Uma Abordagem para a Indução de Árvores de Decisão voltada para Dados de Expressão Gênica

Pedro Santoro Perez

Dissertação apresentada
ao
Programa Interunidades em Bioinformática
da
Universidade de São Paulo
para
obtenção do grau de mestre
em
Ciências

Área de Concentração: **Bioinformática**
Orientador: **Prof. Dr. José Augusto Baranauskas**

Durante a elaboração deste trabalho, o autor recebeu apoio financeiro da CAPES e da FAPESP

– Ribeirão Preto, Maio de 2012 –

Uma Abordagem para a Indução de Árvores de Decisão voltada para Dados de Expressão Gênica

Este exemplar corresponde à redação final
da dissertação de mestrado devidamente
corrigida e defendida por
Pedro Santoro Perez
e aprovada pela comissão julgadora.

Ribeirão Preto, 28 de maio de 2012

Banca Examinadora:

- Prof. Dr. Fabricio Martins Lopes – UTFPR
- Prof. Dr. Renato Tinós – FFCLRP - USP
- Prof. Dr. José Augusto Baranauskas (orientador) – FFCLRP - USP

Dedico este trabalho
a minha esposa, Janaina,
a meus pais, Rita e Carlos,
e a meu orientador, Augusto.

Agradecimentos

Agradeço, pelo apoio e paciência, a todos os meus familiares, em especial: Janaina, minha esposa; Rita, minha mãe; Isabel e Marcelle, minhas irmãs; Eunice e José, meus avós; Carlos, meu pai; e Cinira.

Por todo apoio e efetiva orientação, agradeço ao Augusto, orientador e amigo.

INCT Adapta. O presente trabalho foi financiado pelas agências CNPq (Conselho Nacional de Desenvolvimento Científico e Tecnológico) e FAPEAM (Fundação de Amparo à Pesquisa do Estado do Amazonas), por meio do Programa INCT (Institutos Nacionais de Ciência e Tecnologia), Projeto ADAPTA (Centro de Estudos de Adaptações da Biota Aquática da Amazônia), coordenado por Adalberto Luis Val.

CAPES. O autor foi bolsista da Coordenação de Aperfeiçoamento de Pessoal de Nível Superior pelo Programa de Pós-Graduação em Bioinformática da Universidade de São Paulo, vinculado ao Programa Institutos Nacionais de Ciência e Tecnologia (INCT Adapta - Amazônia), pelo período de 01/01/2010 a 31/07/2010.

FAPESP. O autor foi bolsista da Fundação de Amparo à Pesquisa do Estado de São Paulo (processo 2009/04511-4) de 01/08/2010 a 30/09/2011.

Resumo

Estudos de expressão gênica têm sido de extrema importância, permitindo desenvolver terapias, exames diagnósticos, medicamentos e desvendar uma infinidade de processos biológicos. No entanto, estes estudos envolvem uma série de dificuldades: grande quantidade de genes, sendo que geralmente apenas um pequeno número deles está envolvido no problema estudado; presença de ruído nos dados analisados; entre muitas outras. O projeto de pesquisa deste mestrado consiste no estudo de algoritmos de indução de árvores de decisão; na definição de uma metodologia capaz de tratar dados de expressão gênica usando árvores de decisão; e na implementação da metodologia proposta como algoritmos capazes de extrair conhecimento a partir desse tipo de dados. A indução de árvores de decisão procura por características relevantes nos dados que permitam modelar precisamente um conceito, mas tem também a preocupação com a compreensibilidade do modelo gerado, auxiliando os especialistas na descoberta de conhecimento, algo importante nas áreas médica e biológica. Por outro lado, tais indutores apresentam relativa instabilidade, podendo gerar modelos bem diferentes com pequenas mudanças nos dados de treinamento. Este é um dos problemas tratados neste mestrado. Mas o principal problema tratado se refere ao comportamento destes indutores em dados de alta dimensionalidade, mais especificamente dados de expressão gênica: atributos irrelevantes prejudicam o aprendizado e vários modelos com desempenho similar podem ser gerados. Diversas técnicas foram exploradas para atacar os problemas mencionados, mas este estudo se concentrou em duas delas: *windowing*, que foi a técnica mais explorada e para a qual este mestrado propôs uma série de alterações com vistas à melhoria de seu desempenho; e *lookahead*, que procura construir a árvore levando em considerações passos subsequentes do processo de indução. Quanto ao *windowing*, foram explorados aspectos relacionados ao procedimento de poda das árvores geradas durante a execução do algoritmo; uso do erro estimado em substituição ao erro de treinamento; uso de ponderação do erro calculado durante a indução de acordo com o tamanho da janela; e uso da confiança na classificação para decidir quais exemplos utilizar na atualização da janela corrente. Com relação ao *lookahead*, foi implementada uma versão de um passo à frente, ou seja, para tomar a decisão na iteração corrente, o indutor leva em consideração a razão de ganho de informação do passo seguinte. Os resultados obtidos, principalmente com relação às medidas de desempenho baseadas na compreensibilidade dos modelos induzidos, mostram que os algoritmos aqui propostos superaram algoritmos clássicos de indução de árvores.

Palavras-chave: Aprendizado de Máquina, Árvores de Decisão, Expressão Gênica, Bioinformática, *Windowing*, *Lookahead*.

Abstract

Gene expression studies have been of great importance, allowing the development of new therapies, diagnostic exams, drugs and the understanding of a variety of biological processes. Nevertheless, those studies involve some obstacles: a huge number of genes, while only a very few of them are really relevant to the problem at hand; data with the presence of noise; among others. This research project consists of: the study of decision tree induction algorithms; the definition of a methodology capable of handling gene expression data using decision trees; and the implementation of that methodology as algorithms that can extract knowledge from that kind of data. The decision tree induction searches for relevant characteristics in the data which would allow it to precisely model a certain concept, but it also worries about the comprehensibility of the generated model, helping specialists to discover new knowledge, something very important in the medical and biological areas. On the other hand, such inducers present some instability, because small changes in the training data might produce great changes in the generated model. This is one of the problems being handled in this Master's project. But the main problem this project handles refers to the behavior of those inducers when it comes to high-dimensional data, more specifically to gene expression data: irrelevant attributes may harm the learning process and many models with similar performance may be generated. A variety of techniques have been explored to treat those problems, but this study focused on two of them: windowing, which was the most explored technique and to which this project has proposed some variations in order to improve its performance; and lookahead, which builds each node of a tree taking into consideration subsequent steps of the induction process. As for windowing, the study explored aspects related to the pruning of the trees generated during intermediary steps of the algorithm; the use of the estimated error instead of the training error; the use of the error weighted according to the size of the current window; and the use of the classification confidence as the window update criterion. As for lookahead, a 1-step version was implemented, *i.e.*, in order to make the decision in the current iteration, the inducer takes into consideration the information gain ratio of the next iteration. The results show that the proposed algorithms outperform the classical ones, especially considering measures of complexity and comprehensibility of the induced models.

Keywords: Machine Learning, Decision Trees, Gene Expression, Bioinformatics, *Windowing*, *Lookahead*.

Conteúdo

Lista de Abreviaturas

AD	Árvore de Decisão.
AM	Aprendizado de Máquina.
AUC	*Area Under the (ROC) Curve* ou Área Sob a Curva (ROC).
GUI	*Graphical User Interface* ou Interface Gráfica de Usuário.
INCT Adapta	Instituto Nacional de Ciência e Tecnologia de Adaptações da Biota Aquática da Amazônia.
ROC	*Receiver Operating Characteristic* ou Característica Operativa do Receptor.
SAGE	*Serial Analysis of Gene Expression* ou Análise Serial de Expressão Gênica.

Lista de Algoritmos

Lista de Figuras

Lista de Tabelas

Prefácio

Esta dissertação apresenta o projeto de mestrado do autor, explicitando seus objetivos, motivação, bem como os estudos e atividades efetuados. O trabalho foi desenvolvido durante a participação do autor no programa de mestrado interunidades em Bioinformática da Universidade de São Paulo.

Em linhas gerais, este projeto consiste no estudo de algoritmos de indução de árvores de decisão, visando definir uma metodologia capaz de tratar dados de expressão gênica, bem como na implementação da metodologia proposta como algoritmos capazes de extrair conhecimento a partir desses dados. O mestrado recebeu apoio financeiro da CAPES, de 01/01/2010 a 31/07/2010, e da FAPESP, sob o processo de número 2009/04511-4, com vigência de 01/08/2010 a 30/09/2011.

Capítulo

1

Introdução

1.1 Considerações Iniciais

A indução de árvores de decisão (AD) é um tópico de pesquisa importante em aprendizado de máquina (AM), que tem como objetivo automatizar o processo de aquisição de conhecimento que permita realizar tarefas úteis à sociedade. Considerando esta preocupação com o conhecimento, em contraste a somente predições precisas, a pesquisa aqui proposta enfatiza algoritmos que produzem como saída modelos compreensíveis ao ser humano. Na área médica ou biológica, não é suficiente que um modelo (classificador) seja preciso: ele também deve ser compreendido pelos especialistas, para que eles possam confiar nele e considerá-lo aceitável. Os especialistas frequentemente desejam obter maior compreensão sobre o domínio no qual trabalham, o que só é possível se eles forem capazes de entender o modelo induzido (Tan et al. 2005), facilitando, inclusive, o processo de refinamento interativo. Devido ao fato de os algoritmos de AM, incluindo os de indução de ADs, terem sido projetados para um grande espectro de conhecimento (desde o senso comum até o conhecimento humano profundamente especializado), eles estão focados em representações flexíveis e poderosas. Esta flexibilidade, embora essencial para a área de AM, pode apresentar a desvantagem de fazer com que os indutores sejam muito suscetíveis aos dados de treinamento, produzindo modelos que podem mudar drasticamente com pequenas alterações nos dados. Esta instabilidade mina o propósito de extrair conhecimento a partir dos dados: os especialistas ficam confusos e perdem a confiança nos modelos produzidos (Turney 1995; Dietterich 1996). Este projeto de mestrado tentou tratar o problema, gerando ADs mais estáveis.

1.2 Motivação

Ferramentas de mineração de dados mostraram-se úteis em uma variedade de domínios, incluindo análise de dados genômicos (Statnikov et al. 2005; Wang et al. 2006). Como o volume de dados cresce rapidamente, novas estratégias e abordagens devem ser exploradas nas ferramentas de mineração de dados. Classificação, a separação de dados em classes distintas, é, possivelmente, a tarefa de mineração de dados mais comum e ADs estão entre os classificadores mais populares (Rosenfeld et al. 2008). Pesquisas científicas mostram que a classificação pode ser utilizada, por exemplo, para analisar o efeito de fatores genômicos, clínicos, ambientais e demográficos nas doenças, a resposta a um tratamento e o risco de efeitos colaterais (Risch 2000; Hossain et al. 2008). Portanto, pesquisar e desenvolver algoritmos de indução de ADs eficientes voltados para dados genômicos é uma meta importante para a sociedade na qual estamos inseridos. A medição das taxas de expressão gênica pode contribuir para a descoberta de novos métodos de diagnóstico e prognóstico, tratamento de diversas doenças e desenvolvimento de medicamentos (Li et al. 2006), desde que haja uma maneira adequada para extrair informação desses dados (Slonim 2002). Como consequência, muitos trabalhos têm se concentrado em automatizar a análise de dados de expressão gênica (Golub et al. 1999;

Alizadeh 2000; Gamberger et al. 2004; Tan et al. 2005; Cho & Won 2007; Schaefer et al. 2008).

