, due to the subjective nature of the problem, it is well-known that no criterion can systematically outperform all the others in all scenarios (Vendramin et al., 2010). The variety of relative validity criteria available poses, thus, a challenging decision to practitioners: how to effectively select an adequate relative validity criterion for a given application scenario?

In an attempt at shedding some light upon this question, different studies focused on the assessment of individual relative criteria performance considering particular classes of datasets (usually well-behaved and synthetically generated ones), as, for instance, in the studies from Milligan (1981), Milligan and Cooper (1985), and Vendramin et al. (2009, 2010). These studies

are, however, limited in the sense that: (i) their conclusions cannot be extrapolated to datasets following distributions other than that considered in those papers, unless similar analyses are also performed for other classes of datasets; (ii) performing such analyses for a particular class of datasets is laborious and highly time consuming; and (iii) in practice, one hardly knows in advance which class (*e.g.*, probability distribution function) the dataset in hand belongs to.

In order to overcome such limitations, in this chapter we explore alternatives to the validation of clustering results based on the selection of single relative validity criteria. In particular, we investigate the use of ensembles of relative validity criteria, which, by relying on multiple measures, attempt to obtain more robust evaluations. The rationale behind this approach follows essentially the same intuitive idea of combining multiple experts into a committee so as to get more stable and effective recommendations. In this scenario, when a collection of datasets is considered, the overall performance of a combined criterion is expected to outperform those of its constituent criteria.

Ideally, a good criteria ensemble (combination) should be composed of *complementary* measures. As an example, take the 12 partitions evaluated by the three different relative validity criteria illustrated in Figure 4.1. For the sake of simplicity, in this example each criterion provides a binary evaluation, regarding partitions as *good* (partitions inside the criterion's circle) or *poor* ones (partitions outside the criterion's circle). In this case, Criterion 3 presents the most correct evaluations, with 8 *good* partitions. This example illustrates a scenario where the available criteria are complementary; a mistake of one criterion is balanced by correct evaluations of the other ones. Note that, except for three partitions, when a criterion makes a mistake, the others provide a *good* evaluation. Thus, an ideal combination of such criteria would outperform a single criterion.

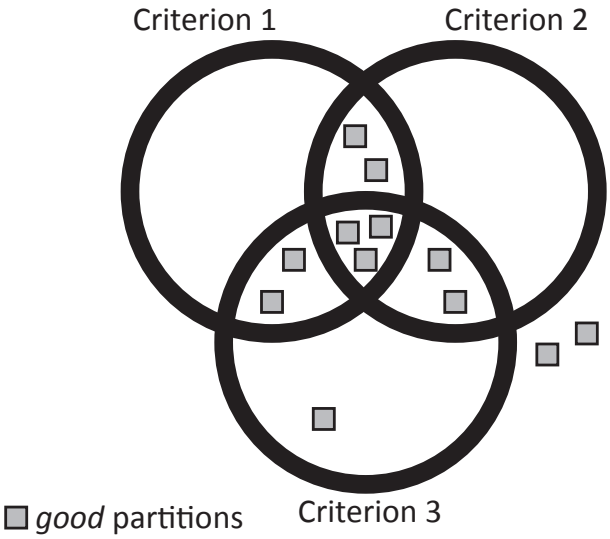


Figure 4.1: Three complementary relative validity criteria in a binary evaluation scenario (for illustration purposes). Partitions inside circles are deemed *good* by the corresponding criteria.

As discussed, *e.g.*, in Rokach (2010), the intuitive idea illustrated in Figure 4.1 finds theoretical support in the jury theorem, which states that a committee of independent voters, each of which

has a probability p_v of being correct when making a binary decision, makes such a decision conjointly with probability $p_c > p_v$ if $p_v > 0.5$. Independence of voters refers to the need of uncorrelated mistakes, which in turn is related to the notion of complementarity discussed above. Even when this condition cannot be formally ensured, the usefulness of committees (ensembles) when composed of elements that are diverse (complementary) enough has already been extensively demonstrated in different areas, such as classification (Rokach, 2010), clustering (Ghosh and Acharya, 2011), and outlier detection (Zimek et al., 2013). When it comes to clustering validation, however, we believe that this topic has not yet been given sufficient attention in the literature. Such lack of attention led to the discussions and developments presented in this chapter.

The remainder of this chapter is organized as follows. In Section 4.1 we review related work regarding ensembles of relative validity criteria. In Section 4.2 we present results based on the random generation of ensembles, *i.e.*, the arbitrary selection of ensemble members. Results from this particular section follow closely previously published work in which the author had participation, see: Vendramin, Jaskowiak and Campello (2013). In Section 4.3 we develop a heuristic for a principled selection of ensemble members and evaluate its results. Results from this section follow closely from published work in which the author of this thesis is the main contributor, see: Jaskowiak et al. (2015). As a final remark, we note that in all the experiments and discussions performed in this chapter the set of 28 relative validity criteria reviewed in Section 2.3.2 was considered as the initial pool of available measures for building ensembles, with a few exceptions that will be clearly identified through the text.

4.1 Related Work

Although ensemble methods have been previously employed in cluster analysis (Ghosh and Acharya, 2011), not much attention has been given to the use of ensembles regarding clustering validation. In this particular scenario, only a handful of research papers have discussed attempts to combine relative validity criteria into ensembles, as we briefly review in the following.

Bolshakova and Azuaje (2003) combined different versions of the Davies-Bouldin criterion (Davies and Bouldin, 1979). Measures were combined by a simple average of their evaluation scores. The authors showed that the resulting combinations were able to correctly determine the number of clusters in a single dataset in which most of the individual indexes had already made the right choice. Given that variants of the Davies-Bouldin criterion are all fairly similar, it is very likely that they will make highly redundant decisions. For this reason, there seems to be little potential for improvements in using this particular kind of ensemble in practice.

Sheng et al. (2005) built an ensemble of six different relative validity criteria to serve as fitness function for a genetic algorithm, which in turn was used to evolve clustering partitions. The values of these six measures were normalized and then combined by a weighted average (actually the authors used a simple average by assigning all the measures the same weight). The authors

showed that the ensemble of measures was able to guide the genetic algorithm to evolve partitions and correctly identify the number of clusters in the datasets used in the experiments. It is worth remarking, however, that their selection of relative validity criteria to compose the ensemble was completely arbitrary. As the authors point out, they believe that the ensemble performance could be improved by "using other more effective combinations of validity functions" (Sheng et al., 2005).

Machado et al. (2007) employed a combination of fuzzy relative validity criteria (which are beyond the scope of this work) to determine the number of clusters for a fuzzy clustering algorithm. The criteria were combined by simply averaging their scores. Even though the combination showed satisfactory results, once again it was purely based on an arbitrary selection of relative validity criteria, which, as the authors acknowledge, was made for illustrative purposes only.

Albalate and Suendermann (2009) introduced an approach to identify the correct number of clusters in a dataset. Their approach is based on the use of a collection of partitions generated from different clustering algorithms and distance measures, each of which is scored by distinct relative validity criteria. The scores of each criterion over all partitions are then assessed for robustness based on statistical reasoning, from which an estimate for the number of clusters is derived. Although based on the combination of different relative validity criteria, the method can only be used to estimate the number of clusters, that is, it cannot be directly used to rank partitions according to their quality or select the best partition from a pool.

Finally, Pihur et al. (2007) proposed a method that aims to combine rankings from different sources, using Monte Carlo simulations and the concept of cross entropy. Although it is in some sense generic and can, in principle, be applied for the combination of rankings from the most diverse sources, it was evaluated considering the clustering validation scenario, for which ten different relative validity criteria were combined. The authors claim that their method can correctly identify the best clustering algorithm as the one that provides the best partition according to the combination of different relative validity criteria. It is important to note, however, that their claim was supported solely on visual inspection from a limited number of datasets and results.

Roughly speaking, in all the aforementioned works in which ensembles were considered, the selection of relative validity criteria that ultimately composed the ensembles was arbitrary. Although in some cases the practitioners experience may have been directly employed in the selection of such criteria, in others the selection resembles just a random process, in which an arbitrary number of, again, randomly selected criteria was employed. In the sequence we evaluate how much of a benefit such a random selection of relative criteria can provide in practice.

4.2 Random Selection of Ensemble Members

Given that no systematic procedure was adopted during the selection of ensemble members in previous works, such selections can be seen as coming from different random procedures. In this section we evaluate how ensembles built on the basis of randomly selected relative validity

criteria perform in practice. We want, therefore, to verify whether or not combining multiple criteria in an arbitrary fashion can be worth it in practical applications, where the best or worst criterion from the pool of available measures is usually unknown. To do so, we evaluate a number of relative criteria combined into committees by means of different combination strategies. Combinations are evaluated against the individual performances of their constituent criteria, in order to verify: (i) if a combination outperforms all its composing criteria; and (ii) if a combination outperforms at least one of its composing criteria. The first scenario is intended to the assessment of eventual improvements on the overall effectiveness of the validation procedure when using a combination (ensemble) rather than individual indexes. The second one is intended to the assessment of possible improvements w.r.t. the worst case, which is particularly relevant in practice as the worst performing criterion is usually unknown and cannot be avoided.

4.2.1 Experimental Setup

Bellow we present the datasets and the evaluation strategy employed during the experiments.

4.2.1.1 Datasets

Both synthetic and real datasets were employed in our analysis. The synthetic datasets are reproductions of those described in Vendramin et al. (2010), following closely the artificial data generator used in the classic studies by Milligan (1981); Milligan and Cooper (1985). In brief, the datasets consist of a total of $n = 500$ objects each, embedded in an $m = 2$, 3, 4, 22, 23, or 24 dimensional Euclidean space. Each dataset contains $k^* = 2$, 4, 6, 12, 14, or 16 clusters whose objects follow (mildly truncated) multivariate normal distributions for which overlap of boundaries is permitted in all but the first dimension. Datasets with three different levels of balance in terms of cluster sizes were generated. The first level refers to approximately balanced clusters. In the second level, one of the clusters contains 10% of the objects and the remaining objects are approximately balanced among the other clusters. In the third level, one of the clusters contains 20% or 60% of the objects, depending on whether $k^* \in \{12, 14, 16\}$ or $k^* \in \{2, 4, 6\}$, respectively. Finally, the three design factors described above were combined to produce a set of 108 (6 no. of clusters $\times$ 6 no. of dimensions $\times$ 3 balances) design cells. For better confidence of the analysis, nine replications were produced per design cell, thus resulting in a total of 972 datasets.

Regarding real data, we considered the Amsterdam Library of Object Images (ALOI) (Geusebroek et al., 2005). ALOI is composed of 1000 categories, each of which contains several pictures of a given object. Although each category has only pictures of a single physical object (*e.g.*, rubber duck, shoe, ball, etc.), the pictures are distinct due to systematic variations of viewing and illumination angle, and illumination color. Moreover, categories have particular characteristics of color, shape, and texture, in such a way that a reasonable description of the images in an appropriate feature space should give rise to data having cluster tendency. Here, we processed the ALOI database in a way similar to (Horta and Campello, 2012).

Specifically, image sets were created by randomly selecting k^* ALOI image categories as class labels 100 times for each number $k^* = 2, 3, 4, 5$, then sampling (without replacement), each time, 25 images from each of the k^* selected categories, thus resulting in 400 sets, each of which contains 2, 3, 4, or 5 clusters and 50, 75, 100, or 125 images (objects), respectively. The images were first represented using seven different descriptors: color moments (144 attributes), texture statistics from the gray-level co-occurrence matrix (88 attributes), Sobel edge histogram (128 attributes), 1st order statistics from the gray-level histogram (5 attributes), gray-level run-length matrix features (44 attributes), gray-level histogram (256 attributes), and single-level discrete 2-D wavelet transform (4 attributes). Then, the final (7-dimensional) representation of the images was obtained by combining the first principal component extracted from each of the 7 descriptors using Principal Component Analysis (PCA). The result is a collection with 400 datasets described in a 7-dimensional space.

4.2.1.2 Clustering Algorithms

Clustering algorithms were systematically applied to the datasets to produce partitions of varied qualities to be evaluated. Such evaluations should be performed over a set of partitions with the number of clusters k limited within an acceptable range of values. Such an acceptable range depends on the application domain, but there are some general rules that can help the less experienced user in practical scenarios. One rule of thumb is to set the upper bound of k to $k_{max} = \lceil \sqrt{n} \rceil$, where n is the number of objects. We adopt this rule in our experiments.

The clustering algorithms used in our experiments are the well-known k-means (MacQueen, 1967) and four variants of Hierarchical Clustering Algorithms (HCAs) (Jain and Dubes, 1988), namely, Single-Linkage, Average-Linkage, Complete-Linkage, and Ward's. For the ALOI datasets we employed these four variants of HCAs in order to generate partitions. Following the above rule of thumb for k_{max}, partitions produced by the HCAs with k between 2 and $k_{max} = 8$, 9, 10, and 12 clusters were considered for the datasets with $n = 50$, 75, 100, and 125 objects (images), respectively. As there are 4 algorithms and 100 datasets per value of n above, we obtained a total of $n_\pi = (7+8+9+11) \times 100 \times 4 = 14,000$ partitions for the ALOI datasets. For the synthetic datasets, k-means was used in lieu of HCAs, primarily due to computational reasons (given the large number of datasets and the quadratic time of HCAs), but also to show that the main conclusions that can be drawn from our experiments are not affected when a different algorithm is considered. k-means was run 20 times with random initial prototypes for each value of $k \in \{2, \cdots, k_{max}\}$, with $k_{max} = \lceil \sqrt{500} \rceil = 23$. The algorithm then produced $n_\pi = (23 - 1) \times 20 = 440$ partitions for each dataset. Since there are 972 datasets, a total of $972 \times 440 = 427,680$ partitions was obtained.

4.2.1.3 Combination Strategies

In order to combine the results obtained with each relative validity criterion we employed four different score-based combination strategies, as follows:

- Mean: Arithmetic mean of evaluation outcomes;

- Harmonic: Harmonic mean of evaluation outcomes. In this approach, partitions with a low score by at least one criterion are more strongly penalized;

- Mean-2: This strategy involves the removal of the most discrepant evaluation (*i.e.*, the value whose sum of absolute differences from the other values is the largest) and then applying the arithmetic mean;

- Median: The median of the evaluation scores.

Note that the range of values for each relative validity criterion subject to combination may be different, therefore, their values must be somehow normalized before applying any of the score-based combination strategies. For such reason we scaled the values of each relative criterion between 0 and 1 (Min-Max Normalization). The investigation of different normalization procedures is beyond the scope of this work (in the next section we examine rank-based combination strategies, which do not require the previous application of normalization procedures). Finally, minimization criteria, such as Davies-Bouldin and C-Index, were converted into maximization criteria by flipping their values around their mean.

4.2.1.4 Evaluation

For this particular evaluation, the 28 relative validity criteria reviewed in Section 2.3.2 were combined into groups of three ($n_c = 3$) and five ($n_c = 5$) using the four different score-based combination strategies. All possible combinations into groups of n_c criteria were evaluated. For instance, in the case of experiments involving combinations of $n_c = 3$ out of 28 candidate criteria, $\binom{28}{3}$ = 3276 combinations were evaluated. To analyze the effectiveness of combining different criteria, we counted the number of combinations that outperformed: (i) all criteria involved in the combination individually; and (ii) at least one of the criteria involved in the combination. To identify improvements from a combination over each individual criterion, both evaluation methodologies described in Section 2.4 were used. For the traditional methodology, an improvement refers to a larger number of datasets for which the combination suggests a partition with the *right* number of clusters (k^*) as the best one. For the alternative methodology, an improvement refers to more consistent correlations w.r.t. the evaluations provided by an external criterion, which can be characterized by a larger mean value of correlation across all datasets and, possibly, a smaller variance as well. In this case, it is important to consider not only the number of combinations that provide an improvement w.r.t. the mean/variance, but also the significance of such an improvement. To do so, we employed the Friedman (Friedman, 1940) and the Brown-Forsythe (Brown and Forsythe, 1974) tests to identify statistically significant differences among means and variances, respectively. Only statistically significant better results (at a 95% confidence level) were counted and reported as improvements.

As a final remark, we note that two external validity criteria and two correlation coefficients were employed in the evaluations using the alternative methodology. Regarding external validity criteria, we employed both the well-known Adjusted Rand Index (ARI) (Jain and Dubes, 1988) and the Jaccard external index (Jain and Dubes, 1988). With respect to correlation coefficients, we employed the well-known Pearson correlation coefficient (Pearson, 1895) and the Weighted Goodman-Kruskal correlation coefficient (Campello and Hruschka, 2009). For the sake of compactness we present here results only for the ARI and Pearson correlation coefficient, given that these are more commonly used in the literature. It is important to note, however, that the conclusions that can be drawn from the different measures do not change the overall picture.

4.2.2 Results and Discussion

We begin our discussion of the results regarding randomly generated ensembles considering the collection of 972 synthetic datasets. For these datasets we consider the results that refer to improvements of each combination over *all* its composing criteria. Table 4.1 shows the number of combinations that resulted in improvements according to the traditional and alternative methodologies, respectively. Note that this number is under 11% according to the traditional methodology. When the alternative methodology is considered, the numbers are even lower, namely, under 2% w.r.t. the correlation mean, under 8% w.r.t. variance, and under 1.5% w.r.t. both. In this case the highest values are obtained with the *harmonic mean* strategy.

Table 4.1: Improvements over all the individual criteria involved in the combination, according to the traditional and the alternative methodologies, for $n_c = 3$. Total of $3,276$ combinations.

Combination Strategy	# Improvements (Percentage)			
	Traditional Methodology	Alternative Methodology		
		Mean	Variance	Both
Mean	315 (9.62)	22 (0.67)	10 (0.30)	4 (0.12)
Harmonic	338 (10.32)	52 (1.58)	239 (7.29)	43 (1.31)
Mean-2	163 (4.98)	3 (0.09)	4 (0.12)	0 (0)
Median	174 (5.31)	21 (0.64)	6 (0.18)	5 (0.15)

The results in Table 4.1 suggest that a combination is unlikely to be more accurate than the most accurate of its elements (the best performing criterion). In other words, the results suggest that using the best composing criterion alone is more likely to be a better choice if such a criterion is known in advance. However, it is important to note that the user can hardly know which criterion is the best for the dataset in hand. Hence, it is also insightful to check whether each combination exhibits better performance when compared with its worst composing criterion, as the user generally cannot discard or avoid the use of such a criterion in practice. For this reason, we also analyzed the number of combinations that provided better performances w.r.t. at least one out of the n_c criteria involved in each combination. The results with $n_c = 3$ are depicted in Table 4.2 for the traditional and alternative evaluation methodologies, respectively. The results for $n_c = 5$ follow the same trend and are not reported here for the sake of simplicity.

Table 4.2: Improvements over at least one of the criteria involved in the combination, according to the traditional and the alternative methodologies, for $n_c = 3$. Total of $3,276$ combinations.

Combination Strategy	# Improvements (Percentage)			
	Traditional Methodology	Alternative Methodology		
		Mean	Variance	Both
Mean	3274 (99.94)	3248 (99.14)	1777 (54.24)	1777 (54.24)
Harmonic	3274 (99.94)	3100 (94.62)	2676 (81.68)	2587 (78.96)
Mean-2	3264 (99.63)	2946 (89.92)	1685 (51.43)	1536 (46.88)
Median	3264 (99.63)	3108 (94.87)	1475 (45.02)	1454 (44.38)

As shown in Table 4.2, the number of combinations outperforming its worst constituent criterion is high according to both evaluation methodologies. In fact, all combination strategies provided improvements in more than 99% of the combinations when considering the traditional methodology. Significant improvements can also be observed for the alternative methodology, for which the number of combinations outperforming its worst constituent criterion is above 89%, 45%, and 44% w.r.t. the mean, variance, and both, respectively. When considering solely the *harmonic mean* combination strategy, these results increase to 94%, 81%, and 78%, respectively. These results support that combining multiple relative validity criteria can be beneficial as a conservative choice for real-world application scenarios where it is almost impossible for the user to determine which one is the best criterion to use or the worst criterion to avoid.

Next, we evaluate results regarding the collection of 400 ALOI datasets. In order to reduce the computational burden associated with our experiments, for the ALOI datasets only the original version of Dunn's index was used, i.e., we considered a set of 11 relative criteria (excluding the 17 Dunn's variants). In Tables 4.3 and 4.4 we depict results w.r.t. combinations of $n_c = 3$ relative criteria. In total, we have $\binom{11}{3} = 165$ different combinations. From Table 4.3, we note that, when considering solely the traditional methodology, there is a more significant improvement over all individual criteria (in comparison to the values observed for synthetic data). For the alternative methodology, however, the number of combinations that improve on the results of *all* its composing criteria is very small (it is actually null if both mean and variance are considered simultaneously). This can be explained due to the fact that the traditional methodology accounts only for the number of clusters found by the clustering algorithm, whereas the alternative methodology is more strict, looking for the quality of all partitions produced by the clustering algorithm. Even in this scenario, the improvements observed with the traditional methodology are low, *i.e.*, one can expect improvements over all criteria that compose the ensemble in one out of four combinations evaluated. Contrastingly, from Table 4.4 we note that the number of combinations that improve on the results of the *worst* composing criterion is very high for both evaluation methodologies (and slightly higher when using the harmonic mean). Both results presented in these tables are quite in accordance to those observed from the previous experiments with synthetic datasets.

The results involving combinations of $n_c = 5$ criteria are shown in Tables 4.5 and 4.6, for improvements over the best and worst criterion from each combination, respectively. In total,

Table 4.3: Improvements over all the three criteria involved in the combination, according to the traditional and the alternative methodologies, for $n_c = 3$. Total of 165 combinations.

Combination Strategy	# Improvements (Percentage)			
	Traditional Methodology	Alternative Methodology		
		Mean	Variance	Both
Mean	46 (27.87)	1 (0.60)	3 (1.81)	0 (0.00)
Harmonic	44 (26.66)	2 (1.21)	26 (15.75)	0 (0.00)
Mean-2	33 (20.00)	0 (0.00)	0 (0.00)	0 (0.00)
Median	34 (20.60)	1 (0.60)	0 (0.00)	0 (0.00)

Table 4.4: Improvements over at least one of the criteria involved in the combination, according to the traditional and the alternative methodologies, for $n_c = 3$. Total of 165 combinations.

Combination Strategy	# Improvements (Percentage)			
	Traditional Methodology	Alternative Methodology		
		Mean	Variance	Both
Mean	165 (100)	165 (100.00)	146 (88.48)	146 (88.48)
Harmonic	165 (100)	164 (99.39)	161 (97.57)	160 (96.96)
Mean-2	165 (100)	163 (98.78)	130 (78.78)	129 (78.18)
Median	165 (100)	165 (100.00)	130 (78.78)	130 (78.78)

we have $\binom{11}{5} = 462$ different combinations. When comparing the results of Tables 4.5 and 4.6, one can draw essentially the same conclusions as before, *i.e.*, it is very unlikely (if possible) that a combination improves on the results of *all* its composing criteria, but it is very likely that it improves on the results of the *worst* one. Furthermore, when comparing the results obtained with $n_c = 5$ (Tables 4.5 and 4.6) against those obtained with $n_c = 3$ (Tables 4.3 and 4.4), one can see that this behavior is even more noticeable for $n_c = 5$ than for $n_c = 3$.

Table 4.5: Improvements over all the five criteria involved in the combination, according to the traditional and the alternative methodologies, for $n_c = 5$. Total of 462 combinations.

Combination Strategy	# Improvements (Percentage)			
	Traditional Methodology	Alternative Methodology		
		Mean	Variance	Both
Mean	111 (24.02)	0 (0)	0 (0)	0 (0)
Harmonic	73 (15.80)	0 (0)	75 (16.23)	0 (0)
Mean-2	102 (22.07)	0 (0)	0 (0)	0 (0)
Median	102 (22.07)	0 (0)	0 (0)	0 (0)

Table 4.6: Improvements over at least one of the criteria involved in the combination, according to the traditional and the alternative methodologies, for $n_c = 5$. Total of 462 combinations.

Combination Strategy	# Improvements (Percentage)			
	Traditional Methodology	Alternative Methodology		
		Mean	Variance	Both
Mean	462 (100)	462 (100.00)	456 (98.70)	456 (98.70)
Harmonic	462 (100)	462 (100.00)	462 (100.00)	462 (100.00)
Mean-2	462 (100)	462 (100.00)	435 (94.15)	435 (94.15)
Median	462 (100)	462 (100.00)	435 (94.15)	435 (94.15)

If a minimum level of complementarity and effectiveness of the composing criteria could be guaranteed when building a combination, it follows that, from the principles of ensemble techniques (where those are basic requirements), one should expect to observe an improvement on the performance of a combination by increasing the number of combined elements. However, when considering the whole universe of possible combinations, built in a completely non-informed (*i.e.*, "blind") manner, our results show that the relative proportion of successful combinations — in terms of improvements on the performance of all their composing criteria — reduces as the number of possible combinations increases (by increasing n_c). The proportion of successful combinations increases only in terms of improvements on the performance of the worst composing criterion. This suggests that, by increasing n_c, it is more likely that a randomly chosen combination will be composed of one or more criteria having superior performance (in relative terms) and others having inferior, possibly non-complementary behavior. These last ones will, with high probability, cause the performance of the combination to deteriorate w.r.t. the best composing criterion, but not enough to be worse than that of the worst criterion that composes the ensemble.

The results obtained in our experiments subsume practical application scenarios where the user or analyst has no clue how to choose validation indexes (to be used individually or combined). The poor results obtained in terms of improvements over the best composing criterion of a combination calls attention to an important issue: the lack of a theoretical and algorithmic apparatus to support a more systematic, non-blind construction of combinations of validity criteria that, if it is not guaranteed to improve on the use of each single criterion individually, at least could do so with high probability. Such an apparatus is already quite well developed in the realm of ensemble techniques for classification, clustering, and outlier detection. Particularly, there are two important questions that are well studied in these areas and that are still open in what concerns the combination of validity criteria. The first one is how to efficiently combine a selected set of criteria. The second question is how to guarantee diversity among the selected set of criteria to be combined, which can be translated into how to select criteria to be combined that are *complementary* in terms of their evaluations. Diversity and complementarity are key related features for a successful combination or ensemble. These features seem to be particularly critical in the context of validity criteria, as most of the existing criteria are based on common conceptions about cluster quality, such as compaction and separation. This may partially explain our results in what concerns the overall inability of the combinations in outperforming their best constituent criteria.

In the next section such questions are further explored, as we develop a heuristic for a systematic, non-blind, construction of relative validity criteria ensembles. This heuristic is based on the selection of both *accurate* and *complementary* ensemble members. Alongside with it, we also explore different combination strategies that can overcome the need of normalization procedures.

4.3 Heuristic Selection of Ensemble Members

The idea behind ensembles is to strategically combine two or more models to solve a particular predictive or descriptive task (Polikar, 2012; Rokach, 2010). Ensemble models were originally developed to improve the performance of a decision making system by decreasing its variance and the chances of selecting a poor single model. The intuitive justification for ensembles is that no single model can always be superior to others in all possible scenarios. Therefore, aggregating outcomes from different single models has the potential to enhance the overall performance of the resulting combined system when compared to its individual constituents. But from the theory of ensembles it is well-known that two basic ingredients are required so that such a potential can be turned into reality: an ensemble learner tends to make more robust and accurate recommendations than its individual base learners if the base learners are fairly *accurate* and *diverse* (Kuncheva and Whitaker, 2003). From this perspective, it is clear that ad-hoc approaches that combine arbitrarily selected validity criteria into ensembles, like those that have been discussed in the related work, have limited chances of success, unless they meet the requirements by chance. Indeed, this was observed experimentally in the previous section. Even though, for ensembles of clustering validity criteria, we do not train and test any models (there is no learning involved in the application of a cluster validation index to a dataset), an analogous reasoning applies when combining cluster validity criteria into ensembles: an ensemble validation criteria that does not have diverse and minimally effective members (*e.g.*, when members are randomly selected, as discussed in the related work), typically offers only limited performance gains over the individual criteria.

In the following, we propose a more principled approach for selecting relative criteria that meets requirements corresponding to minimum effectiveness and diversity for ensembles of learners, in a justified way. We start with the initial set of 28 different cluster validation criteria discussed in Section 2.3.2, that have all been proposed for the evaluation of the same type of clustering model, in which, intuitively, a good clustering is characterized by a higher "within-cluster-similarity" than "between-cluster-similarity". While the considered criteria differ in the details on how they formalize the measure of quality for such a clustering, some criteria differ more from each other than others. In a few cases one can already analytically determine that two criteria are just minor variations of each other and that they can be expected to perform pretty much the same most of the time, and differ only slightly under very specific conditions. When selecting a subset of validation criteria, such dependencies between them can lead to a bias in the final vote, when combining them into an ensemble. This is the equivalent of a lack of diversity in ensemble learners. We propose to avoid this type of bias by selecting methods that perform more independent (or complementary) from each other. Unfortunately, determining the complementarity of the methods analytically is far from obvious, which is why we opt for an "empirical" approach. We use a large number of different datasets (972 in total), varying the dimensionality, the number of clusters, and the characteristics of the clusters, with the intention to cover a wide spectrum of possible "ground truth" scenarios. An even larger number of partitions is then produced by

running diverse clustering algorithms with different parameter settings on these datasets, resulting in a very large variety of partitions suitable to detect correlations between methods, and determine expected minimal effectiveness of each criterion. An external validation index is first applied to each partition to measure the agreement of each clustering result with the ground truth. Then all the cluster validation criteria are applied to the clustering partitions, and the resulting scores for all methods on all partitions are analyzed to determine (i) the correlations between pairs of criteria, to ensure a selection of complementary criteria (corresponding to the diversity requirement for ensemble learners), and (ii) correlation between each criterion and the external validation index, to ensure a selected method has a certain minimum overall effectiveness (corresponding to the minimum effectiveness requirement for ensemble learners).

4.3.1 Combination Strategies

Different combination strategies can be employed to ensemble the results of two or more relative validity criteria. In brief, these strategies can be divided in two main subclasses, namely, score-based and rank-based strategies. Score-based strategies were already employed in Section 4.2 and evaluated alongside with blindly generated ensembles of relative validity criteria. Here we also consider rank-based combination strategies, given that we are not particularly concerned about the absolute scores of each relative criterion or about the absolute score obtained from a particular combination strategy, especially because these scores are usually not commensurable across different criteria. To combine criteria into ensembles we need commensurable values. By adopting rank-based aggregation techniques we get around score normalization problems that are difficult to solve and demand that scores follow particular probability distributions. This is a problem that has been studied previously, *e.g.*, for outlier scores (Kriegel et al., 2011). An important difference to the rank aggregation problem for outlier rankings is that only the top ranks are important for outlier detection; in our experimental evaluation of ensembles, we are interested in the overall correlation of the aggregated ranking with an external measure. For these reasons, we will base our analyses on the relative rankings of partitions according to different validity criteria, focusing on typical practical applications in which one wants to sort a collection of candidate clustering results according to their quality. For this, we assume the following convention: given a *ranked list* τ of n_π items (partitions in our case), the *rank* of an item $x = \tau(i)$ is the *index*, *i.e.*, position i of x in τ (from 1 for the highest ranked item to n for the lowest ranked item).

Before introducing the rank combination strategies we consider for our posterior evaluation, it is worth noticing that, although the combination strategy proposed by Pihur et al. (2007) was preliminary considered in our study, it was not computationally feasible to include it in our final evaluation. This method is based on Monte Carlo simulations and it did not scale well in face of the large number of datasets and partitions used in our experiments (this conclusion is based on experiments with the R implementation provided by the authors (Pihur et al., 2007)). Another

method, designed for the combination of outlier rankings (Lazarevic and Kumar, 2005), was not included since it is sensitive to the order in which the ranked lists are evaluated by the combination strategy. From an ensemble point of view, this method is not interesting since it does not allow the majority of methods to overrule wrong decisions by individual voters (Zimek et al., 2013, 2014).

4.3.1.1 Borda Count Method

The Borda Count method (de Borda, 1781) is a classic voting system that was developed in the late eighteenth century and can be employed to combine rankings from different sources. Given a particular ranking as a sorted list of elements, the method works by assigning a score to each member of the list (partitions evaluated by a relative validity criterion in our case) according to its relative position. Once the method is applied to different rankings, the final aggregated ranking is a sorted list based on the sum of scores of each element. It is easy to verify that this method is equivalent to computing the mean of the ranks, *i.e.*, combining ranks by their mean, and can be easily adapted to cope with both the minimization and the maximization of relative criteria.

4.3.1.2 Median

Another simple yet commonly used approach to combine different rankings is the median (Vendramin et al., 2013). The idea is similar to the Borda Count method, but we take the median of each element's ranks across different sorted lists rather than the sum. Along with the Borda Count method we use the median as a classic representative of rank combination method.

4.3.1.3 Footrule

The Scaled Footrule aggregation, or simply Footrule (Dwork et al., 2001), is an attempt to approximate an optimal aggregation. The optimization aims to find a permutation π according to Equation (4.1), where $\mathcal{R}$ is a collection of ranked lists of partitions given by different relative validity criteria in our case.

$$\text{Footrule}(\mathcal{R}) = \arg\min_{\pi} \left(\sum_{\tau \in \mathcal{R}} d(\tau, \pi) \right) \tag{4.1}$$

In Equation (4.1), $d(\tau_1, \tau_2)$ is a distance between two rankings τ_1 and τ_2 of a set of elements S. A common choice for this distance is the Spearman Footrule distance, which measures the total displacements computed over all the elements ranked in τ_1 and τ_2. Specifically, it is the sum of the individual displacement for each element, which in turn is given by the absolute difference between the ranks of this element in the two lists (τ_1 and τ_2). Formally, given two rankings τ_1 and τ_2, the distance is given by:

$$d(\tau_1, \tau_2) = \sum_{i=1}^{|\tau|} |\tau_1(i) - \tau_2(i)|$$

and can be normalized by its highest possible value, $|\tau|^2/2$, where $|\tau|$ is the size of the lists, *i.e.*, the number of ranked items.

The problem posed by Equation (4.1) can be solved in polynomial time when reduced to the computation of the minimum cost of matching a weighted bipartite graph (Dwork et al., 2001). Let us define a weighted bipartite graph $G = (S, P, W)$, for which S is the collection of items (partitions) being ranked, P is the set of $|\tau|$ positions available to place an item in the final ranking, and W contains the set of weights

$$W(\tau_j(i), p) = \sum_{\tau_j \in \mathcal{R}} |\tau_j(i) - p|,$$

which quantify how far each candidate item i (partition) is from a position p (rank). The Footrule aggregated rank over all lists is then given by a minimum cost perfect matching in the bipartite graph G (Dwork et al., 2001).

4.3.1.4 Condorcet

Rank aggregation methods referred to as Condorcet methods are those that satisfy the so-called Condorcet criterion (Marquis de Condorcet, 1785), which states that if a given alternative x (in our case a partition) is ranked higher in the *majority* of all the rankings, in all possible pairwise comparisons with the other alternatives, then x (if such an item exists) should be ranked first. A straightforward implementation that satisfies the Condorcet criterion is as follows. For a given ranking of n_π elements, consider all possible pairs and assign a "win" to the element that is ranked higher than the other. For each ranked list this can be represented in the form of a matrix, where a value of one indicates that the element in the row "won" over the column element, for that ranked list. Given that this procedure is applied to all ranked lists, at the end a results matrix which adds the results for all lists can be obtained. Based on this final matrix a series of pairwise contests are performed. The result of these contests will provide the final ranking, that is, elements are ranked by the total number of pairwise contests they win. Given that there is no standard procedure to handle ties we solved them randomly.

4.3.1.5 Reciprocal Rank Fusion

The Reciprocal Rank Fusion method (RRF) (Cormack et al., 2009) uses a simple approach to combine rankings from different sources. Given a set $\mathcal{S}$ of elements to be ranked and a collection of ranked lists of these elements, $\mathcal{R}$, the final score of an element $s \in \mathcal{S}$ is given by Equation (4.2):

$$\text{RRF Score}(s) = \sum_{\tau \in \mathcal{R}} \frac{1}{\epsilon + \tau(s)} \tag{4.2}$$

where $\tau(s)$ stands for the rank of element s in ranked list τ, and ϵ is a real-valued constant. The authors suggest a value of $\epsilon = 60$, which we also adopt in our experiments.

The intuition behind Equation (4.2) is that while higher ranks are more important for the overall ranking of elements, the importance of lower ranks in some rankings does not vanish as it would if an exponential function was used. The constant ϵ mitigates the impact of outliers.

4.3.1.6 ULARA

The ranking algorithm proposed by Klementiev et al. (2007) follows the principle that the relative contribution of an individual ordering to the combined ranking should be determined by its tendency to agree with other members of the pool. Based on this idea, the authors introduce an index that measures the inconsistency of a particular criterion when compared to the rest of the pool. The inconsistency for a given criterion c_i, given its ranked list of results τ_{c_i}, is given by Equation (4.3):

$$\text{Inconsistency}(\tau_{c_i}) = \sum_{j=1}^{|\tau|} (\tau_{c_i}(j) - \mu(j))^2 \tag{4.3}$$

where μ is the average ranked list based on the pool and $|\tau|$ denotes the number of ranked elements (size of the list). Based on the inconsistency measure, the authors derive a weight for each ranked list, as defined in Equation (4.4):

$$W(\tau_{c_i}) = \frac{\text{Inconsistency}(\tau_{c_i})}{\sum_{j=1}^{|\tau|} \text{Inconsistency}(\tau_{c_j})} \tag{4.4}$$

This value is used to produce the final ranking combination by performing an weighted average of the ranks across the different lists.

4.3.1.7 Markov Chain

Dwork et al. (2001) proposed four different methods based on the concepts of Markov chains. They conclude that the method referred to as MC4 outperformed the others. Based on these results, we employ MC4 in our evaluation, setting aside the remaining three methods. For MC4 the states of the Markov chain correspond to the elements to be ranked (in our case each state corresponds to a partition). The transitions between any two pair of states are defined as follows. In a given state P of the chain, pick a state Q uniformly from all states available. Chose Q as the next state if $\tau(Q) < \tau(P)$ for the majority of the ranked lists, that is, if for the majority of the ranked lists the partition from state Q has a better rank than the partition from state P. Otherwise, stay in state P. In our study, we employed the MC4 implementation from Schalekamp and van Zuylen (2009).

4.3.1.8 Robust Rank Aggregation

The method called Robust Rank Aggregation (RRA) (Kolde et al., 2012) was introduced to deal with the combination of ranked lists containing noise and outliers. The main idea behind RRA is to identify elements (partitions in our case) that are ranked consistently better than expected

under the null hypothesis of uncorrelated input ranked lists and assign a significance score for each element. Elements are then ranked by their significance scores. For a detailed description on how significance scores are derived, we refer to the original publication (Kolde et al., 2012). In our experiments we employed the implementation provided by the authors under the R Package `RobustRankAggreg` (Kolde et al., 2012).

4.3.2 Selecting Relative Criteria

We employ the previously described collection of 972 synthetically generated datasets (see Section 4.2.1.1) to select a subset of relative validity criteria as future ensemble members. Ensembles built from the selected criteria will be evaluated subsequently using different datasets. Partitions were generated with the very same clustering algorithms, in the same range discussed previously, *i.e.*, $2 \leq k \leq \lceil \sqrt{n} \rceil$. Regarding the evaluation of single relative validity criteria and its resulting ensembles we employ from hereafter only the alternative methodology, since we believe it is more appropriate than the one referred to as traditional. Moreover, our experiments regarding the random generation of ensembles showed a good agreement between the evaluations provided by the two methodologies. Note that the methodology can be used both to evaluate the *effectiveness* of a relative validity criterion, w.r.t. an external criterion, and its *complementarity*, w.r.t. another relative validity criterion. In this sense, a high/low degree of correlation between a relative and an external evaluation can be seen as a high/low degree of effectiveness. Similarly, a high/low degree of correlation between two relative validity measures indicates a high/low redundancy between them. Based on these concepts of effectiveness and redundancy, in the following we develop a systematic procedure for selecting highly effective and complementary ensemble members.

4.3.2.1 Criteria Effectiveness

Given our dataset collection and our measure of effectiveness (the alternative methodology for relative criteria evaluation), we evaluate the effectiveness of each relative validity criterion from our initial set of 28 candidates (see Section 2.3.2). The results are depicted in Figure 4.2, which shows the average Spearman correlation of the relative validity criteria with the external index over all the 972 training datasets. One can see that a relatively large number of measures shows poor correlation with the external index (bellow 0.2). A variant of the Dunn criterion (Dunn 41) even shows a slightly negative correlation (effectiveness) w.r.t. the external index. These results show that such measures cannot properly evaluate the relative qualities of partitions as determined by their closeness to the ground truth. Note that all Dunn's variants (including its original version) display weak correlations with the external index. This is also the case for C/Sqrt(k) and, less prominently, for DB. The remaining measures show a correlation value (effectiveness) of at least 0.5, which can be considered moderate. Note that some of these measures display a correlation greater than 0.7, which is usually regarded as a strong correlation.

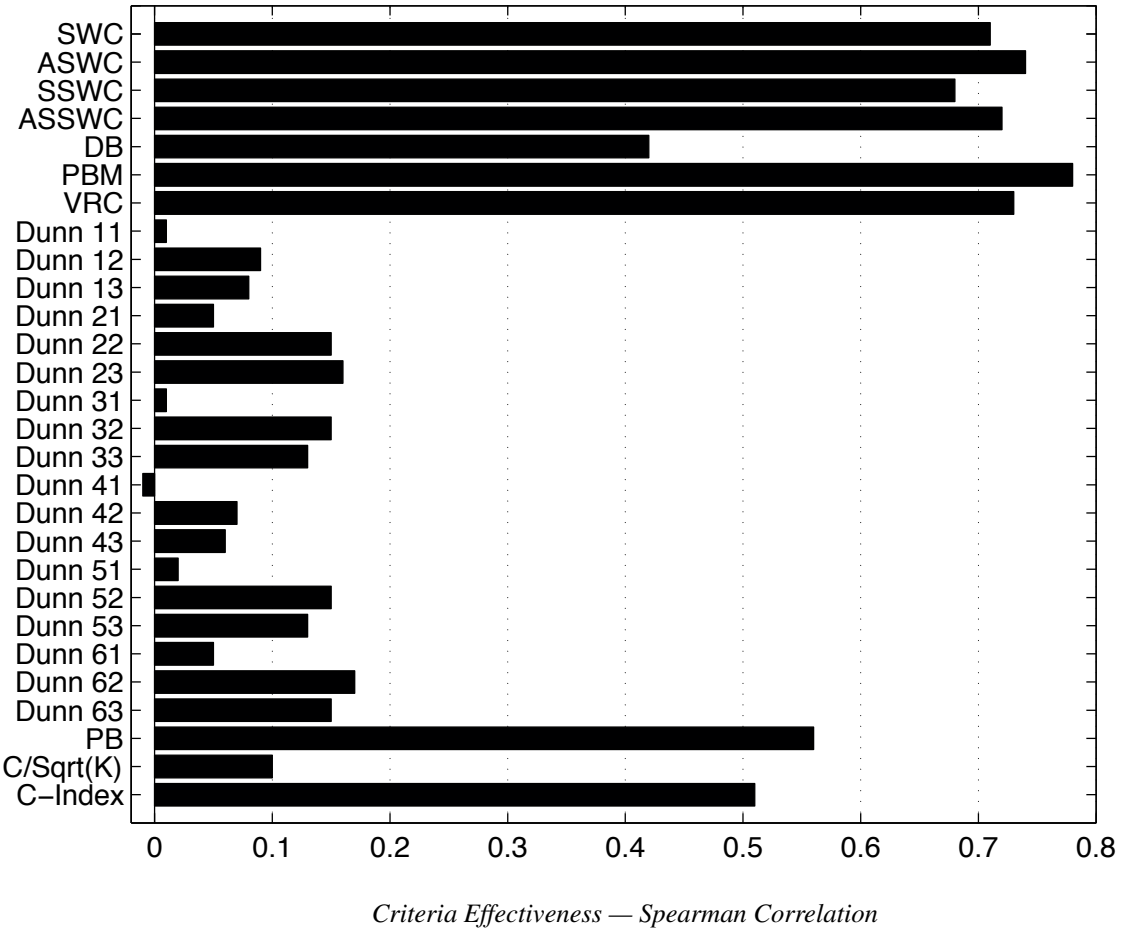


Figure 4.2: Average Spearman correlation (effectiveness) between the scores of each individual relative validity criterion and the scores provided by the external index (Adjusted Rand Index).