O desafio em predizer classes de diagnósticos ou outros fenômenos de interesse utilizando dados de expressão gênica é que o número de genes (atributos) é frequentemente muito maior que o número de amostras de tecidos disponíveis (exemplos) e apenas um subconjunto de genes é relevante em distinguir diferentes classes (Yeung & Bumgarner 2003; Hossain et al. 2008). Neste contexto, abordagens computacionais utilizando AM estão sendo adotadas com sucesso no reconhecimento de padrões de expressão gênica que determinam um fenótipo (Lyons-Weiler et al. 2003). Os sistemas de aprendizado são capazes de adquirir conhecimento de forma automática a partir de grandes volumes de dados e contribuem com a geração de modelos úteis (Baldi & Brunak 2001).

Neste contexto, algoritmos de AM dos paradigmas conexionista (*e.g.*, redes neurais), estatístico (*e.g.*, Naïve Bayes), genético (*e.g.*, algoritmos genéticos) e de memorização (*e.g.*, KNN) não fornecem como saída um modelo simbólico (Monard & Baranauskas 2003a). No contexto deste trabalho, um classificador simbólico é um modelo cuja linguagem de descrição é equivalente a um conjunto de regras, ou seja, o classificador pode ser representado em linguagem lógica proposicional ou relacional (Monard & Baranauskas 2003b). Além disso, tais paradigmas, por não contarem com um processo embutido de seleção de atributos (genes) relevantes, em geral utilizam todos os atributos fornecidos, o que pode impedi-los de identificar determinados fenótipos de interesse, por não focar em genes informativos (Rosenfeld et al. 2008). Isto quer dizer que os paradigmas mencionados acima geram modelos cuja representação é mais complexa e que utilizam todos os atributos disponíveis, fazendo com que sua interpretação seja mais difícil.

Como mencionado anteriormente, um fator importante na área médica ou biológica é que os modelos devem fornecer uma descrição simbólica compreensível do conceito embutido nos dados, assumindo que esses modelos são analisados por seres humanos (Michalski 1983; Michie 1988; Kubat et al. 1998; Tan et al. 2005). Além disso, algoritmos de aprendizado que contribuem para a compreensão do domínio considerado podem produzir conhecimento novo (Dietterich 1986; Matukumalli et al. 2006), algo importante em ambas as áreas citadas.

Embora pesquisas relacionadas sejam encontradas na literatura, ainda há questões importantes não respondidas. Por exemplo, a indução de modelos simbólicos que representem relações significativas a partir de dados com alta dimensionalidade de atributos, mas com poucos exemplos, tipicamente encontrados em análises de expressão gênica, qualquer que seja a técnica utilizada. Em geral, ADs induzidas a partir de dados de expressão gênica tendem a incluir poucos atributos; contudo, devido à extrema abundância de atributos, várias árvores podem ser induzidas, cada qual utilizando um subconjunto diferente de atributos, mas com precisão similar (Netto et al. 2010; Lemos 2007). A indução de modelos mais adequados a este problema de alta dimensão tem se mostrado uma tarefa difícil (Ben-Dor et al. 2000; Dudoit et al. 2002; Dettling & Buhlmann 2003; Wang et al. 2006), mas é crucial para determinar com sucesso o fenômeno de interesse. É preciso estudar, pesquisar e implementar novas abordagens de indução de ADs (ou outros paradigmas simbólicos) voltadas para este problema, que é a proposta deste trabalho.

Nas subseções seguintes, são fornecidos conceitos elementares sobre genes e sua expressão, cujos dados são o foco deste estudo.

1.2.1 Expressão Gênica

Em geral, todas as células de um organismo eucarioto possuem o mesmo material genético em seu núcleo. Esse material genético é representado principalmente por moléculas de DNA, em que se encontram os genes. No entanto, organismos mais complexos apresentam altos graus de especificidade celular, possuindo tipos celulares (ou tecidos) que desempenham funções diferentes. Por exemplo, células epiteliais apresentam forma, funcionamento e função diferentes de células sanguíneas. Essa especialização celular se deve, basicamente, à expressão gênica de cada célula ou tipo celular, já que, apesar de presentes em todas as células, cada gene em uma célula pode ou não ser "utilizado". A expressão gênica de uma célula é, portanto, o conjunto de genes que estão expressos ou ativados nessa célula em um dado momento, o que, em última análise, se traduz no conjunto de proteínas nela produzidas.

Considerando-se que a vida de um organismo complexo começa com uma única célula, tamanha

especificidade dos diferentes tipos celulares se apresenta como uma característica também muito complexa. O processo pelo qual as células passam a ter expressão gênica diferente umas das outras se chama *diferenciação*, que define o destino de uma célula e permite a construção dos diferentes tecidos. Durante o desenvolvimento embrionário, é possível inclusive que um mesmo grupo de células expressem diferentes genes em momentos diferentes de seu desenvolvimento (Nussbaum et al. 2002).

No entanto, a porcentagem de genes expressos que promove todas essas diferenças entre os tipos celulares representa cerca de 20% de todos os genes expressos, sendo denominados genes especializados. Os restantes 80%, os genes de manutenção e outros, estão expressos em todos os tecidos e possuem relação com funções básicas do metabolismo celular, como o transporte de substâncias e a síntese proteica (Nussbaum et al. 2002).

Os fatores que regulam a expressão gênica de uma célula podem ser, entre outros: fatores de transcrição, uma classe de proteínas extremamente importantes para a expressão dos genes; os promotores, uma parte do DNA que fica adjacente ao gene e é importante na determinação do início da transcrição; os silenciadores, que promovem o bloqueio da transcrição de determinados genes; e os acentuadores, que aumentam os níveis de expressão de determinados genes.

O fluxo da expressão de um gene pode ser visualizado na Figura 1.1: primeiramente o gene passa pelo processo de transcrição, em que a molécula de DNA é lida e se produz o transcrito primário; o transcrito primário sofre uma série de alterações, como clivagem da extremidade $3'$ e subsequente adição da cauda poli-A e adição de uma estrutura cap à extremidade $5'$; remoção dos íntrons — porções não codificantes do gene — e junção dos éxons — porções codificantes do gene, formando o mRNA (RNA mensageiro); passagem do mRNA do núcleo para o citoplasma; e, finalmente, produção de proteína a partir do processo de tradução do mRNA. Vale lembrar que mutações podem ocorrer: em uma ou mais das etapas desse fluxo, em que podem também ocorrer diversos tipos de erros; nos fatores de transcrição; ou nos próprios genes.

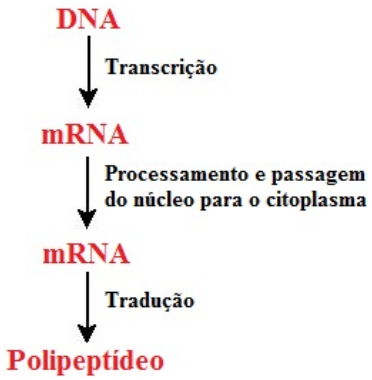


Figura 1.1: *Fluxo da expressão de um gene, partindo da leitura do DNA, passando pelo processamento do RNA e seu transporte para fora do núcleo e terminando na tradução e montagem da cadeia polipeptídica.*

Se o fluxo descrito acima fosse sempre seguido, seria uma tarefa menos difícil estudar o metabolismo e as funções de uma célula e associá-los aos genes responsáveis. No entanto, há uma série de caminhos alternativos naquele fluxo. Esses caminhos alternativos podem ser, entre outros:

- Quando da remoção dos íntrons e junção dos éxons, há oportunidades em que, por exemplo, nem todos os éxons acabam por fazer parte do mRNA produzido, gerando, portanto, um mRNA com uma sequência diferente de nucleotídeos e, possivelmente, uma proteína ou cadeia polipeptídica diferente;
- Após a tradução do mRNA, a proteína produzida pode sofrer processamento, que pode consistir de: dobramento da molécula, gerando uma estrutura tridimensional; combinação de diferentes cadeias polipeptídicas de um mesmo gene ou de diferentes genes, formando uma única molécula; clivagem; entre outros.

O estudo da expressão gênica dos diferentes tecidos tem extrema importância na determinação dos genes responsáveis por cada função metabólica nesses tecidos. Além disso, as mutações descritas

acima podem promover uma série de mudanças no padrão de expressão gênica de uma célula, o que pode acarretar consequências biológicas ou clínicas importantes, podendo, por exemplo, causar uma série de doenças. Uma das possibilidades provenientes desse estudo é o desenvolvimento de terapias de doenças e localização de alvos moleculares (Li et al. 2006).

1.2.2 Quantificação de Expressão Gênica

Para o estudo da expressão gênica, existe uma série de técnicas e ferramentas que desempenham papéis desde a extração dos transcritos até a análise dos dados coletados. A Biologia Molecular é responsável por coletar esses dados. Essa coleta se dá na forma de perfis de expressão gênica, que consistem basicamente de uma lista dos mRNAs presentes em um dado tipo celular juntamente com a quantidade de cada um deles. Há várias técnicas para a extração desses perfis, sendo as mais conhecidas e usadas *Microarray* e *Serial Analysis of Gene Expression* (SAGE). A tarefa dessas técnicas não é fácil, visto que as moléculas de mRNA representam de 1% a 5% de todo RNA de uma célula, além do fato de algumas dessas moléculas de mRNA se apresentarem em baixíssimas concentrações (Müller et al. 2008).

Microarray (Maskos & Southern 1992) é baseado em hibridização de DNA. Basicamente, sondas de DNA são presas a uma matriz. A técnica pode ser utilizada para uma série de fins: medir alterações em níveis de expressão gênica, caracterizar genomas mutantes, detectar SNPs (*Single Nucleotide Polymorphisms*), entre outros. Um dos grandes poderes desta técnica é permitir inúmeros testes moleculares simultaneamente[1]. Na Figura 1.2 é mostrada uma forma de análise do resultado de um experimento de *microarray*. SAGE (Velculescu et al. 1995) é uma técnica que permite "fotografar" a população de mRNA de uma amostra. Para isto, ela utiliza *tags*, cada uma correspondendo a uma pequena sequência de um transcrito (ver Figura 1.3). Uma das diferenças entre SAGE e *microarray* é que o primeiro é baseado no sequenciamento dos transcritos e o segundo é baseado em hibridização dos transcritos a sondas, ou seja, a análise feita pelo SAGE é mais quantitativa, como uma contagem. Além disto, no SAGE as sequências dos transcritos não precisam ser previamente conhecidas, permitindo a descoberta de novos genes e suas variações. Por ser uma contagem direta, SAGE tende a ser mais exato e *microarray* é mais suscetível a ruído. No entanto, *microarray* possui um espectro maior de aplicações e é muito mais barato, sendo o mais usado em grandes projetos[2].

Depois de coletados e preparados, os dados de expressão gênica precisam ser analisados. A Bioinformática e o Aprendizado de Máquina são duas das responsáveis por essa análise, fornecendo ferramentas extremamente úteis para análise de grandes volumes de dados. Exemplos dessas ferramentas podem ser encontrados em (Tan et al. 2005; Li et al. 2006; Amin 2007; Schramm et al. 2007; Rosenfeld et al. 2008; Statnikov et al. 2005; Wang et al. 2006).

1.3 Objetivo Geral

O objetivo principal deste projeto de mestrado consiste em pesquisar métodos de indução de árvores de decisão e aplicá-los em dados de alta dimensionalidade, tipicamente encontrados em análises de expressão gênica. O estudo inclui os algoritmos e sistemas clássicos, *e.g.*, C4.5, J48 e CART, e também as técnicas que compõem o estado da arte nesta área (Rosenfeld et al. 2008). Por meio da análise do perfil de expressão gênica por modelos simbólicos, espera-se contribuir para tornar tal análise mais clara e precisa sob o ponto de vista biológico. Para atingir o objetivo exposto acima, foi definida, implementada e avaliada uma metodologia capaz de tratar dados de expressão gênica usando ADs.

1.4 Objetivos Específicos

Diversos algoritmos de AM foram estudados neste mestrado, mas dois deles foram explorados mais aprofundadamente: *windowing* e *lookahead*. Ambos podem ser considerados meta-algoritmos de AM, pois trabalham sobre os algoritmos propriamente ditos, alterando a forma utilizada por eles

[1] http://en.wikipedia.org/wiki/DNA_microarray
[2] http://en.wikipedia.org/wiki/Serial_analysis_of_gene_expression

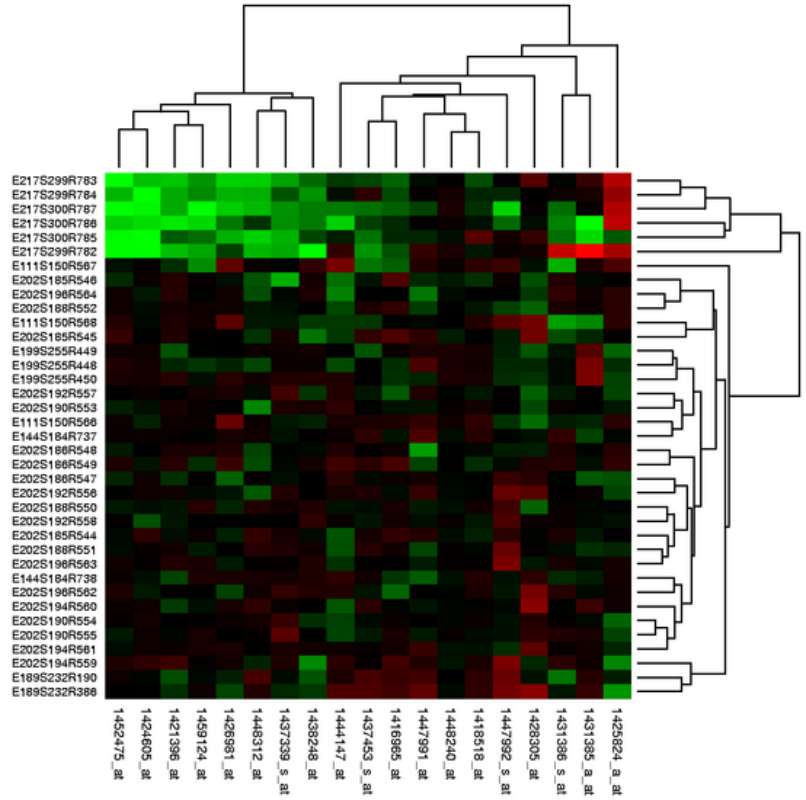


Figura 1.2: *Representação em mapa de calor do resultado de um experimento de microarray. Cada ponto e sua intensidade representam a expressão de um gene (retirado de http://en.wikipedia.org/wiki/File:Heatmap.png).*

para explorar o espaço de soluções.

1.4.1 Windowing

Windowing é uma técnica cuja ideia é encontrar um subconjunto dos exemplos de treinamento que forneça informação suficiente para induzir um classificador e obter resultados melhores ou similares aos obtidos por um modelo construído a partir do conjunto inteiro, reduzindo, assim, a complexidade do problema de aprendizado (Chen 2004). O conceito de "melhor" pode estar relacionado à acurácia do classificador, ao seu tamanho, ou a outros aspectos. Desta forma, *windowing* pode ser considerado uma técnica de subamostragem (Reinartz 2002).

Maiores detalhes sobre a técnica serão dados na Seção 4.2, inclusive com relação a um de seus pontos fracos, que é a perda de desempenho em domínios com ruído, muito comum em dados de expressão gênica. Neste mestrado, o autor propôs uma série de alterações no algoritmo original, na tentativa de minimizar suas limitações e adaptá-lo melhor a aplicações baseadas em análise de perfil de expressão gênica.

1.4.2 Lookahead

Lookahead é uma técnica que pode também ser aplicada a ADs e que visa à minimização de algumas das limitações associadas ao processo de indução desses modelos. Basicamente, ela faz com que a escolha do teste em um determinado nó leve em consideração passos futuros, ou seja, o teste escolhido depende do valor do critério (razão de ganho, por exemplo) em nós descendentes do nó atual, não no nó atual propriamente dito. Esta técnica é apresentada na Seção 4.3.

1.5 Organização do Trabalho

O restante da dissertação está organizado da seguinte forma:

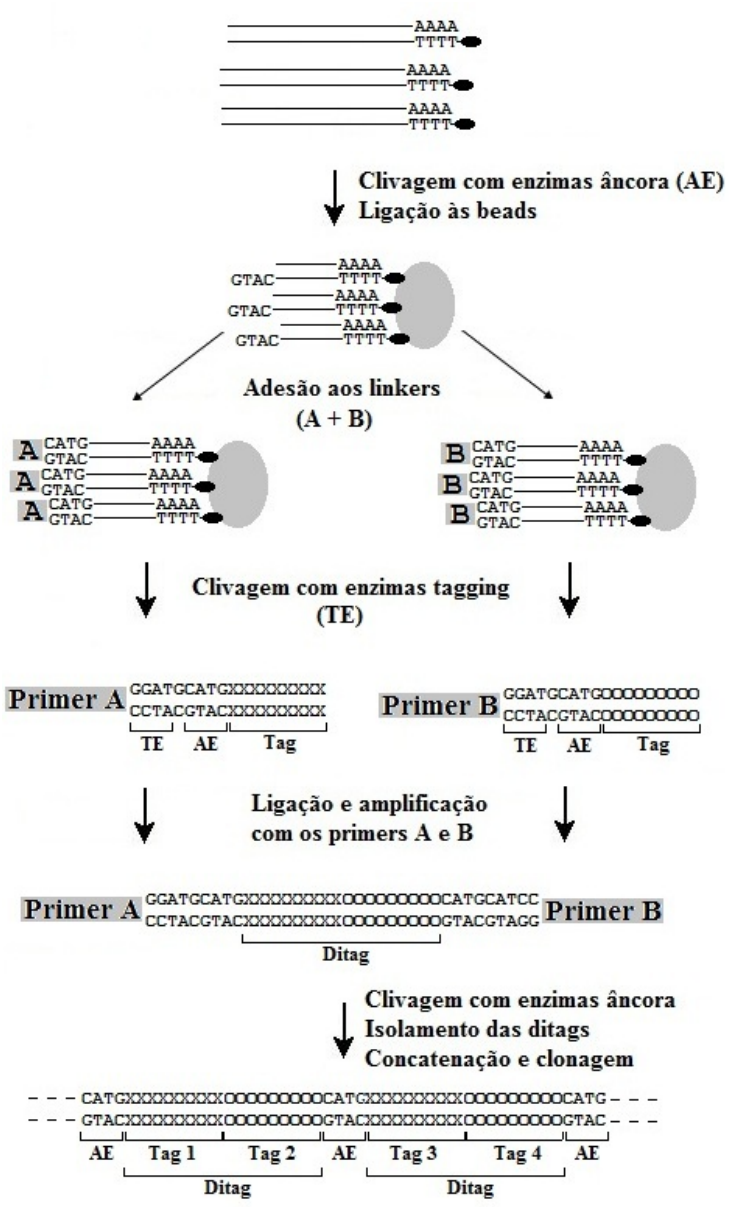


Figura 1.3: *Figura esquemática do procedimento básico seguido pelo SAGE. Baseado em Velculescu et al. (1995).*

- No Capítulo 2, é fornecida uma introdução ao aprendizado de máquina supervisionado, com ênfase nas árvores de decisão;
- No Capítulo 3, são apresentadas as medidas utilizadas neste trabalho para quantificar o desempenho dos indutores. São introduzidas também medidas de compreensibilidade e complexidade de ADs propostas pelo grupo;
- No Capítulo 4, a abordagem de *windowing* é descrita em detalhes, juntamente com todas as alterações propostas neste mestrado. É apresentada também a abordagem de *lookahead*;
- No Capítulo 5, estão os resultados e discussão dos experimentos realizados com a aplicação das duas abordagens anteriores em dados de expressão gênica;
- No Capítulo 6, são apresentadas aplicações práticas diretas de algumas das técnicas exploradas, demonstrando seu potencial na área biológica;

- No Capítulo 7, são apresentadas as conclusões do trabalho. Além disto, apresentam-se também a produção bibliográfica proveniente do mestrado e propostas de trabalhos futuros;
- No Apêndice A, são descritas as ferramentas computacionais utilizadas nas implementações dos algoritmos explorados neste mestrado;
- No Apêndice B, detalhes são fornecidos sobre as bases de dados de expressão gênica utilizadas;
- No Apêndice C, os resultados originais da validação cruzada para cada medida utilizada nos experimentos são fornecidos em tabelas.
- No Apêndice D, são descritos os conceitos utilizados neste trabalho para medir e comparar desempenho de algoritmos;

Capítulo

2

Aprendizado de Máquina Supervisionado

2.1 Considerações Iniciais

Um dos grandes desafios atuais em AM é o de desenvolver e aplicar processos computacionais automáticos de extração de conhecimento a partir de grandes volumes de dados e informações, criando modelos capazes de explicar os fenômenos estudados.

O processo de extração automática de conhecimento pode ser resumido em três etapas elementares (Baranauskas 2001): (i) pré-processamento, (ii) mineração e (iii) pós-processamento. A etapa de pré-processamento, na qual se preparam os dados para o processo de mineração, pode ser entendida como duas subetapas. Na subetapa de preparação, os dados são coletados e transformados para início do processo de mineração. A subetapa de redução, opcional quando a quantidade de dados é moderada, diminui a quantidade de dados de forma a viabilizar a aplicação da etapa de mineração. A etapa de mineração procura por soluções, que podem ter diferentes objetivos e complexidade e na qual são, normalmente, utilizados algoritmos de AM. Por último, a etapa de pós-processamento, ou análise das soluções obtidas, é efetuada, consolidando os resultados obtidos em uma solução final, que será apresentada ao usuário. O trabalho de pesquisa aqui proposto se concentra na etapa de mineração de dados.

2.2 Aprendizado Supervisionado

Uma das categorizações possíveis em AM é dividir o aprendizado em supervisionado, não-supervisionado e semissupervisionado. Em qualquer um dos casos, para que um algoritmo de aprendizado aprenda o conceito desejado, exemplos sobre esse conceito devem ser apresentados a ele, denominados *exemplos de treinamento*.

No *aprendizado supervisionado*, os exemplos de treinamento trazem a "resposta correta" (rótulo), assumindo-se a existência de um professor dizendo ao algoritmo quando ele erra ou acerta (Caruana & Niculescu-Mizil 2006). No *aprendizado não-supervisionado*, os exemplos não trazem consigo a "resposta correta", ou seja, não existe um professor guiando o aprendizado (Duda et al. 2001). Neste caso, o programa de computador deve tentar separar os exemplos em grupos, de acordo com similaridades ou dissimilaridades entre eles. A avaliação e interpretação do agrupamento formado requer a análise por parte do especialista humano. Por outro lado, em muitas situações, existem alguns poucos exemplos rotulados e muitos exemplos não rotulados, já que nem sempre é fácil ou barato analisá-los um a um em busca de tal rótulo. É nestes casos que entra o *aprendizado semissupervisionado*, que utiliza exemplos rotulados e não rotulados na criação do modelo durante o treinamento (Zhou & Li 2010).