4.3.2.2 Criteria Complementarity

To assess diversity, we continue our analysis now to identify a set of complementary measures. Complementarity can also be measured through correlation. We use again the Spearman correlation coefficient, but now we correlate evaluations obtained for all pairs of relative validity criteria. For any given pair of relative validity criteria, its complementarity is defined by the Spearman correlation between their values for different partitions (these correlation values for complementarity can be obtained by employing the same procedure we use to obtain relative validity criteria effectivenesses). The smaller the correlation between two criteria, the more complementary they are. The results of this analysis of complementarity are displayed in Figure 4.3, in a Heatmap. The "hotter" ("colder") a cell is the stronger (weaker) the correlation between the corresponding pair of relative validity criteria.

From Figure 4.3, it is possible to visually identify groups of correlated criteria, such as one consisting of the variants of Dunn's index and another one containing the variants of the Silhouette, each of which is also composed of subgroups with varying degrees of correlation (note that there are several subgroups of measures composed of Dunn's variants). From an ensemble perspective,

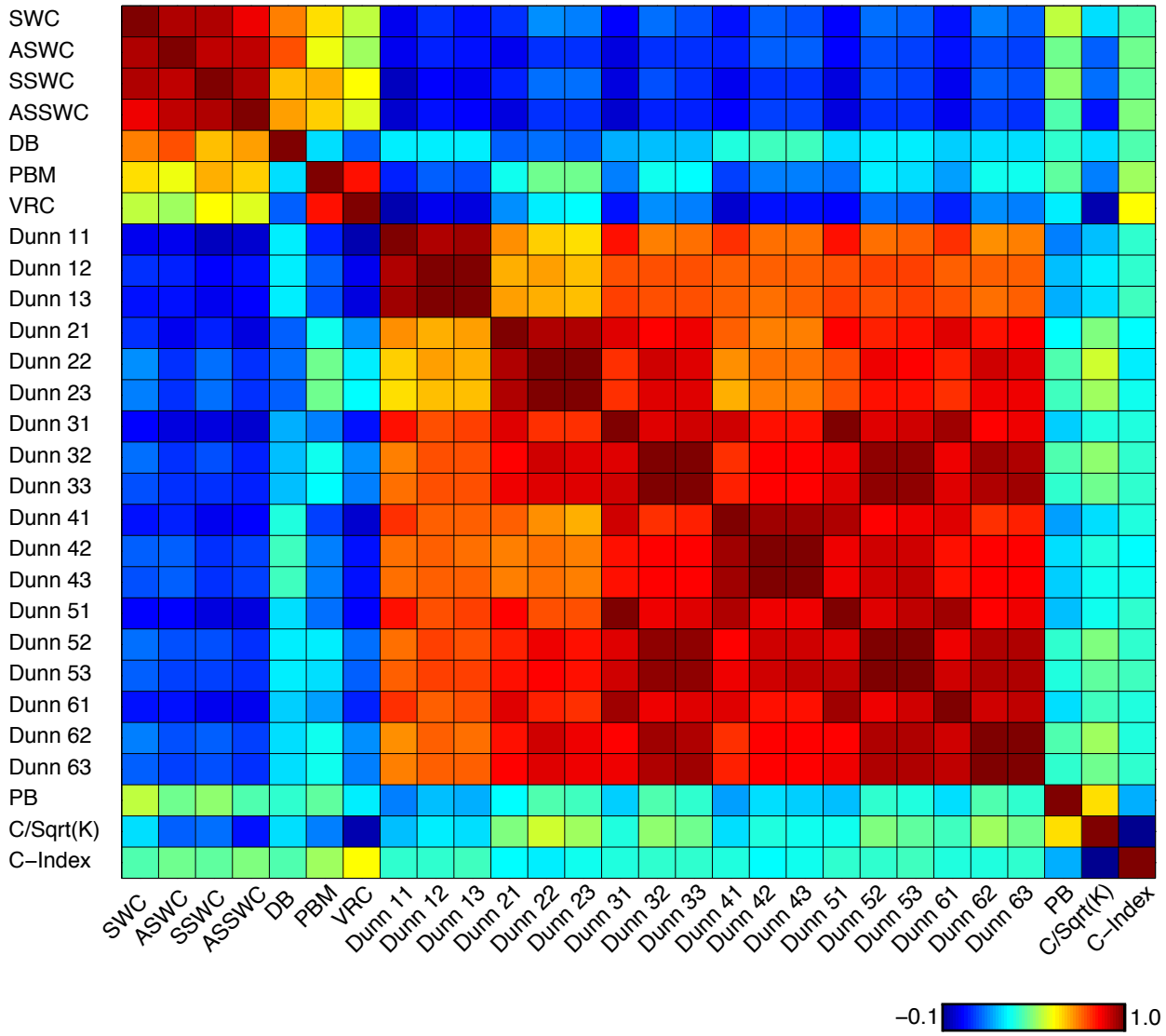


Figure 4.3: Assessing complementarity: pairwise Spearman correlation for the 28 relative criteria. Each cell depicts the average Spearman correlation over the 972 synthetic datasets.

selecting several criteria from the same group is very unlikely to bring any benefit as they provide very similar evaluations, so this should be avoided. Subsets of measures across groups are those that do not produce redundant outcomes and therefore satisfy the requirement of diversity for ensemble construction.

4.3.2.3 Guided Selection of Measures

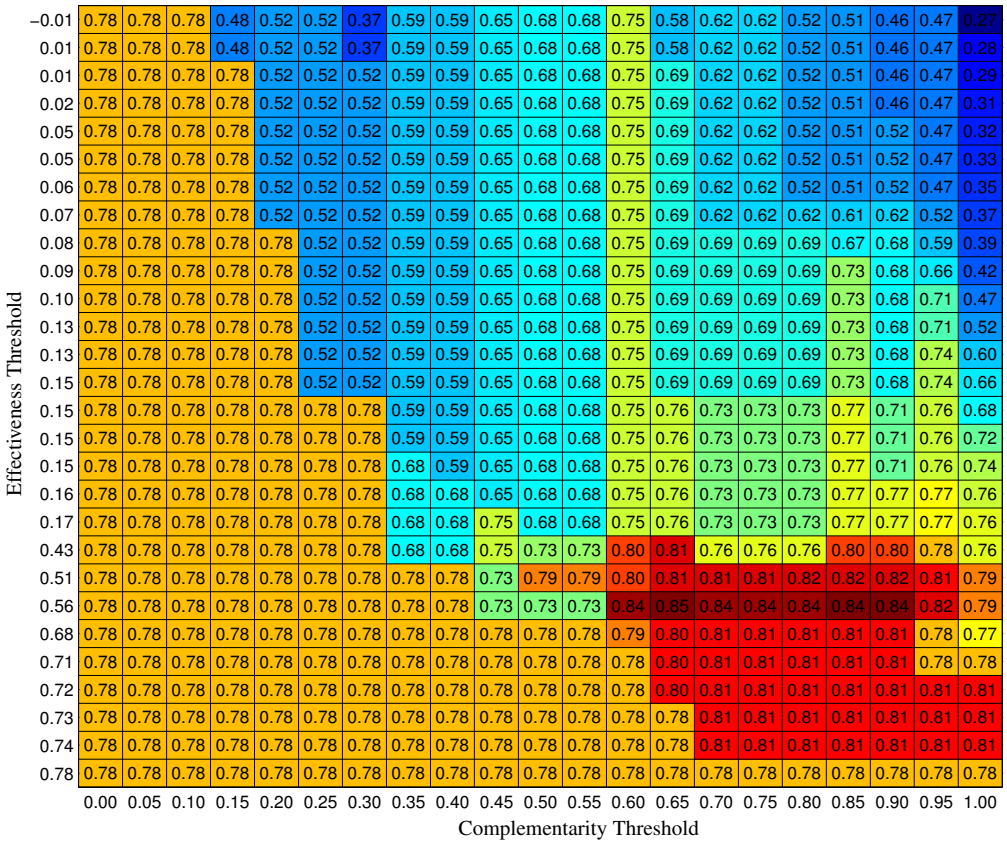
Once information regarding effectiveness and complementarity is available we can proceed and select specific relative validity measures to build ensembles. Given the results from the previous sections we only need to select two threshold values, one for effectiveness and one for complementarity, in order to generate a subset of relative validity criteria that can be used as basis for an ensemble. In brief, the problem we face is how to select the values of thresholds that lead to the best performing ensembles. Note that covering all the search space of criteria subsets

is not a viable alternative, given that it would require us to generate all possible combinations of the 28 relative validity criteria available and evaluate their respective ensembles, a total of $\binom{28}{1} + \binom{28}{2} + \cdots + \binom{28}{27} + \binom{28}{28} = 268,435,455$ combinations.

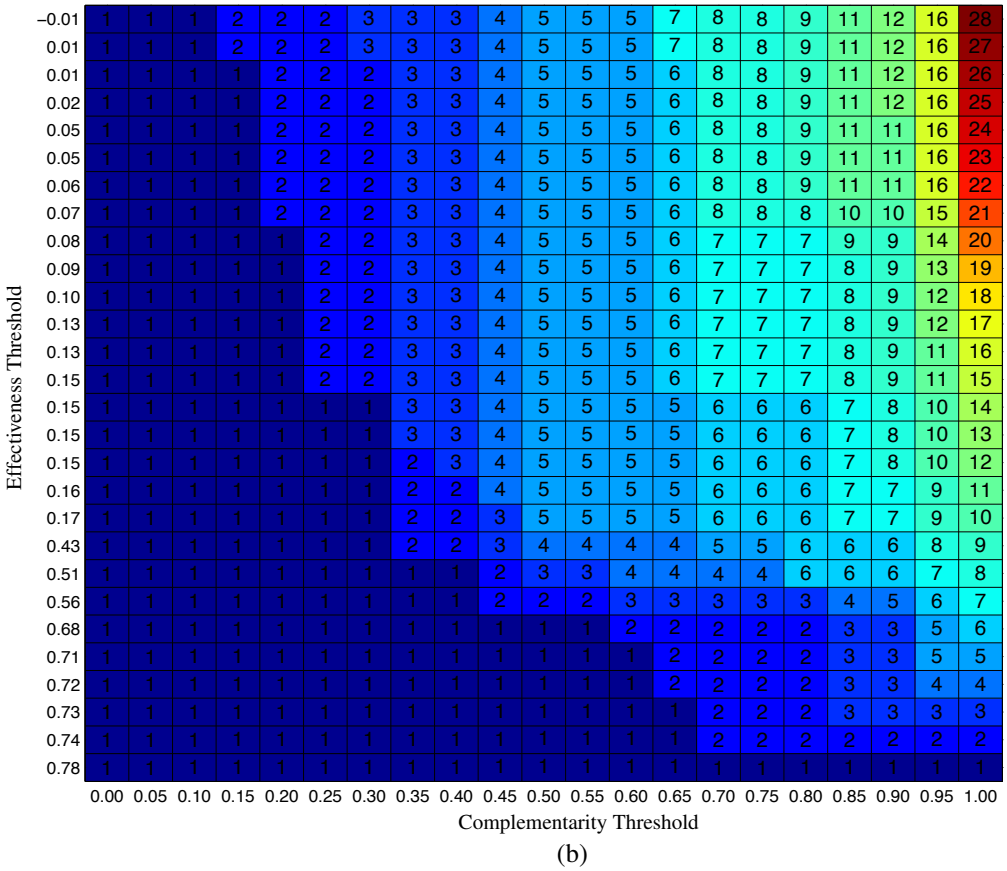
Given the large number of possible combinations we devised a simple heuristic approach to guide our selection and limit the search space. Our heuristic is as follows. First we set a threshold for effectiveness and further consider only measures with effectiveness equal or greater than such a threshold. Given that we have 28 different relative validity criteria, we have 28 different possible thresholds for effectiveness. After narrowing down the number of measures by the effectiveness threshold, we consider different thresholds for complementarity. In this case, we consider values in the $[0, 1]$ interval, with increments of 0.05 (a total of 21). Building all these combinations leads us to $21 \times 28 = 588$ subsets of measures (considering all possible subsets, including repeated ones). Each ensemble subset is obtained as follows: (i) first, the criterion with highest effectiveness is added to the subset; (ii) criteria that do not violate the restriction of complementarity are further added to the subset, one at a time, in an ordered fashion, from the most accurate to the least accurate one. Based on these subsets we can build their respective ensembles (with the strategies discussed in Section 4.3.1) and obtain their effectiveness for our data collection. These will serve as the basis for our selection of ensemble members that will later be evaluated on previously unseen data.

Effectiveness results regarding the ensembles built using the relative validity criteria selected with our heuristic are shown in Figure 4.4(a) alongside with the number of members (criteria) for each respective ensemble, as shown in Figure 4.4(b). In Figure 4.4(a) each cell gives the average effectiveness of the ensembles that were built using relative validity measures with an effectiveness equal or greater than specified and a complementarity (correlation) equal or at most the specified. The same interpretation is valid for the number of ensemble members, in Figure 4.4(b). Note that the effectivenesses shown in Figure 4.4(a) are w.r.t. the 972 synthetic datasets. Furthermore, these values account for the average effectiveness of the different rank-based ensemble strategies presented in Section 4.3.1. Even though we considered the score-based strategies initially in our analysis, they showed considerably inferior results to all rank-based strategies in all cases we considered. Given their poor results and their inherent drawbacks (*e.g.*, the need of normalization) we do not recommend their use and exclude them from the remain of our analysis.

Considering the division of Figure 4.4(a) in quadrants we have the following interpretation. In the top left quadrant we build ensembles for which almost all measures are initially available, since the effectiveness threshold is "relaxed". A small number of measures end up in the final subset, however, due to the complementarity threshold, which prevents the addition of a large number of criteria to the ensemble subset. In the bottom left quadrant we have the greatest restrictions, regarding both effectiveness and complementarity. In the top right quadrant both effectiveness and complementarity threshold are "relaxed", leading to ensemble subsets with more criteria. Finally, in the bottom right quadrant, ensemble subsets are composed of measures with high effectiveness and a moderate to low complementarity (correlation among members ranging from 0.5 to 1.0).



(a)



(b)

Figure 4.4: Results for ensembles built with relative criteria subsets selected with our approach.

From Figure 4.4(a) it is possible to see that the initial subset of criteria (top left corner) consists of a single measure (PBM). The addition of measures with low effectiveness and low level of complementarity among themselves to this set decreases the effectiveness of the ensembles that are generated. In the extreme case in which all the 28 relative validity criteria are included in the set (top right corner) one can observe the lowest overall effectiveness. When certain thresholds are considered, however, one can observe better and somewhat stable results (red region from Figure 4.4(a)). We note that even though some of these ensembles show the same effectiveness and number of members, their constituent members are different because of the threshold restrictions. Therefore, we select different candidates from this region for further evaluation.

Detailed results for these subsets of measures are shown in Table 4.7. The best results are obtained with a subset of 3 measures, which displays the higher complementarity among the subsets we select. Although our main focus is not on the evaluation of the ensemble strategies, we note that the RRA (Robust Rank Aggregation) is consistently the worst strategy. Given its results and the fact that such strategy was developed with a biological setting in mind, considering cases with high levels of noise, we believe it is not fully appropriate for our setting and decide not to consider it further. It is important to note that this decision, along with the one of ruling out score-based methods, was based solely on the information obtained from the synthetic datasets.

Apart from RRA, all the other ensemble strategies perform quite consistently. In fact, on average, all the ensembles from Table 4.7 performed better than the best single relative validity criterion (PBM). Of course, these results are from the actual data that was employed to select the ensemble candidates. Therefore, these results can be expected to be overly optimistic. It is important to note that although the differences between the average ensemble results and the best single relative validity criterion are not large, in practice one does not know which one is the best available criterion. Moreover, the best measure most likely will change for different datasets. In this cases, the use of ensembles can be of great help and benefit, since the user does not have to pick a specific measure. These particular points are further discussed in the next sections during the evaluation of the selected ensembles.

4.3.3 Experimental Setup

We used datasets from different sources and distributions to evaluate the ensembles built on our selection of relative validity criteria. It is important to note that the datasets we describe here were not employed during the selection of ensemble members, which was performed considering solely the 972 synthetic datasets. Having made such considerations, the first group of datasets we employed during our evaluation was obtained from the Amsterdam Library of Object Images (ALOI) (Geusebroek et al., 2005). This collection of datasets was already described in Section 4.2.1.1. Although the ALOI collection provide a considerable number of datasets, they are alike within each collection, that is, they are similar in nature, providing a limited variability within the collection. To better show the behavior of the ensembles of relative validity criteria we

Table 4.7: Results for the selected criteria subsets.

	Selected Thresholds				
Effectiveness	0.56	0.56	0.56	0.51	0.51
Complementarity	0.65	0.85	0.90	0.80	0.95
	Selected Subsets				
Subset Size	3	4	5	6	7
Subset Criteria	ASSWC PB PBM	PB PBM SSWC VRC	CI PB PBM SSWC VRC	ASSWC CI PB PBM SWC VRC	ASSWC CI PB PBM SSWC SWC VRC
	Ensemble Effectiveness				
Borda	0.84	0.86	0.84	0.82	0.83
Condorcet	0.89	0.86	0.86	0.84	0.83
Footrule	0.88	0.86	0.85	0.84	0.83
Median	0.88	0.88	0.85	0.87	0.83
RRF	0.80	0.81	0.83	0.75	0.78
ULARA	0.89	0.89	0.86	0.86	0.85
MC4	0.89	0.88	0.86	0.86	0.83
RRA	0.73	0.73	0.73	0.69	0.70
Best	0.89	0.89	0.86	0.87	0.85
Average	0.85	0.84	0.84	0.82	0.81
Worst	0.73	0.73	0.73	0.69	0.70

also employ eight different datasets, which are not part of a similar collection. These are: (i) The Yeast Galactose (Yeast), with 205 objects, 20 features and 4 clusters, from Yeung et al. (2001b); and seven datasets from UCI (Frank and Asuncion, 2010), namely: (i) E. Coli, with 336 objects, 8 features and 8 clusters; (ii) Glass, with 214 objects, 9 features and 7 clusters; (iii) Iris, with 150 objects, 4 features and 3 clusters; (iv) Control Chart (KDD), with 600 objects, 60 features and 6 clusters; (v) Karhunen, with 2000 objects, 64 features and 10 clusters; (vi) Vehicle, with 946 objects, 18 features and 4 clusters; and (vii) Ionosphere, with 351 objects, 34 features and 2 clusters.

4.3.4 Results and Discussion

We start our discussion by presenting the results obtained with the use of each one of the relative validity criterion individually in Table 4.8. One interesting aspect of these results is the variability, in the sense that, for each dataset, a different criterion is the best choice (the best values for each dataset are highlighted in bold). We believe that such differences in performance, as presented by Table 4.8, are a good indicative that our data come from different distributions and, therefore, are adequate for our evaluation. For the nine datasets (we consider ALOI as a single value, *i.e.*, the average over the 400 datasets) it is clear that one cannot select a single criterion and obtain the top performance over all datasets. For instance, for the Karhunen dataset one can obtain the best overall results with ASSWC. This very same criterion, however, is among the worst performing ones for Ionosphere and Glass datasets. Similarly, C-Index, which is the best

Table 4.8: Effectiveness (correlation w.r.t. external index) of individual relative validity criteria.

Criterion	E. coli	Glass	Iris	KDD	Karhunen	Vehicle	Yeast	Ionosphere	ALOI
SWC	0.77	0.35	0.83	**0.60**	0.68	0.71	**0.92**	0.51	0.40
ASWC	0.68	0.35	0.81	0.59	0.73	0.69	0.87	0.14	0.41
SSWC	0.78	0.33	0.86	**0.60**	0.70	0.73	0.89	0.53	0.43
ASSWC	0.73	0.34	0.84	0.58	**0.78**	0.73	0.85	0.19	**0.48**
DB	0.35	0.30	0.80	0.55	0.55	0.76	0.58	-0.16	0.27
PBM	0.76	**0.52**	0.77	0.44	0.34	**0.81**	0.82	0.66	0.36
VRC	0.68	0.44	0.61	0.49	0.32	0.71	0.82	**0.71**	0.32
Dunn 11	0.10	0.02	-0.11	-0.02	0.01	-0.25	0.80	-0.04	0.14
Dunn 12	0.32	-0.10	0.39	-0.12	0.22	0.32	0.87	0.38	0.20
Dunn 13	0.28	-0.11	0.36	-0.11	0.20	0.22	0.85	0.30	0.20
Dunn 21	0.71	0.34	0.47	0.49	0.05	0.60	0.82	0.51	0.02
Dunn 22	0.77	0.19	0.73	0.50	0.27	0.74	0.87	0.65	0.10
Dunn 23	0.76	0.19	0.72	0.48	0.26	0.71	0.87	0.60	0.11
Dunn 31	0.67	0.35	0.36	0.43	0.18	0.50	0.82	0.50	0.03
Dunn 32	0.75	0.26	0.71	0.44	0.32	0.72	0.89	0.60	0.11
Dunn 33	0.72	0.26	0.67	0.42	0.32	0.69	0.88	0.56	0.12
Dunn 41	0.64	0.35	0.41	0.54	0.31	0.52	0.82	0.47	0.04
Dunn 42	0.72	0.26	0.73	0.53	0.39	0.72	0.89	0.59	0.11
Dunn 43	0.68	0.26	0.69	0.53	0.39	0.69	0.87	0.57	0.12
Dunn 51	0.66	0.35	0.39	0.48	0.21	0.51	0.82	0.48	0.04
Dunn 52	0.74	0.26	0.72	0.49	0.35	0.72	0.89	0.59	0.11
Dunn 53	0.71	0.27	0.68	0.47	0.35	0.69	0.86	0.56	0.12
Dunn 61	0.75	0.34	0.44	0.47	0.26	0.54	0.83	0.25	0.03
Dunn 62	0.78	0.21	0.73	0.48	0.46	0.72	0.88	0.56	0.10
Dunn 63	0.76	0.21	0.70	0.45	0.46	0.69	0.88	0.52	0.12
PB	**0.96**	0.46	**0.87**	0.58	0.61	0.80	**0.92**	0.56	0.24
C/Sqrt(K)	0.81	0.12	0.80	0.34	0.35	0.79	0.85	0.56	0.23
C-Index	0.23	**0.52**	0.21	0.16	0.10	0.64	0.79	0.36	0.21
Best	0.96	0.52	0.87	0.60	0.78	0.81	0.92	0.71	0.48
Average	0.65	0.27	0.61	0.42	0.36	0.62	0.85	0.45	0.18
Worst	0.10	-0.11	-0.11	-0.12	0.01	-0.25	0.58	-0.16	0.02

performing relative validity criterion for the Glass dataset is among the worst ones for Karhunen. Even for the ALOI datasets, which are fairly similar in nature, such observation holds, with different criterion figuring as the best choice for different datasets from the collection. Therefore, given the lack of *a priori* knowledge one cannot select the best relative validity criterion for each dataset in hand. If such a selection is made randomly, that is, by picking one among the 28 relative validity criteria available, one can expect to obtain a certain degree of effectiveness (correlation with the external index) on average, which we show at the bottom of Table 4.8. From the bottom of Table 4.8 also note the large difference between the best and worst performing criterion for each dataset, which shows the difficulty in selecting an appropriate relative validity measure for each dataset under analysis.

Let us now, in comparison, explore the results for the ensemble strategies regarding the same datasets. We show in Table 4.9 the results for the best performing ensemble selected with our approach from the previous section, composed of three criteria, namely: ASSWC, PB, and PBM. It is important to recall that the selection of these measures was made only with the use of the 972 synthetic datasets. Even though we are using ensembles it is possible to note some variability in the results, *i.e.*, the best ensemble method is different for each dataset under analysis. It is interesting to note, however, that the ensemble methods perform much more consistently than the single relative criteria that are used to build them. This can be readily observed if we analyze the differences between the best and worst results for single criteria and ensembles. From these values one can see

that the range of correlation with the external index for the ensemble methods is much narrower than for the single relative validity criteria. In fact, in *all the cases* the worst ensemble method performs better than the average (expected value) of employing a single relative criterion. In brief, we see that, while the choice of a particular rank aggregation method does not matter too much, aggregating the ranks of sufficiently accurate and complementary criteria renders evaluation results much more stable and reliable than picking any of the individual criteria randomly. Furthermore, note that in some cases (Glass, Iris, Karhunen, and Yeast), one can improve over the performance of the best validity criterion by using different ensemble strategies.

Table 4.9: Spearman correlation w.r.t. external index for the best performing criteria subset selected with our heuristic approach (using the 972 synthetic datasets): ASSWC, PB, and PBM.

	E. coli	Glass	Iris	KDD	Karhunen	Vehicle	Yeast	Ionosphere	ALOI
Borda	0.89	0.52	0.89	0.57	0.73	0.81	0.95	0.66	0.42
Condorcet	0.90	0.43	0.89	0.57	0.59	0.80	0.94	0.59	0.42
Footrule	0.90	0.44	0.89	0.56	0.58	0.80	0.94	0.52	0.41
Median	0.90	0.44	0.89	0.56	0.58	0.80	0.94	0.52	0.41
RRF	0.87	0.53	0.87	0.57	0.87	0.80	0.92	0.59	0.44
ULARA	0.90	0.48	0.89	0.57	0.64	0.81	0.95	0.64	0.42
MC4	0.90	0.42	0.89	0.57	0.60	0.80	0.94	0.57	0.42
Best	0.90	0.53	0.89	0.57	0.87	0.81	0.95	0.66	0.44
Average	0.89	0.47	0.89	0.57	0.65	0.80	0.94	0.58	0.42
Worst	0.87	0.42	0.87	0.56	0.58	0.80	0.92	0.52	0.41

Results for all the different ensembles selected from the training datasets (3, 4, 5, 6, and 7 criteria ensembles) are shown in Figure 4.5. These are labeled 3CE, 4CE, 5CE, 6CE, and 7CE for short and are given in tones of blue. Each bar in the figure accounts for the average of the different ensemble strategies, whereas standard deviations are given by the error bars. In addition to the results from the selected ensembles from the training data we also show the average results concerning the use of single criterion (SC) and ensembles generated from the random selection of ensemble members, which are labeled #CRE (# Criteria Random Ensemble), where # is the number of criteria used to build the ensemble. The random ensembles are shown in different red tones. These are included as baseline, in order to verify how well our guided selection of measures perform in comparison to randomly generated ones, *i.e.*, different and arbitrary selection of ensemble members. Note that building and evaluating all possible ensembles for 3, 4, 5, 6, and 7 criteria combinations is computationally prohibitive (it accounts for a total of $268,435,455$ combinations), therefore we report average values of 1,000 randomly built ensembles (along with their standard deviation) for each different combination size. Let us note that the standard deviations for the randomly generated ensembles (red bars) are lower than for the other methods because they account for the deviation of 1,000 randomly generated ensembles. Therefore, although the averages of the ensembles selected with our approach and the averages of the randomly generated ensembles are comparable, their standard deviations are not.

The first aspect we highlight from Figure 4.5 is the variance of the different approaches. It is easy to verify that the use of single criterion over all the datasets leads to the highest variance. In

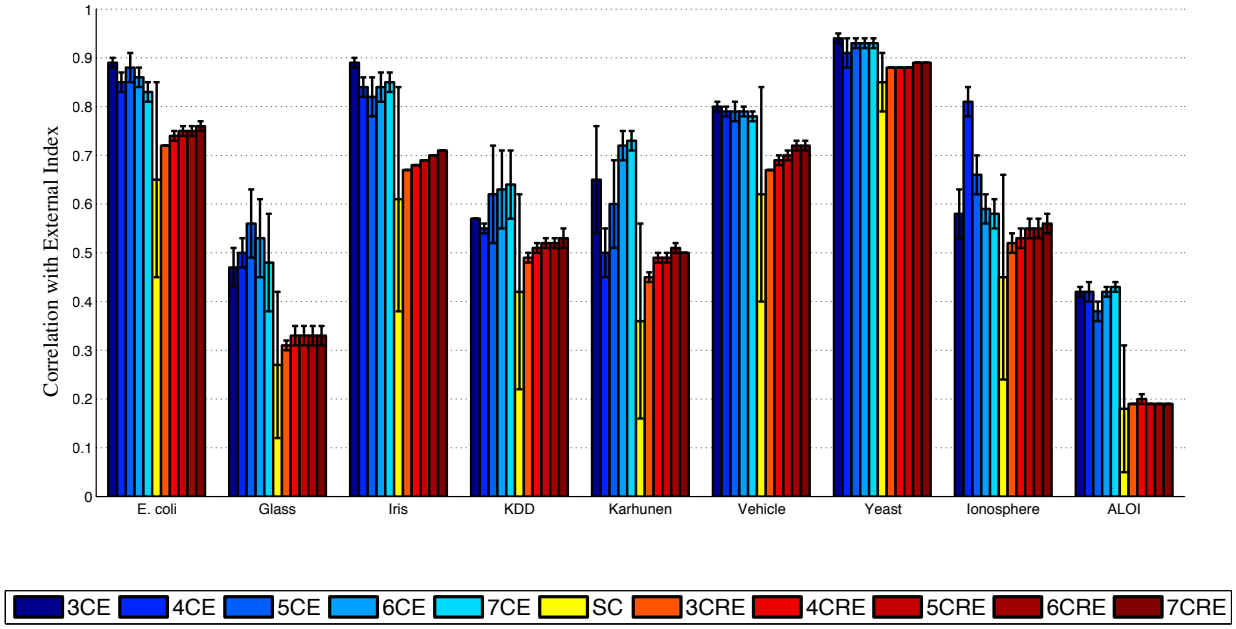


Figure 4.5: Results for ensembles selected with our heuristic, single criterion, and random.

contrast the use of different ensembles, no matter if built on the basis of our guided selection of measures (blue bars) or if randomly (red bars) provide a lower variance, that is, more stable results. It is interesting to note that, on average, all ensemble strategies perform better than the expected results obtained with the use of single relative validity criterion. It is noticeable, however, that the ensembles generated with our guided selection of measures provide, in general, better results than the ones obtained with ensembles built on the basis of random selection of measures (arbitrary selection). In order to provide some reassurance about the validity and non-randomness of our results we applied the Friedman (Friedman, 1937) and the Nemenyi (Nemenyi, 1983) tests for pairwise comparison of the methods, considering their average results on the test datasets. At a 95% confidence level, both tests suggest favorable statistically significant differences for all ensembles built on the basis of our approach when compared against the use of single relative validity criteria. Such differences were not observed when the results of randomly built ensembles were compared against the results of single relative validity criteria.

Finally, for the sake of completeness, we show the results of all the ensembles selected based on the 972 synthetic datasets, using our guided selection of relative validity criteria, in Figure 4.6. In this figure, each cell shows the average result over the 9 test datasets (again, ALOI counts as a single value by obtaining the average). Note that the ensemble members and the number of members are the same as those presented on Figure 4.4, that is, the ensembles are the ones selected from the synthetic data. The results, however, consider only unseen data (w.r.t the selection of ensemble members). It is interesting to see that, although Figure 4.6 is not an exact reproduction of Figure 4.4, the trend in both figures is quite the same. Indeed, the highest effectiveness ensembles (the hotter cells) are located in a similar region in both figures, showing that our approach for

selecting the ensemble members does not only provide good results on the datasets employed in the selection of ensemble members, but provides also good results when one considers new data.

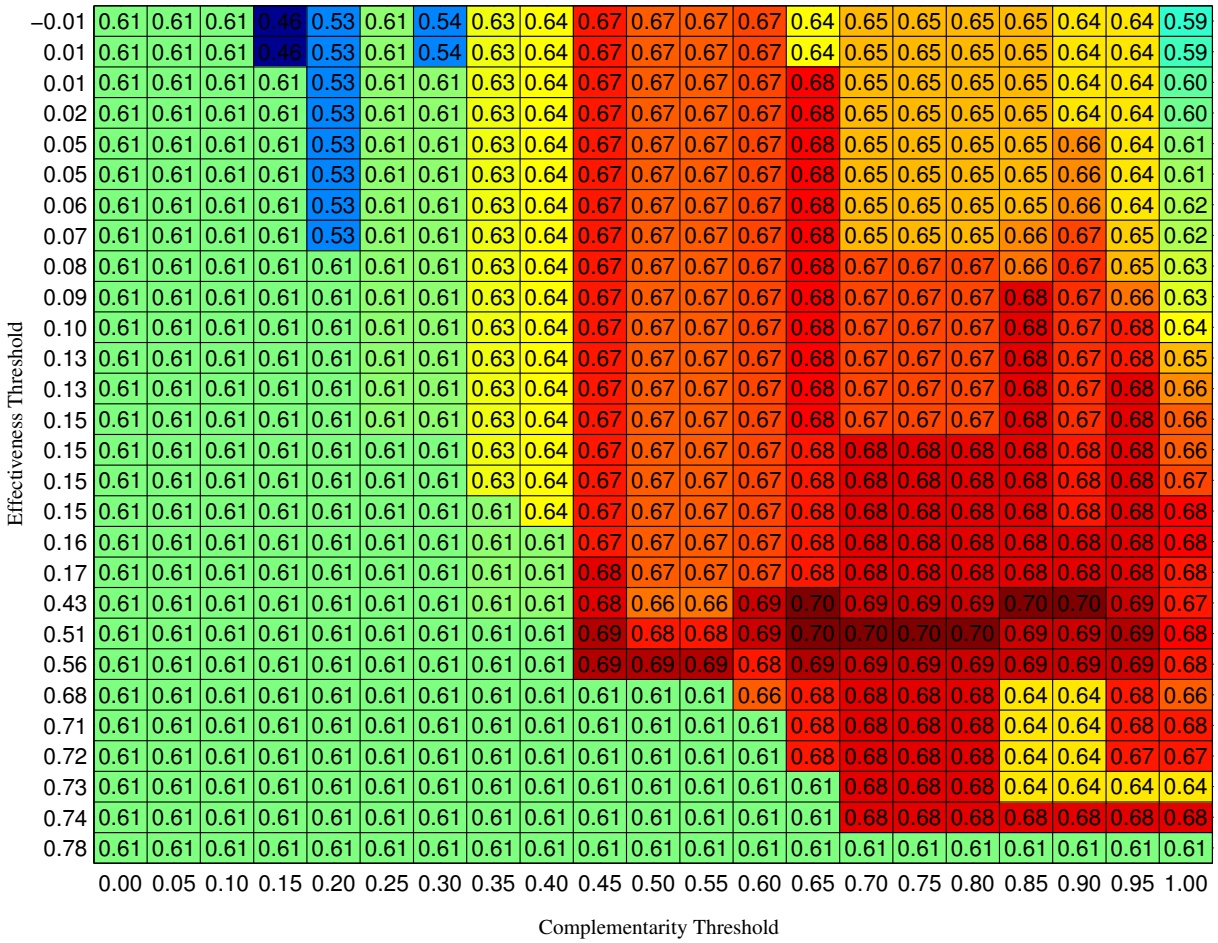


Figure 4.6: Effectiveness (average Spearman correlation with the external index) of all the ensembles built considering different effectiveness and complementarity thresholds.

4.4 Chapter Remarks

In this chapter we explored different aspects involved in the ensemble of criteria for relative evaluation of clustering results. We started our analysis from an initial set of 28 relative validity criteria, which served as candidates for building our ensembles. In the first half of the chapter we discussed ensembles generated on the basis of a random selection of relative validity measures. In summary, our results suggest that the combination strategies studied are not recommended if the user somehow knows the most appropriate individual criterion for the dataset in hand, which is, however, not very realistic in most practical applications. On the other hand, the results suggest that combinations are viable alternatives to improve on the worst case scenarios, which usually cannot be avoided in practice. In the remaining of the chapter we elaborated on a principled selection of ensemble members. To that end, we introduced an heuristic approach to select ensemble members

based on complementarity and effectiveness. Ensemble members were selected on the basis of synthetic datasets and were later evaluated on different datasets. Our results support that effective ensembles can be built from a judicious selection of relative validity criteria, in which one aims to assure the basic requirements established in ensemble theory. We showed empirical evidence that our approach performs consistently well across different combination strategies. Moreover, our ensembles provided results that are more robust and effective than those achieved by simply employing a "blind selection" of measures to build validation ensembles. We believe that our work provides novel insights regarding ensembles of validity indexes and uncovers the potential of ensemble methods for improved cluster evaluation procedures.

CHAPTER

5

Relative Validation of Clustering Results

Relative validity criteria are the measures ultimately employed for the validation of clustering results in most practical application scenarios. This arises due to the fact that apart from the data itself, relative validity measures require only a partitioning of the data, that is, a clustering solution to be evaluated. A number of relative validity criteria have been introduced in the past years, as we previously reviewed in Chapter 2. In this chapter both practical and theoretical contributions to the field are provided. Given the different nature of each one of the contributions, the chapter is divided into two main sections. In the first part, Section 5.1, the use of Receiver Operating Characteristics (ROC) in the validation of clustering results is discussed. The concept of Receiver Operating Characteristics has been successfully employed for the evaluation of results in classification (supervised learning). It has not been employed, however, to the validation of clustering results so far. In this section the Area Under the Curve (AUC) of the ROC Curve is introduced as a relative measure of cluster quality. Properties regarding its expected value for random solutions and its equivalence to the Gamma relative validity criterion (Baker and Hubert, 1975) are theoretically explored. Finally, experimental results show that AUC is a reasonable alternative to the validation of clustering results, with a reduced computational complexity, when compared to Gamma. In the second part of the chapter, Section 5.2, a relative validity criterion specifically designed to the validation of density-based clustering results, named DBCV (Density-based Clustering Validation), is reviewed. This work was performed in collaboration with Davoud Moulavi (main

contributor), during the author's one year internship at the University of Alberta, Edmonton, AB, Canada, under the supervision of Prof. Jörg Sander. The criterion is already published as: *Moulavi, D.;* Jaskowiak, P.A.*; Campello, R.J.G.B.; Zimek, A.; Sander, J. Density-Based Clustering Validation. In: SIAM International Conference on Data Mining (SDM 2014)*. We note that DBCV was originally conceived by its first author Moulavi (2014). To that end, the author of this thesis participated in its theoretical developments and experimental evaluation, which ultimately led to its publication.

5.1 ROC Curves in Clustering Validation

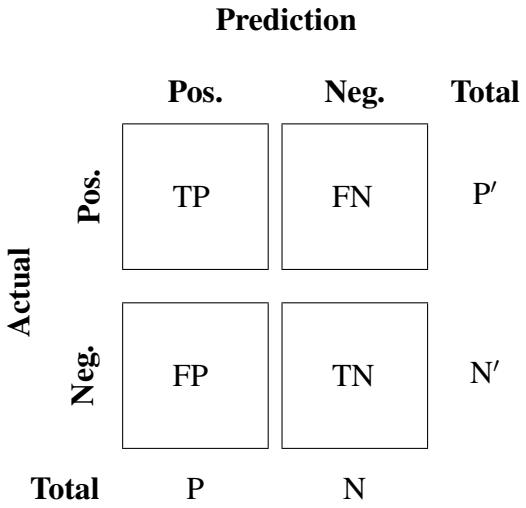
In this Section we explore the use of Receiver Operating Characteristics (ROC) in the validation of clustering results. More specifically, we employ one of the most commonly statistics derived from the ROC Graph, *i.e.*, its Area Under the Curve (AUC), as a relative validity criterion. The remainder of the section is organized as follows. In Section 5.1.1 we review basic concepts from Receiver Operating Characteristic. Section 5.1.2 elaborates on the use of the Area Under the Curve (AUC) as a relative validity criterion. In this section we also discuss theoretical properties of the AUC regarding clustering validation, formally showing that its expected value is 0.5, regardless of the number of clusters from the result partition, and we use this property to normalize the index by chance. In Section 5.1.3 we discuss the relation between Baker and Hubert's Gamma (Baker and Hubert, 1975) and the AUC of a clustering result, showing that these two measures are closely related, but only the latter can be computed in acceptable computational time. Finally, on Section 5.1.4 we provide an empirical evaluation of AUC as a relative validity criterion. To place its results into perspective, we compare AUC against the 28 relative validity criteria discussed in Chapter 2.

5.1.1 Basic Concepts

Receiver Operating Characteristics (ROC) (Fawcett, 2006) is a technique usually employed to evaluate and visualize the performance of predictive models (classifiers) in the context of supervised learning, given that the desired outcome is available for comparison. The technique output is generally depicted as a plot, which is referred to as ROC Graph. The ROC Graph is based on statistics derived from the comparison between two classification solutions, *e.g.*, one obtained with a classifier and an expected one (ground truth). Such statistics are usually displayed in the form of a so-called confusion matrix. Considering a binary classification problem[1], in which we have a positive class (also denoted by the class label 1) and a negative class (also denoted by the class label 0), a confusion matrix can be obtained, as depicted below. In this particular matrix, *Prediction* accounts for the classifier outcome (model under evaluation) and *Actual* accounts for the desired outcome, also referred to as golden standard. The values in each cell account for the

[1] Multiclass ROC Curves do exist (Fawcett, 2006), but they are beyond the scope of this work.

counts of True Positives (TP), False Positives (FP), True Negatives (TN), and False Negatives (FN), with their sum amounting to the total number of objects under evaluation.



The ROC Graph consists of plotting the values of False Positive Rate ($FPR = FP/N$) and True Positive Rate ($TPR = TP/P$), as shown in Figure 5.1. If the model under evaluation produces a discrete classification outcome for each object, a single confusion matrix is obtained, thus resulting in a single point in the ROC Graph. On the other hand, if for each object the model produces as output a probability or a score, representing the degree of membership to a class, a ROC Curve can be derived. In such a case, each point in the curve is associated with a threshold value, which lies within the interval of the scores produced by the classifier. For each threshold, objects are deemed as positive or negative according to their relative value to the threshold and a confusion matrix is obtained. These two cases are illustrated in Figure 5.1. The diagonal line in Figure 5.1 accounts for the performance of a completely random classifier. Points or curves close to the top-left corner of the ROC Graph indicate good models. A model with performance below the diagonal line performs worse than random. If the classification outcomes of such a model are reversed, its performance (point or curve) is a reflection above the diagonal line.

For classifiers that provide as outcome a probability or a score, an important concept that can be derived from a ROC Graph is the Area Under the Curve (AUC). The AUC consists of a single value that can be employed quantitatively in the evaluation of classifiers. AUC values are in the $[0, 1]$ interval. The larger the AUC value, the better the classifier. An AUC value of 0.5 indicates a completely random classifier. Values below 0.5 indicate a worse than random classifier. For computationally efficient ways of obtaining ROC Graphs and AUC values see (Fawcett, 2006).