No caso do aprendizado supervisionado, há ainda uma distinção adicional quanto ao conceito sendo aprendido. Assim, existem os problemas de: *classificação*, em que o conceito sendo aprendido é formado por c classes, que assumem valores discretos e finitos; e *regressão*, em que o conceito sendo

aprendido assume valores numéricos, podendo ser inclusive contínuos (por exemplo, uma determinada função matemática). Por exemplo, se um especialista possuir dados de expressão gênica de pacientes que possuem diferentes graus de uma dada doença, inclusive de pacientes que não possuem a doença, ele pode utilizar estes dados para construir um modelo de AM. Caso seu objetivo seja categorizar pacientes em "NORMAL", "LEVEMENTE DOENTE" ou "GRAVEMENTE DOENTE", ou seja, o conceito é representado por valores discretos e bem definidos, o modelo construído estará resolvendo um problema de classificação; caso a tarefa seja, por exemplo, associar um número que represente a chance de um dado paciente ter a referida doença e este número puder assumir infinitos valores dentro de um intervalo, o problema tratado é de regressão. Vale ressaltar que nem todos os algoritmos de AM conseguem resolver os dois tipos de problemas. Há os que só trabalham com classificação, outros que só trabalham com regressão e há também aqueles que tratam os dois problemas. Os algoritmos utilizados neste mestrado são especializados em classificação.

No aprendizado supervisionado, cada exemplo do conceito sendo aprendido é descrito por um vetor de valores de características, ou atributos, e o rótulo da classe associada (Monard & Baranauskas 2003a). Formalmente, em classificação, um exemplo $\vec{z}$ é um par $\vec{z} = (\vec{x}, y) = (\vec{x}, f(\vec{x}))$ onde $\vec{x}$ é a entrada e $y = f(\vec{x})$ é a saída e onde tanto $\vec{z}$ como $\vec{x}$ são vetores. A tarefa de um algoritmo de AM é, dado um conjunto de exemplos, induzir uma função h que aproxima f, normalmente desconhecida. Neste caso, $h(\cdot)$ é uma hipótese sobre a função objetivo $f(\cdot)$, ou seja, $h(\vec{x}) \approx f(\vec{x})$. Isto pode ser representado por uma tabela em que cada linha traz um exemplo diferente do conceito estudado, totalizando N linhas (ou exemplos). Tal tabela possui $m + 1$ colunas, sendo que m é o número de atributos que descrevem cada exemplo e a coluna adicional representa o rótulo da classe associada: $\vec{z}_i = (x_{i1}, x_{i2}, \ldots, x_{im}, y_i) = (\vec{x}_i, y_i), i \in [1, N]$. Esta tabela é conhecida como *tabela atributo-valor*, representada na Tabela 2.1.

Tabela 2.1: *Conjunto de exemplos no formato atributo-valor. Cada $\vec{z}_i, i \in [1, N]$, representa um exemplo do conjunto de dados; cada $X_j, j \in [1, m]$, representa um atributo utilizado para caracterizar os exemplos; Y representa os rótulos de classe relacionados ao conceito a ser aprendido.*

$\vec{Z}$	X_1	X_2	$\cdots$	X_m	Y
$\vec{z}_1$	x_{11}	x_{12}	$\cdots$	x_{1m}	y_1
$\vec{z}_2$	x_{21}	x_{22}	$\cdots$	x_{2m}	y_2
$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\vdots$
$\vec{z}_N$	x_{N1}	x_{N2}	$\cdots$	x_{Nm}	y_N

Neste projeto de mestrado os problemas abordados são todos de classificação, gerando modelos conhecidos como classificadores. Um classificador é construído a partir da execução de um algoritmo de aprendizado sobre um determinado número de exemplos de treinamento, para os quais o rótulo da classe associada é conhecido. O aprendizado aqui utilizado é do tipo indutivo, já que, a partir de alguns exemplos do conceito sendo estudado, extrapola um modelo para outros exemplos do mesmo conceito. Por este motivo, o algoritmo de aprendizado pode ser chamado também de *indutor*.

Uma vez construído um classificador, ele pode ser usado para predizer a classe de exemplos que siga a mesma representação e distribuição utilizada no treinamento, inclusive daqueles nunca apresentados ao indutor. Dependendo do indutor utilizado, o classificador pode ser simbólico ou não-simbólico. No contexto deste trabalho, um classificador simbólico é um modelo cuja linguagem de descrição é equivalente a um conjunto de regras, ou seja, o classificador pode ser representado em linguagem lógica proposicional ou relacional (Monard & Baranauskas 2003b). Um classificador simbólico de especial interesse nesta pesquisa são as árvores de decisão, que serão descritas com maiores detalhes na próxima seção.

2.3 Árvores de Decisão

Em geral, os algoritmos de indução de AD podem ser entendidos como uma ampla família de algoritmos de AM indutivo conhecida como *Top Down Induction of Decision Trees*. Uma AD é uma

estrutura de dados definida recursivamente como:

- um nó folha que corresponde a uma classe;
- ou um nó de decisão, que contém um teste sobre algum atributo. Para cada resultado do teste, existe uma aresta para uma subárvore. Cada subárvore tem a mesma estrutura da árvore.

A indução de AD tem sido aprimorada há algum tempo. Na década de 1960, Hunt et al. (1966) usaram métodos de busca exaustiva em AD para modelar o aprendizado de conceitos humanos. Já na década de 1970, foi publicado o trabalho de Quinlan (1979) com proto-ID3 (*Induction of Decision Trees*). Na década de 1980, houve a primeira publicação em massa, por Breiman et al. (1984), do software *Classification And Regression Trees* (CART) (presente atualmente em vários produtos comerciais), juntamente com o artigo de Quinlan (1986) sobre ID3. Desde então, uma variedade de melhorias têm sido incorporadas às ADs, tais como tratamento de ruído, tratamento de atributos contínuos e ausentes, árvores oblíquas (não paralelas aos eixos) e heurísticas de controle de *overfitting*.

O indutor ID3 pode ser considerado um algoritmo básico para a construção de AD sem poda, na qual é conduzida uma busca gulosa (*greedy*), ou seja, o algoritmo não reconsidera escolhas anteriores (Quinlan 1986). O algoritmo básico de construção de uma AD é bem simples: utilizando o conjunto de treinamento, um atributo é escolhido de forma a particionar os exemplos em subconjuntos, de acordo com valores desse atributo. Para cada subconjunto, outro atributo é escolhido para particionar novamente cada um deles. Cada escolha de atributo representa um teste realizado em um nó interno da árvore, ou seja, cada nó interno executa um teste em apenas um atributo e tem dois ou mais ramos, cada um representando um possível resultado do teste. Este processo prossegue enquanto um dado subconjunto criado contenha uma mistura de exemplos com relação aos rótulos de classe. Uma vez obtido um subconjunto uniforme — todos os exemplos naquele subconjunto pertencem à mesma classe — um nó folha é criado, sendo rotulado com o mesmo nome da respectiva classe. Um novo exemplo é rotulado da seguinte maneira: começando do nó raiz, testes são realizados e, de acordo com seus resultados, o exemplo caminha para baixo na árvore até encontrar um nó folha, recebendo, então, seu rótulo.

Uma evolução do ID3 encontra-se implementada no algoritmo C4.5 (Quinlan 1993). Muitas extensões foram acrescentadas ao algoritmo básico do ID3, tais como a melhora da eficiência computacional, tratamento de valores desconhecidos e de atributos contínuos, uso de *windowing* e o uso do critério de razão de ganho de informação em substituição ao critério de ganho. A construção da árvore e classificação de novos exemplos é feita de forma similar ao ID3. Adicionalmente, um segundo indutor foi implementado juntamente ao C4.5: o C4.5RULES, que examina a árvore produzida pelo C4.5 e deriva um conjunto de regras na forma $L \rightarrow R$, onde L é uma conjunção de testes nos atributos e R é um rótulo de classe. Todavia, o C4.5RULES não se limita a escrever os possíveis caminhos da árvore na forma de regras. Ele as generaliza, removendo condições irrelevantes (que não afetam a conclusão) sem afetar a acurácia. A última versão publicamente disponível é conhecida como *Release 8*. A partir dela, houve o lançamento da versão comercial, conhecida como C5.0 (http://www.rulequest.com/). O indutor C4.5 tem uma reimplementação Java na biblioteca Weka (Hall et al. 2009), conhecida como J48.

O algoritmo CART induz ADs para classificação ou regressão (Breiman et al. 1984). O classificador oblíquo OC1 é um algoritmo de indução projetado para aplicações nas quais os atributos são numéricos (Murthy et al. 1994), tais como na análise de sequências de DNA (Salzberg 1995a). O algoritmo OC1 constrói ADs que contêm uma combinação linear de atributos em cada nó interno. Essas árvores, portanto, podem particionar o espaço de descrição com hiperplanos paralelos ou oblíquos aos eixos.

Na Figura 2.1, é mostrado um exemplo de AD para dados de expressão gênica obtida por Lemos (2007). Na figura, cada elipse é um teste em um atributo para um dado gene. Cada retângulo representa uma classe: paciente saudável, paciente com leucemia mieloide aguda (LMA) ou paciente com leucemia linfocítica crônica (LLC). Ou seja, cada retângulo representa o fenótipo de interesse ou diagnóstico associado. Para classificar um novo exemplo (paciente), basta começar pela raiz e seguir

um único caminho para baixo, de acordo com os testes presentes nos nós internos visitados, até atingir uma folha. Na Figura 2.2, uma árvore fictícia é mostrada, juntamente com sua representação no espaço de atributos. Neste caso, o problema foi simplificado: existem apenas dois atributos na base de dados, representando os genes A e B; os dois atributos são contínuos e representam o nível de expressão dos seus genes respectivos; os possíveis valores para a classe são *Normal* e Doente. O processo de construção da árvore, juntamente com a delimitação das fronteiras de decisão no espaço de atributos, é também mostrada. Na situação 13, é possível ver os exemplos dispersos no espaço, sendo que as bolas representam pacientes doentes e os losangos representam pacientes normais, mas a árvore não começou a ser construída. Na situação 2, foi escolhido um teste sobre o atributo referente ao gene A: para um nível de expressão acima de 1000, os pacientes podem ser classificados como normais; para um nível de expressão abaixo de 1000, ainda há mistura de classes. Finalmente, na situação 3, foi escolhido um teste sobre o atributo referente ao gene B: para um nível de expressão abaixo de 900, os pacientes são classificados como normais; caso contrário, os pacientes são classificados como doentes. Nesta situação, podem ser visualizadas a árvore final e a divisão do espaço na forma de hiperplanos paralelos aos eixos.