5.1.2 AUC as a Relative Validity Criterion

The use of ROC Curves as an evaluation tool is well-established in the supervised learning community. Its use, however, has not been fully explored in the unsupervised context of cluster

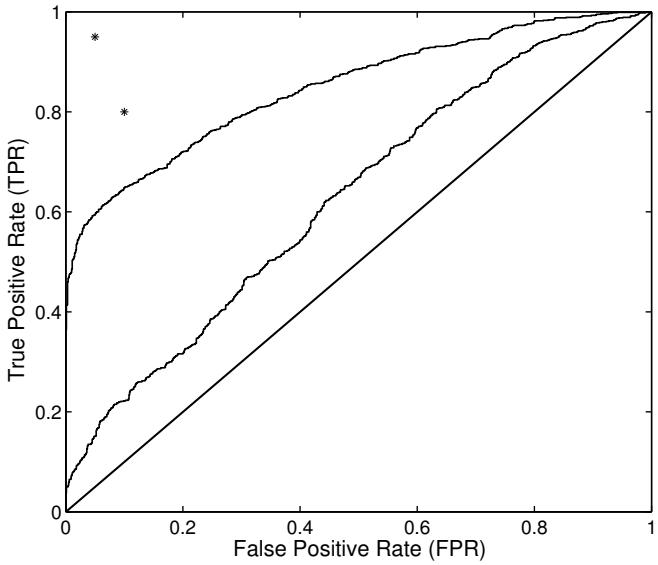


Figure 5.1: An example of ROC Graph for different classifiers.

analysis. So far, the concept of Area Under the Curve (AUC) has been used by Giancarlo et al. (2013) and Jaskowiak et al. (2013) in a very limited scope, in which the AUC was employed to measure the agreement between distance measures and *external labels* (gold standard) of a dataset, in an attempt to determine which distance measures are more adequate in the scenario of gene expression microarray data (more details are given in Chapter 6). Here we explore the use of Receiver Operating Characteristics (ROC) as a relative validity index of clustering results. Particularly, we consider the Area Under the Curve (AUC) as a relative validity criterion.

Let $\mathcal{C} = \{C_i, \ldots, C_k\}$ be a partition obtained by applying a clustering algorithm on dataset $\mathbf{X}$. Also, let $\mathbf{D}$ be a pairwise dissimilarity matrix of the objects from dataset $\mathbf{X}$. It is easy to see that such a clustering result cannot be straightforwardly provided as input to a ROC analysis. If we transform the clustering solution into a *pairwise* representation, however, it can be easily validated with the use of ROC analysis. For each pair of objects (considering a total of n objects, there are $n(n-1)/2$ distinct pairs of objects) we build a *pairwise* clustering solution, assigning the "class" label 1 if the two objects are within the same cluster and 0 otherwise. Such pairwise solution is then provided as input to a ROC evaluation with the corresponding distances for each object pairs, which can be obtained from $\mathbf{D}$. The rationale behind such evaluation is that object pairs belonging to the same cluster should have a smaller distance than those that belong to different clusters, considering a meaningful clustering solution. If one wants a probabilistic interpretation, distances can be transformed into similarities and normalized in the $[0, 1]$ interval. In this case, the greater the value, the greater the probability that the pair of objects should belong to the same cluster.

It is worth noticing that in a traditional ROC analysis the model under evaluation provides the scores, while the desired solution is represented by class labels. In the evaluation of clustering solutions, however, we have a reversed situation. That is, the model under evaluation provides the "class" labels, whilst the desired solution is represented in a continuous rather than binary way,

in the form of distances (or similarities) among objects. Having made such considerations, we show below that, even for this reversed case, the expected AUC of a random clustering solution is equal to 0.5. This value also holds when comparing clustering solutions with different numbers of clusters, that is, the expected value of the AUC does not depend on the number of clusters from the partition under evaluation.

First, let us denote by $P_u = \{p_1, \ldots, p_m\}$, with $p_c \in \{0, 1\}$ and $m = n(n-1)/2$, a pairwise clustering solution derived from a partition $\mathcal{C}$, in which a value of 0 indicates that the pair of objects belongs to different clusters and a value of 1 indicates that the pair of objects belongs to the same cluster. Excluding the non-usual clustering solutions in which (i) every object belongs to a single cluster ($k = 1$) and (ii) every object belongs to a singleton cluster ($k = n$), we have a total of $2^m - 2$ possible pairwise clustering solutions. Let us also define $\mathcal{P} = \{P_1, \ldots, P_{2^m-2}\}$ as the set containing all possible pairwise clustering solutions. For convenience, we can obtain subsets of $\mathcal{P}$ in the following form $\mathcal{P} = \mathcal{P}_+ \cup \mathcal{P}_- \cup \mathcal{P}_=$, with:

$$\mathcal{P}_+ = \{P_u \in \mathcal{P} | \sum_{i=1}^{m} p_i > \frac{m}{2}\}$$
$$\mathcal{P}_- = \{P_u \in \mathcal{P} | \sum_{i=1}^{m} p_i < \frac{m}{2}\}$$
$$\mathcal{P}_= = \{P_u \in \mathcal{P} | \sum_{i=1}^{m} p_i = \frac{m}{2}\}$$

Note that $\mathcal{P}_=$ can be empty in some cases, specifically, for any case in which m is odd. Given these subsets of pairwise solutions, a complement function that simply flips the values of a pairwise solution can be easily derived, as given by Definition 6. On the basis of the previously defined subsets and the complement function we can also define two more subsets of pairwise solutions from $P_=$. They are defined as follows: if $P_=$ is empty, there is nothing to be done, then simply define $P' = P'' = \emptyset$; otherwise, if $P_=$ is not empty, pick $P_i \in P_=$. Note that $P_i \neq P_i^C$. If $P_= = \{P_i, P_i^C\}$ then define, $P' = \{P_i\}$ and $P'' = \{P_i^C\}$. If $P_=$ has more than two elements, then proceed the same way until there is no element left. Note that $P_=$ has necessarily a even number of elements.

Definition 6. The Complement function of a pairwise clustering solution, for a pairwise solution P denoted by $\Psi(P)$ or P^C, can be obtained by simply flipping its values, as given by Ψ below:

$$\Psi : \mathcal{P} \to \mathcal{P}$$
$$P \mapsto \Psi(P) = P^C = \{1 - p_1, \ldots, 1 - p_m\}$$

It is easy to see that complement of a pairwise solution belongs to the set of possible solutions ($\mathcal{P}$). In Proposition 1 we show (sketched proofs) that $\Psi(.)$ has four desired properties. As usual, the symbol $\circ$ accounts for function composition, that is $\Psi(.) \circ \Psi(.) = \Psi(\Psi(.))$.

Proposition 1.

1. $\Psi(.)$ is a bijective function
2. $\Psi^2(.) = \Psi(.) \circ \Psi(.) = Id$
3. $\Psi(\mathcal{P}_-) = \mathcal{P}_+$ and $\Psi(\mathcal{P}_+) = \mathcal{P}_-$
4. $\Psi(\mathcal{P}_=) = \mathcal{P}_=$

Proof. The proof sketch for each item of Proposition 1 is given below.

1. To prove that $\Psi(.)$ is bijective we must show that (i) it is injective and (ii) it is surjective.
   - Injective: let $\Psi(P_i) = \Psi(P_j)$, with $P_i = \{p_1^i, \ldots, p_m^i\}$ and $P_j = \{p_1^j, \ldots, p_m^j\}$, we must show that $P_i = P_j$, which is trivial, since $\{1 - p_1^i, \ldots, 1 - p_m^i\} = \Psi(P_i) = \Psi(P_j) = \{1 - p_1^j, \ldots, 1 - p_m^j\}$. It follows that, $p_c^i = p_c^j, \forall c$, therefore $P_i = P_j$.
   - Surjective: for any given $P_i = \{p_1^i, \ldots, p_m^i\} \in \mathcal{P}$, we must show that there exist a $P_j = \{p_1^j, \ldots, p_m^j\}$, such that $\Psi(P_j) = P_i$. This is straightforward, just take $P_j = P_i^C$.
2. $\Psi \circ \Psi(P) = \Psi(\{1-p_1, \ldots, 1-p_m\}) = \{1-1+p_1, \ldots, 1-1+p_m\} = \{p_1, \ldots, p_m\} = P$.
3. The proof follows immediately from the definition of the function. Given that 1's are turned into 0's (and vice versa) it is easy to see the total number of 1's after applying $\Psi(.)$ will be equal to the number of 0's before applying it (and vice versa).
4. Trivial since $\Psi(P)$ does not change the proportions of values from its input, it simply swaps them. If P has the same proportions of 0's and 1's, they remain the same.

□

Given a clustering solution, it is easy to verify by building its confusion matrices for different thresholds (the threshold correspond to all unique values of pairwise distances from $\mathbf{D}$) that $AUC(P_u) + AUC(\Psi(P_u)) = 1$. This is, the ROC Curve of P_u is a perfect reflection of the ROC Curve of $\Psi(P_u)$ considering the diagonal line of a completely random evaluation. This also follows from the definition of ROC Graphs. From this observation, Theorem 1 can be stated.

Theorem 1. The expected value of Area Under the Curve $E(AUC)$ obtained from the Receiver Operating Characteristics evaluation of a *random* clustering result is equal to 0.5, independently of the number of clusters present in the cluster solution under evaluation.

Proof. Given that all random pairwise clustering solutions have the same probability, the expected value of AUC ($E(AUC)$) is given bellow, where $AUC(.)$ is the Area Under the Curve of a pairwise solution and $\mathbb{P}(.)$ is the probability of a solution.

$$E(AUC) = \sum_{P_i \in \mathcal{P}} AUC(P_i)\mathbb{P}(P_i)$$

Using the subsets previously defined we have:

$$= \sum_{P_i \in \mathcal{P}_+} AUC(P_i)\mathbb{P}(P_i) + \sum_{P_i \in \mathcal{P}_-} AUC(P_i)\mathbb{P}(P_i) + \sum_{P_i \in \mathcal{P}'_=} AUC(P_i)\mathbb{P}(P_i) + \sum_{P_i \in \mathcal{P}''_=} AUC(P_i)\mathbb{P}(P_i)$$

$$= \sum_{P_i \in \mathcal{P}_+} AUC(P_i)\mathbb{P}(P_i) + \sum_{P_i \in \mathcal{P}_+} AUC(P_i^C)\mathbb{P}(P_i) + \sum_{P_i \in \mathcal{P}'_=} AUC(P_i)\mathbb{P}(P_i) + \sum_{P_i \in \mathcal{P}'_=} AUC(P_i^C)\mathbb{P}(P_i)$$

Since the probabilities of the solutions are constant:

$$= \frac{1}{\mathcal{P}} \left(\sum_{P_i \in \mathcal{P}_+} AUC(P_i) + \sum_{P_i \in \mathcal{P}_+} AUC(P_i^C) + \sum_{P_i \in \mathcal{P}'_=} AUC(P_i) + \sum_{P_i \in \mathcal{P}'_=} AUC(P_i^C) \right)$$

$$= \frac{1}{\mathcal{P}} \left(\sum_{P_i \in \mathcal{P}_+} \left(AUC(P_i) + AUC(P_i^C)\right) + \sum_{P_i \in \mathcal{P}'_=} \left(AUC(P_i) + AUC(P_i^C)\right) \right)$$

$$= \frac{1}{\mathcal{P}} \left(\sum_{P_i \in \mathcal{P}_+} 1 + \sum_{C_i \in \mathcal{P}'_=} 1 \right)$$

$$= \frac{1}{\mathcal{P}} \left(|\mathcal{P}_+| + |\mathcal{P}_=|\right)$$

$$= \frac{1}{|\mathcal{P}_+| + |\mathcal{P}_-| + |\mathcal{P}'_=| + |\mathcal{P}''_=|} (|\mathcal{P}_+| + |\mathcal{P}'_=|)$$

Given that $\Psi(.)$ is a bijection, we have:

$$= \frac{1}{|\mathcal{P}_+| + |\mathcal{P}_+| + |\mathcal{P}'_=| + |\mathcal{P}'_=|} (|\mathcal{P}_+| + |\mathcal{P}'_=|)$$

$$= \frac{1}{2|\mathcal{P}_+| + 2|\mathcal{P}'_=|} (|\mathcal{P}_+| + |\mathcal{P}'_=|)$$

$$= 1/2$$

□

As a result, the expected value of AUC for any given clustering solution is equal to 0.5, regardless of its number of clusters. The use of AUC as a clustering relative validity criterion has, therefore, the same desired properties when it is employed in classification tasks.

5.1.3 Equivalence Between AUC and Baker and Hubert's Gamma

In this Section we discuss the equivalence between the AUC of a clustering result and its evaluation with the Gamma Index (Baker and Hubert, 1975). Before showing such an equivalence, let us recall the definition of the Gamma Index. The relative validity criterion know as Gamma is defined by Equation (5.1) :

$$\gamma = \frac{s_+ - s_-}{s_+ + s_-} \tag{5.1}$$

with:

$$s_+ = \frac{1}{2}\sum_{l=1}^{k}\sum_{\substack{\mathbf{x_i},\mathbf{x_j}\in C_l \\ \mathbf{x_i}\neq\mathbf{x_j}}} \frac{1}{2}\sum_{m=1}^{k}\sum_{\substack{\mathbf{x_r}\in C_m \\ \mathbf{x_s}\notin C_m}} \delta(||\mathbf{x_i}-\mathbf{x_j}|| < ||\mathbf{x_r}-\mathbf{x_s}||)$$

$$s_- = \frac{1}{2}\sum_{l=1}^{k}\sum_{\substack{\mathbf{x_i},\mathbf{x_j}\in C_l \\ \mathbf{x_i}\neq\mathbf{x_j}}} \frac{1}{2}\sum_{m=1}^{k}\sum_{\substack{\mathbf{x_r}\in C_m \\ \mathbf{x_s}\notin C_m}} \delta(||\mathbf{x_i}-\mathbf{x_j}|| > ||\mathbf{x_r}-\mathbf{x_s}||)$$

where $\delta(.)$ is equal to 1 if the inequality is satisfied and 0 otherwise. In the equation above s_+ (s_-) indicates the number of *object pairs* from the same cluster that have a smaller (greater) distance than that of *object pairs* that belong to different clusters. Intuitively, s_+ accounts for the number of well placed pairs of objects, whereas s_- accounts for the number of misplaced pairs of objects. It is important to note that the original formulation of Gamma (Baker and Hubert, 1975) is computationally very expensive, turning out to be prohibitive in virtually any practical application. As originally conceived the criterion has complexity $O(n^2m + n^4/k)$, where n is the number of objects, m is the number of features, and k the number of clusters of the solution under evaluation (Vendramin et al., 2010). For this particular reason the Gamma criterion was not discussed in Chapter 2 or employed during the experiments conducted in Chapter 4. Having made such considerations, in Theorem 2 we show the relation between the evaluation of a clustering result with the Gamma relative validity criterion and the Area Under the Curve (AUC).

Theorem 2. The Area Under the Curve (AUC) obtained from the Receiver Operating Characteristics (ROC) evaluation of a clustering result is equal to $(1+\gamma)/2$, where γ is the resulting value from the evaluation of the same clustering result with the Gamma relative validity criterion, proposed by Baker and Hubert (1975).

Proof. Considering the random selection of one positive and one negative "object" that are subject to a Receiver Operating Characteristics (ROC) evaluation (note that in the case of clustering validation, each "object" of the evaluation corresponds in fact to a pair of objects from the clustering result), the AUC has the interesting statistical property of being equivalent to the probability that the model under evaluation will rank the randomly selected positive example *higher* than the negative randomly selected one, as discussed by Fawcett (2006).

Recall from the definition of the Gamma index that s_+ is equal to the *number* of positive pairs from the cluster solution which have a smaller distance (higher similarity) than negative pairs. We can easily define the probabilities of s_+ and s_- by dividing these values by the sum of positive and negative pairs under consideration. Let us denote such probabilities as $P(s_+)$ and $P(s_-)$. Given that the probabilities are obtained by dividing s_+ and s_- by a constant, we can rewrite Gamma as:

$$\gamma = \frac{P(s_+) - P(s_-)}{P(s_+) + P(s_-)}$$

Since $P(s_+) + P(s_-) = 1$, we have that:

$$\begin{aligned}\gamma &= P(s_+) - P(s_-) \\ &= P(s_+) - (1 - P(s_+)) \\ &= P(s_+) - 1 + P(s_+) \\ &= 2P(s_+) - 1\end{aligned}$$

It is easy to see that $P(s_+) = (\gamma + 1)/2$. Since s_+ is the number of positive examples with smaller distance than negative ones, *i.e.*, the number of positive examples ranked correctly (higher) than negative ones, $P(s_+)$ is the probability of ranking a positive example higher than a negative one. This is exactly the interpretation of the AUC value. Therefore, AUC $= (\gamma + 1)/2$. As a byproduct of the proof one can see that the value obtained with the application of the Gamma index is the very same one obtained with the application of the Gini Coefficient (Ceriani and Verme, 2012; Gini, 1912), since AUC $= (\text{Gini} + 1)/2$ (Fawcett, 2006). □

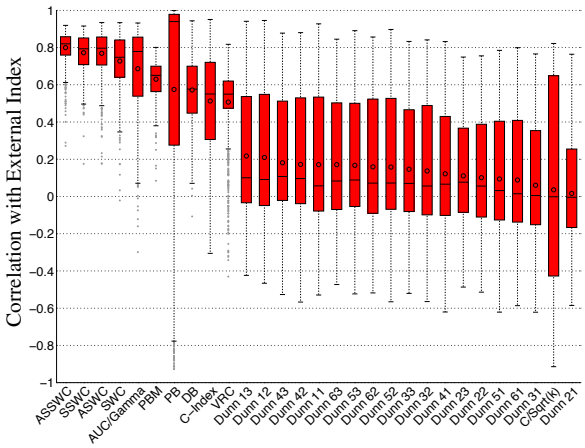
Recall that the original formulation of Gamma (Baker and Hubert, 1975) is computationally very expensive, turning out to be prohibitive in many practical applications. As originally conceived the criterion has $O(n^2m + n^4/k)$, where n is the number of objects, m is the number of features, and k the number of clusters of the solution under evaluation (Vendramin et al., 2010). Such a high complexity has prevented its evaluation in datasets with only 500 objects by Vendramin et al. (2010), for instance. As pointed out by Fawcett (2006), computing the AUC for a binary classification problem has $O(n \log n)$ complexity, given n objects. Note that in the case of clustering evaluation we are dealing with pairs of objects, therefore we have an $O(n^2 \log n)$ time complexity, a considerable reduction in comparison to the original formulation of Gamma.

5.1.4 Experimental Evaluation

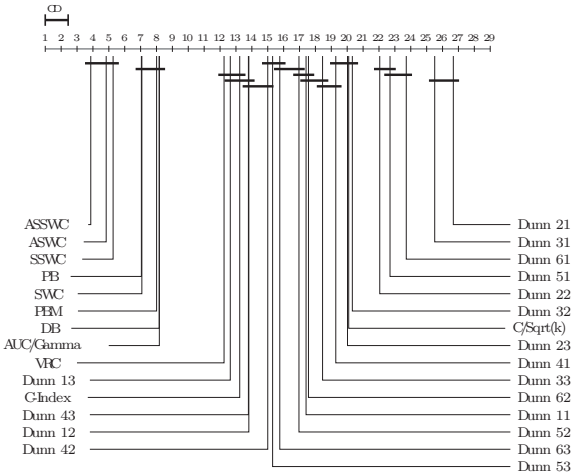
In this section AUC/Gamma is evaluated according to the alternative methodology introduced by Vendramin et al. (2009, 2010), which was already discussed in Section 2.4.2. It is worth noticing that Gamma was previously evaluated by Vendramin et al. (2010). Such evaluation, however, was performed in a scenario with a limited number of objects (50), due to the high computational cost of the original measure. Here AUC/Gamma is evaluated with the collection of 972 synthetic datasets already presented in Section 4.2.1, considering the same clustering algorithms and configurations also defined in this section. In order to put the results of AUC/Gamma into perspective, the results regarding the 28 relative validity criteria reviewed in Chapter 2 are also provided.

Results for this comparison are shown in Figure 5.2, considering both Pearson (top plots) and Spearman (bottom plots). In order to provide some reassurance about the validity of the results, we applied the Friedman statistical test followed by the Nemenyi post-hoc test following

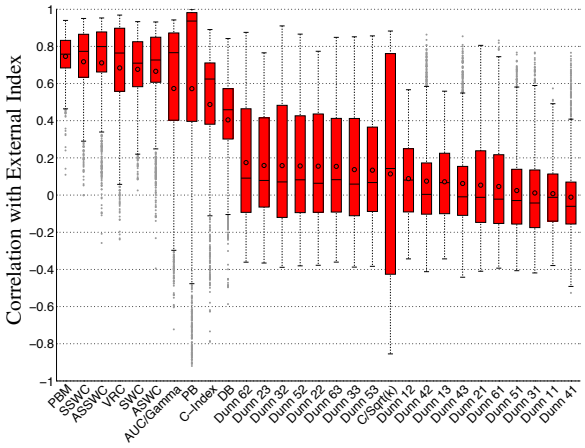
the approach proposed by Demšar (2006). Figure 5.2a and Figure 5.2c present boxplots for both evaluations (regarding Pearson and Spearman), whereas Figure 5.2b and Figure 5.2d, present a summary of the results from the statistical tests at a 95% confidence level.



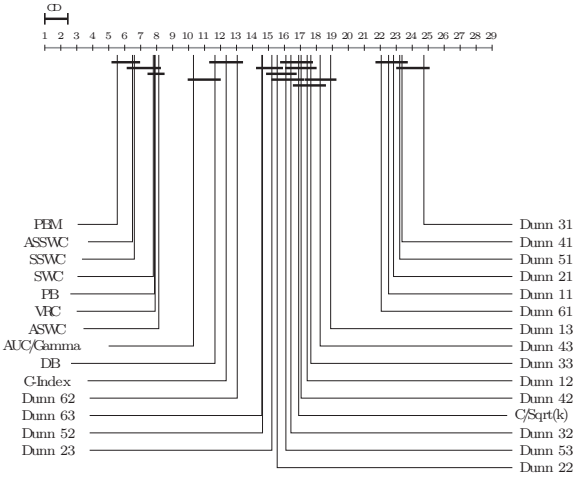
(a) Results for Pearson



(b) Statistical Test Summary for Pearson



(c) Results for Spearman



(d) Statistical Test Summary for Spearman

Figure 5.2: Evaluation of AUC/Gamma and other 28 relative validity criteria, according to the alternative methodology. Boxplots (a) and (c) are sorted with respect to the mean correlation (highest to lowest value). Subfigures (b) and (d) present a summary of Friedman and Nemenyi statistical tests. The vertical line indicates the average rank of each method. Methods connected by a solid vertical black line do not have a significant difference ($p\text{-}value = 0.05$).

For both evaluations, the median values for AUC/Gamma are close to that of the Silhouettes (SWC, SSWC, ASWC, and ASSWC), which accordingly to Vendramin et al. (2010) are amongst the best relative validity criteria available. There is a clear distinction between two groups of measures. In this sense, no matter the correlation employed in the evaluation, most of the Dunn variants and $C/Sqrt(k)$ ($C/\sqrt{k}$) displayed the worst overall results. For both Pearson and Spearman correlations, AUC/Gamma was ranked 8^{th} position, regarding its average

rank. Considering the evaluation with Pearson, AUC/Gamma provided results as good as those obtained with Point-Biserial (PB), Silhouette Width Criterion (SWC), PBM, and Davies-Bouldin (DB), according to the statistical test (note that these are connected in Figure 5.2b, indicating no difference among them). In the case of Spearman correlation, the AUC/Gamma provided results similar to those of Davies-Bouldin (DB). Although AUC/Gamma did not emerge as the best measure available, it provides reasonable results, in some cases comparable to those of top performing measures. To that extent, its reduced computational cost makes it a viable additional alternative for inclusion into ensembles, such as those discussed in Chapter 4.

From the previous section we know that AUC is adjusted by change, that is, its expected value is 0.5 in the case of random cluster solutions. This information is not available for other criteria, however. In order to assess how the different measures behave in this particular case, we arbitrarily selected one dataset from the 972 synthetic data collection employed in our previous evaluation. For this particular dataset we generated partitions by randomly assigning class labels. For each number of clusters in the range of 2 to $\sqrt{n} = 23$ (the dataset has a total of 500 objects) 100 random clustering solutions were obtained. These solutions were evaluated by the relative validity criteria previously described. Results from such an evaluation are provided in Figure 5.3, where error bars account for the standard deviation. We plot results for the top criteria from previous evaluation. One can see that AUC has values close to 0.5, as already expected. Other two measures seem to have a similar behavior when evaluating random partitions, namely PB and VRC, although there is no formal proof to support such an observation. The remaining top performing measures are affected by the number of clusters under and do not provide a constant evaluation value. In a practical scenario, the lack of an expected value for a relative measure can impair evaluation, given that significance values for an evaluation cannot be easily derived for different numbers of clusters.

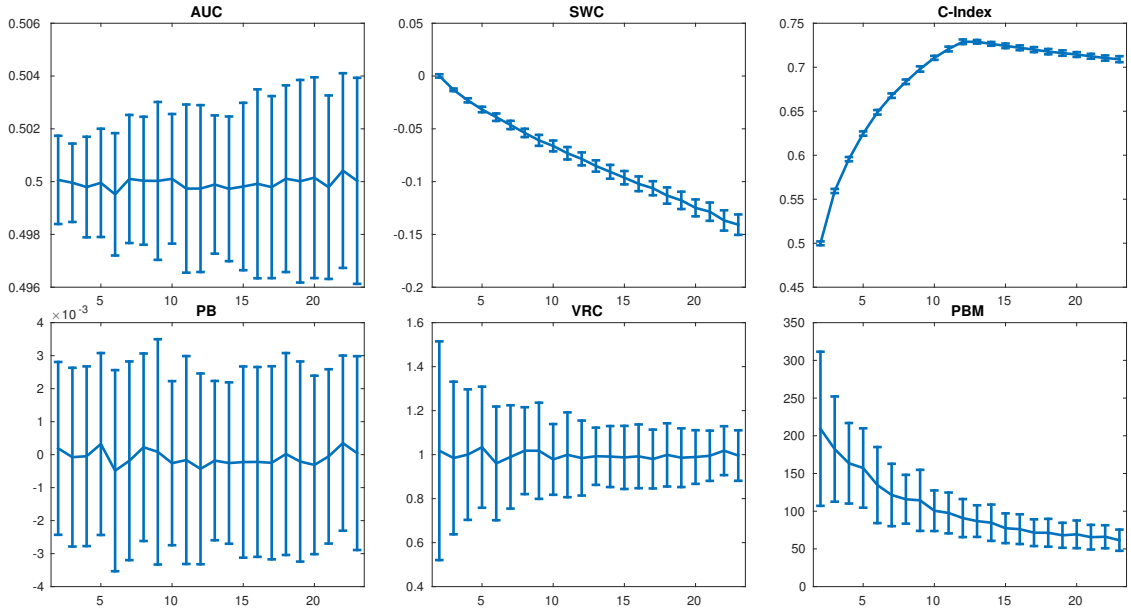


Figure 5.3: Results regarding the evaluation of randomly generated partitions. Results for other variants of the Silhouette (SWC) are suppressed given their similar results to the original measure.

5.2 Validation of Density-Based Clustering Solutions

All the relative validity criteria discussed so far have a bias towards the validation of hyperspherical shaped clusters. In a broad sense, they can be characterized as favoring partitions with a higher "within-cluster-similarity" than "between-cluster-similarity", of course, based on different formulations of such concepts. Given this particular bias, these measures cannot be employed, at least not directly as we shall discuss, to the validation of density-based clustering results, such as those generated by DBSCAN. In this section we discuss the relative validity criterion known as Density-based Clustering Validation (DBCV) (Moulavi, 2014; Moulavi et al., 2014), which was developed with participation of the author, during his one year internship at the University of Alberta, Edmonton, AB, Canada. We begin the discussion by presenting related work in Section 5.2.1. Section 5.2.2 provides details regarding the DBCV criterion, whereas in Section 5.2.3 we elaborate on how to adapt different relative validity criteria so that they can handle noise during their evaluation. Finally, in Section 5.2.4 we provide an empirical evaluation of DBCV and competitors, which are adapted to handle noise.

5.2.1 Related Work

Only a handful of works considered the relative validation of density-based clustering results. Chou et al. (2004) discusses that several existing relative validity criteria are unable to properly deal with the validation of clusters of different densities. Based on such an observation, the authors then propose a new relative validity measure that aims at the validation of clustering solutions composed of clusters with different densities. It is worth noticing that the approach introduced by the authors relies on the same rationale behind Dunn (Dunn, 1974) and Davies-Bouldin (Davies and Bouldin, 1979), and, therefore, is not capable of validating arbitrary shaped clusters. Yet another adaptation of these two relative indices was introduced by Pal and Biswas (1997). In this work, the authors keep the same basic formulations of Dunn and Davies-Boulding, but replace the actual distances between objects. These are replaced with edge weights, as obtained by graphs built on the basis of the clustering solution. Pal and Biswas (1997) consider three different approaches, namely Minimum Spanning Tree (MST), Relative Neighborhood Graph (RNG), and Gabriel Graph (GG). Note that even though graphs are employed to represent the clustering solution, the authors still employ the concept of centroid for Davies-Boudling, which is meaningless when dealing with arbitrary shaped cluster. Moreover, edge weights in the graphs are based on an Euclidean perspective of the data, disregarding density properties.

Pauwels and Frederix (1999) introduce a relative validity measure, which according to the authors is capable of identifying arbitrary shaped clusters. The measure is based on the identification of cluster isolation (analogous to separation), through means of the nearest neighbors of each object, and clustering connectivity (analogous to compactness), which is based on the distance between pairs of objects that belong to the same cluster. The measure has a major

drawback: *i.e.*, both cluster isolation and connectivity rely on the definition of parameters, which are clearly not desirable in a validation measure. Finally, to obtain the cluster connectivity, a randomized procedure is employed, introducing a non-deterministic component to the measure.

The works from Maria Halkidi and colleagues are probably the most well-known in the realm of density-based clustering validation. To that end, the authors introduced three different measures aiming the relative validation of clustering results, namely, SD (Halkidi et al., 2000), $SDbw$ (Halkidi and Vazirgiannis, 2001), and $CDbw$ (Halkidi and Vazirgiannis, 2008). From these, $CDbw$ (Composed Density between and within clusters) is the only measure capable of handling arbitrary shaped clusters. In brief, given a clustering solution, $CDbw$ determines for each cluster a set of representative objects, which aim to capture its structure, regardless of its shape. The initial representative object of each cluster is its centroid (which we note, tends to be meaningless in arbitrary shaped clusters). Further representatives are selected in order to maximize their distance to existing ones, until the desired number of representatives is selected. Based on such representatives, the authors derive three measures, which ultimately result in $CDbw$, namely, separation, compactness, and cohesion. Separation is defined on the basis of a virtual object created between the two closest representatives of two given clusters. Compactness and cohesion are defined solely on the basis of representative objects, aiming at small clusters with low variances in their density, respectively. Although interesting at a first glance, it is clear that the user has to determine the actual desired number of representatives. Apart from this, a shrinking factor must also be determined by the user, resulting in a considerable number of undesired parameters, given that we are dealing with a validation measure. Ultimately, one may start questioning which configuration of parameters yield the best validation result and how to choose it.

5.2.2 Density-Based Clustering Validation

In brief, Density-Based Clustering Validation (DBCV) is based on two main underlying ideas. The first one is to transform the original space in which the clustering solution is embedded, considering the known cluster memberships. The aim of this transformation is to capture the density of each object w.r.t. all other objects that belong to its cluster. The second one is to use the density information to derive a Minimum Spanning Tree (MST) for each cluster from the solution. Given that the MST is built considering the transformed space, both density and shape properties of each cluster are captured by the criterion. Information from the edges of the MSTs (one for each cluster) are then employed to find: (i) regions of low density inside each cluster, which will ultimately account for its density sparseness; and (ii) regions of high density between pairs of clusters, which will account for their density separation. Based on the combination of these two concepts, *i.e.*, density sparseness and density separation, one has DBCV.

More formally, for a dataset $\mathbf{X} = \{\mathbf{x_1}, \ldots, \mathbf{x_n}\}$, with n objects in the form $\mathbf{x_i} = \{x_1, \ldots, x_m\}$, we define a cluster solution $\mathcal{C} = \{C_1, \ldots, C_k\}$ containing k disjoint and non-empty clusters. We also define the set of noise objects as those that do not belong to any cluster, that is $N = \{\mathbf{x} \in$

$\mathbf{X}|\forall i : \mathbf{x} \notin C_i\}$. Based on these, we can obtain for each object a core distance that takes into account all other objects within its cluster, as given by Equation (5.2), where $d(\cdot,\cdot)$ is the Squared Euclidean Distance between objects and $|C|$ gives the size of the cluster. This definition of core distance, which is equivalent to the inverse of the object's density, is closely related to the one previously introduced by Lelis and Sander (2009). In that work, the core distance gives the smallest radius that makes the object in question a core object[2], w.r.t. $MinPts$ (user defined parameter). In the case of clustering validation, in which we have pre-defined cluster memberships, the actual number of objects within the cluster eliminates the necessity of the parameter $MinPts$.

$$CoreDist(\mathbf{x}_i) = \left(\frac{\sum_{\substack{\mathbf{x}_j \in C \\ \mathbf{x}_j \neq \mathbf{x}_i}} \left(\frac{1}{d(\mathbf{x}_i, \mathbf{x}_j)} \right)^m}{|C| - 1} \right)^{-\frac{1}{m}} \tag{5.2}$$

Having made such considerations, the all points core distance of an object captures its density properties w.r.t. all other objects within its cluster. Note, however, that objects close to $\mathbf{x}_i$ have a greater contribution to its core distance than distant ones. In order to amplify this particular effect, we consider the dimensionality of the data under evaluation, which is given by m. Following from the previous definition of all points core distance we define the mutual reachability distance between two objects, in Equation (5.3). Note that the mutual reachability distance (d_{mr}) between two objects is symmetric. The mutual reachability distance provides the basis for the construction of the Minimum Spanning Tree (MST) within each cluster, which is done as follows.

$$d_{mr}(\mathbf{x}_i, \mathbf{x}_j) = \max\{CoreDist(\mathbf{x}_i), CoreDist(\mathbf{x}_j), d(\mathbf{x}_i, \mathbf{x}_j)\} \tag{5.3}$$

Considering all objects from the clustering solution, a mutual reachability graph is built. This is a complete graph (all vertices are connected), in which vertices account for objects, with edges connecting vertices having a weight equal to their mutual reachability distance (d_{mr}). From this graph, a MST can be readily obtained for each cluster, capturing thus its shape. Once all MSTs, one for each cluster, are computed, the density sparseness of a cluster and the density separation between a pair of clusters can be derived. The density sparseness of a cluster (D_{Sp}) is given by the maximum edge value of its corresponding MST, considering only its internal vertices, *i.e.*, vertices that have a degree greater than one. The rationale behind the use of internal vertices is to aim at reducing the effect of objects that lie in the border of clusters, which can, in some cases, have a large distance to its other objects, thus affecting the evaluation. The density separation of two given clusters (D_{Sep}) is given by the minimum mutual reachability distance between the internal vertices of the clusters, capturing, therefore, the region with lowest density between the clusters.

The evaluation of a single cluster is given by the combination of its density sparseness (D_{Sp}) and its density separation (D_{Sep}) w.r.t. its closest cluster (its neighbor cluster), as provided in

[2]We recall from Chapter 2 that a core object is that with at least $MinPts$ in its neighborhood considering a radius ϵ.

Equation (5.4). The denominator in this equation is a normalization term, which makes clustering evaluations lie in the $[-1, 1]$ interval. The $\min$ in D_{Sep} accounts for the neighbor cluster of C_i.

$$V_C(C_i) = \frac{\min\limits_{j \neq i}\left(D_{Sep}(C_i, C_j)\right) - D_{Sp}(C_i)}{\max\left(\min\limits_{j \neq i}(D_{Sep}(C_i, C_j)), D_{Sp}(C_i)\right)} \tag{5.4}$$

The combination of the quality of all clusters defines DBCV, in Equation (5.5). Note that noise is implicitly considered, given that the quality of each cluster is multiplied by its percentage of objects. To that end, a clustering solution with low coverage (large number of noise objects) will be penalized, receiving a small score, even if its clusters have a good overall evaluation.

$$\text{DBCV}(\mathcal{C}) = \sum_{C_i \in \mathcal{C}} \frac{|C_i|}{|\mathbf{X}|} V_C(C_i) \tag{5.5}$$

5.2.3 Adapting Relative Validity Criteria to Handle Noise

Though not a strict rule, several density-based clustering algorithms may label objects as noise. In practice, noise objects are usually represented by a specific and unique class label, even though they may have no relation at all. Given this particular characteristic, it is interesting to note that none of the criteria introduced specifically to the density-based validation scenario account for noise in their validation procedure. Here we discuss different alternatives that can be employed to adapt such criteria, so that they can properly handle partitions containing noise. These adaptations were first envisioned as a way to provide a fair evaluation of DBCV competitors in (Moulavi et al., 2014). We believe, however, that the potential of such adaptations has not been fully discussed in that work. In fact, as we later discuss during our experimental evaluation, by selecting and applying a proper adaptation to traditional relative measures (*i.e.*, Silhouette), reasonable results can be obtained. We envision five ways in which noise can be handled, so that a relative index without noise handling capabilities can be employed to evaluate a cluster solution with noise:

1. **Assign noise to pre-existing clusters:** Following this option, noise is integrated into other clusters that were already detected by the algorithm. A straightforward choice is, therefore, to assign noise to its "closest" cluster. The main drawback from such alternative is that it directly modifies the clustering solution, ultimately leading to a new one. To that end, well separated clusters surrounded by noise can become close to each other, impairing evaluation.

2. **Assign each noise object to a singleton cluster:** This approach artificially inflates the number of clusters in the solution. Although most of the relative validity indices are not directly affected by an increase in the number of clusters from the solution, as soon as singleton noise clusters become close to real ones, the overall separation of the solution tends to decrease. Previously well-separated clusters surrounded by noise, therefore, receive a poor score, given that their closest cluster is now a singleton cluster (a noise object).

3. **Assign all noise objects to the same cluster:** With this option all noise belongs to the very same cluster. From a clustering perspective, the resulting noise cluster is meaningless, given that it is most probably sparse. From the perspective of a relative validity criterion, the noise cluster is likely to have a low level compactness. Moreover, the real clusters (as found by the clustering algorithm) would end up embedded in a single cluster of noise, which for most criteria would result in a poor evaluation, even for solutions with well separated clusters.

4. **Discard all noise objects:** If noise is simply discarded, preference will be given to solutions with a large number of noise objects. As noise objects are removed from the solution (note that these are not necessarily true noise objects), one tends to shrink real clusters, thus increasing both their compactness and separation. In an extreme case, solutions with only a handful of objects would be preferred by the evaluation measures, despite their reduced size.

5. **Discard noise with proportional penalty:** This alternative is based on the previous one, with the addition of a penalty component. By penalizing the removal of noise objects, one prevents the assignment of high scores to partitions with low coverage, *i.e.*, a large number of noise objects. This is basically the approach we adopted in the formulation of DBCV, which accounts for $(|\mathbf{X}| - |N|)/|\mathbf{X}|$, where N is the set of noise objects previously defined.

Given the discussion above, we believe that the last approach, *i.e.*, discarding noise with proportional penalty, is the best alternative available. This approach is, therefore, employed to adapt relative validity criteria that are unable to handle noise in our evaluation. Another aspect that needs to be addressed is how to handle noise during external evaluation, since we are going to perform controlled experiments considering partitions with external labeling (ground truth). Note that external validity indices, such as the Adjusted Rand Index, are not based on measures of compactness and separation, they simply count agreements between partitions. To that end, we considered each noise object as a singleton cluster when performing external evaluation.

5.2.4 Experimental Evaluation

In the following we provide an empirical evaluation of DBCV and its competitors, which were adapted to handle noise as previously specified. In order to evaluate the relative validity criteria we employed the alternative methodology, which was described in Chapter 2. Regarding clustering algorithms, we employed both DBSCAN (Ester et al., 1996) and HDBSCAN* (Campello et al., 2013, 2015). In the case of DBSCAN we considered values of $MinPts \in \{4, 6, \ldots, 18, 20\}$ with 1,000 Epsilon (ϵ) values equally distributed within the range comprising the minimum and maximum distance values for each dataset under consideration. Given that different ϵ values do not guarantee different partitions, we removed from further evaluation successive identical partitions. In the case of HDBSCAN*, we considered $MinPts = MinClSize$, within the same range as for DBSCAN. It is worth noticing that in the original DBCV paper (Moulavi

et al., 2014) successive repeated partitions from DBSCAN were not removed before evaluation with the alternative methodology. Regarding competitors we consider $CDbw$ from (Halkidi and Vazirgiannis, 2008) and 11 relative validity criteria we described in Chapter 2. We exclude here the 17 Dunn variants given their poor performance (see previous section), but do consider its original version, which we call Dunn 11.

Regarding datasets, we considered four synthetically generated datasets with two dimensions, which are depicted in Figure 5.4. These particular datasets are interesting for our evaluation given that they contain arbitrary shaped clusters, which can be later visually inspected. Apart from the synthetic datasets we employed real datasets. These are: (i) Cell237 (Yeung et al., 2003) with 237 objects, 4 cluster, and 17 features; (ii) Cell384 (Yeung et al., 2001a), with 384 objects, 5 clusters, and 17 features; (iii) The Yeast Galactose (Yeast), with 205 objects, 20 features and 4 clusters, from Yeung et al. (2001b); and four datasets from UCI Machine Learning Repository (Frank and Asuncion, 2010), namely: (iv) Iris, with 150 objects, 4 features and 3 clusters; (v) Wine, containing 178 objects, 13 features and 3 clusters; (vi) Glass, composed of 214 objects, 9 features and 7 clusters; and (vii) Control Chart (KDD), with 600 objects, 60 features and 6 clusters.