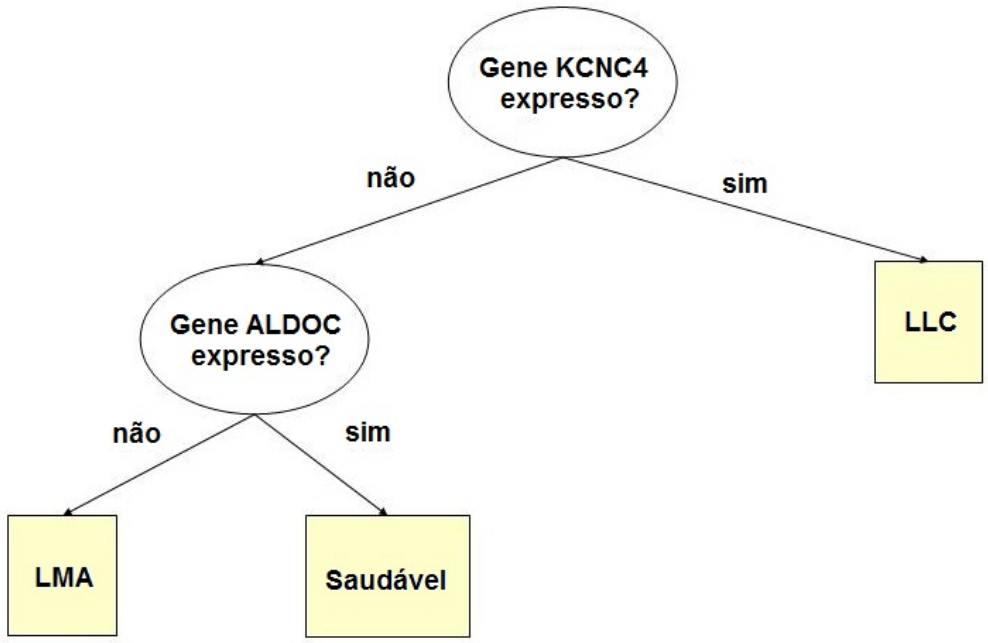


Figura 2.1: *Uma AD em um problema de classificação envolvendo dados de expressão gênica. Cada nó interno utiliza um atributo (neste caso um gene) para testar se o gene em questão está expresso. Os nós folha podem assumir os seguintes rótulos: Saudável, indicando pacientes saudáveis; LMA, que indica pacientes com leucemia mieloide aguda; ou LLC, indicando pacientes com leucemia linfocítica crônica.*

A chave para o sucesso da AD gerada é altamente dependente do critério utilizado para escolher o atributo que particiona o conjunto de exemplos em cada iteração. Os critérios geralmente utilizados tentam escolher um atributo que resulte no menor tamanho esperado das subárvores, entre eles o ganho máximo de informação (Quinlan 1986), índice Gini (Breiman et al. 1984) e razão de ganho de informação (Quinlan 1993).

2.4 Considerações Finais

Neste capítulo foram descritos alguns conceitos importantes em AM supervisionado utilizando AD. Ao invés de considerar apenas medidas como acurácia, ADs podem ser usadas para prover melhor compreensão do domínio considerado quando interpretadas por especialistas, uma característica muito importante nas áreas médica e biológica. Neste contexto, uma vantagem do *windowing* e do *lookahead* em relação a outras técnicas usadas para melhorar o desempenho de classificadores é que seu uso com classificadores simbólicos também produz classificadores simbólicos. No próximo capítulo, serão apresentadas as medidas propostas e analisadas neste mestrado.

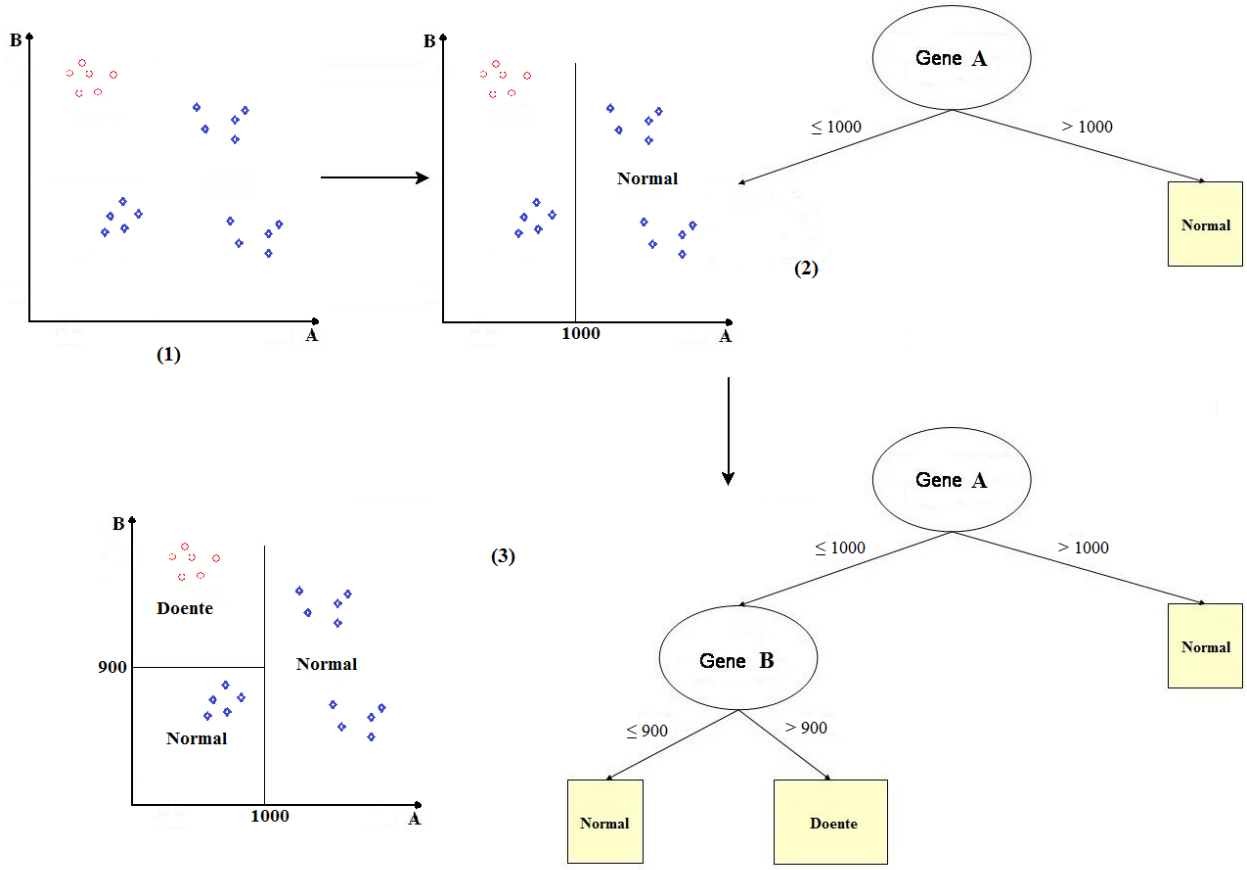


Figura 2.2: *Exemplo de AD de um problema fictício: existem apenas dois atributos na base de dados, representando os genes A e B; os dois atributos são contínuos e representam o nível de expressão dos seus respectivos genes; os possíveis valores para a classe são* Normal *e* Doente*. Todo o processo de construção da árvore e delimitação das fronteiras de decisão no espaço de atributos é mostrado. Na situação 1, os exemplos estão dispersos no espaço, sendo que as bolas representam pacientes doentes e os losangos representam pacientes normais. Neste ponto, não há nós de decisão na árvore e, consequentemente, não há fronteiras no espaço de atributos. Na situação 2, foi escolhido um teste sobre o gene A: para um nível de expressão acima de 1000, os pacientes podem ser classificados como normais; para um nível de expressão abaixo de 1000, ainda há mistura de classes. Finalmente, na situação 3, foi escolhido um teste sobre o gene B: para um nível de expressão abaixo de 900, os pacientes são classificados como normais; caso contrário, os pacientes são classificados como doentes. Nesta situação, a árvore final é mostrada. Neste caso, as fronteiras construídas pela árvore são sempre hiperplanos paralelos aos eixos.*

Capítulo

3

Medidas de Avaliação de Árvores

3.1 Considerações Iniciais

A acurácia, ou taxa de acerto, é definida em classificação como o número de exemplos cuja classe foi predita corretamente pelo modelo dividido pelo número total de predições realizadas. A área sob a curva ROC é apresentada na Seção D.2. As duas medidas anteriores, entre outras, são muito utilizadas na avaliação de modelos, mas, quando se quer ter uma ideia do grau de interpretabilidade do modelo induzido, ou do grau de utilidade do conhecimento trazido pelo modelo, outras medidas devem ser usadas.

No contexto de regras de decisão, muitas dessas medidas já estão definidas (Lavrač et al. 1999). No entanto, não foram encontrados trabalhos que propuseram tais medidas aplicáveis diretamente a ADs. O que os trabalhos existentes fazem é reescrever a árvore na forma de regras de decisão, sendo que cada caminho da raiz até as folhas pode ser representado na forma $L \rightarrow R$, onde L é uma conjunção de testes nos atributos e R é o rótulo prevalente de classe da folha atingida. A partir daí, as medidas aplicáveis a regras podem ser usadas. Regras geradas desta forma possuem alguns problemas, entre eles a redundância de testes. Caso esta redundância seja eliminada, as regras passam a ser dependentes umas das outras, sendo que a ordem em que serão aplicadas passa a ser relevante. Regras de decisão puras são geralmente geradas de maneira independente umas das outras, o que não ocorre no processo de reescrita das ADs.

O grupo de pesquisa do qual o aluno faz parte tem estudado algumas medidas que podem ser aplicadas diretamente a árvores. Algumas das medidas com as quais o grupo tem trabalhado são apresentadas nas próximas seções. Na Figura 3.1, são apresentadas três árvores fictícias, que servirão como base para a exemplificação do cálculo das medidas propostas.

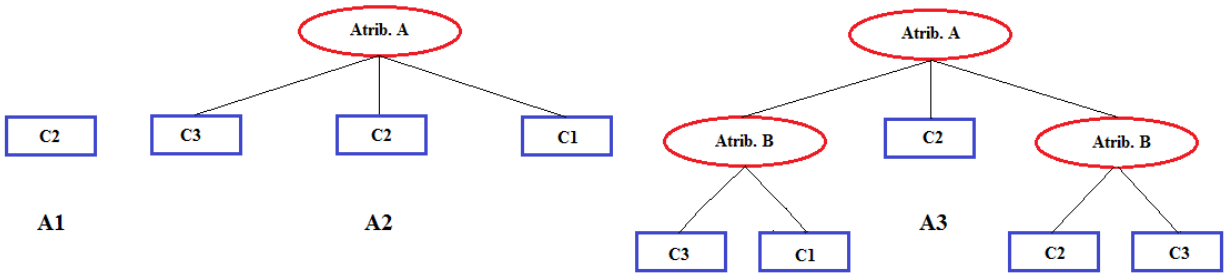


Figura 3.1: *Três árvores fictícias (A1, A2 e A3), que servirão de base para o cálculo de algumas medidas que serão apresentadas nesta seção. As três árvores foram construídas a partir da mesma base de dados: existem oito atributos, A, B, C, D, E, F, G e H, e a classe pode assumir três valores distintos, C1, C2 e C3.*

3.2 Tamanho da Árvore

Neste trabalho, o tamanho da árvore final produzida pelo indutor é medido em termos do número de nós do modelo. Esta é uma medida trivial, mas que, no entanto, fornece um dos aspectos relacionados à interpretabilidade do modelo: assume-se que, quanto menor a árvore, mais facilmente ela pode ser entendida pelos especialista de domínio. Ainda que sejam modelos simbólicos, ADs muito grandes podem dificultar muito sua interpretação. Para as três árvores da Figura 3.1, tem-se:

$$\begin{aligned} \text{Tamanho}_{A1} &= 1 \\ \text{Tamanho}_{A2} &= 4 \\ \text{Tamanho}_{A3} &= 8 \end{aligned} \tag{3.1}$$

3.3 Altura da Árvore

A altura de uma árvore é definida como o número de passos necessários para se caminhar da raiz até o nó folha mais profundo. Este caminho considera apenas a passagem de um dado nó a um de seus descendentes diretos. Esta é outra medida trivial, mas que fornece uma ideia do grau de complexidade da árvore. Quanto maior a altura, mais atributos são necessários para se explicar o conceito, o que torna a árvore mais complexa. Para as três árvores da Figura 3.1, tem-se:

$$\begin{aligned} \text{Altura}_{A1} &= 1 \\ \text{Altura}_{A2} &= 2 \\ \text{Altura}_{A3} &= 3 \end{aligned} \tag{3.2}$$

3.4 Tamanho da Janela

Esta é uma medida proposta pelo grupo, mas que é específica para a análise de modelos produzidos pelo *windowing*. Esta técnica pode produzir bons modelos utilizando apenas parte do conjunto original de treinamento. Por exemplo, o conjunto de treinamento completo pode possuir mil exemplos, mas o algoritmo pode utilizar apenas quinhentos para construir o modelo. A medida é definida então como o número de exemplos presentes na janela que gerou o classificador final produzido pelo algoritmo. Será possível compreendê-la melhor após a apresentação detalhada do algoritmo. Desde que o a medida de desempenho utilizada não tenha prejuízo significativo, quanto menor o tamanho da janela que gerou o melhor classificador, melhor pode ser considerado o algoritmo de janelamento que o produziu. Para o caso das três árvores da Figura 3.1, não é possível calcular o tamanho da janela apenas analisando os modelos gerados. Para isto, seria necessário analisar a execução do algoritmo.

3.5 Coesão da Árvore

A coesão é outra medida proposta pelo grupo. Considerando c, $c > 1$, como o número de diferentes rótulos que a classe do problema estudado pode assumir e f como o número de nós folha presentes no modelo induzido, a coesão pode ser definida como:

$$\text{Coesão} = \frac{c}{f - 1 + c} \tag{3.3}$$

Como se pode notar, o valor calculado para coesão pertence ao intervalo $(0, 1]$. Considerando que a árvore tem pelo menos um nó folha (na pior das hipóteses, a árvore tentará predizer a classe majoritária), o valor 1 será atingido quando o modelo contiver exatamente um nó folha, representando o modelo mais simples possível. Conforme o número de folhas aumenta, o valor da coesão se aproxima de 0 e a árvore vai se tornando mais complexa, pois mais formas de explicar o conceito existem (segundo o modelo, é claro). Um valor mais próximo de 1 indica que o modelo é mais coeso com relação ao número possível de rótulos de classe. O único caso especial é o citado acima, em que o modelo se resume a um nó folha. Apesar de assumir valor 1, geralmente não

poderá ser considerado um bom modelo, pois não descobriu nenhuma informação relevante sobre o problema estudado.

Para as três árvores da Figura 3.1, tem-se:

$$\begin{aligned} \text{Coesão}_{A1} &= \frac{3}{1-1+3} = 1,00 \\ \text{Coesão}_{A2} &= \frac{3}{3-1+3} = \frac{3}{5} = 0,60 \\ \text{Coesão}_{A3} &= \frac{3}{5-1+3} = \frac{3}{7} = 0,43 \end{aligned} \tag{3.4}$$

3.6 Compactação do Conhecimento

Esta é outra medida proposta pelo grupo e que mede o quão compacta a árvore é com relação ao conhecimento envolvido no problema tratado. Considerando a_a, $a_a \geq 0$, como o número de diferentes atributos efetivamente utilizados no modelo e a_b, $a_b > 0$, como o número total de atributos presentes na base de dados, a compactação é definida como:

$$\text{Compactação} = \frac{a_a}{a_b} \tag{3.5}$$

O valor desta medida pertence ao intervalo $[0, 1]$. O valor 1 indica que cada atributo da base apareceu pelo menos uma vez no modelo; o valor 0 indica que nenhum atributo apareceu no modelo. O conceito de compactação aqui se refere a quantos atributos o modelo considerou relevantes para descrever o conceito, relativamente ao número total de atributos presentes. Um valor próximo a 0 indica que o conhecimento é mais compacto (segundo o modelo gerado), sendo o modelo considerado mais simples que um modelo cuja compactação se aproxima de 1, caso em que foram necessários muitos atributos para representar o conceito.

Esta medida é principalmente interessante nos problemas envolvendo expressão gênica, onde geralmente existem milhares de atributos, mas apenas alguns deles são efetivamente relevantes. É claro que o fato de uma árvore possuir poucos atributos e, assim, indicar que o conhecimento envolvido é compacto, não significa que ela tenha escolhido os atributos corretos (os mais relevantes para o problema). Para as três árvores da Figura 3.1, tem-se:

$$\begin{aligned} \text{Compactacao}_{A1} &= \frac{0}{8} = 0,00 \\ \text{Compactacao}_{A2} &= \frac{1}{8} = 0,13 \\ \text{Compactacao}_{A3} &= \frac{2}{8} = 0,25 \end{aligned} \tag{3.6}$$

3.7 Coesão-Compactação

Pode-se dizer que coesão e compactação estão ligados. Se há poucas folhas na árvore, a tendência é que haja poucos atributos e vice-versa. Ou seja, quanto maior a coesão, menor a compactação. O contrário também é verdadeiro. Uma forma de relacioná-los é fazer a média geométrica entre (1 − Coesão) e Compactação. Geometricamente, o valor resultante seria o tamanho do lado do quadrado com a mesma área do retângulo cujos lados têm os tamanhos iguais aos valores de cada uma das medidas. Quanto maior esse quadrado, mais complexo poderia ser considerado o conceito estudado. O uso de (1 − Coesão), e não simplesmente Coesão, se deu para que as medidas se tornassem diretamente proporcionais.

$$\text{Coesão-Compactação} = \sqrt{(1 - \text{Coesão}) * \text{Compactação}} \tag{3.7}$$

Para as três árvores da Figura 3.1, tem-se:

$$\begin{aligned}
\text{Coesão-Compactação}_{A1} &= \sqrt{(1-1)*0} = 0,00 \\
\text{Coesão-Compactação}_{A2} &= \sqrt{(1-\frac{3}{5})*\frac{1}{8}} = 0,22 \\
\text{Coesão-Compactação}_{A3} &= \sqrt{(1-\frac{3}{7})*\frac{2}{8}} = 0,38
\end{aligned} \tag{3.8}$$

3.8 Desvio Padrão

Indutores de AD são tidos como instáveis, já que pequenas alterações na base de dados podem produzir grandes alterações no modelo induzido. Um dos objetivos deste mestrado é propor uma forma de minimizar este problema. Para verificar se os algoritmos propostos estão atingindo tal objetivo, algumas das variáveis consideradas nos testes estatísticos, além de terem sido comparadas em termos de sua média, foram comparadas também em termos de seu desvio padrão. É claro que tal medida não considera todos os aspectos do problema mencionado, mas fornece uma estimativa razoável. É claro também que o desvio padrão de uma variável não é uma medida específica de árvores e pode ser usado com qualquer outro tipo de indutor.

3.9 Considerações Finais

Medidas clássicas, como acurácia, e as propostas acima foram utilizadas na análise de alguns dos modelos produzidos neste mestrado. No próximo capítulo, são apresentados o *windowing*, cuja análise utilizou as medidas definidas anteriormente, e o *lookahead*.

Capítulo

4

Windowing & Lookahead

4.1 Considerações Iniciais

Neste capítulo, as duas abordagens exploradas no mestrado são apresentadas: *windowing* e *lookahead*. Além disto, uma série de alterações realizadas na técnica de *windowing* são descritas, na tentativa de melhorar seu desempenho em aplicações envolvendo construção de modelos de AM a partir de dados de expressão gênica.

4.2 Windowing

4.2.1 Versão Original

Windowing foi proposto por Quinlan (1979) no contexto de ADs como uma maneira de lidar com restrições de memória impostas pelos computadores do final da década de 1970. Alguns conjuntos de exemplos eram muito grandes para serem inteiramente carregados na memória principal. Assim como os recursos computacionais evoluem, as bases de dados têm crescido, havendo algumas com tamanho da ordem de tera ou até petabytes (http://www.archive.org/). Mesmo que memória não mais representasse um problema, Quinlan (1993) argumenta que *windowing* continuaria interessante, por duas razões:

i) Em alguns casos, especialmente em problemas livres de ruído, ele pode diminuir o tempo necessário para se treinar um classificador (por exemplo, quando a base de dados é muito grande e um modelo que classifique perfeitamente todos os exemplos de treinamento seja alcançado nas primeiras iterações do algoritmo). No entanto, para a maioria dos casos, a técnica torna o tempo de treinamento maior;

ii) *Windowing* pode produzir classificadores melhores, visto que explora um pouco mais o espaço de soluções, no sentido de produzir várias árvores, a partir de diferentes subconjuntos da base de treinamento.

Alguns estudos têm explorado a técnica no contexto de indução de regras (Fürnkranz 1997; Domingos 1996), já que foi notado que ela produz melhores resultados com aquele tipo de indutor: regras são aprendidas independentemente umas das outras e são menos suscetíveis a mudanças na distribuição de classes. Apesar da experiência que Quinlan teve com *windowing*, aqueles estudos afirmam que a técnica não é adequada para o uso com AD, especialmente em domínios ruidosos. Possivelmente por isto, pouco tem sido estudado com o intuito de melhorar a combinação entre *windowing* e AD. No entanto, ainda há alguns aspectos que poderiam ser mais explorados, por exemplo: a aplicação da técnica em domínios com distribuição balanceada e desbalanceada de classes; alteração nos critérios de desempenho das árvores intermediárias obtidas durante a execução

do algoritmo; alteração nos critérios de parada do algoritmo; tratamento de ruído, entre outros. Um exemplo de uso recente da técnica é o de Moon & Marwala (2008). Os autores usaram *windowing* para melhorar a acurácia e economizar memória virtual da construção do instrumento que usaram para estimar dados ausentes.

No Algoritmo 1 é mostrado o pseudo-código do *windowing*, baseado em Quinlan (1993), onde N representa o número de exemplos no conjunto de treinamento e $\vec{x}_i$ e y_i $(i = 1, \ldots, N)$ representam um vetor com os valores de atributos e a classe do exemplo i, respectivamente. O operador $\| E \|$ retorna 1, se E é verdadeiro, ou 0, caso contrário. Primeiramente, uma amostra do conjunto de treinamento é escolhida, formando a janela inicial (Linha 2), a partir da qual um modelo é induzido (Linha 5); o modelo é usado para predizer a classe considerando todos os exemplos de treinamento (dentro e fora da janela), o que pode produzir classificações erradas (Linhas 6-7); se o total de erros encontrados for menor que o total de erros do melhor modelo encontrado até então (inicialmente, $N + 1$), o modelo atual é guardado como o melhor (Linhas 9-12); se houve erros fora da janela, esta é atualizada (Linhas 13-16) e utilizada para treinar outro classificador; o modelo resultante é testado novamente e o processo se repete até que nenhum erro seja encontrado fora da janela.