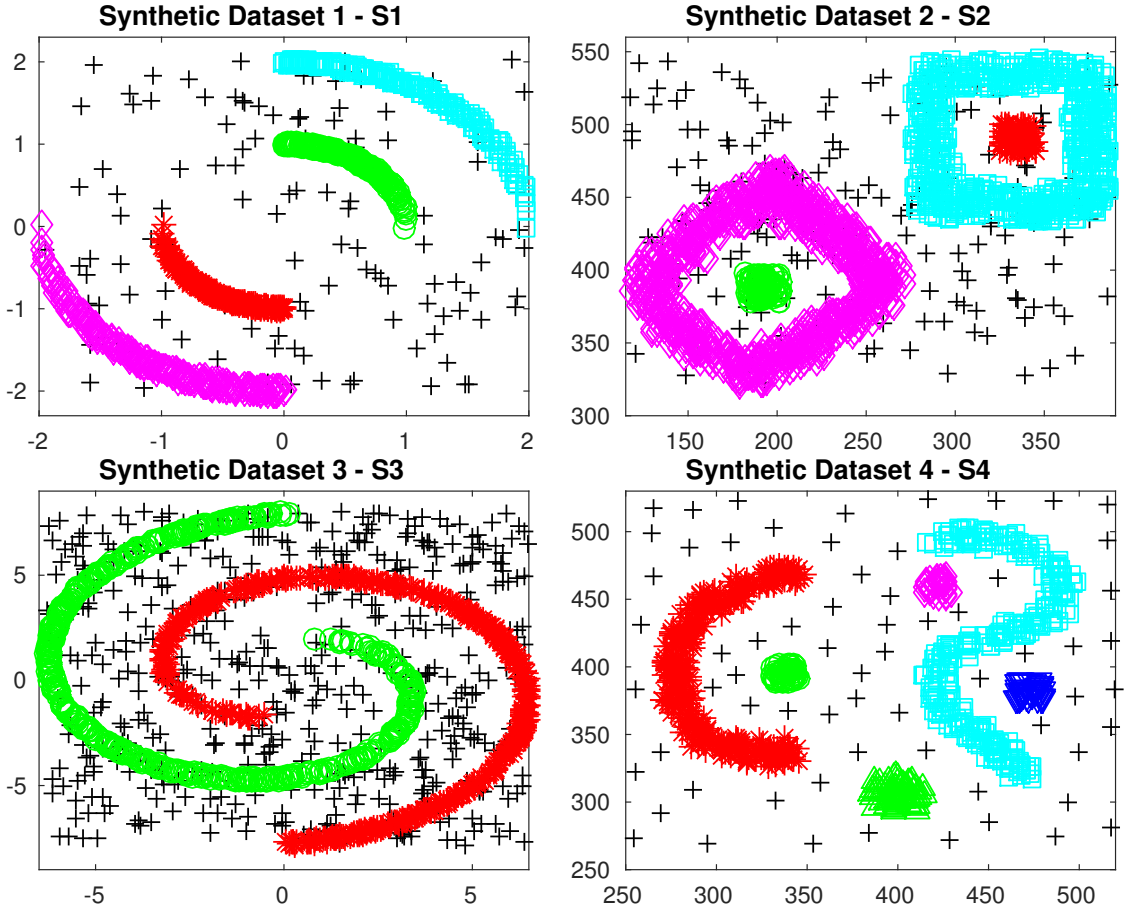


Figure 5.4: Synthetic datasets employed during DBCV evaluation (noise in black).

5.2.4.1 Results and Discussion

Before proceeding to the presentation of the results, we note that some differences are observed w.r.t. the results presented in the original DBCV paper (Moulavi et al., 2014). These arise due to the fact that the OPTICS clustering algorithm (Ankerst et al., 1999) was not employed in the evaluation we present here. Moreover, for DBSCAN we now removed successive repeated partitions before the actual evaluation. This removal can particularly affect the evaluation performed with the alternative methodology, which is based on the computation of a correlation value between the values for each relative criterion and the external validity criterion. To that end, repeated partitions provide "flat" regions with the same evaluation value, for both relative and external measures. These regions can artificially inflate the final correlation, clearly an undesired effect. It is also worth noticing that here we compare DBCV against a larger number of competitors than initially considered in its original publications (Moulavi, 2014; Moulavi et al., 2014).

Having made such considerations, we provide in Table 5.1 the best Adjusted Rand Index (ARI), as identified by each relative criterion for both synthetic and real datasets. For synthetic data, DBCV produces the best overall results, finding partitions with high ARI values. Note that even with the adaptation of other relative validity indices to handle noise, they still cannot handle arbitrary shaped partitions well. It is interesting to see that $CDbw$, which theoretically can handle arbitrary shaped clusters, has a performance similar to that of the Silhouettes for the synthetic datasets. Even with the noise adaptation the measure fails to recognize the arbitrary shapes present in the datasets. This becomes clear with the best partition found for each measure, as presented in Figure 5.5 (we limit the plots to the 4 best performing measures for the case of synthetic datasets).

Table 5.1: Best Adjusted Rand Index (ARI) value found for each relative validity criterion.

Criteria	Synthetic Datasets				Real Datasets						
	S1	S2	S3	S4	Cell237	Cell384	Yeast	Iris	Wine	Glass	KDD
DBCV	**0.93**	**0.91**	**0.74**	**0.97**	0.42	**0.39**	0.87	0.55	0.00	**0.28**	0.07
CDbw	0.43	0.73	0.02	0.59	0.09	0.32	0.72	0.52	0.25	0.26	0.07
SWC	0.43	0.71	0.17	0.34	0.55	0.33	0.85	**0.56**	**0.28**	0.24	0.36
ASWC	0.39	0.70	0.17	0.54	0.08	0.33	0.91	**0.56**	**0.28**	0.24	0.36
SSWC	0.45	0.71	0.00	0.41	0.08	0.33	0.79	0.52	0.04	0.19	0.36
ASSWC	0.01	0.70	0.00	0.02	0.08	0.33	0.79	**0.56**	0.04	0.24	0.36
DB	0.00	0.00	0.00	0.00	0.00	0.01	0.21	0.11	0.00	0.00	0.00
PBM	0.00	0.00	0.00	0.00	0.08	0.34	0.64	0.26	0.00	0.27	0.56
VRC	0.02	0.73	0.00	0.00	0.59	0.33	0.74	0.53	0.00	0.27	0.36
Dunn 11	0.00	0.00	0.00	0.00	0.08	0.36	0.38	0.41	0.00	0.16	0.58
PB	0.42	0.70	0.37	0.71	0.56	**0.39**	0.87	0.55	**0.28**	0.24	0.41
C/Sqrt(K)	0.40	0.70	0.63	0.70	**0.64**	0.33	**0.92**	**0.56**	**0.28**	0.24	0.36
C-Index	0.41	0.71	0.17	0.70	0.08	0.33	0.87	**0.56**	**0.28**	0.24	**0.61**

Note that $CDbw$ provides in some cases worse results than those obtained with SWC and C/Sqrt(K), as is the case for Synthetic Dataset 3, presented in the bottom left of Figure 5.5. In the case of real datasets, although DBCV provides in general reasonable evaluations, we note that some of the traditional relative validity measures perform better than it in particular cases. We believe that this arises due to two main reasons: (i) first, the adaptation of these measures in

order to handle noise, making them capable of evaluating density-based partitions; (ii) secondly, differently from the synthetic datasets, the real ones we employ may in fact be composed of hyperspherical clusters, favoring the original formulations of those measures. To that end, applying these particular measures with the appropriate adaptation to handle noise seems to be sufficient in several of the cases under evaluation. Of course, for arbitrary-shaped data, this is not sufficient.

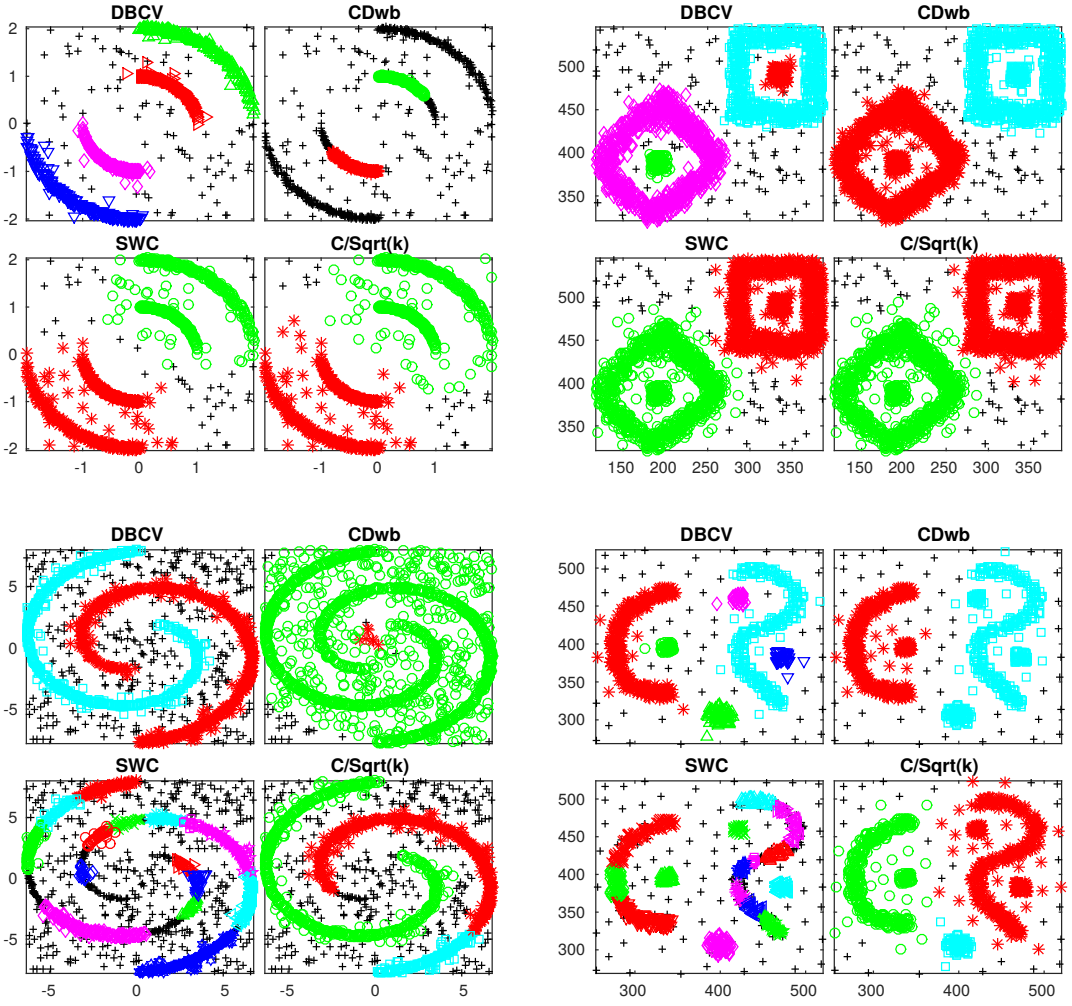


Figure 5.5: Results for the best measures regarding synthetic datasets (noise in black).

Apart from the best Adjust Rand Index (ARI) found for each pair of criterion and dataset, we also provide results regarding the evaluation of each relative criterion considering the alternative methodology (employing the Spearman correlation coefficient), as described in Chapter 2. Regarding this particular evaluation, for the synthetic datasets DBCV is among the top measures, although C/Sqrt(K) provides better results than it in some cases. In the case of the real datasets, DBCV also provides consistent correlation values w.r.t the external index, but for the Cell237 dataset. Once again, the adaptation of the relative validity indices described in Chapter 2 to handle noise seems to provide great benefits in the case of real datasets. Exceptions here are $CDbw$, Dunn 11, and DB, which provide a poor overall correlation with the external validity index.

Table 5.2: Spearman correlation with respect to the external validity index (ARI).

Criteria	Synthetic Datasets				Real Datasets						
	S1	S2	S3	S4	Cell237	Cell384	Yeast	Iris	Wine	Glass	KDD
DBCV	**0.87**	0.86	0.70	0.96	-0.02	0.34	0.61	0.87	0.78	0.77	0.62
CDbw	0.39	0.86	0.44	0.86	-0.40	-0.68	0.52	0.60	0.67	0.74	0.33
SWC	0.65	-0.59	-0.36	0.09	0.79	0.91	**0.97**	0.90	0.81	0.63	0.82
ASWC	0.71	-0.42	-0.35	0.49	0.88	0.96	0.95	0.92	0.82	0.54	0.78
SSWC	0.08	0.17	-0.72	0.71	0.90	0.87	0.76	0.83	0.64	**0.79**	0.70
ASSWC	-0.34	-0.47	-0.74	0.10	0.88	0.86	0.72	0.86	0.70	**0.79**	0.71
DB	-0.90	-0.97	-0.93	-0.97	-0.89	-0.86	-0.90	-0.91	-0.77	-0.15	-0.24
PBM	0.23	-0.27	-0.91	-0.77	0.75	-0.13	0.24	0.36	0.35	0.72	0.67
VRC	0.40	0.21	-0.85	-0.47	0.75	0.76	0.43	0.76	0.36	0.73	0.80
Dunn 11	-0.25	0.34	0.01	0.05	0.11	-0.33	-0.31	-0.12	-0.28	0.74	0.76
PB	0.73	**0.99**	0.43	0.95	0.87	**0.98**	0.92	0.93	**0.92**	0.64	**0.85**
C/Sqrt(K)	0.70	**0.99**	**0.89**	**0.97**	**0.94**	0.96	0.94	0.92	0.89	0.46	0.79
C-Index	0.73	0.95	0.61	0.89	0.91	0.94	0.93	**0.96**	0.86	0.53	0.76

5.3 Chapter Remarks

In this Chapter we presented contributions to the validation of clustering results. In the first half of the chapter (Section 5.1) we proposed and discussed the use of ROC Curves in the validation of clustering results. More specifically, we considered the Area Under the Curve (AUC) of a ROC evaluation as a relative validity criterion. We showed that the expected value for AUC is 0.5, regardless of the number of clusters under evaluation. Moreover, we showed that the AUC of a clustering result is closely related to the Gamma relative validity criterion from Baker and Hubert (1975). Although similar, the AUC of a clustering result has, however, a much lower computational complexity than that of Gamma. In the second half of the chapter (Section 5.2) we explored the validation of density-based clustering results with arbitrary shapes. The work presented in this section had as principal investigator Davoud Moulavi (Moulavi et al., 2014), (Moulavi, 2014), and was performed with the participation of this thesis' author, during his one year internship at the University of Alberta. The resulting relative validity criterion, DBCV, is capable of handling both arbitrary shapes and noise during the validation of density-based clustering results. Here we also explored the adaptation of other relative validity criteria in order to handle noise, showing that they can evaluate density-based results well, given that the data is composed of hyperspherical clusters.

CHAPTER

6

Distances for Clustering Gene Expression Data

Microarray technology has allowed researchers to gather huge amounts of data from the most diverse biological phenomena, changing the way in which biological experiments are conducted (Jiang et al., 2004; Zhang, 2006). A number of limitations from microarrays have been surpassed with the rise of RNA-Seq (Ozsolak and Milos, 2011; Wang et al., 2009; Zhang et al., 2015). Ultimately, with the fall of the cost associated with sequencing technologies, it is expected that RNA-Seq technology will replace that of microarrays in a number of applications, given its advantages (Wang et al., 2009). Notwithstanding, data collection is only the first step towards the laborious path that comprehends gene expression data analysis. In order to transform data into knowledge, efficient and effective computational methods are required. Among all the techniques employed to the analysis of gene expression data, clustering has played an important role. Due to its unsupervised nature, it is one of the first procedures applied to extract information from gene expression data, no matter its source. As an exploratory tool, clustering can help researchers to formulate new hypotheses, which ultimately can improve our understanding of gene expression.

In this particular domain of analysis, clustering has two major application scenarios. The first one is obtained when biological samples are clustered together. In this application scenario the main objective is to detect previously unknown clusters of biological samples, which are usually associated with unknown types of cancer (de Souto et al., 2008). The second clustering application is found when genes that show similar expression patterns are clustered together. In this particular

application scenario, different experiments are usually performed with the same biological sample in different time instants for a given process of interest, *e.g.*, cell cycle. Such experiments have also been employed to the study of cell responses to different types of stress conditions, *e.g.*, starvation, as well as to drug treatments, *e.g.*, Gasch et al. (2000). Since features are different time points, genes are regarded as short time-series experiments, for which distinct sampling frequencies and time resolutions may apply. The clustering of gene time-series may help, for instance, to identify genes that share the same regulatory mechanisms or functions (D'haeseleer, 2005; Heyer et al., 1999).

Taking into account the peculiarities of each one of the aforementioned scenarios, a number of clustering algorithms have been employed or developed specifically to the analysis of gene expression data (Freyhult et al., 2010; Jiang et al., 2004; Zhang, 2006). Given the plethora of clustering algorithms, a user usually faces the question: which clustering algorithm is more suited to my analysis? To answer such a question numerous theoretical and empirical studies have been conducted, *e.g.*, Costa et al. (2004); Datta (2003); D'haeseleer (2005); Freyhult et al. (2010); Kerr et al. (2008); Pirooznia et al. (2008); de Souto et al. (2008); Thalamuthu et al. (2006). Albeit important, the choice of the clustering algorithm itself is not the only factor determining the quality of clustering results. As a matter of fact, the choice of an appropriate proximity measure, whether in the form of distance or similarity employed between pairs of objects, is often regarded as a central issue in cluster analysis (D'haeseleer, 2005; Tan et al., 2006; Xu and Wunsch II, 2009; Zhang, 2006).

Bearing that in mind, in this chapter we elaborate on the selection of distance measures to the clustering of gene expression data from both microarray and RNA-Seq technologies. The work presented in this chapter extends significantly that already performed by the author during his Master's Degree (Jaskowiak, 2011). In that work, different correlation coefficients were compared with the use of four classical clustering algorithms, for both the clustering of cancer samples and genes coming from microarray experiments. In the present work we consider: (i) more distance measures; (ii) the effect of different noise levels in the performance of the measures; and (iii) datasets obtained with RNA-Seq technology (previously we only considered microarray data). Moreover, we introduce a methodology to assess the quality of different distance measures without the bias of a particular clustering algorithm when comparing different genes. This is built on the basis of external information provided by the Gene Ontology (Ashburner et al., 2000).

The remaining of this chapter is organized as follows. In Section 6.1 we discuss related work, whereas in Section 6.2 we review the distance measures that were considered during our evaluation. Different strategies for the evaluation of distance measures are discussed in Section 6.3. In this particular section we introduce an evaluation methodology based on external information from the Gene Ontology (GO) (Ashburner et al., 2000), which aims to evaluate distance measures considering pairs of genes, without the influence (bias) of any particular clustering algorithm. Finally, experimental results from the evaluation of distance measures are provided in Section 6.4 and Section 6.5, considering microarray and RNA-Seq data, respectively.

6.1 Related Work

There is no doubt that a suitable clustering algorithm is needed to achieve good quality clustering results. However, selecting a clustering algorithm is one of several *parameters* that comprise the clustering procedure. Provided that most clustering algorithms are based on distance calculations, *i.e.*, clusters are defined on the basis of distances between objects, selecting the distance between pairs of objects to be employed by the clustering algorithm is at least as important as selecting the clustering algorithm itself (Brazma and Vilo, 2000; Jain and Dubes, 1988; Jaskowiak et al., 2010; Priness et al., 2007; Steuer et al., 2002). Yet, the distance *parameter* has often been overlooked in what concerns the clustering of gene expression data, as pointed by Brazma and Vilo (2000); Priness et al. (2007); Steuer et al. (2002). If on one hand diverse studies addressed the issue of clustering algorithm selection, on the other hand just a few tried to provide guidelines regarding the selection of distances for gene expression data. Thus, when the question "which distance is more suited to my analysis?" is asked, there is no precise answer.

In view of gene expression data characteristics, objects are deemed similar if they exhibit trend or shape similarity, both for the clustering of samples and genes (Heyer et al., 1999). Although this somehow limits the number of choices from the whole universe of distance measures, there is still a considerable variety of measures capable of identifying trend similarity available in the general clustering literature. Additionally, some distances have been specifically introduced aiming the clustering of gene time-series, *e.g.*, Heyer et al. (1999), Balasubramaniyan et al. (2005), Möller-Levet et al. (2005), and Son and Baek (2008), taking into account the temporal characteristics involved in the experiments. It is worth noticing that despite the variety of distance measures available for the clustering of gene expression data, only a handful of works tried to provide guidelines concerning their choice, as we review in the following.

In the realm of microarrays, theoretical reviews highlighting the importance of selecting appropriate distances were presented by D'haeseleer (2005) and Gentleman et al. (2005). The first empirical studies concerned with the comparison of distance measures were conducted by Costa et al. (2004) and Gibbons and Roth (2002), with focus on the clustering of short gene time-series. Neither, however, considered distances that were specifically proposed to this scenario[1]. Considering the clustering of cancer samples, a handful of distance measures were evaluated by de Souto et al. (2008) and Freyhult et al. (2010), although the authors were primarily interested in the comparison of clustering algorithms rather than the distances themselves. In the works presented by Giancarlo et al. (2010) and Giancarlo et al. (2011), the authors considered a small number of distances and datasets, without any distinction between the clustering of cancer samples and the clustering of gene time-series, which are fairly different problems by nature. Once more, distance measures specifically designed for gene time-series data were not evaluated.

[1] In fact, most distance measures specifically designed for gene time-series were introduced after these studies.

Given the distinct nature of RNA-Seq and microarray data, the previous studies, and consequently their conclusions, cannot be straightforwardly transferred to RNA-Seq data. There are so far few works addressing clustering methods tailored for RNA-Seq data, not necessarily focusing on the issue of distance measure selection. In this context, Si et al. (2013) and Rau et al. (2015) examined the use of Poison and Negative Binomial distribution on a model based clustering framework. Both works aim at the problem of clustering of genes and performed small case study analysis on a few datasets. The problem of gene clustering is also evaluated by Sîrbu et al. (2012), where the performance of two clustering algorithms on three datasets measured with single and double channel microarrays, and RNA-Seq are compared. A review of clustering algorithms with illustrative examples of their application to RNA-Seq data is provided by Liu and Si (2014). Common to these works is the focus on the clustering of genes, which contrasts with lack of information regarding methods to the clustering of samples obtained with RNA-Seq.

6.2 Distance Measures

Any given object, *i.e.*, a gene or a cancer sample, can be regarded as a real valued sequence $\mathbf{x} = (x_1, \ldots, x_m)$, composed of m features. Having made such a consideration, we review in the sequel the 15 proximity measures evaluated in this chapter. First, we describe 4 "classical" proximity measures from the literature. Then, 6 correlation coefficients are reviewed. Finally, we discuss in detail 5 measures specifically proposed for the clustering of gene time-course data.

6.2.1 Classical Measures

In the sequel we review four "classical" proximity measures, all with linear time complexity.

6.2.1.1 Cosine Distance

The cosine similarity (D'haeseleer, 2005) is given by Equation (6.1) and can be regarded as the normalized inner product between objects. It is sometimes referred to as uncentered correlation or angular separation, given its relation to the Pearson correlation (D'haeseleer, 2005). Cosine measures the angle between two data objects w.r.t. the origin, whereas Pearson measures this angle considering their mean. Cosine distance (COS) is given by $COS(\mathbf{x}, \mathbf{y}) = 1 - s_c(\mathbf{x}, \mathbf{y})$.

$$s_c(\mathbf{x}, \mathbf{y}) = \frac{\sum_{i=1}^{m} x_i y_i}{\sqrt{\sum_{i=1}^{m} (x_i)^2} \sqrt{\sum_{i=1}^{m} (y_i)^2}} \tag{6.1}$$

6.2.1.2 Minkowski Distance

One of the most popular proximity indices that measures dissimilarity between two data objects is the Minkowski distance (Jain and Dubes, 1988), defined by Equation (6.2).

$$d_p(\mathbf{x}, \mathbf{y}) = \left(\sum_{i=1}^{m} |x_i - y_i|^p \right)^{1/p} \tag{6.2}$$

Note that the Minkowski dissimilarity is parametric, *i.e.*, for different values of parameter p different dissimilarity measures are obtained. In our experiments, we considered three realizations of the Minkowski dissimilarity measure (three different values of the parameter p), which are the most commonly employed ones. These measures are commonly known as Manhattan distance (MAN) for $p = 1$, Euclidean distance (EUC) for $p = 2$, and Supreme distance[2] (SUP) for $p = \infty$.

6.2.2 Correlation Coefficients

Considering gene expression data, two objects (genes or samples) are usually regarded as similar if they exhibit similarity in shape (trend), rather than in absolute differences from their values. Correlation coefficients have been widely used, given their ability to capture such a type of similarity. These measures provide values within $[-1, 1]$ and were adapted to distances as bellow:

$$\text{distance}(\mathbf{x}, \mathbf{y}) = 1 - \text{correlation coefficient}(\mathbf{x}, \mathbf{y}).$$

6.2.2.1 Pearson

Pearson correlation (PE) (Pearson, 1895) allows the identification of linear correlations between sequences. It is given in Equation (6.3), where $\bar{x}$ and $\bar{y}$ stand for the mean of the sequences. Pearson may be sensitive to outliers, thus producing false positives, *i.e.*, sequence pairs that are not alike, but receive a high correlation value (Zhang, 2006). It has $O(m)$ time complexity.

$$PE(\mathbf{x}, \mathbf{y}) = \frac{\sum_{i=1}^{m}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{m}(x_i - \bar{x})^2}\sqrt{\sum_{i=1}^{m}(y_i - \bar{y})^2}} \tag{6.3}$$

6.2.2.2 Spearman

Spearman correlation (SP) (Spearman, 1904) can be seen as a particular case of Pearson, provided that values of both sequences are replaced by their ranks. Therefore, Spearman can also be defined by Equation (6.3). As only the ranks of the sequences are considered, it is more robust to outliers than it Pearson (Zhang, 2006). Spearman has also been employed to gene expression

[2]Supreme distance is sometimes referred to as Chebyshev distance. Given two objects, it corresponds to the greatest of their differences along any given feature, *i.e.*, the maximum difference among their features.

data (Kerr et al., 2008; Zhang, 2006), though less often than Pearson. Its has $O(m\ log\ m)$ time complexity.

6.2.2.3 Goodman-Kruskal

Goodman-Kruskal (GK) (Goodman and Kruskal, 1954) takes into account only the ranks of sequences. It is defined according to the number of concordant (s_+), discordant (s_-), and neutral pairs of elements in the sequences. In a concordant pair, the same relative order applies to both sequences, *i.e.*, $x_i < x_j$ and $y_i < y_j$ or $x_i > x_j$ and $y_i > y_j$. For discordant pairs, the inverse relative order applies, *i.e.*, $x_i < x_j$ and $y_i > y_j$ or $x_i > x_j$ and $y_i < y_j$. All other pairs are deemed neutrals. The measure is given by Equation (6.4), having a $O(m\ log\ m)$ time complexity.

$$GK(\mathbf{x}, \mathbf{y}) = \frac{s_+ - s_-}{s_+ + s_-} \tag{6.4}$$

6.2.2.4 Kendall

The Kendall correlation coefficient (KE) (Kendall, 1938) is based on the same building blocks used by GK. Kendall is defined in Equation (6.5), where $m(m-1)/2$ is the total number of pairs of elements in the sequences. Note that, differently from GK, extreme correlation values are obtained only in the absence of neutrals. As for GK, Kendall can be computed in $O(m\ log\ m)$ time.

$$KE(\mathbf{x}, \mathbf{y}) = \frac{s_+ - s_-}{m(m-1)/2} \tag{6.5}$$

6.2.2.5 Rank-Magnitude

The Rank-Magnitude correlation was proposed by Campello and Hruschka (2009) as an asymmetric measure, for cases in which one of the sequences is composed of ranks and the other by real values. It is given by Equation (6.6), with $R(x_i)$ denoting the rank of the i^{th} position for sequence $\mathbf{x}$, $r^{min} = \sum_{i=1}^{m}(m+1-i)y_i^s$ and $r^{max} = \sum_{i=1}^{m} i y_i^s$. Value y_i^s corresponds to the i^{th} element of the sequence obtained by rearranging sequence $\mathbf{y}$ in ascending order.

$$r_m(\mathbf{x}, \mathbf{y}) = \frac{2\sum_{i=1}^{m} R(x_i)y_i - r^{max} - r^{min}}{r^{max} - r^{min}} \tag{6.6}$$

We employ a symmetric version of the measure, given by $RM(\mathbf{x}, \mathbf{y}) = (r_m(\mathbf{x}, \mathbf{y}) + r_m(\mathbf{y}, \mathbf{x}))/2$. This version can be used to compare two real valued sequences, taking both their magnitudes and ranks into consideration. As the original measure, RM has $O(m\ log\ m)$ time complexity.

6.2.2.6 Weighted Goodman-Kruskal

The Weighted Goodman-Kruskal correlation (WGK) (Campello and Hruschka, 2009) is given in Equation (6.7). It takes into account ranks and magnitudes of both sequences. Term $\hat{w}_{ij}$ is

defined by Equation (6.8). Terms $\hat{w}_{ij}^{\mathbf{x}}$ and $\hat{w}_{ij}^{\mathbf{y}}$ in Equation (6.8) are defined by Equation (6.9) and represent the signed percentage differences between the values of the i^{th} and j^{th} elements of the corresponding sequences. Term w_{ij} in Equation (6.7) is defined by Equation (6.10), with $w_{ij}^{\mathbf{x}} = sign(x_i - x_j)$ and $w_{ij}^{\mathbf{y}} = sign(y_i - y_j)$. The measure has $O(m^2)$ time complexity.

$$WGK(\mathbf{x}, \mathbf{y}) = \frac{\sum_{i=1}^{m-1} \sum_{j=i+1}^{m} \hat{w}_{ij}}{\sum_{i=1}^{m-1} \sum_{j=i+1}^{m} |w_{ij}|} \tag{6.7}$$

$$\hat{w}_{ij} = \begin{cases} min\left\{\frac{\hat{w}_{ij}^{\mathbf{x}}}{\hat{w}_{ij}^{\mathbf{y}}}, \frac{\hat{w}_{ij}^{\mathbf{y}}}{\hat{w}_{ij}^{\mathbf{x}}}\right\} & \text{if } \hat{w}_{ij}^{\mathbf{x}}\ \hat{w}_{ij}^{\mathbf{y}} > 0 \\ max\left\{\frac{\hat{w}_{ij}^{\mathbf{x}}}{\hat{w}_{ij}^{\mathbf{y}}}, \frac{\hat{w}_{ij}^{\mathbf{y}}}{\hat{w}_{ij}^{\mathbf{x}}}\right\} & \text{if } \hat{w}_{ij}^{\mathbf{x}}\ \hat{w}_{ij}^{\mathbf{y}} < 0 \\ 1 & \text{if } \hat{w}_{ij}^{\mathbf{x}} = \hat{w}_{ij}^{\mathbf{y}} = 0 \\ 0 & \text{otherwise} \end{cases} \tag{6.8}$$

$$\hat{w}_{ij}^{\mathbf{x}} = \begin{cases} \frac{x_i - x_j}{x_{max} - x_{min}} & \text{if } x_{max} \neq x_{min} \\ 0 & \text{otherwise} \end{cases} \tag{6.9}$$

$$w_{ij} = \begin{cases} w_{ij}^{\mathbf{x}} / w_{ij}^{\mathbf{y}} & \text{if } w_{ij}^{\mathbf{y}} \neq 0 \\ 1 & \text{if } w_{ij}^{\mathbf{x}} = w_{ij}^{\mathbf{y}} = 0 \\ 0 & \text{otherwise} \end{cases} \tag{6.10}$$

6.2.3 Time-Course Specific Measures

We review measures specifically proposed for clustering gene time-series. For these $\mathbf{t} = (t_1, \ldots, t_m)$ gives the time instants at which each feature (sample) is measured (taken) for a gene.

6.2.3.1 Jackknife

The idea behind the Jackknife correlation (Heyer et al., 1999) is to minimize the effect of single outliers on the final correlation value by removing one single feature at a time from both sequences. If the sequences do not contain outliers, their correlation value remains stable, otherwise, feature removal causes a decrease in their correlation, indicating that the sequences were correlated partly due to the presence of outliers. Jackknife is given by Equation (6.11), where $PE^i(\mathbf{x}, \mathbf{y})$ is the Pearson correlation between $\mathbf{x}$ and $\mathbf{y}$ with their i^{th} values removed. It has $O(m^2)$ time complexity.

$$JK(\mathbf{x}, \mathbf{y}) = min\{PE^1(\mathbf{x}, \mathbf{y}), \ldots, PE^m(\mathbf{x}, \mathbf{y}), PE(\mathbf{x}, \mathbf{y})\} \tag{6.11}$$

6.2.3.2 Short Time-Series Dissimilarity

The Short Time-Series dissimilarity (STS) was proposed by Möller-Levet et al. (2005) and measures the distance between the $m-1$ slopes that compound two gene time-series. For two genes

$\mathbf{x}$ and $\mathbf{y}$, STS is given by Equation (6.12). The greater the interval between the measurements, the lower its impact on the dissimilarity. The measure has a linear time complexity.

$$STS(\mathbf{x},\mathbf{y}) = \sqrt{\sum_{i=1}^{m-1}\left(\frac{y_{i+1}-y_i}{t_{i+1}-t_i} - \frac{x_{i+1}-x_i}{t_{i+1}-t_i}\right)^2} \tag{6.12}$$

6.2.3.3 Local Shape-based Similarity

Based on the observation that biological relationships between genes may be present in the form of local and possibly shifted similarity patterns, Balasubramaniyan et al. (2005) introduced the concept of Local Shape-based Similarity (LSS). LSS seeks the most similar subsequences of size s in sequences $\mathbf{x}$ and $\mathbf{y}$. It is defined by Equations (6.13) and (6.14), where Sim_{base} is the base similarity employed between subsequences of a given size s. The minimum subsequence size is given by s_{min}, which is usually set to $m-2$, allowing for two time instant shifts (Balasubramaniyan et al., 2005). Note that although subsequences must have the same sizes, they do not have to be aligned, thus allowing locally shifted similarity patterns.

$$LSS(\mathbf{x},\mathbf{y}) = \max_{s_{min}\leq s\leq m} SIM_s(\mathbf{x},\mathbf{y}) \tag{6.13}$$

$$SIM_s(\mathbf{x},\mathbf{y}) = \max_{1\leq i,j\leq m-s+1} Sim_{base}(\mathbf{x}[i,i+s-1],\mathbf{y}[j,j+s-1]) \tag{6.14}$$

In Equation (6.13) LSS is given by the maximum similarity among sequences of different sizes. Since for shorter sequences greater similarities are more likely, LSS is derived from the probability of obtaining a specific similarity value for a given subsequence size. Similarity Sim_{base} is given by the probability of obtaining a value of Spearman correlation for the subsequences under comparison, as further detailed in (Balasubramaniyan et al., 2005). LSS time complexity is $O(m^3)$, but it can be reduced by the use of approximations (Balasubramaniyan et al., 2005).

6.2.3.4 YR1 and YS1 Dissimilarities

Based on the presumption that correlations may not capture all information contained in gene time-series, Son and Baek (2008) introduced two dissimilarities that combine different types of information with correlation values. The first information taken into account concerns the agreement between the $m-1$ slopes that compose the two gene series under comparison, as given by Equation (6.15). Function L is given by Equation (6.16). Function $\mathcal{I}$ returns 1 in case of agreement and 0 otherwise. The slope of each gene $\mathbf{x}$, for a given interval, is given by Equation (6.17).

$$A(\mathbf{x},\mathbf{y}) = \sum_{i=1}^{m-1}\frac{\mathcal{I}(L(\mathbf{x},i)=L(\mathbf{y},i))}{m-1} \tag{6.15}$$

$$L(\mathbf{x},i) = \begin{cases} 1 & \text{if } \textit{slope}(\mathbf{x},i) > 0 \\ -1 & \text{if } \textit{slope}(\mathbf{x},i) < 0 \\ 0 & \text{if } \textit{slope}(\mathbf{x},i) = 0 \end{cases} \quad (6.16) \qquad \textit{slope}(\mathbf{x},i) = \frac{x_{i+1} - x_i}{t_{i+1} - t_i} \quad (6.17)$$

The second information concerns the agreement of maximum (t^{max}) and minimum (t^{min}) expression levels of both genes ($\mathbf{x}$ and $\mathbf{y}$), as given by Equation (6.18).

$$M(\mathbf{x},\mathbf{y}) = \begin{cases} 1 & \text{if } t^{min}_{\mathbf{x}} = t^{min}_{\mathbf{y}} \text{ and } t^{max}_{\mathbf{x}} = t^{max}_{\mathbf{y}} \\ 0.5 & \text{if } t^{min}_{\mathbf{x}} = t^{min}_{\mathbf{y}} \text{ or } t^{max}_{\mathbf{x}} = t^{max}_{\mathbf{y}} \\ 0 & \text{if } t^{min}_{\mathbf{x}} \neq t^{min}_{\mathbf{y}} \text{ and } t^{max}_{\mathbf{x}} \neq t^{max}_{\mathbf{y}} \end{cases} \quad (6.18)$$

The two proposed measures combine Equations (6.15) and (6.18) with correlation coefficients, as given by Equations (6.19) and (6.20). For *YR*1 and *YS*1 the authors also consider, respectively, PE and SP, in the forms: $R(\mathbf{x},\mathbf{y}) = (PE(\mathbf{x},\mathbf{y}) + 1)/2$ and $S(\mathbf{x},\mathbf{y}) = (SP(\mathbf{x},\mathbf{y}) + 1)/2$.

$$YR1(\mathbf{x},\mathbf{y}) = \omega_1 R(\mathbf{x},\mathbf{y}) + \omega_2 A(\mathbf{x},\mathbf{y}) + \omega_3 M(\mathbf{x},\mathbf{y}) \quad (6.19)$$

$$YS1(\mathbf{x},\mathbf{y}) = \omega_1 S(\mathbf{x},\mathbf{y}) + \omega_2 A(\mathbf{x},\mathbf{y}) + \omega_3 M(\mathbf{x},\mathbf{y}) \quad (6.20)$$

In Equations (6.19) and (6.20), the weight terms ω_1, ω_2, and ω_3 must satisfy the condition $\sum_{i=1}^{3} \omega_i = 1$. Son and Baek (2008) proposed a way to estimate their values, but their approach is highly computationally demanding. To compare the two measures we set weights in the following form: $\omega_1 = 0.5$, $\omega_2 = 0.25$, and $\omega_3 = 0.25$, which are values employed by the authors in (Son and Baek, 2008). *YR*1 has $O(m)$ time complexity whilst *YS*1 has $O(m\ log\ m)$ time complexity.

6.3 Distance Measures Evaluation

In order to assess the behavior of the different distance measures previously reviewed we take into account two different approaches. The first one tries to evaluate distances disregarding the effect of any particular clustering algorithm. This approach, which we refer to as clustering algorithm independent, is detailed in Section 6.3.1. The second approach, which we refer to as clustering algorithm dependent, takes into account both the bias of the clustering algorithm and the distance measure. It is reviewed in Section 6.3.2.

6.3.1 Clustering Algorithm Independent

To evaluate distance measures independently of the bias of clustering algorithms we resort to methodologies based on their intrinsic separation abilities. More specifically, two different methodologies are employed. The first one, referred to as Intrinsic Separation Ability (ISA)

is employed when class labels (ground truth) are available. This particular methodology was introduced by Giancarlo et al. (2011, 2010). The second methodology, which we refer to as Intrinsic Biological Separation Ability (IBSA), was developed by the author of the thesis in order to circumvent the need of class labels from ISA. Although, at first, ISA can be employed to evaluate distances for both samples and genes, we note that class labels are usually available only for the case of samples. In the case of genes, even for controlled experiments, ground truth partitions cannot be easily obtained. This motivated the development of IBSA, which employs information from the Gene Ontology (GO) (Ashburner et al., 2000), avoiding the need of class labels. In brief, IBSA assigns class labels according to semantic similarities among genes extracted from the GO (biological information), which explains the addition of the term biological in its name. In the sequel we describe ISA in detail and introduce our methodology, namely IBSA.

6.3.1.1 Intrinsic Separation Ability (ISA)

The Intrinsic Separation Ability (ISA) of a distance indicates how well it can separate samples without the influence of a clustering algorithm (Giancarlo et al., 2011, 2010). Given a dataset with n objects (samples in this case) $\mathbf{x}_1, \ldots, \mathbf{x}_n$, we build a distance matrix $\mathbf{D}$, where $\mathbf{D}(i,j) = distance(\mathbf{x}_i, \mathbf{x}_j)$, with $1 \leq i, j \leq n$. Assuming that all the values of $\mathbf{D}$ are in the $[0,1]$ interval (if they are not, they must be normalized), we proceed and build a binary classifier that assigns a pair of objects (samples) to a given class according to Equation (6.21), where ϕ_1 is a given threshold in the $[0,1]$ interval. By Applying Equation (6.21) to all pairs of objects from a dataset with a *fixed* threshold we obtain a predicted solution, based solely on object distances.

$$I_{\phi_1}(\mathbf{x}_i, \mathbf{x}_j) = \begin{cases} 1 & \text{if } \mathbf{D}(i,j) \leq \phi_1 \\ 0 & \text{otherwise} \end{cases} \tag{6.21}$$

Provided that we are dealing with labeled data in the case of ISA, we can proceed and build a desired solution for the classifier previously described. The desired solution is built upon the golden standard partition of each dataset, given by Equation (6.22) for all $\mathbf{x}_i$ and $\mathbf{x}_j$.

$$J(\mathbf{x}_i, \mathbf{x}_j) = \begin{cases} 1 & \text{if } \mathbf{x}_i \text{ and } \mathbf{x}_j \text{ belong} \\ & \text{to the same cluster} \\ 0 & \text{otherwise} \end{cases} \tag{6.22}$$

Note that by setting a threshold ϕ_1 and applying Equations (6.21) and (6.22) to all object pairs we have a predicted and a desired solution, respectively. For Equation (6.21), however, the predicted solution is not unique, as different threshold values are possible. We consider all possible values[3] of ϕ_1 in the $[0,1]$ interval, generating a set of all possible predicted solutions for a given distance, *i.e.*, one for each different value of ϕ_1. To evaluate ISA for a given distance, we have to compare its

[3]The possible values of ϕ_1 are those in the finite set of values contained in matrix $\mathbf{D}$.

predicted solutions against the expected one. In brief, we have multiple comparisons to perform, one per each value of ϕ_1. These comparisons can be addressed by employing the well established concept of Receiver Operating Characteristics analysis, ROC analysis for short (Giancarlo et al., 2011, 2010; Hand and Till, 2001). The whole procedure is as follows. First, given a threshold (ϕ_1), we compute the number of True Positives (TP), False Positives (FP), True Negatives (TN), and False Negatives (FN). More specifically, these values are given by:

$$\begin{aligned} TP = & \sum_{\forall i,j,i\neq j} I_{\phi_1}(\mathbf{x}_i, \mathbf{x}_j) J(\mathbf{x}_i, \mathbf{x}_j) \\ FP = & \sum_{\forall i,j,i\neq j} I_{\phi_1}(\mathbf{x}_i, \mathbf{x}_j)(1 - J(\mathbf{x}_i, \mathbf{x}_j)) \\ TN = & \sum_{\forall i,j,i\neq j}(1 - I_{\phi_1}(\mathbf{x}_i, \mathbf{x}_j))(1 - J(\mathbf{x}_i, \mathbf{x}_j)) \\ FN = & \sum_{\forall i,j,i\neq j}(1 - I_{\phi_1}(\mathbf{x}_i, \mathbf{x}_j)) J(\mathbf{x}_i, \mathbf{x}_j) \end{aligned}$$

Then we compute the False Positive Rate, given by $FPR = FP/(FP + TN)$, and the True Positive Rate, given by $TPR = TP/(TP + FN)$. Values of FPR and TPR are computed for all the values of ϕ_1. With these values in hand, we plot an ROC Curve (FPR vs TPR), which is summarized by its Area Under the Curve (AUC). AUC values are in the $[0, 1]$ interval. In this particular case, an AUC value of 1 indicates a distance measure that perfectly separates samples according to the desired solution. On the other hand, an AUC value close to or lower than 0.5 indicates a distance measure that fails to separate objects according to the desired solution. Having made such considerations, Giancarlo et al. (2010) defines the ISA of a distance as its AUC.

6.3.1.2 Intrinsic Biological Separation Ability (IBSA)

The Intrinsic Separation Ability can be computed only for datasets with a golden standard partition, *i.e.*, datasets for which class labels are available. Note that for most gene clustering problems, as time-series datasets, no class labels are available. Therefore, we take advantage of the information provided by the Gene Ontology (GO) (Ashburner et al., 2000) to overcome the lack of labeled data and devise a new procedure to evaluate the ISA of a distance regarding the clustering of genes. This new procedure is called *Intrinsic Biological Separation Ability* (IBSA). Instead of using class labels, our methodology employs external biological information (semantic similarities among genes) extracted from the GO. Note that since IBSA employs information from the GO to evaluate a particular proximity measure, it tends to favor proximity measures that are in agreement with GO external information. If the user is interested in finding a different type of structure in the data (not related with GO), another methodology should be selected and employed.