A janela inicial não é escolhida de maneira totalmente aleatória. Primeiramente o conjunto de exemplos é embaralhado, processo este aleatório; em seguida o algoritmo tenta construir uma janela tão uniforme quanto possível, *i.e.*, a distribuição de classes deve ser tão balanceada quanto possível. Considere c como o número de rótulos de classes existentes e $\mathcal{E} = W/c$ como o número esperado de exemplos representando cada classe na janela inicial. Para cada rótulo de classe possível, se o número de exemplos com este rótulo for pelo menos igual a $\mathcal{E}$, ele será representado por $\mathcal{E}$ exemplos na janela inicial; caso contrário, todos os exemplos contendo aquele rótulo serão adicionados à janela. Isto geralmente leva a resultados melhores, especialmente em problemas com desbalanceamento de classes (Quinlan 1993), e contribui também para uma melhor escolha de pontos de corte para atributos contínuos (Catlett 1991). Devido ao embaralhamento aleatório, a janela inicial pode ser diferente a cada execução do algoritmo, levando possivelmente a classificadores finais diferentes.

Como pode ser visto no Algoritmo 1, pelo menos metade dos exemplos fora da janela que o classificador corrente errou é adicionada a ela a cada iteração (Linhas 13-16), para que o algoritmo possa convergir mais rapidamente (Quinlan 1993). Na Linha 16, a janela é atualizada com exemplos classificados erroneamente fora dela. A quantidade de exemplos a ser adicionada à janela é igual a `incremento`.

O processo pode ser repetido muitas vezes, sendo que cada repetição é denominada *trial* e começa com uma janela inicial diferente, o que geralmente gera um classificador final diferente. Por padrão, o C4.5 usa dez *trials*. O melhor classificador de todas as repetições do algoritmo é retornado ao usuário ou o sistema pode guardar o melhor classificador de cada *trial* e combiná-los em um único classificador baseado em regras de decisão.

4.2.2 Alterações Propostas

O autor estudou a técnica de *windowing* aprofundadamente, inclusive seu código-fonte mais recente (Quinlan 1993), e realizou uma série de experimentos com a técnica. Com esta experiência acumulada, alguns problemas foram identificados e o autor, juntamente com o grupo de pesquisa em que está inserido, desenvolveu algumas alterações, que visam a minimizar tais problemas.

No algoritmo original, apenas a melhor árvore de cada *trial* é podada (Linha 42 do Algoritmo 2). Todas as árvores intermediárias geradas dentro de um determinado *trial* são mantidas sem poda. Para decidir qual é o melhor classificador entre os *trials*, o algoritmo poda a árvore retornada pelo *trial* atual e calcula o erro estimado desse classificador, que é então comparado ao erro estimado do melhor classificador encontrado nos *trials* já processados. Na implementação realizada neste mestrado, existe a opção de não se usar a Linha 42, mas sim a Linha 8 (Algoritmo 2), de forma que todas as árvores são podadas. Esta alteração foi proposta porque a poda de ADs é considerada importante, já que ajuda a árvore a melhorar seu erro de generalização (Monard & Baranauskas 2003b). Apesar de acarretar um esforço computacional maior, desejava-se verificar se a alteração traria algum benefício em comparação com a versão original da técnica e com a árvore produzida da maneira tradicional.

Algoritmo 1 Windowing.

Pré-condição: Exemplos: um conjunto de N exemplos rotulados $\{(\vec{x}_i, y_i), i = 1, 2, \ldots, N\}$
W: o tamanho inicial da janela, com valor padrão $W \leftarrow \max\{0, 2N; 2\sqrt{N}\}$
I_0: o tamanho requerido do incremento, com valor padrão $I_0 \leftarrow \max\{0, 2W; 1\}$
Pós-condição: melhorClassificador: o melhor classificador encontrado

1: $N \leftarrow$ obterTamanho(Exemplos)
2: janela $\leftarrow$ amostrar(W, Exemplos)
3: errosMelhorClassificador $\leftarrow N + 1$
4: **repita**
5: classificador $\leftarrow$ induzirClassificador(janela)
6: errosJanela $\leftarrow \sum_{\vec{x}_i \in \text{janela}} \parallel \text{classificar}(\vec{x}_i, classificador) \neq y_i \parallel$ *// Erros na janela*
7: errosTeste $\leftarrow \sum_{\vec{x}_i \notin \text{janela}} \parallel \text{classificar}(\vec{x}_i, classificador) \neq y_i \parallel$ *// Erros fora da janela*
8: errosTotal $\leftarrow$ errosJanela + errosTeste
9: **se** (errosTotal $<$ errosMelhorClassificador) **então**
10: melhorClassificador $\leftarrow$ classificador
11: errosMelhorClassificador $\leftarrow$ errosTotal
12: **fim se**
13: $I \leftarrow$ obterMax(obterMin(errosTeste, I_0), errosTeste /2)
14: $I \leftarrow$ obterMin(I, N - obterTamanho(Exemplos $-$ janela))
15: incremento $\leftarrow$ I primeiros exemplos teste classificados incorretamente
16: janela $\leftarrow$ janela + incremento
17: **até** (errosTeste $= 0$)
18: melhorClassificador $\leftarrow$ podar(melhorClassificador)
19: **retorne** melhorClassificador

Domingos (1996) afirma que *windowing* tem seu desempenho deteriorado em domínios com ruído, porque acaba adicionando todos os exemplos ruidosos à janela, já que eles são normalmente classificados erroneamente mesmo por bons modelos (Nettleton et al. 2010). Para tratar este problema, outra alteração aqui proposta baseia-se na confiança na classificação realizada e teve sua inspiração no aprendizado semissupervisionado: a técnica de *co-training* (Blum & Mitchell 1998) utiliza a confiança na classificação como critério para adicionar exemplos não rotulados ao conjunto de exemplos rotulados. Seguindo esta ideia, o autor implementou a opção de utilização do critério de confiança para atualizar a janela (Linha 26 do Algoritmo 2).

O indutor J48 confere um vetor de valores a cada predição que realiza. Cada posição do vetor corresponde a uma das classes possíveis e traz a distribuição desta classe no nó folha atingido. A classe predita será aquela com a maior distribuição no nó folha. Estes valores podem ser vistos como um fator de confiança do modelo na predição que acabou de realizar.

Em uma dada iteração do algoritmo de *windowing*, podem existir exemplos classificados incorretamente fora da janela. Na alteração proposta, tais exemplos são ordenados de acordo com a confiança com que o modelo corrente prediria a classe verdadeira. Lembrando que, neste caso, a classe predita foi diferente da verdadeira, pois o exemplo foi classificado incorretamente, mas a classe verdadeira possui uma confiança associada. Estes valores de confiança (da classe verdadeira) são colocados em ordem decrescente. A heurística utilizada aqui é a seguinte: apesar de o exemplo ter sido classificado incorretamente, quanto maior a confiança associada pelo modelo à classe verdadeira, mais perto o modelo está de acertar a classificação do exemplo e menor é a chance de tal exemplo ser ruído. Por exemplo: em um problema em que a classe pode assumir os valores "Sim" e "Não", o modelo construído por uma dada iteração do algoritmo classificou incorretamente dois exemplos fora da janela. Um deles tinha a classe "Sim" como verdadeira, sendo que o modelo atribuiu uma confiança de 0,40 a ela e 0,60 à classe "Não" (que foi a classe predita pelo modelo). O outro exemplo tinha a classe "Não" como verdadeira, sendo que o modelo atribuiu uma confiança de 0,05 a ela e 0,95 à classe "Sim" (que foi a classe predita pelo modelo). Apesar de, em ambos os casos, o modelo ter cometido erros, pode-se dizer que ele está mais próximo de acertar no primeiro caso.

O algoritmo calcula a cada iteração o tamanho do incremento usado para atualizar a janela. No

momento em que a janela for atualizada, os exemplos a serem adicionados primeiro serão aqueles com maior confiança associada. No entanto, ao utilizar o critério de confiança, o algoritmo faz algumas alterações também no cálculo do incremento. Quando da ordenação dos exemplos pelo valor de confiança, o algoritmo conta quantos valores de confiança foram maiores do que zero, número a ser designado por N_0 (Linha 28 do Algoritmo 2). É verificado, então, se N_0 foi maior que a metade do incremento. Em caso positivo, o incremento é atualizado como sendo N_0, desde que $N_0 <$ incremento. Caso contrário, o incremento é dividido pela metade. Isto é feito por dois motivos: o fato de N_0 ser pequeno pode se dever à presença de ruído, fazendo com que não seja interessante acrescentar muitos exemplos à janela neste momento; o algoritmo está tendo alguma dificuldade em aprender o conceito, fazendo com que seja interessante fazer ajustes mais finos no incremento. Assim, o algoritmo teria convergência mais lenta, mas não correria o risco de acrescentar uma série de exemplos que talvez não trouxessem nenhuma informação realmente útil.

Quinlan (1993) afirma que há três critérios de parada para o *windowing*: (i) em uma dada iteração, não houve exemplos fora da janela classificados incorretamente; (ii) todos os exemplos já foram adicionados à janela; e (iii) se o algoritmo verificar que nenhum progresso significativo está sendo realizado, mesmo que ainda haja exemplos a adicionar à janela. No entanto, após análise detalhada do código-fonte original, o terceiro critério não foi encontrado. Um *e-mail* foi enviado ao autor, para que ele pudesse esclarecer a dúvida. O autor original da técnica respondeu que a implementação do terceiro critério não tinha sido realmente disponibilizada no sistema. A partir disto, este mestrado propôs uma forma de acrescentar tal possibilidade: se N_0 tiver valor 0 por mais de três iterações consecutivas, *i.e.*, se os modelos produzidos nas últimas iterações não tiveram um valor de confiança maior que 0 para nenhum exemplo, a parada ocorre (Linha 32 do Algoritmo 2). Este número de iterações consecutivas (três) foi definido empiricamente. Outros valores foram testados, mas este foi o que apresentou os melhores resultados.

A terceira proposta implementada se refere ao cálculo do erro das árvores intermediárias (aquelas produzidas dentro de um *trial*). No algoritmo original, o erro de uma dada árvore dentro de um *trial* é dado pela soma dos erros encontrados dentro da janela (erros de treinamento) e dos encontrados fora dela (exemplos não vistos durante o treinamento). Porém, o erro das árvores construídas nas últimas iterações de um *trial* se aproxima do erro de ressubstituição, já que a janela fica cada vez maior conforme as iterações acontecem. O erro de ressubstituição é um erro otimista, fazendo com que as árvores construídas no início de um dado *trial* fiquem em desvantagem, pois nesse momento a janela é menor e o número de exemplos não vistos é maior. Para tentar tornar mais justa a avaliação das árvores, este mestrado propõe, no lugar do erro encontrado na janela, o uso do erro estimado (Linha 11 do Algoritmo 2), proposto por Quinlan (1993), que o usa como uma estimativa pessimista do erro de generalização do modelo. Esta estimativa é baseada na distribuição binomial e também leva em consideração a distribuição de classes dos nós folha. Este erro estimado é o mesmo utilizado pelo algoritmo original para avaliar as melhores árvores de cada *trial*. Quinlan propôs o cálculo do erro estimado da árvore seguindo novamente uma definição recursiva, em que o erro estimado de um nó não folha é a soma dos erros de seus descendentes; já o erro estimado de um nó folha é calculado assim: considerando que o nó folha cubra M exemplos de treinamento, E dos quais incorretamente, Quinlan comparou esta situação à observação de E eventos em M tentativas. Se os M casos forem vistos como uma amostra, poderia ser calculada uma probabilidade *a posteriori* que daria uma ideia da probabilidade de erro do nó folha. Esta probabilidade é então calculada na forma de um intervalo de confiança. Uma vez definido um nível de confiança, que é um dos parâmetros do indutor, o intervalo de confiança desta probabilidade pode ser calculado por meio do intervalo de confiança de uma distribuição binomial. Como o interesse aqui é a predição do erro de generalização da árvore, só interessa o limite superior do intervalo, que fornece uma estimativa pessimista para tal erro. Tal aproximação representa uma heurística que em geral produz boas estimativas, segundo Quinlan (1993).