Given a dataset with n objects (genes, in this particular case) $\mathbf{x}_1, \ldots, \mathbf{x}_n$ we can build a distance matrix $\mathbf{D}$, where $\mathbf{D}(i, j) = distance(\mathbf{x}_i, \mathbf{x}_j)$, with $1 \leq i, j \leq n$. Assuming that all the values of $\mathbf{D}$ are in the $[0, 1]$ interval, all pairs of objects can be distinguished by the same binary classifier previously described by Equation (6.21). In brief, object pairs are assigned to class 1 if the distance between them is smaller than or equal to a given threshold ϕ_1 in the $[0, 1]$ interval and 0 otherwise.

By applying this equation to all object pairs from a given dataset (with a *fixed* threshold) we obtain a predicted solution based solely on the distances between object pairs.

In order to build a desired solution for this classifier, the first step of our methodology consists in obtaining biological dissimilarities for all *pairs of genes* from the dataset in hand, devising a biological dissimilarity matrix ($\mathbf{B}$). Considering the Gene Ontology, several proximity measures can be employed to quantify the degree of concordance between the *sets of terms* that annotate any two genes. By combining dissimilarities that operate between *sets of terms* it is possible to measure the degree of concordance between any two genes (Pesquita et al., 2009). Note that the methodology presented here is the same regardless of the biological similarity employed between genes. We elaborate on the choice of the biological measure after describing the full procedure, at the end of this Section. Once a biological dissimilarity matrix is available, it can be interpreted as external information and fill the gap left by the lack of class labels. For a given biological dissimilarity matrix ($\mathbf{B}$) with values in the $[0, 1]$ interval we proceed and build a desired biological solution, as specified by Equation (6.23), where ϕ_2 is a threshold in the $[0, 1]$ interval. By applying Equation (6.23) to all pairs of objects from a given dataset (with a *fixed* threshold) we obtain a desired biological solution, based on external information from the GO.

$$J_{\phi_2}(\mathbf{x}_i, \mathbf{x}_j) = \begin{cases} 1 & \text{if } \mathbf{B}(i,j) \leq \phi_2 \\ 0 & \text{otherwise} \end{cases} \tag{6.23}$$

Note that by fixing thresholds ϕ_1 and ϕ_2 and applying Equations (6.21) and (6.23) to all object pairs we have a predicted and a desired biological solution, respectively. As we are dealing with two thresholds, there exist two sets of solutions: (i) a set of predicted solutions, obtained when Equation (6.21) is applied to all pairs of objects with ϕ_1 taking all its possible values, and (ii) a set of desired biological solutions, obtained by applying Equation (6.23) to all pairs of objects with all possible values for ϕ_2. All the possible values for ϕ_1 and ϕ_2 are the values of the elements in matrices $\mathbf{D}$ and $\mathbf{B}$, respectively. To evaluate the IBSA of a given distance its set of predicted solutions must be compared against the set of expected ones, obtained from the Gene Ontology.

Since two thresholds are employed in our methodology, multiple ROC analyses must be performed for the different values of ϕ_1 and ϕ_2. First, we fix a value for ϕ_2 obtaining a desired biological solution. With a desired solution in hand, the analysis is performed as previously described for ISA, *i.e.*, (i) with a fixed biological solution, we set a value for ϕ_1 and obtain the values of True Positives, False Positives, True Negatives, and False Negatives, which are given by:

$$\begin{aligned} TP = &\ \textstyle\sum_{\forall i,j,i\neq j} I_{\phi_1}(\mathbf{x}_i, \mathbf{x}_j) J_{\phi_2}(\mathbf{x}_i, \mathbf{x}_j) \\ FP = &\ \textstyle\sum_{\forall i,j,i\neq j} I_{\phi_1}(\mathbf{x}_i, \mathbf{x}_j)(1 - J_{\phi_2}(\mathbf{x}_i, \mathbf{x}_j)) \\ TN = &\ \textstyle\sum_{\forall i,j,i\neq j}(1 - I_{\phi_1}(\mathbf{x}_i, \mathbf{x}_j))(1 - J_{\phi_2}(\mathbf{x}_i, \mathbf{x}_j)) \\ FN = &\ \textstyle\sum_{\forall i,j,i\neq j}(1 - I_{\phi_1}(\mathbf{x}_i, \mathbf{x}_j)) J_{\phi_2}(\mathbf{x}_i, \mathbf{x}_j) \end{aligned}$$

We then compute the False Positive Rate ($FPR = FP/(FP + TN)$) and the True Positive Rate ($TPR = TP/(TP + FN)$) for different values of ϕ_1; (ii) we plot an ROC Curve and obtain its corresponding AUC value. This procedure is repeated for all different values of ϕ_2, which correspond to the set of desired biological solutions, as provided by the GO. Finally, distances are evaluated by the average of all their AUCs, which lie in the $[0, 1]$ interval. By applying such a methodology we can verify whether the IBSA of a particular distance is in agreement with the biological distance extracted from the GO. Note that this evaluation method also avoids the bias of employing a particular clustering algorithm. We refer to this novel methodology as the *Intrinsic Biological Separation Ability* of a proximity measure (distance), IBSA for short.

Recall that IBSA requires the definition of a biological proximity measure between genes to generate a biological dissimilarity matrix for each dataset. Here we employ the measure proposed by Resnik (1999), as previous work has shown that it correlates best with both gene expression similarity patterns (Sevilla et al., 2005) and gene sequence pattern similarities (Pesquita et al., 2008, 2009). We used the Best-Match Average of the Resnik measure, as it provides better results than other approaches employed to combine ontology term similarities (Pesquita et al., 2008).

Given two Gene Ontology terms t_1 and t_2, Resnik's similarity is given by Equation (6.24), where $S(\mathrm{t}_1, \mathrm{t}_2)$ is the set of terms that subsume both t_1 and t_2, *i.e.*, common ancestors of both t_1 and t_2. For a given common ancestor t, $p(\mathrm{t})$ is the probability of annotating a gene with it, whereas $[-\log p(\mathrm{t})]$ is usually referred to as the Information Content (IC) of term t. For our experiments $p(\mathrm{t})$ (probability values) were estimated with the `GOSim` R Package from Frohlich et al. (2007). Such an estimation is based on empirical observations regarding the number of times that a GO term annotates a gene (Frohlich et al., 2007). Resnik's similarity seeks for the common ancestor t with greatest IC. In fact, Resnik's similarity between two terms is equal to the greatest Information Content (IC) amongst the term's common ancestors.

$$\mathrm{Resnik}_{\mathrm{Sim}}(\mathrm{t}_1, \mathrm{t}_2) = \max_{\mathrm{t} \in S(\mathrm{t}_1, \mathrm{t}_2)} [-\log p(\mathrm{t})] \tag{6.24}$$

Note that Resnik's similarity is computed only between two terms. As one gene may be annotated with a set of terms, we need to obtain gene similarities based on the sets of terms that annotate each gene under consideration. To obtain *gene similarities*, we combined term similarities, obtained with Equation (6.24), employing the Best-Match Average (BMA) (Pesquita et al., 2008) of the Resnik similarity, as given by Equation (6.25). In this equation $\mathcal{S}_1$ and $\mathcal{S}_2$ are the sets of GO terms that annotate genes g_1 and g_2, respectively. For use with IBSA, BMA dissimilarities can be easily obtained from $\mathrm{BMA}_{\mathrm{Dis}} = 1 - \mathrm{BMA}_{\mathrm{Sim}}$.

$$\begin{aligned} \mathrm{BMA}_{\mathrm{Sim}}(g_1, g_2) = & \frac{1}{2} \mathrm{avg}_{i \in \mathcal{S}_1} (\max_{j \in \mathcal{S}_2} \mathrm{Resnik}_{\mathrm{Sim}}(\mathrm{t}_i, \mathrm{t}_j)) + \\ & \frac{1}{2} \mathrm{avg}_{j \in \mathcal{S}_2} (\max_{i \in \mathcal{S}_1} \mathrm{Resnik}_{\mathrm{Sim}}(\mathrm{t}_i, \mathrm{t}_j)) \end{aligned} \tag{6.25}$$

Finally, we computed IBSA for time-course datasets considering the Molecular Function (MF) and Biological Process (BP) ontologies separately. Cellular Component (CC) ontology was left out as it usually shows little correlation with gene expression data (Bolshakova et al., 2005, 2006b).

6.3.2 Clustering Algorithm Dependent

In the clustering algorithm dependent approach we consider the combined effect of both clustering algorithm and distance measure. In this particular evaluation we employ different combinations of clustering algorithms and distance measures and evaluate their results externally. Given that different clustering algorithms provide different biases, it is virtually impossible to evaluate proximity measures considering all the different clustering algorithms proposed in the literature. To that end, we limit ourselves to four well-known clustering algorithms, namely Single-Linkage (SL), Average-Linkage (AL), Complete-Linkage (CL), and k-medoids (these were described in Chapter 2). Here we employ k-medoids in order to avoid convergence problems with k-means, given that different distances are being employed (this is discussed in detail in Chapter 2).

In the case of sample clustering the Adjusted Rand Index (ARI) can be employed as a measure of quality, given the availability of datasets with ground truth. We consider three different evaluation scenarios. In the first one, we generate partitions containing the same number of clusters as defined by the reference partition, *i.e.*, the original labeling of each dataset. Resulting partitions are then compared based on their ARI values, which evaluate the capability of each distance in recovering partitions in conformity with the structure defined in the ground truth. In the second evaluation scenario we choose for further comparison partitions that provide the best ARI values, regardless of their number of clusters. For each dataset we generate partitions within the interval $[2, \lceil\sqrt{n}\rceil]$, where n stands for the number of objects. Note that partitions with number of clusters different from those found in the reference partition may, in certain cases, contain more natural clusters than those found in a partition with the "correct" number of clusters. Finally, in the third evaluation scenario, we simulate a real application in which the user does not have any *a priori* information about the number of clusters in the data. For each dataset we generate partitions within the interval $[2, \lceil\sqrt{n}\rceil]$. Differently from the previous scenario, however, the best partition for each pair of cluster algorithm and distance is chosen by the Silhouette Width Criterion (SWC).

In the case of gene clustering we recall that external labels are usually unavailable. To that end we once more resort to the Gene Ontology (Ashburner et al., 2000) in order to validate results. Here we consider only the third evaluation scenario (estimated number of clusters). Given that we do not have a reference partition we cannot employ an external criterion to evaluate the quality of clustering results, *i.e.*, in this case we cannot employ ARI to validate the results. To compare the results obtained with the different pairs of clustering algorithms and distances, we adopt a heuristic similar to the one used by Ernst and Bar-Joseph (2006) and Costa et al. (2008). The procedure is as follows. For each clustering result we perform a gene enrichment analysis (Beißbarth and Speed,

2004) and obtain the respective list of enriched terms that have a $p\text{-}value \leq 0.05$ within each cluster (the Fisher's Exact Test is employed to determine whether the difference is significant or not (Beißbarth and Speed, 2004)). To perform the gene enrichment analysis we use the well-known *GOstat* tool from (Beißbarth and Speed, 2004). For two results r_1 and r_2, we count the number of times that r_1 provided enrichments with smaller $p\text{-}values$ than r_2 and the number of times that r_2 provided enrichments with smaller $p\text{-}value$ than r_1, these are then combined as bellow.

$$\text{Comparison}(r_1, r_2) = \log\left(\frac{\#(p\text{-}value\ r_1 < p\text{-}value\ r_2)}{\#(p\text{-}value\ r_2 < p\text{-}value\ r_1)}\right) \tag{6.26}$$

Note that changing the order of the results under comparison (r_1,r_2) or (r_2,r_1) changes only the sign of the result, not its absolute value. For this comparison procedure, positive values mean that r_1 is better than r_2, whereas negative values means the opposite. In brief, the evaluation procedure for gene time-series data is as follows: (i) the best partition for each pair of clustering algorithm and distance (as chosen by the Silhouette) is selected for further comparison; (ii) we evaluate all pairs of results obtained based on the previous heuristic. Such an evaluation is made on the basis of Equation (6.26); (iii) finally, we compare the values obtained for all pairs of results from step (ii).

6.4 Experiments on Microarray Data

In the following we provide results regarding the evaluation of distances for microarray data.

6.4.1 Experimental Setup

For the evaluation regarding microarray datasets we employ two different data collections. Regarding the clustering of samples, we considered the 35 benchmark datasets introduced by de Souto et al. (2008). In brief, this publicly available benchmark collection encompasses 35 microarray datasets from cancer gene expression experiments and comprehend the two platforms in which the technology is generally available, *i.e.*, Affymetrix (21 datasets) and cDNA (14 datasets) (Zhang, 2006). All datasets are already preprocessed, with the most significant preprocessing performed by de Souto et al. (2008) being related to the removal of uninformative genes (genes that are not differentially expressed across samples). A summary of the 35 cancer datasets is shown in Table 6.1. Further details are provided in (de Souto et al., 2008).

For the clustering of genes, we considered time-course datasets from three different sources (the three sources are shown in Table 6.2). These datasets are from cDNA microarrays experiments regarding the *Saccharomyces cerevisiae* organism (yeast). Two data sources comprehend multiple time-series experiments and are further divided into single experiment datasets. Overall there is a total of 17 datasets. Before performing the experiments, we removed genes for which 10% or more expression values are missing. After this removal no gene with missing values remained. Finally, for further analysis, we selected genes that displayed a difference

Table 6.1: Cancer microarray datasets used in the experiments.

	Name	# Samples	# Classes	Class Distribution	# Genes
cDNA	alizadeh-v1	42	2	21,21	1095
	alizadeh-v2	62	3	42,9,11	2093
	alizadeh-v3	62	4	21,21,9,11	2093
	bittner	38	2	19, 19	2201
	bredel	50	3	31,14,5	1739
	chen	180	2	104,76	85
	garber	66	4	17,40,4,5	4553
	khan	83	4	29,11,18,25	1069
	lapointe-v1	69	3	11,39,19	1625
	lapointe-v2	110	4	11,39,19,41	2496
	liang	37	3	28,6,3	1411
	risinger	42	4	13,3,19,7	1771
	tomlins-v1	104	5	27,20,32,13,12	2315
	tomlins-v2	92	4	27,20,32,13	1288
Affymetrix	armstrong-v1	72	2	24,48	1081
	armstrong-v2	72	3	24,20,28	2194
	bhattacharjee	203	5	139,17,6,21,20	1543
	chowdary	104	2	62,42	182
	dyrskjot	40	3	9,20,11	1203
	golub-v1	72	2	47,25	1877
	golub-v2	72	3	38,9,25	1877
	gordon	181	2	31,150	1626
	laiho	37	2	8,29	2202
	nutt-v1	50	4	14,7,14,15	1377
	nutt-v2	28	2	14,14	1070
	nutt-v3	22	2	7,15	1152
	pomeroy-v1	34	2	25,9	857
	pomeroy-v2	42	5	10,10,10,4,8	1379
	ramaswamy	190	14	11,10,11,11,22,10,11,10,30,11,11,11,11,20	1363
	shipp	77	2	58,19	798
	singh	102	2	58,19	339
	su	174	10	26,8,26,23,12,11,7,27,6,28	1571
	west	49	2	25,24	1198
	yeoh-v1	248	2	43,205	2526
	yeoh-v2	248	6	15,27,64,20,79,43	2526
	Name	**# Samples**	**# Classes**	**Class Distribution**	**# Genes**

of at least l-fold in at least c samples from their mean expression level (Faceli et al., 2004). We considered $c = 1$ and adjusted the value of l in order to select about 1000 genes, number employed in several studies (*e.g.*, Gasch et al. (2000); Qin (2006); de Souto et al. (2008); Tamayo et al. (1999)). A summary of the 17 datasets is provided in Table 6.2, indicating their characteristics.

Table 6.2: Time-course microarray datasets used in the experiments.

Name	Source	# Samples	# Genes (Original)	# Genes (Filtered)
alpha factor	Spellman *et al.* (1998)	18	6178	1099
cdc 15		24	6178	1086
cdc 28		17	6178	1044
elutriation		14	6178	935
1mM menadione	Gasch *et al.* (2000)	9	6152	1050
1M sorbitol		7	6152	1030
1.5mM diamide		8	6152	1038
2.5mM DTT		8	6152	991
constant 32nM H2O2		10	6152	976
diauxic shift		7	6152	1016
complete DTT		7	6152	962
heat shock 1		8	6152	988
heat shock 2		7	6152	999
nitrogen depletion		10	6152	1011
YPD 1		12	6152	1011
YPD 2		10	6152	1022
yeast sporulation	Chu *et al.* (1998)	7	6118	1171

To evaluate robustness to noise we artificially added noise to some datasets (performing noise experiments on all datasets was impractical, due to the high computational costs associated). To that end we selected eight datasets, four from each collection previously presented. Since we did not know a priori the amount of noise originally present in each dataset, as a first step, we

chose eight datasets (four for cancer experiments and four for time-course experiments) in which differences among the performances of the distances were as low as possible, *i.e.*, datasets in which different distances provided the closest results without the presence of noise. With this selection we intended to provide a fair starting point and comparison among the distances as the noise is added. After choosing appropriate datasets, we added noise to each dataset as follows: (i) we randomly chose $\alpha\%$ *points* in the dataset (we call *point* a specific expression level regarding one gene and one sample, *i.e.*, a cell of the data matrix) and; (ii) replaced each point with a randomly generated value between the minimum value and the maximum value observed in the data matrix. This procedure was performed for values of α ranging from 1% to 10%, considering an increment step of 1%. To increase the reliability of the results, 100 different noisy datasets were generated for each different percentage of noise (as indicated in Jain and Dubes (1988), at least 100 replications should be considered for obtaining statistical significance). The replications were required to avoid bias in the errors introduced in particular datasets. The noise experiments were based on a usual Monte Carlo analysis proposed by Milligan (1980). Replications were sampled independently of each other.

6.4.2 Results and Discussion

In the following we provide results for both evaluations regarding microarray data.

6.4.2.1 Clustering Algorithm Independent

Cancer Datasets

We start our discussion by considering the 35 datasets that have external information for samples. Results for these are provided in Figure 6.1 regarding the ISA (Intrinsic Separation Ability) of the distances, for each dataset type. For cDNA datasets, PE provided the best results, followed by RM and COS. When considering Affymetrix data, RM, COS, and PE stood out. Note that, regardless of the type of data, rank-based measures, *i.e.*, GK, KE, and SP, provided, along with WGK, worse results than the other correlation coefficients. This behavior may be observed due to the loss of information intrinsic to rank-based measures. It is worth noticing, however, that measures, such as GK and KE, rarely employed in the gene expression literature, appear as good alternatives to the commonly employed SP. Except for COS, which is closely related to PE, all other "classical" distance functions (EUC, MAN, and SUP) provided the worst overall ISAs.

We employed Friedman and Nemenyi statistical tests (at 95% confidence level) separately for cDNA and Affymetrix data following the procedure suggested by Demšar (2006). Regarding cDNA, the tests suggest that PE and RM provided better ISA than EUC, MAN, and SUP distances. Considering Affymetrix, RM and GK provided superior results than SUP.

Next we proceed and report results regarding noise evaluation. Figure 6.2 depicts results considering different levels of noise regarding data from the first collection of datasets (35 cancer

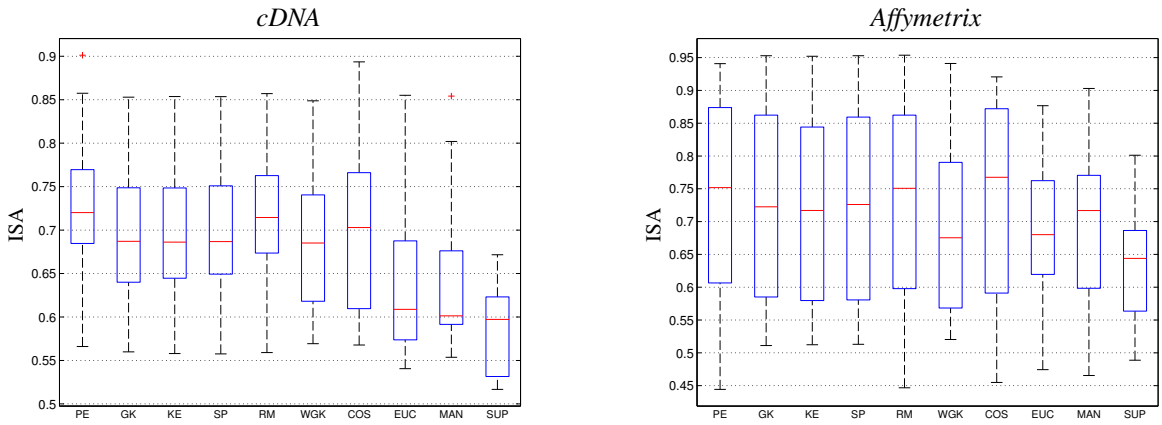


Figure 6.1: Intrinsic Separation Ability (ISA) for each one of the evaluated distances.

datasets). Results are concerned with four datasets, namely, *bittner* and *chen* from cDNA and *golub V2* and *gordon* from Affymetrix. For simplicity, we analyzed the three distances that produced superior or competitive results in the previous evaluation, *i.e.*, PE, RM, and COS. We also included the commonly employed EUC and SP (KE and GK provided similar results when compared to SP and were not included to keep the plots clearer). According to the results from Figure 6.2, SP stands out as the distance less affected by the presence of noise for both cDNA and Affymetrix datasets. These results are not surprising, since SP is a rank-based correlation and has been widely used due to its known robustness. As it takes into account only the ranks of the values for each sample compared, small disturbances in the data tend to vanish in the final comparison. Although RM is more sensitive to the presence of noise than SP, it provided better results than the other distances. PE and COS provided similar results among themselves, whereas EUC showed the worst results. It is worth noting that, although SP provided the best results w.r.t. different levels of noise, it provided inferior mean results when compared to PE, RM, and COS in the previous evaluation scenarios (no noise added). To this extent, RM appears to be one of the best alternatives among the compared measures, as it (i) figures amongst the top two measures in almost all cases in the previous evaluation scenario and (ii) is more robust to noise than PE and COS. In brief, Rank-Magnitude provides a good compromise between accuracy and robustness to noise.

Time-Series Datasets

Results regarding the 17 time-course datasets, for which distances were evaluated regarding their Intrinsic Biological Separation Ability (IBSA), are shown in Figure 6.3, for MF and BP ontologies. The best results were obtained with YS1, which was specifically proposed to clustering gene expression time-course data. Jackknife (JK), another measure proposed specifically to this particular scenario, also provided slightly better results than PE. YR1, YRM1 (YR1 measure with Symmetric Rank Magnitude replacing Pearson), COS, and WGK also stood out as good choices regarding IBSA. Note that YRM1 provided better results than its base measure (RM).

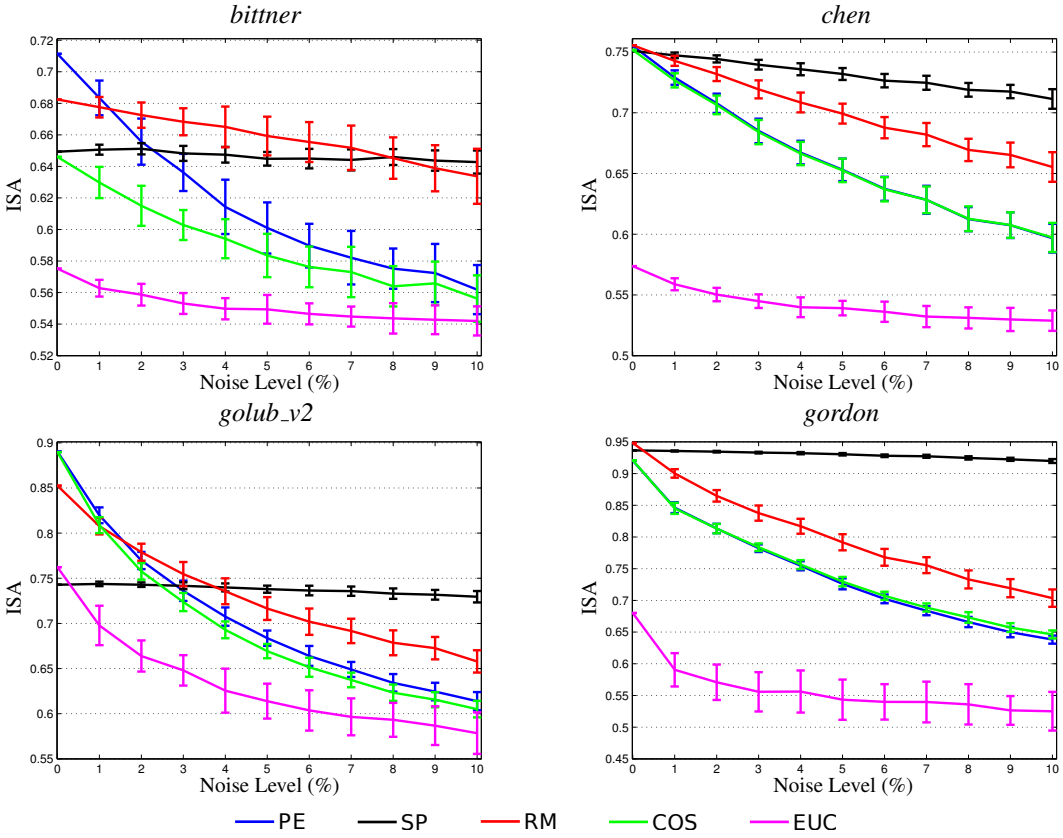


Figure 6.2: Intrinsic Separation Ability (ISA) regarding different noise levels. Lines account for averages, whereas bars account for the standard deviation considering the 100 replicates.

Considering rank-based measures, the best results were obtained by GK, which provided in some cases even better results than PE and JK. Regarding the remaining distances, some trends can be observed: (i) classical distances (except COS) provided inferior results when compared to the other measures; (ii) RM showed inferior results when compared to the remaining correlations; and (iii) STS and LSS did not provide good IBSA values, regardless of the ontology employed (MF or BP).

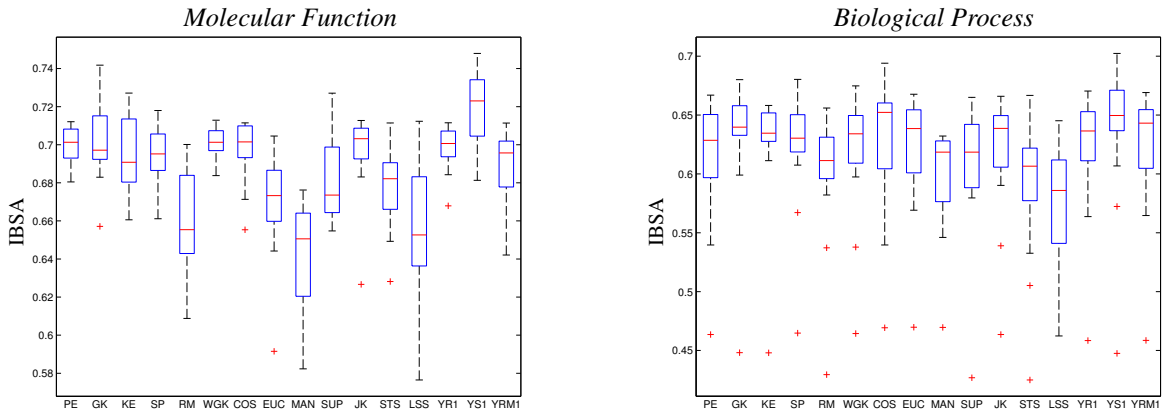


Figure 6.3: Intrinsic Biological Separation Ability (IBSA) for each one of the evaluated distances.

We employed Friedman and Nemenyi statistical tests (at 95% confidence level) separately for the results obtained with MF and BP ontologies. The results are shown in Table 6.3. Once again, PE and JK exhibited very similar results. Both measures provided statistically better results than EUC for the MF ontology. When considering both ontologies, YS1 provided a greater number of statistically significant differences than PE and JK with respect to other proximity measures.

Table 6.3: Statistical Test Summary - MF and BP Ontologies.

	PE	GK	KE	SP	RM	WGK	COS	EUC	MAN	SUP	JK	STS	LSS	YR1	YS1	YRM1
PE	—															
GK		—														
KE			—												∗	
SP				—												
RM	∗	⊛		□	—	∗	∗				∗				⊛	
WGK						—										
COS							—									
EUC	∗					∗	∗	—			∗				∗	
MAN	∗	⊛	∗	∗		∗	∗		—		∗			∗	⊛	∗
SUP										—					⊛	
JK											—					
STS		□		□			□					—			⊛	□
LSS	∗	⊛	⊛	⊛		⊛	⊛	□			∗		—	⊛	⊛	□
YR1														—		
YS1															—	
YRM1																—

Symbols in each cell denote that the measure in the column outperformed the one in the row regarding: ∗ MF ontology, □ BP ontology, ⊛ both.

In Figure 6.4 we provide results regarding IBSA of the distances when different levels of noise were considered. We selected four datasets for evaluation as previously discussed, namely, *alpha factor* and *cdc 15*, for the MF and *1.5mM diamide* and *YPD 1*, for the BP ontology. To provide clearer plots, we selected, based on previous experiments, seven proximity measures for this evaluation: PE, SP, GK, EUC, JK, YR1, and YS1. Considering the results in Figure 6.4, two distinct groups of proximity measures can be observed: one composed of rank-based measures and another formed by the remaining proximity measures. Rank-based proximity measures were less affected by noise, exhibiting slower reductions in their IBSA values when compared to other proximity measures. Note that JK and YR1 provided little or no improvement when compared to PE (their base measure). Within the two groups of measures, it is difficult to determine whether one measure provides better results than the others, given the similar trends among their declines in IBSA. For this particular scenario, YS1, which provided one of the top results regarding the comparison performed in the original datasets, was also one of the least affected measures subjected to different levels of noise. Although SP also showed good to moderate robustness to noise, it provided worse results in comparison to YS1 in the original datasets. Therefore, YS1 stands out as one of the best measures regarding gene time-course data, providing not only accurate results (w.r.t. IBSA), but also a reasonable robustness in the presence of noise.

The results obtained with YS1 were achieved by combining slope and range information with time-series correlation, giving rise to a measure that considers temporal dependencies between time points, *i.e.*, their order. On the one hand, the additional information employed by the measure seems to improve accuracy; on the other hand the use of Spearman correlation as a base measure

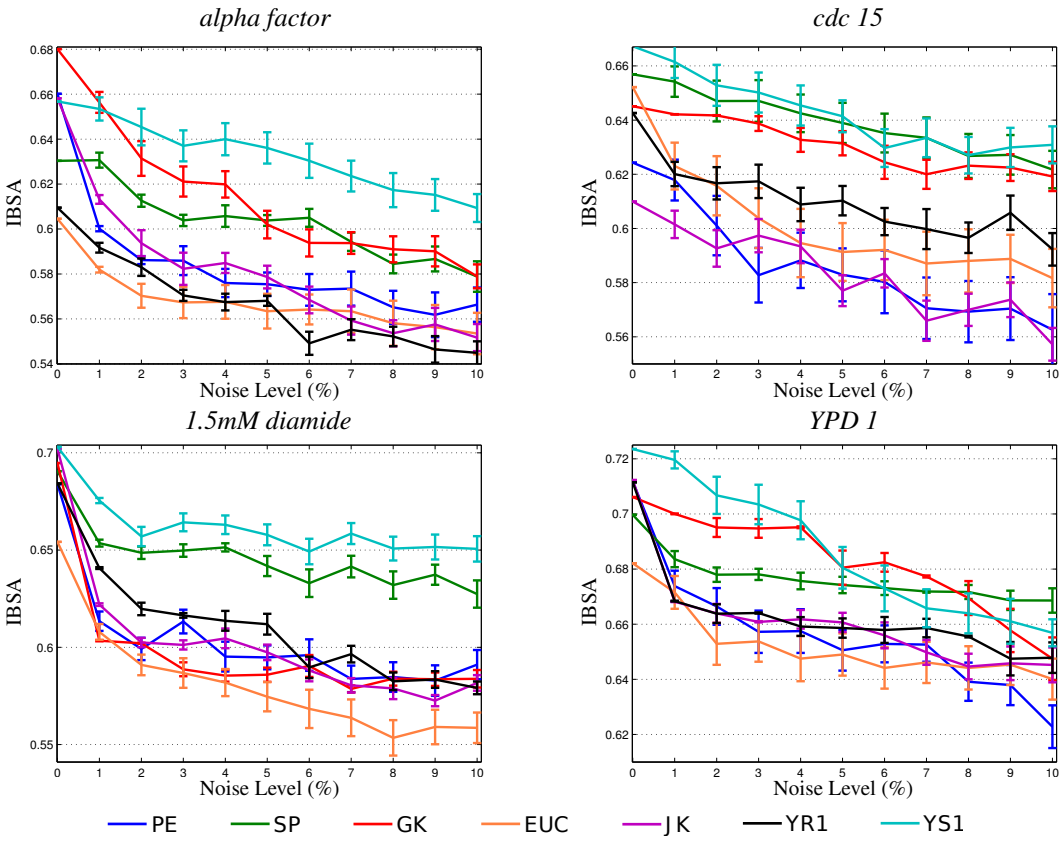


Figure 6.4: Intrinsic Biological Separation Ability (IBSA) regarding different noise levels. Lines account for averages, whereas bars account for the standard deviation considering 100 replicates.

affords robustness to noise. We observed that both YR1 and YS1 provided better results than their "base" proximity measures, *i.e.*, Pearson and Spearman. We call attention to the fact that both YR1 and YS1 are parametrized and that the fine tuning of their parameters for a particular dataset[4] may provide even better results than the ones observed here.

Regarding the other three proximity measures specifically proposed for the clustering of time-course data, we noted that Jackknife provided little improvement in comparison to Pearson. If any, an improvement comes with the price of a quadratic time complexity, which however may not be an issue for *short* time-course. Local Shape-based Similarity (LSS) and Short Time-Series dissimilarity (STS) figured amongst the worst proximity measures under evaluation. Although time shifts are important, considering them in series of limited size may not provide reliable information. Possibly for this reason, poor results were observed with LSS. As for STS, its simple formulation based solely on slope differences may be the root cause for the poor results observed. Note that when this information is combined with others, better results arise, as is the case for both YR1 and YS1, which employ the slope information as a *part* of their formulation.

[4]As previously discussed in Section 6.2.3.4, we employed fixed parameters (weights) for these two proximity measures, using the values originally employed by their authors (Son and Baek, 2008). Given the number of datasets and the experimental setting adopted, analyzing the effect of different parameter values was impractical.

6.4.2.2 Clustering Algorithm Dependent

Cancer Datasets

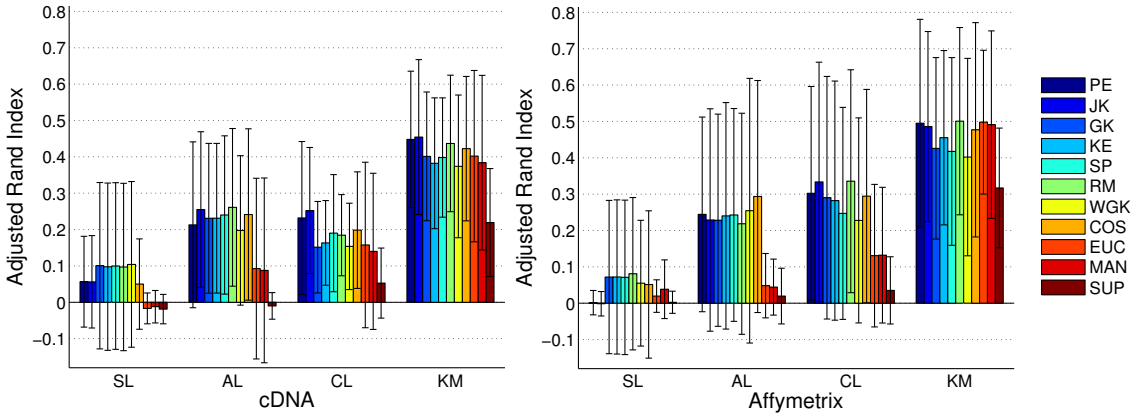
Results regarding clustering algorithm dependent evaluation for the clustering of samples are provided in Figure 6.5. The figure depicts the three evaluation scenarios, namely: (i) fixed number of clusters, (ii) unconstrained number of clusters, and (iii) estimated number of clusters. The first observation we make is with respect to the Single-Linkage clustering algorithm, which as previously reported by de Souto et al. (2008), led to the poorest recovery rates among the clustering algorithms employed. To that end, our results support and reinforce the results presented by de Souto et al. (2008), because even with the use of different distances, the Single-Linkage (SL) algorithm clearly does not stand as a good choice for the sample clustering scenario.

Before proceeding to the comparison of the distances per se, we provide some remarks regarding the three evaluation scenarios under consideration. First, it is worth noticing that ARI results for the second scenario (variable number of clusters) are in general higher in comparison to the first one (fixed number of clusters). This behavior is in agreement with the assumption that a partition with the "wrong" number of clusters may be better than one containing the same number of clusters as defined by the reference partition (Vendramin et al., 2010). When considering the third scenario (number of clusters estimated with the use of the Silhouette), it is quite interesting to observe that k-medoids does not provide, in real applications (as simulated by this scenario), significant differences when compared to hierarchical methods. Despite the similar behavior shown by clustering algorithms in this scenario, different distances do provide different results.

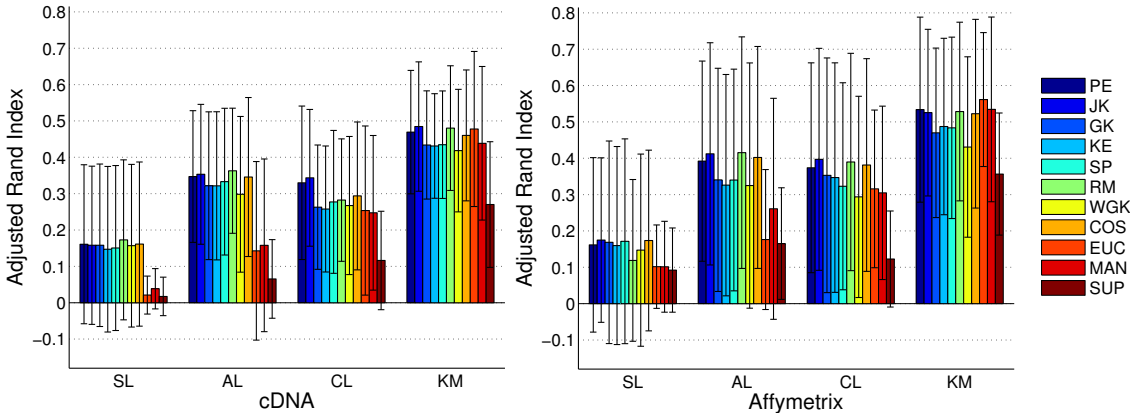
Considering the Average-Linkage (AL) clustering algorithm and Affymetrix data, Pearson (PE), Jackknife (JK), Rank-Magnitude (RM), and Cosine (COS) provide in general the top results. When a fixer number of clusters is considered, Cosine outperforms the aforementioned methods. In the case of cDNA data, JK and RM present the best mean results, followed by COS. Still regarding this type of data, WGK and PE provide the worst results among the correlation coefficients. For both cDNA and Affymetrix datasets, rank-based measures, *i.e.*, Goodman-Kruskal (GK), Kendall (KE) and Spearman (SP), present similar behavior among themselves, whereas "classical" distances provide the worst results. Note, that even the correlation that provided the worst mean results (WGK) stands as a better alternative than the three "classical" distances.

For Complete-Linkage and k-medoids clustering algorithms PE, JK, COS, and RM stand out among the other correlation coefficients, except for cDNA datasets with Complete-Linkage, for which RM shows poorer results than PE and JK. Regarding rank-based correlation coefficients, both GK and KE, which are measures not widely used in gene expression data analysis, show in some cases superior mean results when compared to the also rank-based SP. Among the "classical" distances, SUP provides the worst results. For the k-medoids algorithm, COS, EUC and MAN provide competitive but slightly worse results than the top distances (PE, JK and RM).

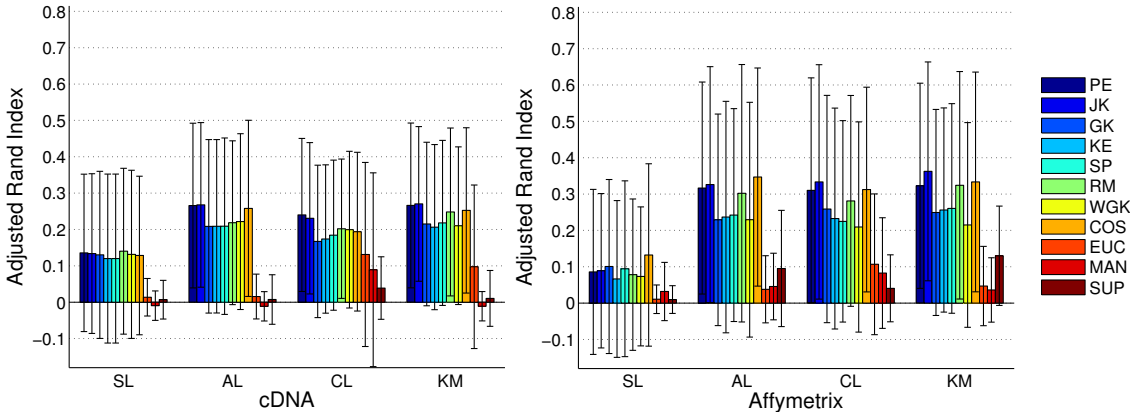
Following Demšar (2006), we applied Friedman and Nemenyi statistical tests (at a 95% confidence level), separately for each clustering algorithm and evaluation scenario. Regarding the



(a) Fixed number of clusters — same number of clusters as defined by the reference partitions.



(b) Variable number of clusters — regardless of the number of clusters defined by the reference partitions.



(c) Estimated number of clusters — partition chosen by the Silhouette is compared against the reference partitions.

Figure 6.5: Class recovery obtained for cancer datasets regarding the three evaluation scenarios under consideration, subfigures (a), (b), and (c). Bars display mean results for each pair of clustering algorithm and distance in different types of datasets: cDNA (left) and Affymetrix (right).

first scenario, for both cDNA and Affymetrix, considering AL, CL, and KM clustering algorithms, PE, JK, and RM provide better results than SUP in virtually all cases. For the second evaluation scenario, for cDNA and KM the tests suggest a statistically significant difference in favor of RM over WGK. Still regarding cDNA, regardless of the clustering algorithm, all correlations are superior to SUP, whereas for the AL algorithm PE, JK, and RM are superior to EUC and MAN.

Regarding Affymetrix the tests suggest that RM, PE (only for KM), and JK (except for KM) are statistically superior to SUP. Finally, regarding the third evaluation scenario, for AL, regarding cDNA, PE and JK are superior to MAN and SUP, whereas for Affymetrix, PE, JK, SP, and COS are superior to EUC. Considering CL, for both data types PE and JK are superior to SUP. For KM and cDNA data, all correlations and COS provide better results than MAN and SUP, whereas for KM and Affymetrix datasets, PE, JK, and RM provide better results than EUC.

Next we present the results regarding noise evaluation, considering the clustering algorithm dependent setting. These are shown in Figure 6.6. We show results for *Bittner* and *Chen* datasets, from cDNA, and *Golub V2* and *Gordon* datasets, from Affymetrix. For simplicity, we analyze results concerning the four distances that produced superior or competitive results during previous evaluation, *i.e.*, PE, JK, RM, and COS. We also include, for a fair comparison, the commonly employed EUC and the rank-based SP. Once more we note that KE and GK provided similar results when compared to SP. These measures were not included to keep the plots clearer.

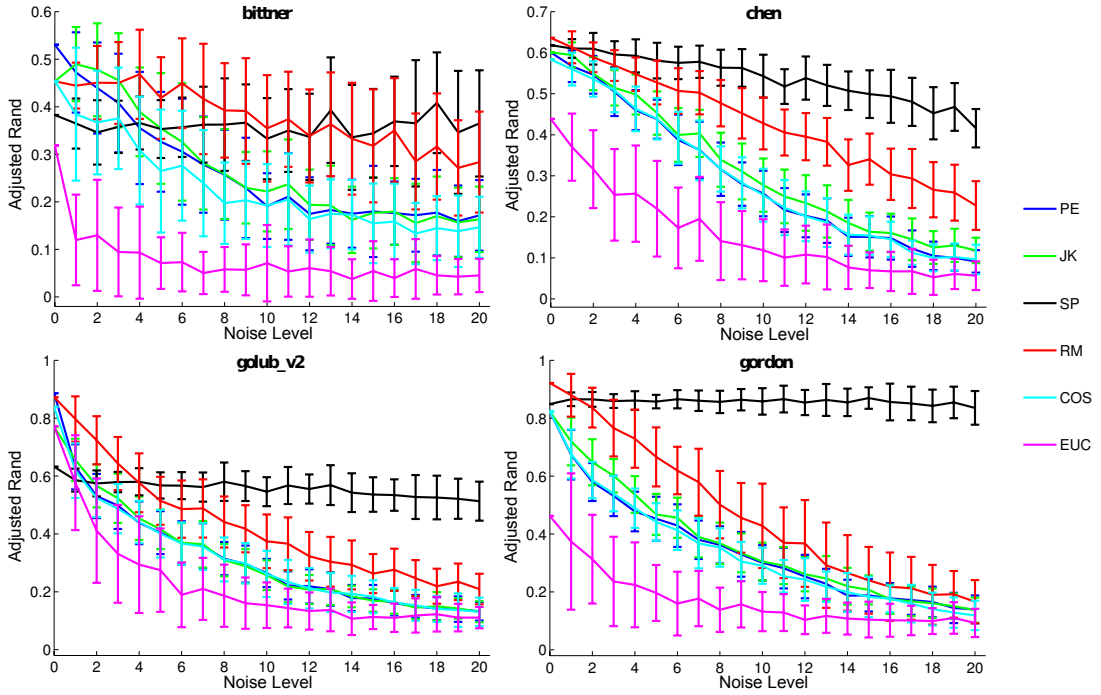


Figure 6.6: Robustness to noise regarding cancer datasets. Figure depicts ARI values for different noise levels (%) regarding PE, JK, SP, RM, COS and EUC. Plots correspond to the mean ARI values for runs performed in 100 different noise datasets with the same amount (%) of noise points.

For both cDNA and Affymetrix datasets, SP appears to be the distance less affected by noise. These results are not surprising, since SP is a rank-based correlation coefficient. As it takes into account only the ranks of the values for each sample compared, small disturbances in the data tend to vanish in its final comparison. Although RM is more sensitive than SP to the presence of noise, it provides better results than the other distances compared. PE, JK and COS provide similar results among themselves, whereas the commonly employed EUC displays the worst results. Finally, we note that, although SP provides the best results with respect to different levels of noise, this measure

also provides inferior mean results when compared to PE, JK, RM, and COS in the previous evaluation scenarios. To this extent, RM appears to be one of the best alternatives among the compared measures, given that it provides a good compromise between accuracy and robustness.

Time-Series Datasets

Next we provide results regarding the comparison of distances for the clustering of gene time-series data. In this case we consider only the third evaluation scenario (estimated number of clusters) given that class labels are not available. Performing noise experiments in such datasets is also impractical, due to: (i) lack of class labels, (ii) the type of evaluation employed (pairwise), which makes comparison among measures for different noise levels not straightforward, and (iii) the amount of time required to biologically evaluate all partitions.

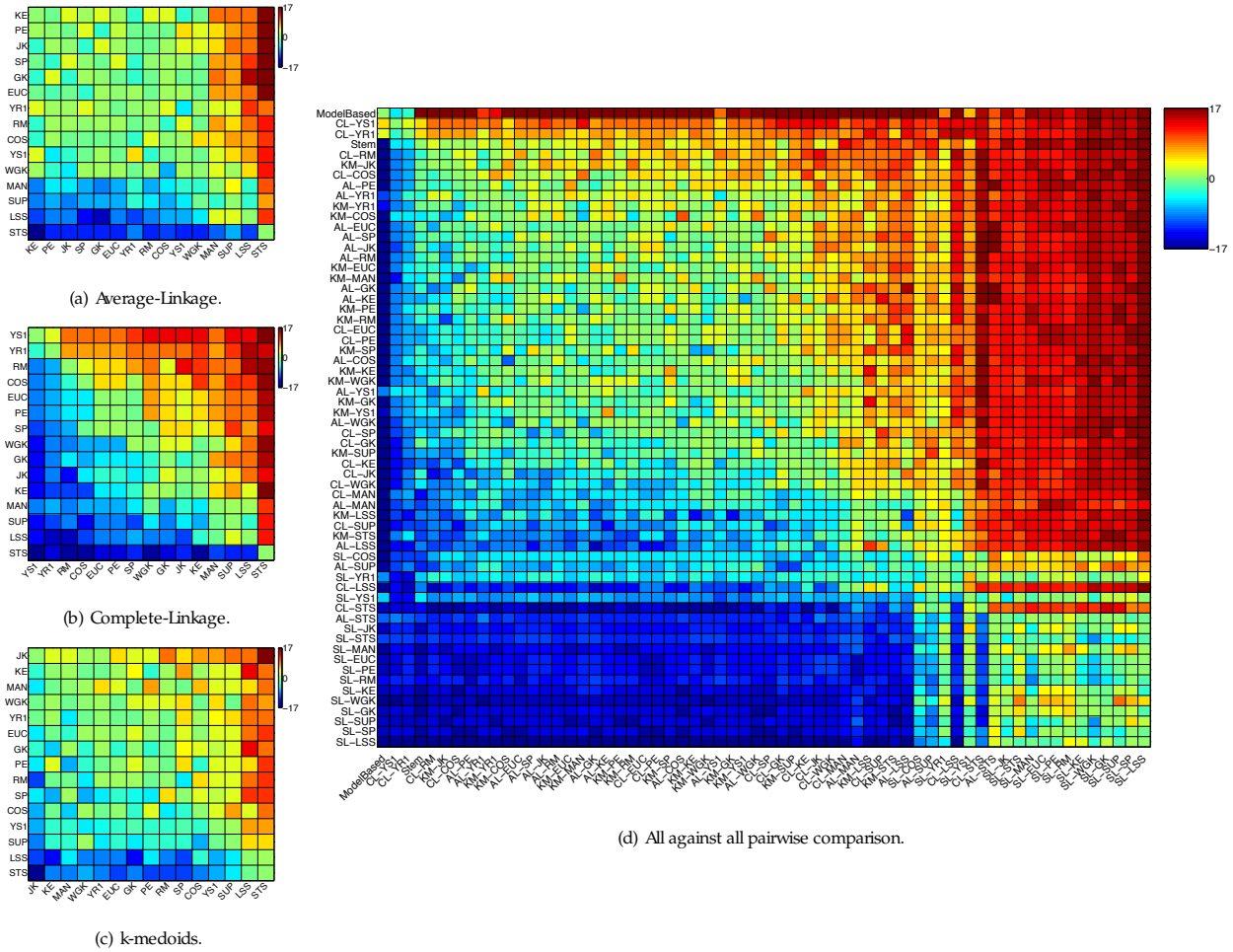
Before comparing the distance measures themselves, we assess the results of clustering algorithms, regardless of the distance measure adopted. These results are shown in Table 6.4, which summarizes results for SL, AL, CL and KM regardless of the distance adopted for the 17 gene time-series datasets. In each table cell we show the number of Wins/Ties/Losses for the row algorithm with respect to the column one. Each table cell comprises $3,825$ pairwise comparisons. For each cell we have two clustering algorithms, each of which is evaluated with 15 distance measures in 17 datasets, *i.e.*, $15 \times 15 \times 17 = 3,825$ pairwise comparisons between any two algorithms. In this scenario the best results are displayed by KM, which is closely followed by AL and CL. These three algorithms provide quite competitive results among each other, whereas the worst overall results are provided by SL, which is excluded from further analysis.

Table 6.4: Wins/Ties/Losses for 15 distances and 17 datasets.

	SL	AL	CL	KM
SL	—	531/370/2924	378/384/3063	385/323/3117
AL	2912/406/507	—	1903/93/1829	1710/80/2035
CL	3063/386/376	1821/106/1898	—	1803/17/2005
KM	3117/323/385	2032/80/1713	2001/18/1806	—

Having made such considerations, we depict in 6.7(a) results regarding the use of different distances for AL. For this algorithm KE, PE and JK displayed similar results, providing better enrichments than the remaining measures in 71% of the cases under comparison. For AL, none of the measures is consistently better than the others, with different measures appearing as the top ones, depending on the dataset. It is interesting to note that LSS and STS, two measures specifically proposed for the gene clustering scenario, figured as the worst choices (alongside MAN and SUP).

For Complete-Linkage, Figure 6.7(b), differences among distance measures become more evident. YS1 and YR1, which were specifically designed for the clustering of gene time-series provided better enrichment results than virtually all the other measures. In fact, YS1 and YR1 provided better enrichments than other distances in 94% and 87% of the evaluated cases,



(a) Average-Linkage.

(b) Complete-Linkage.

(c) k-medoids.

(d) All against all pairwise comparison.

Figure 6.7: Results obtained for the clustering of gene time-series data. Figures (a), (b) and (c) depict pairwise comparison of different distances for each clustering algorithm. Figure (d) depicts an all against all pairwise comparison. Each cell account for the number of datasets in which the method from the row obtained a better enrichment than the method from the column. The "hotter"/"colder" the cell the better/worst is the row method in comparison to the column one.

respectively. Another distance that showed good results for CL was RM, which provided better enrichments than the other measures in 80% of the cases. These results are better than the ones produced by distances commonly employed for gene clustering, such as PE, EUC, and SP, which provided better results than other distances in 72%, 70% and 65% of the cases, respectively.

We show in Figure 6.7(c) evaluation results regarding KM. For this clustering algorithm JK provided the best results, showing better enrichments than other distance measures in 77% of the cases under comparison, which is 12% above those found with the second ranked measure (KE). Good results were also shown by MAN, which performed better than other distances in 60% of the cases under comparison. For this algorithm, at least five distance measures provided better enrichments than the commonly employed PE, EUC and SP in the 17 datasets studied.

To present an overview of clustering algorithms and distance measure pairs we conducted an all against all pairwise comparison, as shown in Figure 6.7(d). Note that in this case we have to take into account the pair clustering algorithm-distance measure to include both biases in the

comparison. To give an idea about the general quality of the results found we also include two state of the art clustering algorithms regarding the clustering of gene time-series, *i.e.*, Stem from Ernst and Bar-Joseph (2006) and Model Based clustering (Costa et al., 2005). Regarding Stem, the number of clusters is automatically determined, so we select for comparison the significant clusters it finds. Considering Model Based clustering, the Bayesian Information Criteria (BIC) (Schwarz, 1978) is employed to estimate the final number of clusters. As one might expect, state of the art Stem and Model based figured among the top results. It is worth noticing that CL, when employed with YR1 and YS1, produced, in general, better enrichments than Stem and in some cases Model Based. From this comparison it is possible to note that for a particular clustering algorithm, the choice of an appropriate distance may provide the difference between an average result and a result close (or better) than those produced by state of the art clustering algorithms, such as Stem.

6.4.2.3 Relation Between Evaluation Methodologies

So far we employed two distinct methodologies to the evaluation of distance measures, one based on their intrinsic separation abilities and another based on their cluster recoveries. Despite their inherent differences, in general, we observed an overall agreement between their evaluations. Based on this observation, we highlight in the following the main results from our analysis w.r.t. the selection of distances for the clustering of samples and genes from microarray data.

For the clustering of cancer samples Pearson and Jackknife displayed, in most of the cases, superior or competitive results when compared to the remainder distance measures. Cosine similarity, which is closely related to Pearson, also figured amongst the best measures. It is important to note here that Jackknife has quadratic computational complexity, in contrast to linear time complexity of Pearson and Cosine. The minor improvements obtained with Jackknife over Pearson and Cosine do not seem to compensate for its computational cost. Another interesting alternative in this particular scenario is Rank-Magnitude. In addition to the good results provided for cancer datasets, Rank-Magnitude is also less sensitive to noise than Pearson, Jackknife, Cosine and Euclidean distance, though more sensitive than Spearman. As Rank-Magnitude has shown, in general, to be more accurate than Spearman, it emerges as a one of the best alternatives for cancer datasets, exhibiting a good compromise between accuracy and robustness.

Regarding the clustering of gene time-series data the best results were obtained with YS1, a measure specifically proposed for the clustering of short gene expression time-series, when combined with the Complete-Linkage algorithm. These results may be due to the fact that YS1 (like YR1) combines a correlation coefficient with other information extracted from the series under evaluation, thus providing a comparison based on more information than the ones performed by any of the other measures considered. By internally employing Spearman, YS1 stands out as a better and more robust option than YR1 (which is based on Pearson) when noise is present. In this particular scenario, given the small number of features, Jackknife should be be preferred to both Pearson and Cosine, as it provided better enrichments than both in most cases.

6.5 RNA-Seq Data

In this section we provide results regarding distance measure evaluation for RNA-Seq data.

6.5.1 Experimental Setup

RNA-Seq datasets were obtained from The Cancer Genome Atlas (TCGA)[5] using the `TCGA-Assembler` R Package (Zhu et al., 2014). The traversal date (data snapshot) was from May 30, 2014. Data was collected from RNA-Seq Version 2, regarding Level 3 using the normalized expression values (see https://wiki.nci.nih.gov/display/TCGA/RNASeq+Version+2). We selected for further analysis datasets that had clinical information for a significant number of patients. In this step, datasets with a strong skew in class distribution were also discarded. This filtering led to a total of 15 datasets, as presented in Table 6.5. For two class datasets the class labels are divided into normal and cancerous, whereas for multi-class datasets the class labels encompass normal and cancer sub-types, as derived from the clinical file from TCGA. Originally, datasets were obtained considering RSEM quantification (Li and Dewey, 2011), for different feature spaces or resolutions, namely: (i) genes, (ii) exons, and (iii) isoforms. Provided that the results for genes were far superior than for exons and isoforms we also gathered data regarding genes with RPKM quantification (Mortazavi et al., 2008) for further evaluation. These were filtered as stated above and have the same characteristics presented in Table 6.5.

Table 6.5: TCGA Datasets Summary. Main characteristics of the datasets under analysis.

Data Name	# Samples	# Classes	Class Distribution
COAD-V1	304	2	41,263
KIHC	91	2	25,66
KIRC	48	2	9,39
KIRP	228	2	30,198
LIHC	249	2	50,199
UCEC-V1	183	2	24,159
LAML	171	8	16,18,35,3,42,39,16,2
LGG	375	3	142,129,104
SARC	84	4	31,46,5,2
THCA-V1	486	3	352,99,35
COAD-V2	301	3	228,32,41
LUAD	517	2	459,58
LUSC	460	2	410,50
THCA-V2	545	4	352,59,99,35
UCEC-V2	140	4	40,67,9,24

The analysis of RNA-Seq data is still in its infancy if compared to that of microarrays. For this reason, for RNA-Seq data we considered different experimental factors that can affect the final clustering results and were not previously systematically evaluated. For each one of the datasets obtained, no matter in which feature space, we selected different numbers of features for further analysis, in order to verify their influence in the clustering results. We considered $1,000$ (1K),

[5] http://cancergenome.nih.gov/

$2,000$ (2K), $3,000$ (3K), $4,000$ (4K), and $5,000$ (5K) features. These were selected using the `varFilter` method from the `GeneFilter` R Package (Gentleman et al., 2014). We employed the Inter Quartile Range (IQR) as variance-filter, which is the package default option.

After obtaining different dataset sizes (by different number of features) we also considered log transformation of the data. For that, given an expression level e_{ij} for the i^{th} feature and the j^{th} sample, we obtained the transformed datasets by $e^t_{ij} = \log_2(e_{ij} + 1)$, where e^t_{ij} is the transformed expression for the same feature and sample. Note that by adding 1 to the original expression values we not only avoid $\log_2(0)$, but we also keep zero counts unchanged, singe $\log_2(1) = 0$. This particular transformation scheme follows from previous work from Lee et al. (2011).

6.5.2 Results and Discussion

Given the good agreement between the two methodologies (clustering algorithm independent and dependent) that we previously employed to evaluate distances in the case of microarrays, for RNA-Seq we chose to employ only the clustering algorithm dependent one. Besides their good agreement, this was motivated due to the fact that for RNA-Seq data we perform more detailed experiments, considering different experimental factors, which results in a higher computational cost. These experiments were not necessary in the case of microarrays, given its already well-established literature. As for the case of datasets, we consider only the case of clustering of samples, given that gene clustering datasets were hard to find. Having made such considerations, in the following we present results of our evaluation regarding RNA-Seq data. We start by describing the experimental factors we considered in Section 6.5.2.1. In Section 6.5.2.2 we discuss the effect of pre-processing in the analysis of RNA-Seq data. Finally, in Section 6.5.2.3 we discuss the clustering of RNA-Seq data, and how the selection of different clustering algorithms and distances may affect its results.

6.5.2.1 Definition of Experimental Factors

As previously discussed, we gathered and compiled a total of 15 RNA-Seq datasets from the TCGA portal. These datasets were selected based on their number of objects (samples) and class distribution. We evaluated the datasets w.r.t. 5 experimental factors, namely: (1) expression estimates, (2) number of features, (3) data transformation, (4) clustering algorithms, and (5) distance methods (see Figure 6.8). Experimental factors 1 to 3 comprehend data pre-processing steps, whereas experimental factors 4 and 5 are the ones directly related to cluster analysis.

We first discuss experimental factors related to pre-processing. Considering expression estimates, we evaluated 4 different types, *i.e.*, exons (RSEM), isoforms (RSEM), and genes (RSEM and RPKM). For number of features, 5 different parameterizations of the filtering were performed yielding datasets with 1K, 2K, 3K, 4K, and 5K features. Concerning data transformation, we further evaluated the option of log-transforming the data, resulting in 2 more variants for our analysis, *i.e.*, data with no transformation ("untransformed" data) and $\log_2$ transformed data.

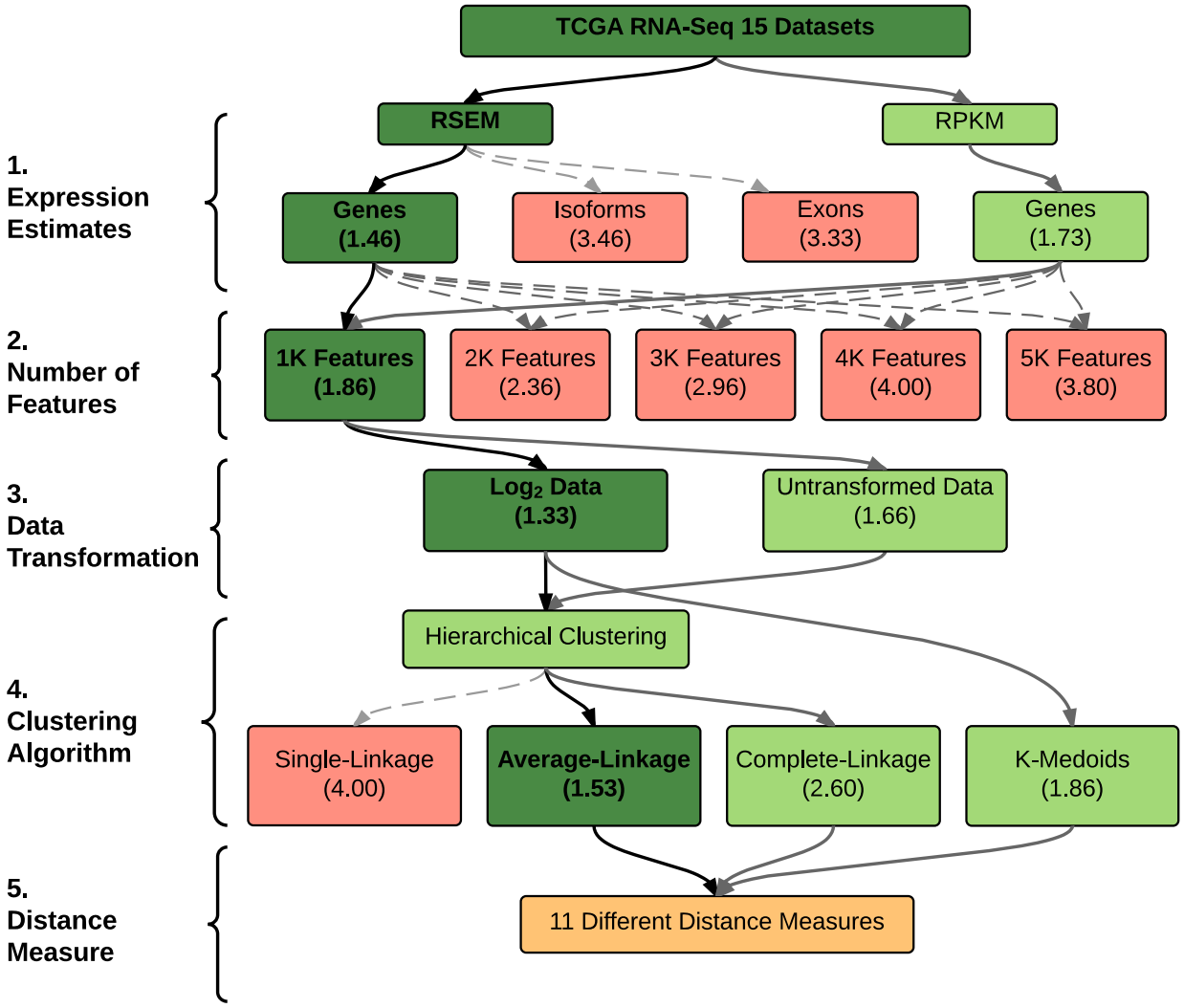


Figure 6.8: RNA-Seq clustering decision pipeline. Each node represents an option for a given experimental factor in our analysis (experimental factors are numbered from 1 to 5). At the bottom of each node we indicate its average rank for the given experimental factor (considering the selection of the best alternative at each one of the previous factors). Dark green (light red) nodes indicate the options with best (worst) empirical results for a given experimental factor. Light green colored nodes represent good scoring options, which are also considered in subsequent factors. In the first experimental factor, for example, we select among four different expression estimate alternatives. At this factor, the best average ranking is provided by RSEM Genes (1.46), with the worst average results coming from RSEM Exons (3.33), which is excluded from further analysis.

Regarding cluster analysis, we considered Single-Linkage (SL), Average-Linkage (AL), Complete-Linkage (CL), and k-medoids (KM), all described in Chapter 2. We combined these four clustering methods with 11 different distance measures (or correlation coefficients adapted into distances), *i.e.*, Pearson (PE), Jackknife (JK), Kendall (KE), Goodman-Kruskal (GK), Spearman (SP), Symmetric Rank-Magnitude (SRM), Weighted Goodman-Kruskal (WGK), Cosine (COS), Euclidean (EUC), Manhattan (MAN), and Supreme (SUP). Experiments were conducted considering a fixed number of clusters, namely, the number of clusters as defined by the external partition (class labels), as well as considering an unconstrained number of clusters.

In the latter case, we varied k in the range $[2, \lceil\sqrt{n}\rceil]$, where n equals the number of objects in the dataset. The solution with highest Adjusted Rand Index was selected for further evaluation. The experimental factors described above resulted in a total of 3,520 different clustering solutions (different combinations). All of these were evaluated externally w.r.t. their Adjusted Rand Index.

Given the large number of solutions under evaluation, it was impractical to evaluate all combinations of experimental factors separately. For this reason, for a given experimental factor, we chose the best performing option(s), which were then used in the next experimental factor evaluation. At a given factor level, we computed the rank of each option, for all the cases and datasets. These were then averaged, resulting in a single ranking value for each option at a given level. The Friedman and Nemenyi statistical tests were employed to evaluate whether differences in ranking were significant (Demšar, 2006). An overview of the evaluation is presented in Figure 6.8.

6.5.2.2 Evaluation of Pre-Processing Factors

Considering the first experimental factor, estimates based on genes (RSEM and RPKM) are in general better than those of exons and isoforms. The Friedman statistical test rejected the null hypothesis, *i.e.*, there is a statistically significant difference between the expression estimates, with a $p\text{-}value = 9.08 \times 10^{-18}$. These statistical differences were further confirmed by the Nemenyi post-hoc test at a 95% confidence level. The Friedman statistical test did not indicate any differences between genes (RSEM) and genes (RPKM), or between isoforms and exons.

Regarding the choice of number of features (second experimental design factor), there is a trend towards the use of as few features as possible (1,000). The use of 1K features lead to better rankings than 4K and 5K features (Friedman $p\text{-}value = 1.44 \times 10^{-4}$). Lastly, there is no statistical difference between the use or not of log transformation, which corresponds to our experimental factor 3. Note that we have observed that this choice can however impact particular distance measures and will be evaluated on the combinations of all clustering/distance measures.

6.5.2.3 Evaluation of Clustering Factors

In the case of clustering algorithms (experimental factor 4), we observed a very poor performance of the Single-Linkage (SL) clustering algorithm. To that end, the Friedman statistical test rejected the null hypothesis of no difference among the methods ($p\text{-}value = 5.88 \times 10^{-23}$). The Nemenyi post-hoc test indicated that all other clustering algorithms provided superior results than SL (at a 95% confidence level). Given such poor results, SL was removed from further analysis

Next, we evaluate in detail clustering performed on the previously selected experimental factors. These are two data transformation strategies ($\log_2$ transformed *vs* "untransformed" count data), two quantification strategies (RSEM and RPKM at gene level), Average-Linkage, Complete-Linkage and k-medoids clustering algorithms, and 11 distance measures. We start with the results for genes (RSEM), which are presented in Figure 6.9. In this figure we show results for

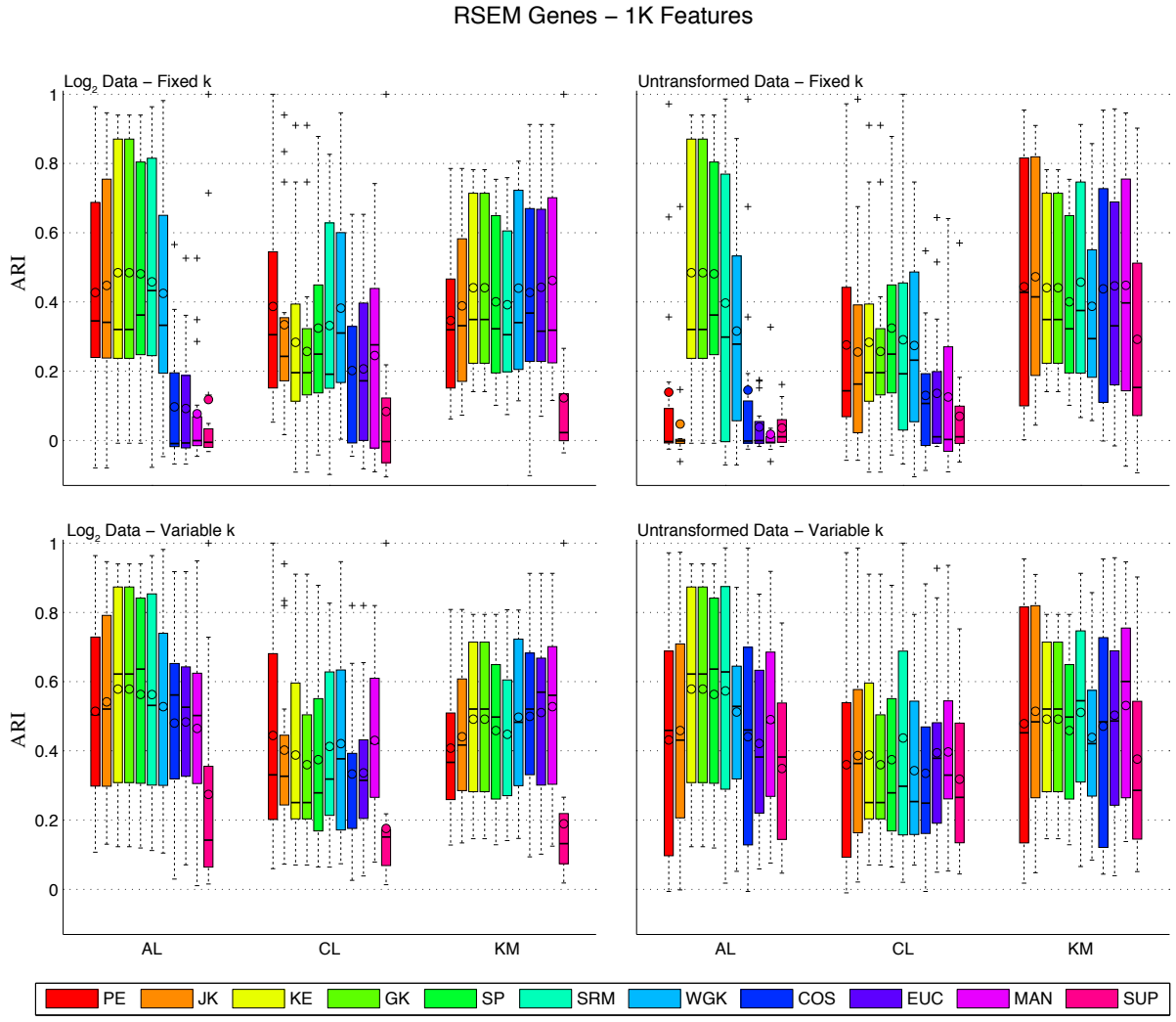


Figure 6.9: Figure depicts detailed results for genes (RSEM), considering 1K features. Boxplots show individual results regarding different distance measures, clustering algorithms, data transformations, and number of clusters. The symbol ∘ accounts for the mean over 15 datasets.

both $\log_2$ and untransformed data (first and second column), considering scenarios with the actual number of clusters and with variable number of clusters (top and bottom plots).

First let us consider the case of fixed number of clusters (Figure 6.9 top). Clearly, a few distance measures were affected positively by $\log_2$ transformation, particularly for Average-Linkage (AL) and Complete-Linkage (CL) clustering algorithms. Pearson (PE) and Jackknife (JK), which displayed a poor performance considering untransformed data have a large improvement in their performance after $\log_2$ transformation. For the k-medoids (KM) clustering algorithm there is not much difference among the distance measures compared, except for Supreme distance (SUP), which in all the cases under analysis provided the worst results. It is worth noticing that rank-based measures (KE, GK, and SP), which are invariant to the magnitude of the variables (features), are not affected by log-transformation. Considering the case of variable k, that is, unconstrained number of clusters, the effects of $\log_2$ transformation are less clear than before. In general, however, there is still an increase in ARIs for AL and a reduction in variability for KM. This is due to the fact

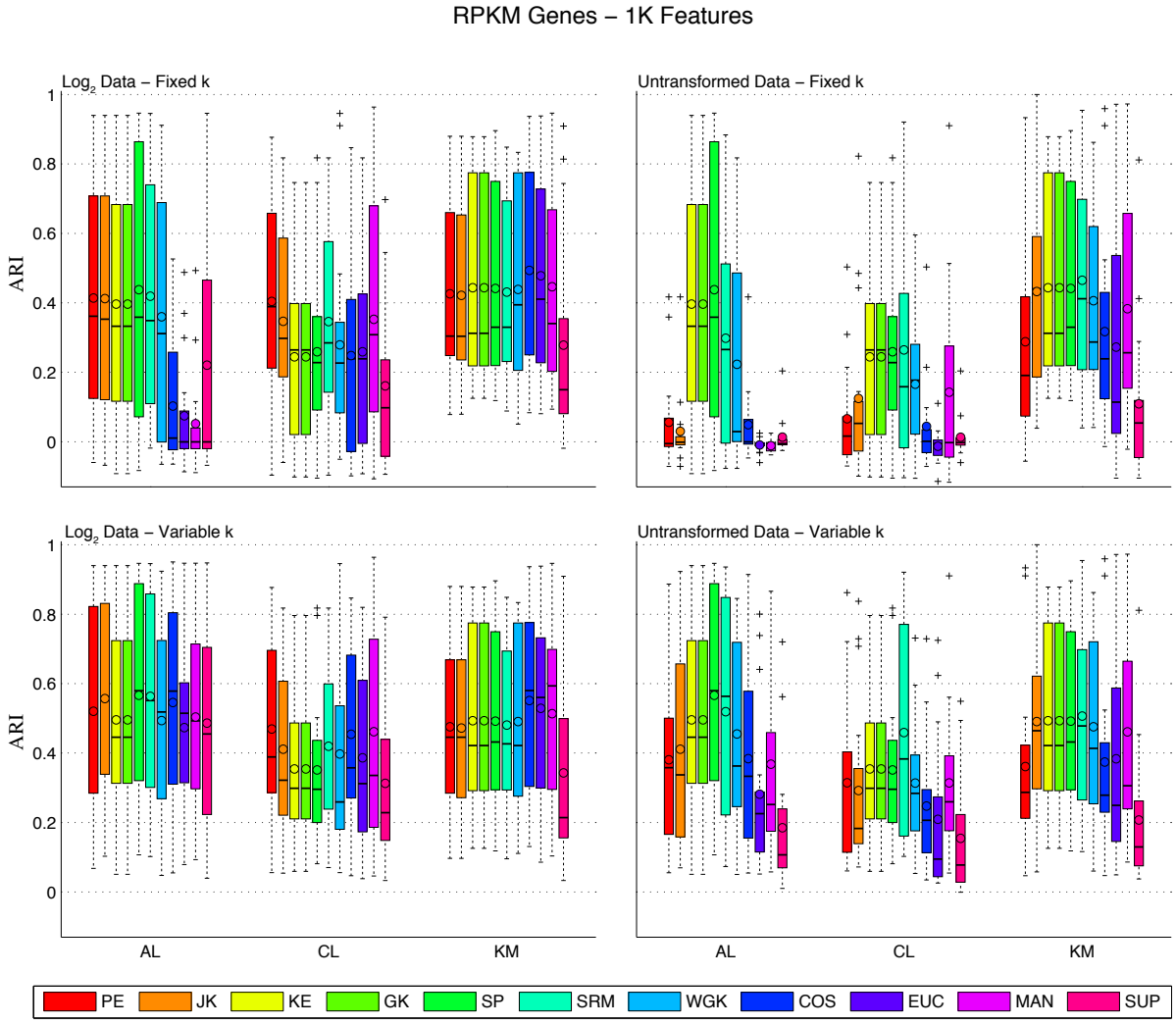


Figure 6.10: Figure depicts detailed results for genes (RPKM), with 1K features. Boxplots show individual results regarding different distance measures, clustering algorithms, data transformations, and number of clusters. The symbol ∘ accounts for the mean over 15 datasets.

that in this particular scenario the results account for the best Adjusted Rand Index (ARI) found, which in some sense represents an upper bound for the Adjusted Rand Index considering each pair of clustering algorithm and distance measure under evaluation.

Given its best overall results, we consider the case of log-transformed data with fixed number of clusters for further discussions (top-left plot). In this case, on average, the best results were obtained with the Average-Linkage clustering algorithm. For this particular method, Symmetric Rank-Magnitude provided the best results, regarding its median. For Complete-Linkage, Pearson and Weighted Goodman-Kruskal displayed the best mean results, whereas for k-medoids, Euclidean and Manhattan displayed good results, on average, alongside with rank-based measures. These results show a contrast with those of microarray data, for which Pearson was, in most of the cases, the best measure, seconded by rank-based correlation coefficients, as we discussed in the previous Section. We applied the Friedman and Nemenyi statistical tests within each clustering algorithm. These indicated that most distance measures outperformed Supreme distance.

Results for genes (RPKM) are shown in Figure 6.10. In this case, the benefits of $\log_2$ transformation seem to be clearer than for the case of RSEM data. Indeed, even for the case of variable number of clusters, performing such a transformation provides considerable improvements. If no transformation is applied to the data, classical distance measures, *i.e.*, Cosine (COS), Euclidean (EUC), Manhattan (MAN), and Supreme (SUP) display a poor recovery for all clustering algorithms, except for k-medoids (KM). Other commonly employed measures, *i.e.*, Pearson (PE) and Jackknife (JK) also show a poor recovery, unless data transformation takes place. Although some changes can be observed, k-medoids seems to be the clustering algorithm least affected by log-transformation in the data. As before, Complete-Linkage (CL) provides worse results than the other clustering algorithms, in general. Here it is important to emphasize the importance of selecting an appropriate distance measure. If AL, which in general provides good recoveries, is employed with COS or EUC distance measures, one can expect poorer recoveries than those obtained with CL and KM, no matter the distance these clustering algorithms employ (see Figure 6.10, top left plot). Bearing that in mind, still considering the case of log-transformed data and fixed number of clusters, Spearman appears as one of the best choices for Average-Linkage. In the case of Complete-Linkage, Pearson provides the best results, whereas for k-medoids, COS provides the best recoveries. Once more, we applied statistical tests within each clustering algorithm, which indicated that most distances outperformed Supreme distance.

6.5.2.4 Final Remarks Regarding RNA-Seq data

We observed a clear advantage on the use of gene level quantifications in comparison to that of exons and isoforms (first experimental factor). One possible explanation for this is the inability of the non-specific filtering to select relevant features in large dimensional exon and isoform feature spaces. Note that annotation from TCGA had $20,531$ genes, $239,322$ exons and $73,599$ isoforms, originally. Moreover, isoform quantification is a much more complex task than exon or gene level quantification. Not surprisingly, it obtains the lowest overall ranking. Clustering samples in the feature space of genes with RSEM quantification provided, in general, a better recovery than those of genes with RPKM quantification. This can also be observed in the specific cases we evaluated. The differences, however, are small. Given its simple computational estimation, we can state that RPKM is a viable alternative to gene expression quantification previous to clustering.

Regarding the second experimental factor, that is, the number of features, we recommend the use of about 1K features. Besides providing the best overall ranking, using less features is usually preferable, given that it provides simpler models and lower computational costs than those obtained with the use of large feature subsets. The preference towards small feature subsets also complies with the practice that is already employed in the microarray literature of selecting around one thousand features for further analysis, as was done by de Souto et al. (2008).

Data transformation (third factor) seems to be a key issue in the analysis of RNA-Seq. The $\log_2$ transformation of data decreases the high variability observed in read counts, benefiting

commonly employed correlation coefficients such as Pearson and Jackknife. This is particularly true when RPKM is used. If a rank-based correlation is to be employed, then no previous data transformation is required. Note that this arises due to the fact that these measures are invariant to such a transformation, therefore it causes no changes at all. Such results are also in agreement with those observed in the microarray literature, as microarray data is usually log-transformed.

Regarding the selection of clustering algorithms, as for microarrays, Single-Linkage should be avoided at all costs also in the case of RNA-Seq data (Jaskowiak et al., 2012, 2014; de Souto et al., 2008). If one intends to employ a hierarchical clustering algorithm, Average-Linkage should be the method of choice, given it was also superior to Complete-Linkage, in general. We note that k-medoids is also a sound alternative, with a higher stability across different distance measures than Average-Linkage. It does not have, however, the visual appeal of a dendrogram, which can be useful in particular cases.

As for the last experimental factor, previous studies that analyzed the case of clustering algorithms and distance measures for microarray data indicated the use of Pearson, Jackknife, and Symmetric Rank-Magnitude in the case of sample clustering (previous section). In the case of RNA-Seq, rank-based measures (Kendall, Goodman-Kruskal, and Spearman) presented themselves as a sounder option than these measures in most of the cases. Therefore, we believe that Spearman should be the measure of choice, given the previous choices already discussed.

6.6 Chapter Remarks

In this chapter we provided an overview regarding the selection of distance measures for clustering gene expression data. We started by providing a review of different distances that can be employed to the clustering of gene expression data in Section 6.2. We then discussed different strategies that can be employed in the evaluation of distance measures, namely, clustering algorithm independent and dependent, in Section 6.3. In that particular section we introduced a new methodology to assess the quality of distance measures regardless of the bias from clustering algorithms, employing information from the Gene Ontology (GO), namely Intrinsic Biological Separation Ability (IBSA). Results regarding the evaluation of distance measures for data obtained with microarray technology were presented in Section 6.4. As for RNA-Seq data, Section 6.5, apart from the comparison of distances themselves, we covered different experimental factors that can affect clustering results. These factors were not previously addressed in the RNA-Seq literature.

CHAPTER

7

Biological Validation of Gene Clustering Results

In previous chapters we discussed relative validity criteria in the general context of clustering validation. By general we mean that relative measures and their ensembles were introduced with no particular application domain in mind. This also holds in the case of DBCV, reviewed in Chapter 5, which aims at the validation of density-based clustering results. In this chapter we discuss the evaluation of clustering results in the specific domain of gene expression data. In this case, as we discussed in Chapter 3, clustering has two major applications, namely the clustering of samples and genes. Due to their distinct characteristics, these two application scenarios were considered separately in Chapter 6, where we evaluated how different distance measures can affect clustering outcomes. In this chapter, we once more acknowledge the differences between these two scenarios, as we focus specifically on the evaluation of results coming from the clustering of genes.

In the gene clustering scenario, genes are regarded as objects, whereas biological samples constitute the features of the problem. In this case features are generally associated with different time instants regarding the same biological process and, for such a reason, each gene is regarded as a short time-series. The reduced length of the series under investigation usually prevents the adoption of traditional techniques from time-series analysis (Zhang, 2006). Another important characteristic in this scenario is the absence of labeled data. Even though labels will not be available in a real clustering application, they can help in the controlled evaluation and selection of algorithms, which based on their performance, can be later preferred or avoided. As an example,

take the evaluation of distance measures we performed in Chapter 6. In that analysis, the use of class labels provided information regarding the behavior of distances on different datasets, which we expect to reflect, at least in general terms, the results observed in an actual clustering scenario encountered in practice.

Despite the lack of class labels, biological information regarding genes and their relationships has been collected in the past years, as in the case of the Gene Ontology (GO) (Ashburner et al., 2000). In this Chapter we explore the use of this particular type of information in the evaluation of clustering results. More specifically, we make use of biological information in the form of semantic similarities among genes. These are then employed with two different relative validity criteria, in order to evaluate clustering results. Initially, we considered the use of semantic similarities and the combination of semantic similarities with statistical ones[1], obtained from gene expression data.

The remaining of the chapter is organized as follows. In Section 7.1 we discuss related work on the evaluation of gene clustering results, with focus on measures that make use of biological information. In Section 7.2 we elaborate on the use of biological information, in the form of semantic similarities, with relative validity criteria and present results from an empirical evaluation. In Section 7.3, we discuss some undesired properties from a commonly employed validity index with biological bias, namely Biological Homogeneity Index (BHI) (Datta, 2006a).

7.1 Related Work

Although traditional validity criteria do provide an indication of clustering quality, they completely disregard biological information, which is available in the case of gene clustering. In this sense, Handl et al. (2005) argues that a valuable clustering solution should be consistent not only from the perspective of traditional validation procedures, but also from a biological point of view. Bearing this in mind, a handful of works attempted to incorporate biological information from the GO during the validation of gene clustering results, as we review in the following.

The first attempts in this direction were based on the individual assessment of gene clusters. This type of evaluation consists in performing enrichment analysis on different gene lists, one for each cluster (Beißbarth and Speed, 2004). In this context, some tools or methods have been proposed in the literature, *e.g.*, (Beißbarth and Speed, 2004; Boyle et al., 2004; Ernst et al., 2005). Although interesting at first glance, these methods do not provide a final value which accounts for cluster quality, but rather return a list of GO terms and their corresponding p-values. In this case, the evaluation is performed for each cluster, rather than the complete partition. In addition, these assume independence between terms, which does not occur in practice (Costa et al., 2007).

The measure introduced by Gibbons and Roth (2002) was, probably, one of the first to consider biological information and to provide a quantitative evaluation of clustering solutions.

[1]In this chapter biological similarities refer to semantic similarities derived from the Gene Ontology, whereas statistical similarities refer to similarities derived directly from the data. Evaluations based on semantic similarities are denoted GO Based, whereas traditional evaluation, based on statistical similarities, is referred to as Data Based.

The criterion proposed by the authors is based on the mutual information between the labels of a clustering solution and all the terms from the Gene Ontology (GO) that annotate its genes. Here it is important to recall that when a gene is annotated by a term it is also annotated by all of its ancestors, which account for more general concepts than it. By considering all the terms from the GO that annotate the clustering result, the measure introduced by the authors completely ignore redundancies among terms, which can lead to a bias in the process.

Datta (2006b) proposed two different criteria that aim at the evaluation of gene clustering results, namely Biological Stability Index (BSI) and Biological Homogeneity Index (BHI). In the case of BSI, the stability of a clustering solution is assessed with respect to biological information, through the removal of features. Cluster solutions for which such removal has little effect are preferred, as they are recognized as stable. In the case of BHI, clustering results are assessed regarding their homogeneity with respect to biological terms extracted from the GO. Common to the two measures is the need to define, beforehand, the set of GO terms that will be employed during validation. Given that the choice of such terms has direct impact on the final evaluation, and since there is no procedure to guide their selection, the set of terms employed become a key parameter in the process. Although based on different concepts, the measure proposed by Loganantharaj et al. (2006) suffers from the same drawback. Finally, as we shall later discuss, the Biological Homogeneity Index (BHI), which is one of the most popular measures in this context, has a bias towards favoring solutions (partitions) with large numbers of clusters.

In an attempt to mitigate the problems related to term redundancy, Costa et al. (2007) introduced an index based on the selection of relevant and non-redundant terms from the Gene Ontology. The measure is based on the same concepts previously employed by Gibbons and Roth (2002). The difference here is the fact that the authors attempt to select only non-redundant terms for the final evaluation. To accomplish this, their measure favors the selection of general GO terms that can be employed to summarize specific ones, which are then discarded. Although this reduces redundancy, as several specific terms are summarized by a single general one, detailed information is lost. Finally, the index requires the generation of random partitions in order to compute the actual significance of the final evaluation, which can be computationally expensive in some cases.

The works from Bolshakova et al. (2006a) and Bolshakova et al. (2006b) were the first to consider semantic similarities from the Gene Ontology in the evaluation of clustering solutions. We believe, however, that the potential of such measures hasn't been fully explored in these studies. First, in both works a single dataset was employed to evaluate the behavior of semantic similarities. Indeed, Bolshakova et al. (2006b) reduced the number of genes to a total of 63 before performing clustering and evaluation of its results. This clearly does not represent a typical application in the clustering of gene time-series, as usually 1.000 genes (at least) are clustered. To that end, results based on a very limited number of genes may not reflect the ones actually encountered in practice, when a large number of genes is considered.

7.2 Gene Ontology Similarities in Relative Validation

The works of Bolshakova et al. (2006a) and Bolshakova et al. (2006b) evaluated the use of semantic similarities in a very limited context, considering a small number of genes, which are not representative of real applications. Here we evaluated the use of semantic similarities with a broader collection of datasets. Our initial plan was to employ semantic similarities in conjunction with statistical ones, in order to evaluate clustering results from both perspectives. With respect to semantic similarities extracted from the Gene Ontology, we employed the same measure described in Chapter 6, namely the Best-Match Average (BMA) (Pesquita et al., 2008) of the Resnik semantic similarity (Resnik, 1999). As in the previous chapter, we considered, for the same reasons already discussed, the Molecular Function (MF) and Biological Process (BP) ontologies. Regarding expression similarities, we employed the Pearson correlation coefficient, due to its good results and widespread use (we refrain from evaluating distinct distances here).

For this experiments we considered the same collection of datasets already described in Chapter 6. More specifically, given that we are concerned with the evaluation of clusterings of genes, we employed the data collection described in Table 6.2, which comprise 17 datasets of short gene time-series. These datasets are from cDNA microarrays experiments regarding the *Saccharomyces cerevisiae* organism (yeast). Recall that each one of the processed datasets has around 1.000 objects (genes). In order to generate partitions we adopted the very same clustering algorithms previously employed in Chapter 6, namely Single-Linkage (SL), Average-Linkage (AL), Complete-Linkage (CL), and k-medoids (KM). The range for number of clusters was, once more, set to $[\lceil 2, \sqrt{n} \rceil]$, where n is the number of genes (objects).

Regarding relative validity criteria, we employed two measures, namely: the Silhouette Width Criterion (SWC) and the Area Under the Curve (AUC). The selection of these two measures for our experimental evaluation is based on their good results in the evaluation performed in Chapter 5. Moreover, note that although we have the actual data objects, for which statistical similarities can be derived, in the case of semantic similarities we are dealing with a relational clustering scenario, where only similarities among objects are available. This, thus, prevents the direct use of relative validity criteria based on centroid calculations, for example.

7.2.1 Results and Discussion

We start our discussion by providing a summary of the overall performance for AUC and SWC in the 17 datasets regarding the use of semantic similarities. These are provided in Figure 7.1 considering the Molecular Function (MF) ontology and different numbers of clusters. In this figure each boxplot summarizes the performance on the 17 datasets, for the corresponding number of clusters. The first aspect we note regarding the use of AUC (Figure 7.1(a)) is that, overall, the evaluation results stay close to 0.5, regardless of the clustering algorithm employed. To that end, k-medoids provides a slight improvement over the other algorithms. In the case

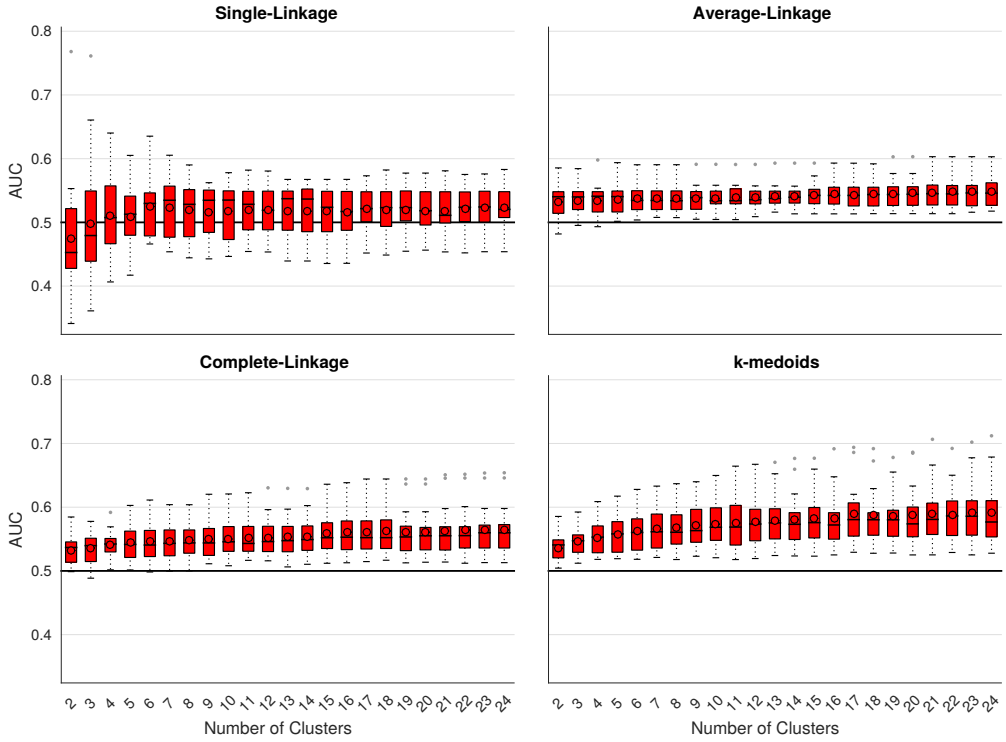
of SWC (Figure 7.1(b)), all clustering algorithms provide poor recoveries when employing semantic similarities with negative values. This contrasts with the results previously presented by Bolshakova et al. (2006b), in which SWC values close to 1.0 were obtained (its maximum value). We believe that this arises due to the limited number of objects considered by the authors in their analysis (63 genes), which oversimplified the problem and its evaluation. We omit results regarding Biological Process (BP) ontology, given that the overall picture does not change.

As previously mentioned, initially our plan was to employ semantic similarities from the Gene Ontology in conjunction with statistical ones, as obtained from gene expression data. Given the poor recoveries obtained with semantic similarities, however, in a first moment we attempted to improve these particular evaluations. To that end, we tried different approaches to improve GO evaluations, such as filter genes with limited or no annotations, and employ annotations based only on specific evidence codes. None of our attempts improved the overall final evaluations though.

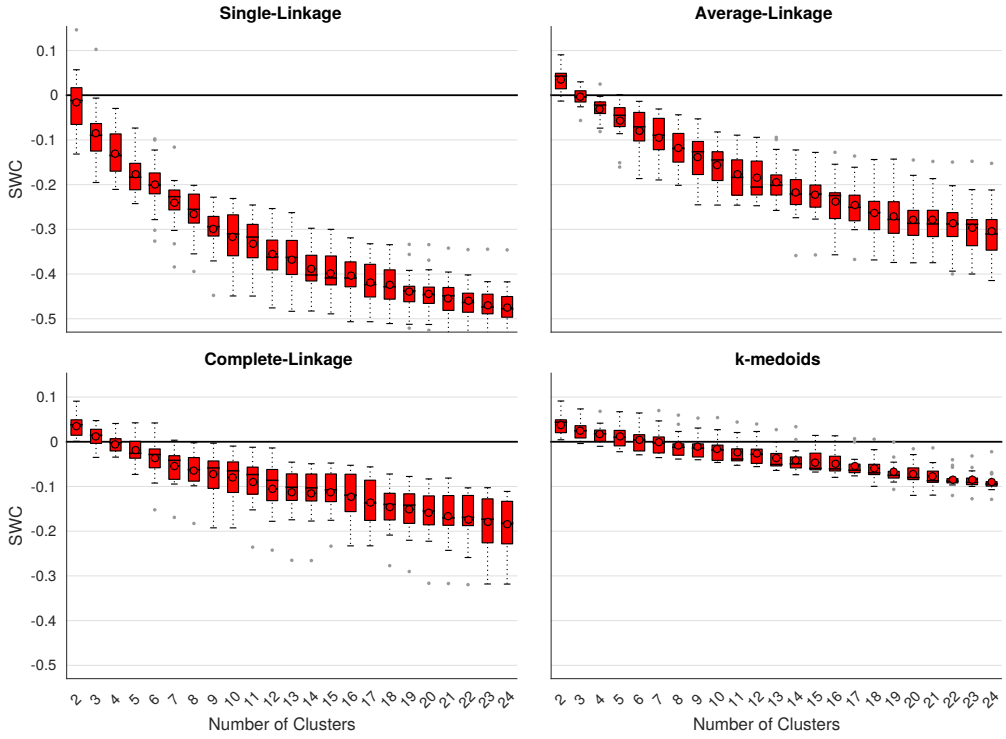
To place the biological evaluation results into perspective, we also performed evaluations based on gene expression similarities (as in a traditional relative evaluation procedure). We further combined such evaluations with the ones obtained from biological information. These results are provided in Figure 7.2, for the *elutriation* dataset, which will serve as an illustrative case (results for the remaining datasets followed similar trends). The first aspect we note is that the y axes in this figure are in a different scale for each plot. Note that data based evaluations provided higher recoveries than those obtained with semantic similarities extracted from the GO. In the last column of this figure we display evaluations based on the combination of biological and statistical similarities. To obtain these values we combined the biological and statistical similarity matrices into a single matrix by their average. This was then supplied to the relative validity criteria.

In the case of biological evaluations, once again, we note that poor recoveries were obtained, no matter if SWC or AUC was used. It is interesting to note, however, that there is some distinction among clustering algorithms, even in this scenario. Specifically, SL results were lower than for those the remaining clustering algorithms. Although the poor performance in the case of SL is no surprise, this shows that there is at least some meaning in the results from such evaluation. In the case of statistical similarities (Data Based) this distinction is clearer than before. In both cases, results from SWC and AUC seem to diverge, *i.e.*, their value trends are different for increasingly number of clusters. Finally, we note that the evaluation based on the combination of statistical and biological similarities seems to be dominated by statistical information, given the similar trends between them.

Based on the previous results, we believe that semantic similarities alone are not enough to provide a reliable estimate of the number of clusters in the case of gene clustering. These results support, however, that the performance of different clustering algorithms can be distinguished by biological information, by the evaluating the overall trend of their results. Given this particular results, we believe that further explorations are necessary in this particular direction and that results from Bolshakova et al. (2006a) and Bolshakova et al. (2006b) are overoptimistic.



(a) Results regarding AUC.



(b) Results regarding SWC.

Figure 7.1: Results regarding relative evaluation based on the GO, AUC left and SWC right. Each boxplot depicts the results regarding 17 datasets, considering a particular number of clusters.

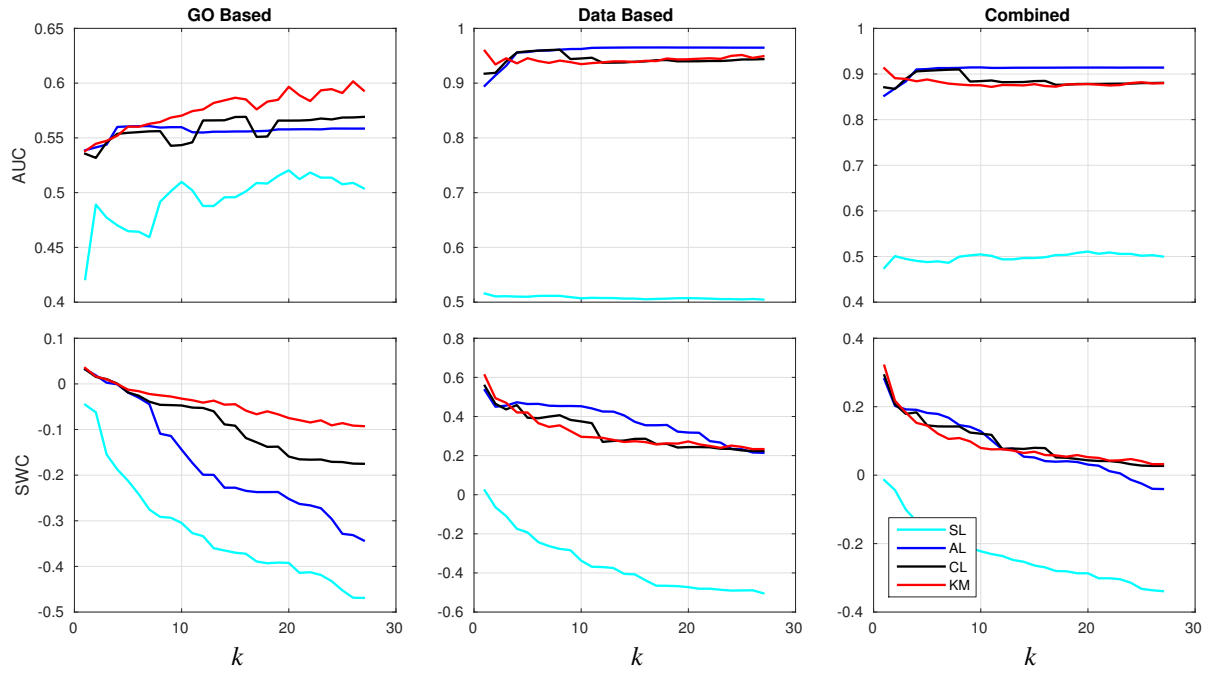


Figure 7.2: Relative validation: biological, statistical, and combined for *elutriation* dataset. In the top plots we depict evaluation results regarding AUC, whereas in the bottom ones SWC values are provided. The evaluations consider different numbers of clusters (k), provided in the x axis.

7.3 Undesired Properties of the BHI

In this section we briefly discuss some undesired properties of the Biological Homogeneity Index (BHI), which was introduced by Datta (2006a). Given a clustering result, BHI measures its homogeneity according to biological functional classes extracted from the Gene Ontology (GO). Given a partition $\mathcal{C} = \{C_1, \ldots, C_k\}$, with n genes and k clusters obtained by applying a clustering algorithm and a set of functional classes $\mathcal{B} = \{B_1, \ldots, B_f\}$ with f functional classes obtained from the GO, the Biological Homogeneity Index (BHI) is defined by Equation (7.1). In that equation $I(B(i), B(j))$ is equal to 1 if two genes have an agreement w.r.t. their functional classes (any single match will suffice), *i.e.*, $B(i) = B(j)$, and 0 otherwise. Intuitively, the BHI criterion verifies if genes in the same cluster have the same biological function. Note that some genes may not be associated with any functional annotations and thus may not be related with a functional class $\mathcal{B}$. This is taken into account with term n_c, which is the number of genes in cluster C with at least one functional annotation. Finally, the number of functional concordances is normalized by the total number of possible concordances and the number of clusters. With such a normalization the criterion produces values in the $[0, 1]$ interval, favoring large values.

$$BHI(\mathcal{C}, \mathcal{B}) = \frac{1}{k} \sum_{c=1}^{k} \frac{1}{n_c(n_c - 1)} \sum_{i \neq j \in \mathcal{C}_c} I(B(i), B(j)) \tag{7.1}$$

As we already discussed, in the case of BHI one has to determine *a priori* which biological classes will be employed during the validation process. This is itself already problematic, as there is no consensus on which biological classes to select, but the measure has also other drawbacks. Note that BHI is an external criterion in essence, as it evaluates a clustering result w.r.t. external labels. In this scenario, Amigó et al. (2009) discuss different desired properties for external measures. From these, two are fundamental to the external evaluation of clustering results:

- **Cluster Homogeneity**: According to the homogeneity property an external validity criterion should prefer homogeneous clustering solutions, that is, solutions for which objects within the same cluster belong to the same class, as defined by the external labeling.
- **Cluster Completeness**: The completeness property states that objects from the same class, as defined by external labels, should belong to the same cluster in a result partition.

These two properties are exemplified by Figure 7.3, considering two cluster solutions.

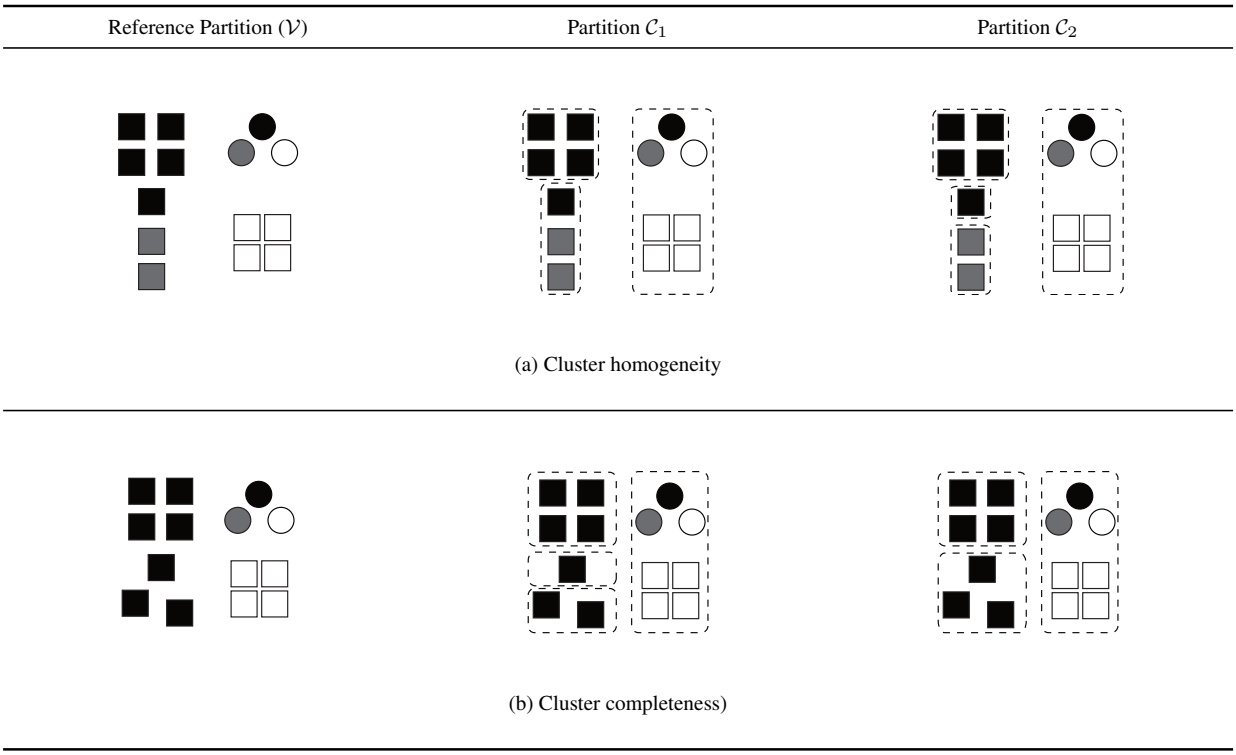


Figure 7.3: Examples regarding cluster homogeneity and completeness properties. The shape and color of each object determine its class, as given by the external partition ($\mathcal{V}$). Each dashed rectangle defines a cluster. According to Amigó et al. (2009), one should have that: $\mathcal{Q}(\mathcal{C}_1, \mathcal{V}) < \mathcal{Q}(\mathcal{C}_2, \mathcal{V})$, that is, considering an external criterion $\mathcal{Q}$, partition $\mathcal{C}_2$ should receive a better evaluation score than $\mathcal{C}_1$, considering the reference partition $\mathcal{V}$. Figure was adapted from (Souto et al., 2012).

It is easy to verify that although the Biological Homogeneity Index (BHI) satisfies the homogeneity condition, it fails to meet the completeness property. In fact, by simply splitting (dividing) an already homogeneous cluster into a number of sub clusters one can easily obtain better results according to the criterion. In an extreme case, an algorithm that produces only

singleton clusters would receive the highest score from the criterion. Although one can argue that this seems unlikely, such a problem actually occurs in practical applications. In the work of Martin et al. (2010), for example, the authors employ BHI in order to measure the quality of partitions generated from a hierarchical clustering algorithm. In that work the authors analyze five proximity measures for the alignment of DNA sequences. In one of the analysis, the proximity measures in question are evaluated by the homogeneity of the groups they generate, according to BHI. This particular analysis is shown in figure 3 of the author's paper, where this behavior is apparent.

Based on these observations we believe that the BHI should not be employed to estimate the final number of clusters in biological applications, although its authors use it for that end (Datta, 2006a). Furthermore, given that BHI is basically an external criterion, we believe that other measure that do not suffer from the same problem discussed here can be employed in its place. For the selection of external measures, the work of Amigó et al. (2009) is a valuable resource.

7.4 Chapter Remarks

In this chapter we considered the evaluation of gene clustering results by incorporating biological information. In the first part of the chapter we explored the use of semantic similarities extracted from the Gene Ontology (GO) in the relative evaluation of gene clustering results. In this particular direction we believe that further explorations are necessary. Although information from the Gene Ontology seems to help in the discrimination between different clustering algorithms, it cannot be used to help in the estimation of the number of clusters for a particular algorithm. Finally, we discussed undesired properties of the Biological Homogeneity Index (BHI). Although the selection of classes is still a problem in this context, we believe that other external measures can be adopted in place of BHI, avoiding thus its bias towards a large number of clusters.

CHAPTER

8

Conclusions

In this thesis we explored different aspects regarding the evaluation of clustering results. In the general domain of clustering evaluation these included: (i) the development of relative measures and (ii) the proposal of ensembles of relative validity criteria. In the specific domain of gene expression data clustering we considered the incorporation of biological information from the Gene Ontology in: (i) the evaluation of distance measures and (ii) the evaluation of gene clustering results. Bellow we provide our concluding remarks regarding each one of these topics.

Regarding measures, we proposed the use of the Area Under the Curve (AUC) of the Receiver Operating Characteristics (ROC) curve as a relative validity criterion for clustering solutions. We showed that the AUC has an expected value of 0.5, regardless of the number of clusters under evaluation. We also showed that the AUC is closely related to the relative validity criterion known as Gamma (Baker and Hubert, 1975). The AUC has, however, a reduced computational cost in comparison to that of Gamma. An empirical evaluation of AUC showed that it ranks well in comparison to measures regarded as state of the art in the literature. These results suggest that the AUC is now a viable and effective alternative to the validation of clustering results. Given that only distances among objects are needed as input to the measure (along with a clustering solution, of course), it has also the potential to be directly applied in the validation of relational clusterings, for which the actual data objects may not be available (Horta, 2010). We believe that these results corroborate our first research hypothesis, regarding the use of AUC in clustering validation. Still with respect to measures for the validation of clustering results, we reviewed a relative validity criterion, named Density-Based Clustering Validation (DBCV) (Moulavi et

al., 2014). This criterion was proposed with participation of the author of this thesis, under the supervision of Prof. Jörg Sander, during his on year internship at University of Alberta.

In the realm of ensembles of relative validity criteria we explored two different scenarios. In the first one, we evaluated the effect of an ad-hoc selection of ensemble members. For this particular scenario we showed that, in practice, little benefits are to be expected. Indeed, with the use of ad-hoc ensembles one can, in general, improve only regarding the worst criterion that constitutes the ensemble. This motivated us to explore ensembles built on the basis of a principled selection of members, our second scenario. To do so, we developed a heuristic that selects ensemble members based on their effectiveness and complementarity levels. In the second scenario we observed an overall improvement in effectiveness and robustness with respect to use of both single criteria and ad-hoc ensembles. We believe that the results regarding ensembles of relative validity criteria support the two research hypotheses we formulated with respect to them.

In the remaining chapters of the thesis we considered the use of semantic similarities from the Gene Ontology to the evaluation of distance measures and clustering results. In the case of distance measures, we developed a systematic approach, namely Intrinsic Biological Separation Ability (IBSA), to evaluate distances regardless of a particular clustering algorithm. Results provided by IBSA were in general accordance with further evaluations, which considered the bias of both clustering algorithm and distance measures, allowing us to distinguish their overall quality. We believe that these results support our research hypothesis regarding the incorporation of biological information to the evaluation of distance measures. In what concerns the biological evaluation of clustering results, we believe that further explorations are necessary. In this particular case there seems to be little agreement between biological information and clustering solutions. Nevertheless, the biological information seems to help to distinguish among the performance of different clustering algorithms, though it cannot be used to help estimating the actual number of clusters from the data. Still regarding the biological evaluation of clustering results, we discussed undesired properties of the Biological Homogeneity Index (BHI), suggesting the use of other well established external indices from the literature as effective alternatives to it.

8.1 Future Work

In this thesis we proposed the Area Under the Curve (AUC) of the Receiver Operating Characteristics (ROC) curve as a relative measure of clustering quality. Although we showed that its expected value is independent of the number of clusters from the solution under evaluation, further developments are necessary towards adjusting the measure for chance, as performed by Marques et al. (2015), for instance (this work is, however, unrelated to the AUC). Given its compelling results, we believe that the application of other measures from the supervised literature deserve further investigation in the clustering domain. Particularly, it would be interesting to assess the potential of Precision-Recall curves (and their corresponding metric, the Area Under

Precision-Recall — AUPR), which in the literature of classification are deemed more informative in the evaluation of highly skewed class distributions (Davis and Goadrich, 2006).

In the case of density-based clustering validation, as pointed out by Moulavi (2014), there are two main components involved in the case of DBCV, these are: (i) the density estimation for each point, regarding the objects of its cluster, and (ii) the estimation of density for paths between object pairs. We believe that it would be interesting to investigate how these two components behave independently of each other. One possibility here is to explore how the density estimation kernel introduced with DBCV, and consequently its distance between objects (mutual reachability distance), behave when employed with other relative validity measures. Graph models, other than the Minimum Spanning Tree (MST), can also be empirically and theoretically evaluated.

Given the variety of relative validity criteria from the literature we believe that efforts towards the meta validation of clustering results are required. This would contemplate the study and development of methods to the automatic selection of a particular relative validity criterion, which is more appropriate to the scenario in hand. Potential benefits are envisioned to the evaluation of clustering results from different paradigms, *e.g.*, distance based and density based. Further benefits can also be explored in the construction of ensembles of relative validity criteria, which can be built on the fly, in an automated fashion, considering the input from the meta validation procedure.

As for the analysis we performed regarding the clustering RNA-Seq data, given the large dimensionality of exon and isoform feature spaces, we believe that further investigation is required. To that end, as we already discussed, their poor recovery rates may be associated with the naiveness of the feature selection we employed. The efforts towards elucidating this question will most probably involve the use and evaluation of more elaborated feature selection algorithms. Another aspect that we believe to deserve further attention is that of estimating the number of clusters in RNA-Seq data. The unconstrained number of cluster scenario we presented provided an upper bound of the qualities one may expect in practical applications. Given that it uses the ARI to select the best partition, it is not appropriate for practical applications. In this direction, the investigation of different relative validity indices in the context of RNA-Seq is necessary, as in other applications.

Finally, we believe that further investigations are necessary regarding the validation of gene clusterings coming from gene expression data. Given the results obtained with the application of relative criteria and semantic similarities extracted from the Gene Ontology, this would potentially include the investigation of external measures. In this direction, we believe that adaptations of the Biological Homogeneity Index (BHI) (Datta, 2006a), which can prevent an artificial increase of its values in the case of clusters splits, are a good starting point that can provide practical benefits.

8.2 Publications

Throughout the PhD, I have published journal and conference papers. Some of these are directly related to this thesis, whereas others are partially related, as presented bellow.

Journal Papers

- Jaskowiak, P.A.; Moulavi D.; Furtado, A.C.S.; Campello, R.J.G.B.; Zimek, A.; Sander, J. *On Strategies for Building Effective Ensembles of Relative Clustering Validity Criteria*. Knowledge and Information KAIS (Systems) — Accepted — In Print.

- de Souto, M.C.P.; Jaskowiak, P.A.; Costa, I. G. *Impact of missing data imputation methods on gene expression clustering and classification*. BMC Bioinformatics, v. 16, p. 09, 2015.

- Barros, R.C.; Jaskowiak, P.A.; Cerri, R.; Carvalho, A.C.P.L.F. *A framework for bottom-up induction of oblique decision trees*. Neurocomputing (Amsterdam), v. 135, p. 3-12, 2014.

- Jaskowiak, P. A.; Campello, R.J.G.B.; Costa, I.G. *On the selection of appropriate distances for gene expression data clustering*. BMC Bioinformatics, v. 15, p. S2, 2014.

- Jaskowiak, P.A.; Campello, R.J.G.B.; Costa, I.G. *Proximity Measures for Clustering Gene Expression Microarray Data: A Validation Methodology and a Comparative Analysis*. IEEE/ACM Trans. on Computational Biology and Bioinformatics, v.10, p. 845-857, 2013.

Conference Papers

- Jaskowiak, P.A.; Campello, R.J.G.B. *A Cluster Based Hybrid Feature Selection Approach*. In: Brazilian Conference on Intelligent Systems (BRACIS 2015), 2015, Natal - Rio Grande do Norte, Brazil. Proceedings of the Brazilian Conference on Intelligent Systems (BRACIS), 2015, p. 1-6.

- Moulavi, D.; Jaskowiak, P.A.; Campello, R.J.G.B.; Zimek, A.; Sander, J. *Density-Based Clustering Validation*. In: SIAM International Conference on Data Mining (SDM 2014), 2014, Philadelphia, Pennsylvania, US. Proceedings of the 14th SIAM International Conference on Data Mining, 2014. p. 1-9.

- Vendramin, L.; Jaskowiak, P.A.; Campello, R.J.G.B. *On the Combination of Relative Clustering Validity Criteria*. In: 25th International Conference on Scientific and Statistical Database Management (SSDBM 2013), 2013, Baltimore. Proceedings of the 25th International Conference on Scientific and Statistical Database Management - SSDBM. New York: ACM Press, 2013. p. 1-12.

- Jaskowiak, P.A.; Campello, R.J.G.B.; Costa, I.G. *Evaluating Correlation Coefficients for Clustering Gene Expression Profiles of Cancer*. In: VII Brazilian Symposium on Bioinformatics (BSB 2012), 2012, Campo Grande. Proceedings of the 7th Brazilian Symposium on Bioinformatics. Berlin / Heidelberg: Springer, 2012. v. 7409. p. 120-131.

- Barros, R.C.; Cerri, R.; Jaskowiak, P.A.; Carvalho, A.C.P.L.F. *A Bottom-Up Oblique Decision Tree Induction Algorithm*. In: International Conference on Intelligent Systems Design and Applications (ISDA 2011), 2011, Córdoba. Proceedings of the 11th International Conference on Intelligent Systems Design and Applications, 2011. p. 450-456.

- Jaskowiak, P.A.; Campello, R. J. G. B. *Comparing Correlation Coefficients as Dissimilarity Measures for Cancer Classification in Gene Expression Data*. In: VI Brazilian Symposium on Bioinformatics (BSB2011), 2011, Brasília. Proceedings of the 6th Brazilian Symposium on Bioinformatics, 2011. p. 1-8.

Finally, we note that some parts of this thesis are still in process of publication.

References

ALBALATE, A.; SUENDERMANN, D. A combination approach to cluster validation based on statistical quantiles. In: *Proceedings of the International Joint Conference on Bioinformatics, Systems Biology and Intelligent Computing (IJCBS), Shanghai, China*, pp. 549–555, 2009. (Cited on page 38)

ALBERTS, B.; JOHNSON, A.; LEWIS, J.; MORGAN, D.; RAFF, M.; ROBERTS, K.; WALTER, P. *Molecular biology of the cell*. 6 edn.. Garland Science, 2014. (Cited on pages 23, 24, and 25)

ALIZADEH, A. A.; EISEN, M. B.; DAVIS, R. E.; MA, C.; LOSSOS, I. S.; ROSENWALD, A.; BOLDRICK, J. C.; SABET, H.; TRAN, T.; YU, X.; POWELL, J. I.; YANG, L.; MARTI, G. E.; MOORE, T.; HUDSON, J.; LU, L.; LEWIS, D. B.; TIBSHIRANI, R.; SHERLOCK, G.; CHAN, W. C.; GREINER, T. C.; WEISENBURGER, D. D.; ARMITAGE, J. O.; WARNKE, R.; LEVY, R.; WILSON, W.; GREVER, M. R.; BYRD, J. C.; BOTSTEIN, D.; BROWN, P. O.; STAUDT, L. M.; JR, J. H. Distinct types of diffuse large B-cell lymphoma identified by gene expression profiling. *Nature*, vol. 403, no. 6769, pp. 503–511, 2000. (Cited on page 31)

ALON, U.; BARKAI, N.; NOTTERMAN, D. A.; GISH, K.; YBARRA, S.; MACK, D.; LEVINE, A. J. Broad patterns of gene expression revealed by clustering analysis of tumor and normal colon tissues probed by oligonucleotide arrays. *Proceedings of the National Academy of Sciences of the United States of America*, vol. 96, no. 12, pp. 6745–6750, 1999. (Cited on page 31)

AMIGÓ, E.; GONZALO, J.; ARTILES, J.; VERDEJO, F. A comparison of extrinsic clustering evaluation metrics based on formal constraints. *Information Retrieval*, vol. 12, no. 5, p. 613, 2009. (Cited on pages 15, 126, and 127)

ANKERST, M.; BREUNIG, M. M.; KRIEGEL, H.-P.; SANDER, J. OPTICS: Ordering points to identify the clustering structure. In: *Proceedings of the ACM International Conference on Management of Data (SIGMOD), Philadelphia, PA*, pp. 49–60, 1999. (Cited on page 80)

ASHBURNER, M.; BALL, C. A.; BLAKE, J. A.; BOTSTEIN, D.; BUTLER, H.; CHERRY, J. M.; DAVIS, A. P.; DOLINSKI, K.; DWIGHT, S. S.; EPPIG, J. T.; HARRIS, M. A.; HILL, D. P.; ISSEL-TARVER, L.; KASARSKIS, A.; LEWIS, S.; MATESE, J. C.; RICHARDSON, J. E.; RINGWALD, M.; RUBIN, G. M.; SHERLOCK, G. Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nature Genetics*, vol. 25, no. 1, pp. 25–29, 2000. (Cited on pages 3, 5, 32, 84, 92, 93, 96, and 120)

BAKER, F. B.; HUBERT, L. J. Measuring the power of hierarchical cluster analysis. *Journal of the American Statistical Association*, vol. 70, no. 349, pp. 31–38, 1975. (Cited on pages 4, 5, 63, 64, 69, 70, 71, 82, and 129)

BALASUBRAMANIYAN, R.; HULLERMEIER, E.; WESKAMP, N.; KAMPER, J. Clustering of gene expression data using a local shape-based similarity measure. *Bioinformatics*, vol. 21, pp. 1069–1077, 2005. (Cited on pages 3, 32, 85, and 90)

BEISSBARTH, T.; SPEED, T. P. GOstat: find statistically overrepresented Gene Ontologies within a group of genes. *Bioinformatics*, vol. 20, no. 9, pp. 1464–1465, 2004. (Cited on pages 96, 97, and 120)

BELK, C.; BORDEN, V. *Biology: Science for Life*. Prentice Hall, 2003. (Cited on page 23)

BEN-DOR, A.; SHAMIR, R.; YAKHINI, Z. Clustering Gene Expression Patterns. *Journal of Computational Biology*, vol. 6, no. 3-4, pp. 281–297, 1999. (Cited on page 31)

BEZDEK, J. C. *Pattern recognition with fuzzy objective function algorithms*. Kluwer Academic Publishers, 1981. (Cited on page 8)

BEZDEK, J. C.; EHRLICH, R.; FULL, W. FCM: The fuzzy c-means clustering algorithm. *Computers & Geosciences*, vol. 10, no. 2, pp. 191–203, 1984. (Cited on page 8)

BEZDEK, J. C.; PAL, N. R. Some new indexes of cluster validity. *IEEE Transactions on Systems, Man and Cybernetics, Part B*, vol. 28, no. 3, pp. 301–315, 1998. (Cited on pages 3 and 17)

BILENKO, M. Integrating constraints and metric learning in semi-supervised clustering. *Proceedings of the 21st International Conference on Machine Learning, (ICML-2004)*, pp. 81–88, 2004. (Cited on page 7)

BISHOP, C. M. *Pattern recognition and machine learning*. Springer, 2006. (Cited on pages 9 and 10)

BOLSHAKOVA, N.; AZUAJE, F. Cluster validation techniques for genome expression data. *Signal Processing*, vol. 83, no. 4, pp. 825–833, 2003. (Cited on page 37)

BOLSHAKOVA, N.; AZUAJE, F.; CUNNINGHAM, P. A knowledge-driven approach to cluster validity assessment. *Bioinformatics*, vol. 21, no. 10, pp. 2546–2547, 2005. (Cited on page 96)

BOLSHAKOVA, N.; AZUAJE, F.; CUNNINGHAM, P. Incorporating Biological Domain Knowledge into Cluster Validity Assessment. In: *Applications of Evolutionary Computing*, vol. 3907 of *Lecture Notes in Computer Science*, Berlin, Heidelberg: Springer Berlin Heidelberg, pp. 13–22, 2006a. (Cited on pages 121, 122, and 123)

BOLSHAKOVA, N.; ZAMOLOTSKIKH, A.; CUNNINGHAM, P. Comparison of the Data-based and Gene Ontology-Based Approaches to Cluster Validation Methods for Gene Microarrays. In: *Computer-Based Medical Systems, 2006. CBMS 2006. 19th IEEE International Symposium on*, pp. 539–543, 2006b. (Cited on pages 96, 121, 122, and 123)

BOYLE, E. I.; WENG, S.; GOLLUB, J.; OTHERS GO::TermFinder - Open source software for accessing Gene Ontology information and finding significantly enriched Gene Ontology terms associated with a list of genes. *Bioinformatics*, vol. 20, no. 18, pp. 3710–3715, 2004. (Cited on page 120)

BRAZMA, A.; VILO, J. Gene expression data analysis. *FEBS Letters*, vol. 480, no. 1, pp. 17–24, 2000. (Cited on pages 23, 30, and 85)

BROWN, M. B.; FORSYTHE, A. B. Robust tests for the equality of variances. *Journal of the American Statistical Association*, vol. 69, no. 346, pp. 364–367, 1974. (Cited on page 41)

CALINSKI, R.; HARABASZ, J. A dentrite method for cluster analysis. *Commun Stat*, vol. 3, pp. 1–27, 1974. (Cited on page 16)

CAMPELLO, R. J. G. B.; HRUSCHKA, E. R. On comparing two sequences of numbers and its applications to clustering analysis. *Information Sciences*, vol. 179, no. 8, pp. 1025–1039, 2009. (Cited on pages 42 and 88)

CAMPELLO, R. J. G. B.; MOULAVI, D.; SANDER, J. Density-based clustering based on hierarchical density estimates. In: *Proceedings of the 17th Pacific-Asia Conference on Knowledge Discovery and Data Mining (PAKDD), Gold Coast, Australia*, pp. 160–172, 2013. (Cited on pages 8, 9, 13, and 78)

CAMPELLO, R. J. G. B.; MOULAVI, D.; ZIMEK, A.; SANDER, J. Hierarchical density estimates for data clustering, visualization, and outlier detection. *ACM Transactions on Knowledge Discovery from Data*, vol. 10, no. 1, 2015. (Cited on pages 9, 13, and 78)

CERIANI, L.; VERME, P. The origins of the gini index: extracts from variabilità e mutabilità (1912) by corrado gini. *The Journal of Economic Inequality*, vol. 10, no. 3, pp. 421–443, 2012. (Cited on page 71)

CHARTRAND, G. *Introductory graph theory*. Dover Publications, 1985. (Cited on page 33)

CHOU, C.-H.; SU, M.-C.; LAI, E. A new cluster validity measure and its application to image compression. *Pattern Analysis and Applications*, vol. 7, no. 2, pp. 205–220, 2004. (Cited on page 74)

CORMACK, G. V.; CLARKE, C. L. A.; BUETTCHER, S. Reciprocal rank fusion outperforms Condorcet and individual rank learning methods. In: *Proceedings of the 32nd International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '09, pp. 758–759, 2009 (*SIGIR '09*,). (Cited on page 49)

COSTA, I. G.; DE CARVALHO, F. A. T.; DE SOUTO, M. C. P. Comparative analysis of clustering methods for gene expression time course data. *Genetics and Molecular Biology*, vol. 27, no. 4, pp. 623–631, 2004. (Cited on pages 84 and 85)

COSTA, I. G.; ROEPCKE, S.; HAFEMEISTER, C.; SCHLIEP, A. Inferring differentiation pathways from gene expression. *Bioinformatics*, vol. 24, no. 13, pp. i156–i164, 2008. (Cited on page 96)

COSTA, I. G.; SCHÖNHUTH, A.; SCHLIEP, A. The graphical query language: a tool for analysis of gene expression time-courses. *Bioinformatics*, vol. 21, no. 10, pp. 2544–2545, 2005. (Cited on pages 31 and 109)

COSTA, L. D. F.; RODRIGUES, F. A.; TRAVIESO, G.; VILLAS BOAS, P. R. Characterization of complex networks: A survey of measurements. *Advances in Physics*, vol. 56, no. 1, pp. 167–242, 2007. (Cited on pages 120 and 121)

DALMA-WEISZHAUSZ, D. D.; WARRINGTON, J.; TANIMOTO, E. Y.; MIYADA, C. G. The affymetrix GeneChip platform: an overview. *Methods in Enzymology*, vol. 410, pp. 3–28, 2006. (Cited on pages 27 and 28)

DATTA, S. Evaluation of clustering algorithms for gene expression data. *BMC Bioinformatics*, vol. 7, no. Suppl 4, p. S17, 2006a. (Cited on pages 120, 125, 127, and 131)

DATTA, S. S. Comparisons and validation of statistical clustering techniques for microarray gene expression data. *Bioinformatics*, vol. 19, no. 4, pp. 459–466, 2003. (Cited on page 84)

DATTA, S. S. Methods for evaluating clustering algorithms for gene expression data using a reference set of functional classes. *BMC Bioinformatics*, vol. 7, no. 1, p. 397, 2006b. (Cited on page 121)

DAVIES, D.; BOULDIN, D. A cluster separation measure. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 1, pp. 224–227, 1979. (Cited on pages 18, 37, and 74)

DAVIS, J.; GOADRICH, M. The relationship between precision-recall and roc curves. In: *Proceedings of the 23rd International Conference on Machine Learning*, ICML '06, New York, NY, USA: ACM, pp. 233–240, 2006 (*ICML '06*,). (Cited on page 131)

DE BORDA, J.-C. Mémoire sur les élections au scrutin. *Histoire de l'Academie Royale des Sciences*, pp. 657–665, 1781. (Cited on page 48)

DE SOUTO, M. C. P.; COELHO, L. V.; FACELI, K.; SAKATA, T. C.; COSTA, I. G. A comparison of external clustering evaluation indices in the context of imbalanced data sets. In: *2012 Brazilian Symposium on Neural Networks (SBRN)*, pp. 49–54, 2012. (Cited on page 15)

DEMŠAR, J. Statistical Comparisons of Classifiers over Multiple Data Sets. *Journal of Machine Learning Research*, vol. 7, pp. 1–30, 2006. (Cited on pages 72, 99, 104, and 113)

D'HAESELEER, P. How does gene expression clustering work? *Nature Biotechnology*, vol. 23, no. 12, pp. 1499–1501, 2005. (Cited on pages 31, 84, 85, and 86)

DUDOIT, S.; FRIDLYAND, J. A prediction-based resampling method for estimating the number of clusters in a dataset. *Genome biology*, vol. 3, no. 7, pp. 0036.1–0036.21, 2002. (Cited on page 14)

DUGGAN, D. J.; BITTNER, M.; CHEN, Y.; MELTZER, P.; TRENT, J. M. Expression profiling using cDNA microarrays. *Nature Genetics*, vol. 21, pp. 10–14, 1999. (Cited on pages 27 and 28)

DUNN, J. Well separated clusters and optimal fuzzy partitions. *Journal of Cybernetics*, vol. 4, pp. 95–104, 1974. (Cited on pages 16 and 74)

DWORK, C.; KUMAR, R.; NAOR, M.; SIVAKUMAR, D. Rank aggregation methods for the web. In: *Proceedings of the 10th International World Wide Web Conference (WWW), Hong Kong, China*, pp. 613–622, 2001. (Cited on pages 48, 49, and 50)

ERNST, J.; BAR-JOSEPH, Z. Stem: a tool for the analysis of short time series gene expression data. *BMC Bioinformatics*, vol. 7, no. 1, p. 191, 2006. (Cited on pages 96 and 109)

ERNST, J.; NAU, G. J.; BAR-JOSEPH, Z. Clustering short time series gene expression data. *Bioinformatics*, vol. 21, pp. i159–i168, 2005. (Cited on pages 31 and 120)

ESTER, M.; KRIEGEL, H.-P.; SANDER, J.; XU, X. A density-based algorithm for discovering clusters in large spatial databases with noise. In: *Proceedings of the 2nd ACM International Conference on Knowledge Discovery and Data Mining (KDD), Portland, OR*, pp. 226–231, 1996. (Cited on pages 9, 12, and 78)

ESTEVES, G. H. *Métodos Estat\'{i}sticos para a Análise de Dados de c{DNA} microarray em um Ambiente Computacional Integrado*. Ph.D. thesis, Universidade de São Paulo, 2007. (Cited on page 28)

ESTIVILL-CASTRO, V. Why so many clustering algorithms – a position paper. *ACM SIGKDD Explorations*, vol. 4, no. 1, pp. 65–75, 2002. (Cited on page 1)

EVERITT, B. *Cluster analysis*. Heinemann Educational for the Social Science Research Council London, 122 pp., 1974. (Cited on page 9)

FACELI, K.; CARVALHO, A. A. C. D. A. C. P. L. F.; SILVA JR, W. A. Evaluation of gene selection metrics for tumor cell classification. *Genetics and Molecular Biology*, vol. 27, no. 4, pp. 651–657, 2004. (Cited on page 98)

FAN, J.; HAN, F.; LIU, H. Challenges of big data analysis. *National Science Review*, vol. 1, no. 2, pp. 293–314, 2014. (Cited on page 23)

FAWCETT, T. An introduction to roc analysis. *Pattern Recognition Letters*, vol. 27, no. 8, pp. 861–874, 2006. (Cited on pages 2, 64, 65, 70, and 71)

FAYYAD, U.; PIATETSKY-SHAPIRO, G.; SMYTH, P. Knowledge discovery and data mining: Towards a unifying framework. In: *Proceedings of the 2nd ACM International Conference on Knowledge Discovery and Data Mining (KDD), Portland, OR*, pp. 82–88, 1996. (Cited on page 1)

FINOTELLO, F.; DI CAMILLO, B. Measuring differential gene expression with rna-seq: challenges and strategies for data analysis. *Briefings in Functional Genomics*, 2014. (Cited on page 23)

FRANK, A.; ASUNCION, A. UCI machine learning repository. http://archive.ics.uci.edu/ml, 2010. (Cited on pages 57 and 79)

FREYHULT, E.; LANDFORS, M.; ONSKOG, J.; HVIDSTEN, T.; RYDEN, P. Challenges in microarray class discovery: a comprehensive examination of normalization, gene selection and clustering. *BMC Bioinformatics*, vol. 11, no. 1, p. 503, 2010. (Cited on pages 84 and 85)

FRIEDMAN, M. The Use of Ranks to Avoid the Assumption of Normality Implicit in the Analysis of Variance. *Journal of the American Statistical Association*, vol. 32, no. 200, pp. 675–701, 1937. (Cited on page 60)

FRIEDMAN, M. A comparison of alternative tests of significance for the problem of m rankings. *The Annals of Mathematical Statistics*, vol. 11, no. 1, pp. 86–92, 1940. (Cited on page 41)

FROHLICH, H.; SPEER, N.; POUSTKA, A.; BEISSBARTH, T. Gosim - an r-package for computation of information theoretic go similarities between terms and gene products. *BMC Bioinformatics*, vol. 8, no. 1, p. 166, 2007. (Cited on page 95)

GASCH, A. P.; SPELLMAN, P. T.; KAO, C. M.; CARMEL-HAREL, O.; EISEN, M. B.; STORZ, G.; BOTSTEIN, D.; BROWN, P. O. Genomic Expression Programs in the Response of Yeast Cells to Environmental Changes. *Molecular Biology of the Cell*, vol. 11, no. 12, pp. 4241–4257, 2000. (Cited on pages 84 and 98)

GENTLEMAN, R.; CAREY, V.; HUBER, W.; HAHNE, F. *genefilter: methods for filtering genes from high-throughput experiments*. R package version 1.48.1, 2014. (Cited on page 111)

GENTLEMAN, R.; DING, B.; DUDOIT, S.; IBRAHIM, J. Distance Measures in DNA Microarray Data Analysis. In: *Bioinformatics and Computational Biology Solutions Using R and Bioconductor*, Statistics for Biology and Health, Springer New York, pp. 189–208, 2005. (Cited on page 85)

GEUSEBROEK, J. M.; BURGHOUTS, G. J.; SMEULDERS, A. W. M. The Amsterdam Library of Object Images. *International Journal of Computer Vision*, vol. 61, no. 1, pp. 103–112, 2005. (Cited on pages 39 and 56)

GHOSH, J.; ACHARYA, A. Cluster ensembles. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, vol. 1, no. 4, pp. 305–315, 2011. (Cited on pages 3 and 37)

GIANCARLO, R.; BOSCO, G. L.; PINELLO, L.; UTRO, F. The Three Steps of Clustering in the Post-Genomic Era: A Synopsis. In: *Computational Intelligence Methods for Bioinformatics and Biostatistics*, vol. 6685 of *Lecture Notes in Computer Science*, Springer Berlin / Heidelberg, pp. 13–30, 2011. (Cited on pages 85, 92, and 93)

GIANCARLO, R.; LO BOSCO, G.; PINELLO, L. Distance Functions, Clustering Algorithms and Microarray Data Analysis. In: BLUM, C.; BATTITI, R., eds. *Learning and Intelligent Optimization*, vol. 6073 of *Lecture Notes in Computer Science*, Springer Berlin / Heidelberg, pp. 125–138, 2010. (Cited on pages 85, 92, and 93)

GIANCARLO, R.; LO BOSCO, G.; PINELLO, L.; UTRO, F. A methodology to assess the intrinsic discriminative ability of a distance function and its interplay with clustering algorithms for microarray data analysis. *BMC Bioinformatics*, vol. 14, no. Suppl 1, p. S6, 2013. (Cited on page 66)

GIBBONS, F. D.; ROTH, F. P. Judging the Quality of Gene Expression-Based Clustering Methods Using Gene Annotation. *Genome Research*, vol. 12, no. 10, pp. 1574–1581, 2002. (Cited on pages 85, 120, and 121)

GINI, C. *Variabilità e mutabilità*. Tipogr. di P. Cuppini, 1912. (Cited on page 71)

GOLUB, T. R.; SLONIM, D. K.; TAMAYO, P.; HUARD, C.; GAASENBEEK, M.; MESIROV, J. P.; COLLER, H.; LOH, M. L.; DOWNING, J. R.; CALIGIURI, M. A.; BLOOMFIELD, C. D.; LANDER, E. S.; OTHERS Molecular Classification of Cancer: Class Discovery and Class Prediction by Gene Expression Monitoring. *Science*, vol. 286, no. 5439, pp. 531–537, 1999. (Cited on page 31)

GOODMAN, L.; KRUSKAL, W. Measures of association for cross-classifications. *Journal of the American Statistical Association*, vol. 49, pp. 732–764, 1954. (Cited on page 88)

GORDON, A. D. *Classification*. New York: Chapman and Hall, 1999. (Cited on page 13)

HALKIDI, M.; VAZIRGIANNIS, M. Clustering validity assessment: finding the optimal partitioning of a data set. *Proceedings 2001 IEEE International Conference on Data Mining*, pp. 187–194, 2001. (Cited on page 75)

HALKIDI, M.; VAZIRGIANNIS, M. A density-based cluster validity approach using multi-representatives. *Pattern Recognition Letters*, vol. 29, pp. 773–786, 2008. (Cited on pages 75 and 79)

HALKIDI, M.; VAZIRGIANNIS, M.; BATISTAKIS, Y. Quality scheme assessment in the clustering process. *PKDD 2000*, pp. 265–276, 2000. (Cited on page 75)

HAND, D. J.; TILL, R. J. A Simple Generalisation of the Area Under the ROC Curve for Multiple Class Classification Problems. *Machine Learning*, vol. 45, no. 2, pp. 171–186, 2001. (Cited on page 93)

HANDL, J.; KNOWLES, J.; KELL, D. B. Computational cluster validation in post-genomic data analysis. *Bioinformatics*, vol. 21, no. 15, pp. 3201–3212, 2005. (Cited on page 120)

HARRINGTON, C. A.; ROSENOW, C.; RETIEF, J. Monitoring gene expression using DNA microarrays. *Current Opinion in Microbiology*, vol. 3, no. 3, pp. 285–291, 2000. (Cited on page 27)

HARTIGAN, J. A. *Clustering algorithms*. John Wiley&Sons, 1975. (Cited on page 1)

HESTILOW, T. J.; HUANG, Y. Clustering of Gene Expression Data Based on Shape Similarity. *EURASIP Journal on Bioinformatics and Systems Biology*, vol. 2009, p. 12, 2009. (Cited on page 31)

HEYER, L. J.; KRUGLYAK, S.; YOOSEPH, S. Exploring expression data: Identification and analysis of coexpressed genes. *Genome Research*, vol. 9, no. 11, pp. 1106–1115, 1999. (Cited on pages 31, 32, 84, 85, and 89)

HILL, R. S. A stopping rule for partitioning dendrograms. *Botanical Gazette*, vol. 141, pp. 321–324, 1980. (Cited on page 19)

HORTA, D. *Abordagens evolutivas para agrupamento relacional de dados*. Master's thesis, Universidade de São Paulo, 2010. (Cited on page 129)

HORTA, D.; CAMPELLO, R. J. G. B. Automatic aspect discrimination in data clustering. *Pattern Recognition*, vol. 45, no. 12, pp. 4370–4388, 2012. (Cited on page 39)

HRUSCHKA, E. R.; CAMPELLO, R. J. G. B.; CASTRO, L. N. Improving the efficiency of a clustering genetic algorithm. In: *Ibero-American Conference on Artificial Intelligence — IBERAMIA*, vol. 3315, pp. 861–870, 2004. (Cited on page 15)

HUBERT, L.; ARABIE, P. Comparing partitions. *Journal of Classification*, vol. 2, no. 1, pp. 193–218, 1985. (Cited on page 14)

HUBERT, L. J.; LEVIN, J. R. A general statistical framework for assessing categorical clustering in free recall. *Psychological Bulletin*, vol. 10, pp. 1072–1080, 1976. (Cited on page 19)

HUMAN GENOME PROGRAM Primer on Molecular Genetics. 1992. (Cited on page 25)

JAIN, A. K. Data clustering: 50 years beyond k-means. *Pattern Recognition Letters*, vol. 31, no. 8, pp. 651–666, 2010. (Cited on page 1)

JAIN, A. K.; DUBES, R. C. *Algorithms for clustering data*. Englewood Cliffs: Prentice Hall, 1988. (Cited on pages 1, 2, 7, 9, 13, 31, 35, 40, 42, 85, 87, and 99)

JASKOWIAK, P. A. *Estudo de coeficientes de correlação para medidas de proximidade em dados de expressão gênica*. Master's thesis, University of São Paulo, São Carlos, São Paulo, Brazil, 2011. (Cited on pages 32 and 84)

JASKOWIAK, P. A.; CAMPELLO, R. J. G. B.; COSTA, I. G. Evaluating correlation coefficients for clustering gene expression profiles of cancer. In: *7th Brazilian Symposium on Bioinformatics (BSB2012)*, Springer / Berlin Heidelberg, pp. 120–131, 2012 (*LNCS*, vol.7409). (Cited on page 117)

JASKOWIAK, P. A.; CAMPELLO, R. J. G. B.; COSTA, I. G. On the selection of appropriate distances for gene expression data clustering. *BMC bioinformatics*, vol. 15 Suppl 2, no. Suppl 2, p. S2, 2014. (Cited on pages 5 and 117)

JASKOWIAK, P. A.; CAMPELLO, R. J. G. B.; COSTA FILHO, I. G. Proximity measures for clustering gene expression microarray data: A validation methodology and a comparative analysis. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, vol. 10, no. 4, pp. 845–857, 2013. (Cited on pages 5 and 66)

JASKOWIAK, P. A.; CAMPELLO, R. J. G. B.; COVÕES, T. F.; HRUSCHKA, E. R. A Comparative Study on the Use of Correlation Coefficients for Redundant Feature Elimination.

In: *11th Brazilian Symposium on Neural Networks (SBRN 2010)*, São Bernardo do Campo, São Paulo, Brazil, pp. 13–18, 2010. (Cited on page 85)

JASKOWIAK, P. A.; MOULAVI, D.; FURTADO, A. C. S.; CAMPELLO, R. J. G. B.; ZIMEK, A.; SANDER, J. On strategies for building effective ensembles of relative clustering validity criteria. *Knowledge and Information Systems (KAIS)*, 2015. (Cited on pages 5 and 37)

JIANG, D.; TANG, C.; ZHANG, A. Cluster analysis for gene expression data: A survey. *IEEE Transactions on Knowledge and Data Engineering*, vol. 16, no. 11, pp. 1370–1386, 2004. (Cited on pages 8, 11, 28, 83, and 84)

KANEHISA, M.; GOTO, S.; FURUMICHI, M.; TANABE, M.; HIRAKAWA, M. {KEGG} for representation and analysis of molecular networks involving diseases and drugs. *Nucleic Acids Research*, vol. 38, no. suppl 1, pp. D355—-D360, 2010. (Cited on page 32)

KAO, J.; SALARI, K.; BOCANEGRA, M.; CHOI, Y.-L.; GIRARD, L.; GANDHI, J.; KWEI, K. A.; HERNANDEZ-BOUSSARD, T.; WANG, P.; GAZDAR, A. F.; MINNA, J. D.; POLLACK, J. R. Molecular profiling of breast cancer cell lines defines relevant tumor models and provides a resource for cancer gene discovery. *PLoS ONE*, vol. 4, no. 7, p. e6146, 2009. (Cited on page 31)

KAUFMAN, L.; ROUSSEEUW, P. J. *Finding groups in data: An introduction to cluster analyis*. John Wiley&Sons, 1990. (Cited on page 8)

KENDALL, M. G. A new measure of rank correlation. *Biometrika*, vol. 30, no. 1–2, pp. 81–93, 1938. (Cited on page 88)

KERR, G.; RUSKIN, H. J.; CRANE, M.; DOOLAN, P. Techniques for clustering gene expression data. *Computers in Biology and Medicine*, vol. 38, no. 3, pp. 283–293, 2008. (Cited on pages 31, 84, and 88)

KLEMENTIEV, A.; ROTH, D.; SMALL, K. An unsupervised learning algorithm for rank aggregation. In: *Proceedings of the 18th European Conference on Machine Learning (ECML), Warsaw, Poland*, pp. 616–623, 2007. (Cited on page 50)

KOLDE, R.; LAUR, S.; ADLER, P.; VILO, J. Robust rank aggregation for gene list integration and meta-analysis. *Bioinformatics*, vol. 28, no. 4, pp. 573–580, 2012. (Cited on pages 50 and 51)

KRIEGEL, H.-P.; KRÖGER, P.; SCHUBERT, E.; ZIMEK, A. Interpreting and unifying outlier scores. In: *Proceedings of the 11th SIAM International Conference on Data Mining (SDM), Mesa, AZ*, pp. 13–24, 2011. (Cited on page 47)

KUNCHEVA, L. I.; WHITAKER, C. J. Measures of diversity in classifier ensembles and their relationship with the ensemble accuracy. *Machine Learning*, vol. 51, pp. 181–207, 2003. (Cited on page 46)

KUO, W. P.; KIM, E.-Y.; TRIMARCHI, J.; JENSSEN, T.-K.; VINTERBO, S. A.; OHNO-MACHADO, L. A primer on gene expression and microarrays for machine learning researchers. *Journal of Biomedical Informatics*, vol. 37, no. 4, pp. 293–303, 2004. (Cited on pages 26 and 27)

LAPOINTE, J.; LI, C.; HIGGINS, J. P.; VAN DE RIJN, M.; BAIR, E.; MONTGOMERY, K.; FERRARI, M.; EGEVAD, L.; RAYFORD, W.; BERGERHEIM, U.; EKMAN, P.; DEMARZO, A. M.; TIBSHIRANI, R.; BOTSTEIN, D.; BROWN, P. O.; BROOKS, J. D.; POLLACK, J. R. Gene Expression Profiling Identifies Clinically Relevant Subtypes of Prostate Cancer. *Proceedings of the National Academy of Sciences of the United States of America*, vol. 101, no. 3, pp. 811–816, 2004. (Cited on page 31)

LAZAREVIC, A.; KUMAR, V. Feature bagging for outlier detection. In: *Proceedings of the 11th ACM International Conference on Knowledge Discovery and Data Mining (SIGKDD), Chicago, IL*, pp. 157–166, 2005. (Cited on page 48)

LEE, S.; SEO, C. H.; LIM, B.; YANG, J. O.; OH, J.; KIM, M.; LEE, S.; LEE, B.; KANG, C.; LEE, S. Accurate quantification of transcriptome from rna-seq data by effective length normalization. *Nucleic Acids Research*, vol. 39, no. 2, p. 9, 2011. (Cited on page 111)

LELIS, L.; SANDER, J. Semi-supervised density-based clustering. In: *Proceedings of the 9th IEEE International Conference on Data Mining (ICDM), Miami, FL*, pp. 842–847, 2009. (Cited on page 76)

LESK, A. M. *Introduction to Bioinformatics*. 3 edn.. Oxford University Press, 2008. (Cited on page 24)

LI, B.; DEWEY, C. Rsem: accurate transcript quantification from rna-seq data with or without a reference genome. *BMC Bioinformatics*, vol. 12, no. 1, p. 323, 2011. (Cited on page 110)

LIU, P.; SI, Y. Cluster analysis of rna-sequencing data. In: DATTA, S.; NETTLETON, D., eds. *Statistical Analysis of Next Generation Sequencing Data*, Frontiers in Probability and the Statistical Sciences, Springer International Publishing, pp. 191–217, 2014. (Cited on page 86)

LOCKHART, D. J.; DONG, H.; BYRNE, M. C.; FOLLETTIE, M. T.; GALLO, M. V.; CHEE, M. S.; MITTMANN, M.; WANG, C.; KOBAYASHI, M.; NORTON, H.; BROWN, E. L. Expression monitoring by hybridization to high-density oligonucleotide arrays. *Nature Biotechnology*, vol. 14, pp. 1675–1680, 1996. (Cited on page 27)

LODISH, H.; BERK, A.; KAISER, C.; KRIEGER, M.; BRETSCHER, A.; PLOEGH, H.; AMON, A.; SCOTT, M. *Molecular cell biology*. Freeman, 2012. (Cited on pages 23, 24, 26, and 27)

LOGANANTHARAJ, R.; CHEEPALA, S.; CLIFFORD, J. Metric for Measuring the Effectiveness of Clustering of DNA Microarray Expression. *BMC Bioinformatics*, vol. 7, no. Suppl 2, p. S5, 2006. (Cited on page 121)

MACHADO, J. B.; CAMPELLO, R. J. G. B.; AMARAL, W. C. Design of OBF-TS fuzzy models based on multiple clustering validity criteria. In: *Proceedings of the 19th IEEE International Conference on Tools with Artificial Intelligence (ICTAI), Patras, Greece*, pp. 336–339, 2007. (Cited on page 38)

MACQUEEN, J. Some methods for classification and analysis of multivariate observations. In: *5th Berkeley Symposium on Mathematics, Statistics, and Probabilistics*, pp. 281–297, 1967. (Cited on pages 9, 10, 31, and 40)

MARQUES, H. O.; CAMPELLO, R. J. G. B.; ZIMEK, A.; SANDER, J. On the internal evaluation of unsupervised outlier detection. In: *Proceedings of the 27th International Conference on Scientific and Statistical Database Management*, SSDBM '15, New York, NY, USA: ACM, pp. 7:1–7:12, 2015 (*SSDBM '15*,). (Cited on page 130)

MARQUIS DE CONDORCET, M. J. A. N. C. *Essai sur l'application de l'analyse à la probabilité des décisions rendues à la pluralité des voix*. Paris: L'Imprimerie Royale, 1785. (Cited on page 49)

MARTIN, J.; ANAMIKA, K.; SRINIVASAN, N. Classification of protein kinases on the basis of both kinase and non-kinase regions. *PloS one*, vol. 5, no. 9, p. e12460, 2010. (Cited on page 127)

MCLACHLAN, G. J.; BEAN, R. W.; PEEL, D. A mixture model-based approach to the clustering of microarray expression data. *Bioinformatics*, vol. 18, no. 3, pp. 413–422, 2002. (Cited on page 31)

MEILA, M. Comparing clusterings – an axiomatic view. In: *Proceedings of the 22nd International Conference on Machine Learning (ICML), Bonn, Germany*, pp. 577–584, 2005. (Cited on page 15)

METZKER, M. L. Sequencing technologies [mdash] the next generation. *Nat Rev Genet*, vol. 11, no. 1, pp. 31–46, 2010. (Cited on pages 28 and 30)

MILLIGAN, G. W. An examination of the effect of six types of error perturbation on fifteen clustering algorithms. *Psychometrika*, vol. 45, no. 3, pp. 325–342, 1980. (Cited on page 99)

MILLIGAN, G. W. A monte carlo study of thirty internal criterion measures for cluster analysis. *Psychometrika*, vol. 46, no. 2, pp. 187–199, 1981. (Cited on pages 19, 35, and 39)

MILLIGAN, G. W.; COOPER, M. C. An examination of procedures for determining the number of clusters in a data set. *Psychometrika*, vol. 50, no. 2, pp. 159–179, 1985. (Cited on pages 2, 20, 21, 22, 35, and 39)

MILLIGAN, G. W.; COOPER, M. C. A Study of the Comparability of External Criteria for Hierarchical Cluster Analysis. *Multivariate Behavioral Research*, vol. 21, no. 4, pp. 441–458, 1986. (Cited on pages 14 and 15)

MÖLLER-LEVET, C.; KLAWONN, F.; CHO, K.-H.; YIN, H.; WOLKENHAUER, O. Clustering of unevenly sampled gene expression time-series data. *Fuzzy Sets and Systems*, vol. 152, pp. 49 – 66, 2005. (Cited on pages 3, 32, 85, and 89)

MORTAZAVI, A.; WILLIAMS, B. A.; MCCUE, K.; SCHAEFFER, L.; WOLD, B. Mapping and quantifying mammalian transcriptomes by RNA-seq. *Nature Methods*, vol. 5, no. 7, pp. 621–628, 2008. (Cited on page 110)

MOULAVI, D. *Finding, Evaluating and Exploring Clustering Alternatives Unsupervised and Semi-supervised*. Ph.D. thesis, University of Alberta, 2014. (Cited on pages 64, 74, 80, 82, and 131)

MOULAVI, D.; JASKOWIAK, P. A.; CAMPELLO, R. J. G. B.; ZIMEK, A.; SANDER, J. Density-based clustering validation. In: *Proceedings of the 14th SIAM International Conference on Data Mining (SDM), Philadelphia, PA*, pp. 839–847, 2014. (Cited on pages 2, 5, 35, 74, 77, 78, 80, 82, and 129)

NALDI, M. C.; CARVALHO, A. C. P. L. F.; CAMPELLO, R. J. G. B. Cluster ensemble selection based on relative validity indexes. *Data Mining and Knowledge Discovery*, vol. 27, no. 2, pp. 259–289, 2013. (Cited on page 14)

NEMENYI, P. B. *Distribution-Free Multiple Comparisons*. Ph.D. thesis, Princeton University, 1983. (Cited on page 60)

OZSOLAK, F.; MILOS, P. M. RNA sequencing: advances, challenges and opportunities. *Nature Reviews - Genetics*, vol. 12, no. 2, pp. 87–98, 2011. (Cited on pages 27, 28, and 83)

PAKHIRA, M. K.; BANDYOPADHYAY, S.; MAULIK, U. Validity index for crisp and fuzzy clusters. *Pattern Recognition*, vol. 37, pp. 487–501, 2004. (Cited on page 18)

PAL, N.; BISWAS, J. Cluster validation using graph theoretic concepts. *Pattern Recognition*, vol. 30, no. 6, pp. 847 – 857, 1997. (Cited on page 74)

PAUWELS, E.; FREDERIX, G. Finding salient regions in images: Nonparametric clustering for image segmentation and grouping. *Computer Vision and Image Understanding*, vol. 75, no. 1, pp. 73–85, 1999. (Cited on page 74)

PEARSON, K. Contributions to the mathematical theory of evolution. iii. regression, heredity, and panmixia. *Proceedings of the Royal Society of London*, vol. 59, pp. 69–71, 1895. (Cited on pages 22, 42, and 87)

PESQUITA, C.; FARIA, D.; BASTOS, H.; FERREIRA, A. E. N.; FALCÃO, A. O.; COUTO, F. M. Metrics for GO based protein semantic similarity: a systematic evaluation. *BMC Bioinformatics*, vol. 9 Suppl 5, no. Suppl 5, p. S4, 2008. (Cited on pages 34, 95, and 122)

PESQUITA, C.; FARIA, D.; FALCÃO, A. O.; LORD, P.; COUTO, F. M. Semantic similarity in biomedical ontologies. *PLoS computational biology*, vol. 5, no. 7, p. e1000443, 2009. (Cited on pages 34, 94, and 95)

PIHUR, V.; DATTA, S.; DATTA, S. Weighted rank aggregation of cluster validation measures: a Monte Carlo cross-entropy approach. *Bioinformatics*, vol. 23, no. 13, pp. 1607–1615, 2007. (Cited on pages 38 and 47)

PIROOZNIA, M.; YANG, J.; YANG, M. Q.; DENG, Y. A comparative study of different machine learning methods on microarray gene expression data. *BMC Genomics*, vol. 9, no. Suppl 1, p. S13, 2008. (Cited on page 84)

POLIKAR, R. Ensemble learning. In: ZHANG, C.; MA, Y., eds. *Ensemble Machine Learning*, Springer, pp. 1–34, 2012. (Cited on page 46)

PRINESS, I.; MAIMON, O.; BEN-GAL, I. Evaluation of gene-expression clustering via mutual information distance measure. *BMC Bioinformatics*, vol. 8, no. 1, p. 111, 2007. (Cited on page 85)

QIN, Z. S. Clustering microarray gene expression data using weighted Chinese restaurant process. *Bioinformatics*, vol. 22, no. 16, pp. 1988–1997, 2006. (Cited on page 98)

RAMASWAMY, S.; ROSS, K. N.; LANDER, E. S.; GOLUB, T. R. A Molecular Signature of Metastasis in Primary Solid Tumors. *Nature genetics*, vol. 33, no. 1, pp. 49–54, 2003. (Cited on page 31)

RAND, W. M. Objective criteria for the evaluation of clustering methods. *Journal of the American Statistical Association*, vol. 66, no. 336, pp. 846–850, 1971. (Cited on page 14)

RATKOWSKY, D. A.; LANCE, G. N. A criterion for determining the number of groups in a classification. *Australian Computer Journal*, vol. 10, pp. 115–117, 1978. (Cited on page 19)

RAU, A.; MAUGIS-RABUSSEAU, C.; MARTIN-MAGNIETTE, M.-L.; CELEUX, G. Co-expression analysis of high-throughput transcriptome sequencing data with poisson mixture models. *Bioinformatics*, vol. 31, no. 9, pp. 1420–1427, 2015. (Cited on page 86)

REIS-FILHO, J. S. Next-generation sequencing. *Breast Cancer Research*, vol. 11, no. Suppl 3, 2009. (Cited on page 27)

RESNIK, P. Semantic Similarity in a Taxonomy: An Information-Based Measure and its Application to Problems of Ambiguity in Natural Language. *Journal of Artificial Intelligence Research*, vol. 11, pp. 95–130, 1999. (Cited on pages 95 and 122)

ROKACH, L. Ensemble-based classifiers. *Artificial Intelligence Review*, vol. 33, pp. 1–39, 2010. (Cited on pages 3, 36, 37, and 46)

ROUSSEEUW, P. J. Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. *Journal of computational and applied mathematics*, vol. 20, pp. 53–65, 1987. (Cited on page 15)

SCHALEKAMP, F.; VAN ZUYLEN, A. Rank aggregation: Together we're strong. In: *Proceedings of the Workshop on Algorithm Engineering and Experiments (ALENEX) SIAM, New York, NY*, pp. 38–51, 2009. (Cited on page 50)

SCHENA, M.; SHALON, D.; DAVIS, R. W.; BROWN, P. O. Quantitative Monitoring of Gene Expression Patterns with a Complementary DNA Microarray. *Science*, vol. 270, no. 5235, pp. 467–470, 1995. (Cited on page 27)

SCHULZE, A.; DOWNWARD, J. Navigating gene expression using microarrays [mdash] a technology review. *Nature Cell Biology*, vol. 3, no. 8, pp. E190–E195, 2001. (Cited on page 29)

SCHWARZ, G. Estimating the Dimension of a Model. *The Annals of Statistics*, vol. 6, no. 2, pp. 461–464, 1978. (Cited on page 109)

SETUBAL, J.; MEIDANIS, J. *Introduction to computational molecular biology*. PWS Publishing Company, 1997. (Cited on pages 24 and 25)

SEVILLA, J. L.; SEGURA, V.; PODHORSKI, A.; GURUCEAGA, E.; MATO, J. M.; MARTINEZ-CRUZ, L. A.; CORRALES, F. J.; RUBIO, A. Correlation between gene expression and GO semantic similarity. *IEEE/ACM transactions on computational biology and bioinformatics*, vol. 2, no. 4, pp. 330–8, 2005. (Cited on page 95)

SHARAN, R.; MARON-KATZ, A.; SHAMIR, R. CLICK and EXPANDER: a system for clustering and visualizing gene expression data. *Bioinformatics*, vol. 19, no. 14, pp. 1787–1799, 2003. (Cited on page 31)

SHENG, W.; SWIFT, S.; ZHANG, L.; LIU, X. A weighted sum validity function for clustering with a hybrid niching genetic algorithm. *IEEE Transactions on Systems, Man, and Cybernetics*, vol. 35, no. 6, pp. 1156–1167, 2005. (Cited on pages 37 and 38)

SHERLOCK, G. Analysis of large-scale gene expression data. *Brief. in Bioinform.*, vol. 2, no. 4, pp. 350–362, 2001. (Cited on pages 23 and 30)

SI, Y.; LIU, P.; LI, P.; BRUTNELL, T. P. Model-based clustering for rna-seq data. *Bioinformatics*, 2013. (Cited on pages 31 and 86)

SÎRBU, A.; KERR, G.; CRANE, M.; RUSKIN, H. J. Rna-seq vs dual- and single-channel microarray data: Sensitivity analysis for differential expression and clustering. *PLoS ONE*, vol. 7, no. 12, p. e50986, 2012. (Cited on page 86)

SON, Y. S.; BAEK, J. A modified correlation coefficient based similarity measure for clustering time-course gene expression data. *Pattern Recognition Letters*, vol. 29, pp. 232 – 242, 2008. (Cited on pages 30, 32, 85, 90, 91, and 103)

DE SOUTO, M.; COSTA, I.; DE ARAUJO, D.; LUDERMIR, T.; SCHLIEP, A. Clustering cancer gene expression data: a comparative study. *BMC Bioinformatics*, vol. 9, no. 1, p. 497, 2008. (Cited on pages 31, 83, 84, 85, 97, 98, 104, 116, and 117)

DE SOUTO, M. C.; JASKOWIAK, P. A.; COSTA, I. G. Impact of missing data imputation methods on gene expression clustering and classification. *BMC Bioinformatics*, vol. 16, no. 1, p. 64, 2015. (Cited on page 30)

SOUTO, M. C. P.; COELHO, L. V.; FACELI, K.; SAKATA, T. C.; COSTA, I. G. A comparison of external clustering evaluation indices in the context of imbalanced data sets. In: *SBRN*, 2012. (Cited on page 126)

SPEARMAN, C. The proof and measurement of association between two things. *The American Journal of Psychology*, vol. 15, no. 1, pp. 72–101, 1904. (Cited on pages 22 and 87)

STEINLEY, D. K-means clustering: A half-century synthesis. *British Journal of Mathematical and Statistical Psychology*, vol. 59, pp. 1–34, 2006. (Cited on page 10)

STEUER, R.; KURTHS, J.; DAUB, C. O.; WEISE, J.; SELBIG, J. The mutual information: detecting and evaluating dependencies between variables. *Bioinformatics (Oxford, England)*, vol. 18 Suppl 2, no. suppl_2, pp. S231–40, 2002. (Cited on page 85)

TAMAYO, P.; SLONIM, D.; MESIROV, J.; ZHU, Q.; KITAREEWAN, S.; DMITROVSKY, E.; LANDER, E. S.; GOLUB, T. R. Interpreting patterns of gene expression with self-organizing maps: Methods and application to hematopoietic differentiation. *Proceedings of the National Academy of Sciences of the United States of America*, vol. 96, no. 6, pp. 2907–2912, 1999. (Cited on page 98)

TAN, P.-N.; STEINBACH, M.; KUMAR, V. *Introduction to data mining*. Addison Wesley, 2006. (Cited on pages 1, 7, and 84)

TARCA, A. L.; CAREY, V. J.; CHEN, X.-W.; ROMERO, R.; DR\UAGHICI, S. Machine Learning and Its Applications to Biology. *PLoS Computational Biology*, vol. 3, no. 6, p. e116, 2007. (Cited on page 27)

THALAMUTHU, A.; MUKHOPADHYAY, I.; ZHENG, X.; TSENG, G. C. Evaluation and comparison of gene clustering methods in microarray analysis. *Bioinformatics*, vol. 22, pp. 2405–2412, 2006. (Cited on page 84)

THE GENE ONTOLOGY CONSORTIUM Gene ontology consortium: going forward. *Nucleic Acids Research*, vol. 43, no. D1, pp. D1049–D1056, 2015. (Cited on page 33)

USCHOLD, M.; GRÜNINGER, M. Ontologies: principles, methods, and applications. *Knowledge Engineering Review*, vol. 11, no. 2, pp. 93–155, 1996. (Cited on page 32)

VANGUILDER, H. D.; VRANA, K. E.; FREEMAN, W. M. Twenty-five years of quantitative PCR for gene expression analysis. *BioTechniques*, vol. 44, no. 5, pp. 619–626, 2008. (Cited on pages 26 and 27)

VENDRAMIN, L. *Study and development of fuzzy clustering algorithms in centralized and distributed scenarios*. Master's thesis, Universidade de São Paulo, 2012. (Cited on page 8)

VENDRAMIN, L.; CAMPELLO, R. J. G. B.; HRUSCHKA, E. R. On the comparison of relative clustering validation criteria. In: *Proceedings of the 9th SIAM International Conference on Data Mining (SDM), Sparks, NV*, pp. 733–744, 2009. (Cited on pages 2, 20, 21, 35, and 71)

VENDRAMIN, L.; CAMPELLO, R. J. G. B.; HRUSCHKA, E. R. Relative clustering validity criteria: A comparative overview. *Statistical Analysis and Data Mining*, vol. 3, no. 4, pp. 209–235, 2010. (Cited on pages 2, 20, 21, 35, 39, 70, 71, 72, and 104)

VENDRAMIN, L.; JASKOWIAK, P. A.; CAMPELLO, R. J. G. B. On the combination of relative clustering validity criteria. In: *Proceedings of the 25th International Conference on Scientific and Statistical Database Management (SSDBM), Baltimore, MD*, pp. 4:1–12, 2013. (Cited on pages 5, 37, and 48)

WANG, Z.; GERSTEIN, M.; SNYDER, M. RNA-seq: a revolutionary tool for transcriptomics. *Nature reviews - Genetics*, vol. 10, no. 1, pp. 57–63, 2009. (Cited on pages 27, 28, 30, and 83)

WU, X.; KUMAR, V.; ROSS; GHOSH, J.; YANG, Q.; MOTODA, H.; MCLACHLAN, G.; NG, A.; LIU, B.; YU, P.; ZHOU, Z.-H.; STEINBACH, M.; HAND, D.; STEINBERG, D. Top 10 algorithms in data mining. *Knowledge and Information Systems (KAIS)*, vol. 14, no. 1, pp. 1–37, 2008. (Cited on page 10)

XU, R.; WUNSCH II, D. *Clustering*. IEEE Press, 2009. (Cited on pages 1, 2, 8, 10, 13, and 84)

YEUNG, K.; MEDVEDOVIC, M.; BUMGARNER, R. Clustering gene-expression data with repeated measurements. *Genome Biol.*, vol. 4, no. 5, 2003. (Cited on page 79)

YEUNG, K. Y.; FRALEY, C.; MURUA, A.; RAFTERY, A. E.; RUZZO, W. L. Model-based clustering and data transformations for gene expression data. *Bioinformatics*, vol. 17, no. 10, pp. 977–987, 2001a. (Cited on page 79)

YEUNG, K. Y.; HAYNOR, D. R.; RUZZO, W. L. Validating clustering for gene expression data. *Bioinformatics*, vol. 17, no. 4, pp. 309–318, 2001b. (Cited on pages 57 and 79)

ZHANG, A. *Advanced analysis of gene expression microarray data*. 1 edn.. World Scientific Publishing Company, 2006. (Cited on pages 1, 3, 8, 11, 23, 27, 28, 83, 84, 87, 88, 97, and 119)

ZHANG, W.; YU, Y.; HERTWIG, F.; THIERRY-MIEG, J.; ZHANG, W.; THIERRY-MIEG, D.; WANG, J.; FURLANELLO, C.; DEVANARAYAN, V.; CHENG, J.; DENG, Y.; HERO, B.; HONG, H.; JIA, M.; LI, L.; LIN, S.; NIKOLSKY, Y.; OBERTHUER, A.; QING, T.; SU, Z.; VOLLAND, R.; WANG, C.; WANG, M.; AI, J.; ALBANESE, D.; ASGHARZADEH, S.; AVIGAD, S.; BAO, W.; BESSARABOVA, M.; BRILLIANT, M. Comparison of rna-seq and microarray-based models for clinical endpoint prediction. *Genome Biology*, vol. 16, no. 1, p. 133, wenqian Zhang, Ying Yu, Falk Hertwig, Jean Thierry-Mieg and Wenwei Zhang contributed equally to this work., 2015. (Cited on pages 28 and 83)

ZHAO, S.; FUNG-LEUNG, W.-P.; BITTNER, A.; NGO, K.; LIU, X. Comparison of rna-seq and microarray in transcriptome profiling of activated t cells. *PLoS ONE*, vol. 9, no. 1, p. e78644, 2014. (Cited on page 28)

ZHU, Y.; QIU, P.; JI, Y. Tcga-assembler: open-source software for retrieving and processing tcga data. *Nature Methods*, vol. 11, pp. 599–600, 2014. (Cited on page 110)

ZIMEK, A.; CAMPELLO, R. J. G. B.; SANDER, J. Ensembles for unsupervised outlier detection: Challenges and research questions. *ACM SIGKDD Explorations*, vol. 15, no. 1, pp. 11–22, 2013. (Cited on pages 3, 37, and 48)

ZIMEK, A.; CAMPELLO, R. J. G. B.; SANDER, J. Data perturbation for outlier detection ensembles. In: *Proceedings of the 26th International Conference on Scientific and Statistical Database Management (SSDBM), Aalborg, Denmark*, pp. 13:1–12, 2014. (Cited on page 48)