Pelo mesmo motivo apresentado no parágrafo anterior, uma outra alteração foi implementada: uso de ponderação do erro das árvores intermediárias. Esta ponderação é baseada no tamanho da janela corrente. Quanto maior a janela, mais dados o algoritmo tem para o treinamento e, assim,

maior é a sua chance de acertar. A proposta é "punir" árvores construídas a partir de janelas maiores e aproximar as condições a que as árvores construídas estão submetidas. O erro total calculado é multiplicado por $1 + J/N$, sendo J o tamanho da janela atual e N o número total de exemplos (Linha 18 do Algoritmo 2).

4.3 Lookahead

A indução de AD segue uma abordagem gulosa, que parte da raiz e caminha até as folhas. Decisões são tomadas com base nas informações que se tem no momento da escolha de um teste em um dado nó e decisões anteriores não são revistas. O resultado deste tipo de indução é composto por árvores subótimas. Uma forma de tentar melhorar esse aprendizado guloso é a aplicação da técnica denominada *lookahead* (Sarkar et al. 1994), em que, a cada decisão a ser tomada em um dado nó, o algoritmo leva em consideração um número finito de passos adiante. Por exemplo, se o *lookahead* for de um passo, o algoritmo, ao escolher o teste em um determinado nó, verifica qual é a escolha que proporcionará o melhor resultado no passo seguinte, não no passo atual. Esta é uma operação dispendiosa computacionalmente, fazendo com que o *lookahead*, quando usado, considere apenas alguns poucos passos adiante. Ainda assim, para alguns problemas, está comprovado que *lookahead* fornece melhores limites (*bounds*) para a solução encontrada do que o algoritmo guloso em que ele é aplicado (Sarkar et al. 1994).

Um dos problemas clássicos da computação é o *XOR*. No caso do *XOR* bidimensional (dois atributos), em que tanto os atributos quanto a classe são *booleanos* (*Verdadeiro* ou *Falso*), a classe assume valor *Verdadeiro* quando os valores dos atributos são diferentes; caso contrário, assume valor *Falso*. Uma variação deste problema, o *XOR* contínuo, apresenta-se um pouco mais difícil para uma AD. Neste caso, a classe continua sendo *booleana*, mas os atributos são números reais. Como é mostrado na Figura 4.1, não existe, quando da definição da primeira partição a ser escolhida, um ponto de corte em nenhum dos dois atributos que faça com que o critério considerado seja melhorado. Caso se defina um corte do atributo X em torno de $0,5$, a mistura de classes nas partições criadas se mantém a mesma da partição original. O mesmo ocorre com o atributo Y. Neste caso, a árvore tende a parar o treinamento precipitadamente, já que ela não tem mecanismos para perceber que, caso ela escolha o ponto de corte citado acima, na próxima iteração ela terá resolvido o problema. Problemas de outros domínios podem apresentar este tipo de situação, que o *lookahead* resolveria com certa facilidade. Por exemplo, se for usado *lookahead* de um nível no problema do *XOR* contínuo bidimensional, a técnica resolverá o problema facilmente, pois, ao escolher o atributo do teste atual, ela verifica o resultado um passo à frente.

No Algoritmo 3, está representado o processo básico de indução *top-down* de AD com *lookahead*. A única diferença dele com o processo clássico é que, ao invés de simplesmente calcular, para cada atributo, o critério utilizado, uma chamada ao *lookahead* é feita (Linha 10). Na realidade, um caso particular ocorre quando o parâmetro P da chamada ao algoritmo recebe o valor 0, fazendo com que seja reduzido à versão clássica de indução de árvores. Quando da primeira chamada a *induzirArvore*, C é uma árvore vazia. Um ponto a ressaltar no algoritmo é a variável A, que sempre conterá os atributos ainda disponíveis para a indução do modelo. No contexto deste mestrado, sempre que o atributo escolhido para um dado nó teste for nominal, ele será retirado do conjunto A, pois não há mais sentido em mantê-lo, já que cada possível partição resultante do teste conterá exemplos de apenas um determinado valor para o atributo; sempre que o atributo escolhido para um dado nó teste for numérico, ele será mantido no conjunto A, pois pode ainda vir a ser usado novamente (tais atributos sofrerão sempre cortes binários). Quando da chamada a *criarTeste* (Linha 12 do Algoritmo 3), a^* representa não somente o atributo escolhido para teste, mas também os pontos de corte para a produção das partições. Ainda no mesmo algoritmo (Linha 11), o operador *oti* retorna o atributo com o maior (ou menor, dependendo do critério utilizado) valor calculado para o critério.

Outro ponto a ressaltar nos Algoritmos 3 e 4 é que o operador de adição, quando aplicado simplesmente a um vetor e não a uma posição específica deste vetor, indica que o valor atribuído será adicionado à última posição do vetor, como um item sendo adicionado a um conjunto. É o caso, por exemplo, da variável *criterioParticoes*, que guarda o valor calculado do critério para uma

Algoritmo 2 Windowing com as alterações propostas.

Pré-condição: Exemplos: um conjunto de N exemplos rotulados $\{(\vec{x}_i, y_i), i = 1, 2, \ldots, N\}$
W: o tamanho inicial da janela, com valor padrão $W \leftarrow \max\{0, 2N; 2\sqrt{N}\}$
I_0: o tamanho requerido do incremento, com valor padrão $I_0 \leftarrow \max\{0, 2W; 1\}$
Pós-condição: melhorClassificador: o melhor classificador encontrado

1: $N \leftarrow$ obterTamanho(Exemplos)
2: janela $\leftarrow$ amostrar(W, Exemplos)
3: errosMelhorClassificador $\leftarrow N + 1$
4: contadorConfianca $= 0$
5: **repita**
6: classificador $\leftarrow$ induzirClassificador(janela)
7: **se** (Podar árvores intermediárias) **então**
8: classificador $\leftarrow$ podar(classificador)
9: **fim se**
10: **se** (Usar erro estimado) **então**
11: errosJanela $\leftarrow$ estimarErro(classificador) *// Erro estimado (para a janela)*
12: **senão**
13: errosJanela $\leftarrow \sum_{x_i \in \text{janela}} \parallel \text{classificar}(\vec{x}_i, classificador) \neq y_i \parallel$ *// Erros na janela*
14: **fim se**
15: errosTeste $\leftarrow \sum_{x_i \notin \text{janela}} \parallel \text{classificar}(\vec{x}_i, classificador) \neq y_i \parallel$ *// Erros fora da janela*
16: errosTotal $\leftarrow$ errosJanela + errosTeste
17: **se** (Ponderar erro) **então**
18: errosTotal $\leftarrow$ ponderarErro(errosTotal)
19: **fim se**
20: **se** (errosTotal $<$ errosMelhorClassificador) **então**
21: melhorClassificador $\leftarrow$ classificador
22: errosMelhorClassificador $\leftarrow$ errosTotal
23: **fim se**
24: $I \leftarrow$ obterMax(obterMin(errosTeste, I_0), errosTeste /2)
25: $I \leftarrow$ obterMin(I, N - obterTamanho(Exemplos $-$ janela))
26: **se** (Utilizar confiança) **então**
27: Ordenar exemplos teste pela confiança
28: $N_0 \leftarrow$ # exemplos teste com confiança 0
29: **se** ($N_0 = 0$) **então**
30: contadorConfianca $\leftarrow$ contadorConfianca +1
31: **se** (contadorConfianca > 3) **então**
32: **retorne** melhorClassificador
33: **fim se**
34: **senão**
35: contadorConfianca $\leftarrow 0$
36: **fim se**
37: **fim se**
38: incremento $\leftarrow$ I primeiros exemplos teste classificados incorretamente
39: janela $\leftarrow$ janela + incremento
40: **até** (errosTeste $= 0$)
41: **se** (**não** Podar árvores intermediárias **e** Podar árvore final) **então**
42: melhorClassificador $\leftarrow$ podar(melhorClassificador)
43: **fim se**
44: **retorne** melhorClassificador

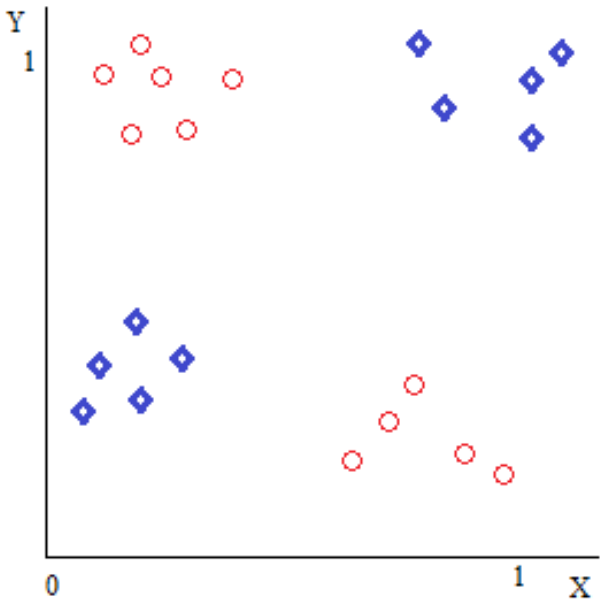


Figura 4.1: *XOR contínuo. O círculo representa a classe Verdadeiro e o losango representa a classe Falso.*

ou mais partições.

Pode-se notar que, no Algoritmo 4, deve-se calcular o valor do critério para todos os possíveis particionamentos de todos os atributos disponíveis (Linhas 7 e 8). Além disto, o algoritmo "abre" as possibilidades de particionamento dos atributos conforme a profundidade definida pelo parâmetro P. Quando este se torna 0, o algoritmo simplesmente calcula, para cada particionamento possível de cada atributo disponível, o valor do critério utilizado. Notadamente, o parâmetro P é decrementado de 1 a cada chamada recursiva feita (Linha 13). A diferença entre as funções *calcularCriterioParticao* (Linha 15) e *calcularCriterio* (Linha 19) é que a primeira calcula o critério (*e.g.*, alguma função baseada em entropia) para uma dada partição e a segunda calcula o critério para um particionamento (conjunto de partições resultantes de um teste em um atributo).

4.4 Considerações Finais

Para verificar o desempenho das duas técnicas descritas neste capítulo em dados de expressão gênica utilizando as medidas propostas no capítulo anterior, foram realizados experimentos com uma série de bases de dados. Os resultados são apresentados no próximo capítulo.

Algoritmo 3 Indução de árvore utilizando lookahead.

Pré-condição: E: um conjunto de N exemplos rotulados $\{(\vec{x}_i, y_i), i = 1, 2, \ldots, N\}$
P: a profundidade a ser considerada no *lookahead*, $P \in \mathbb{N}$
C: a raiz da (sub)árvore sendo construída
algoritmo induzirArvore(E, P, C**)** {
1: A: o conjunto de atributos ainda disponíveis
2: a^*: um atributo, com ou sem pontos de particionamento definidos
3: a: um atributo
4: S: um particionamento (conjunto de partições criadas por um teste)
5: Q: uma partição
6: criterioAtributos: vetor com o valor, para cada atributo disponível, do critério utilizado
7: **se** (Condição de parada atingida) **então**
8: criarFolha(C, E)
9: **senão**
10: criterioAtributos ← lookahead(E, P, A)
11: a^* ← oti$_{a \in A}$(criterioAtributos[a])
12: criarTeste(C, E, a^*)
13: S ← particionarExemplos(E, a^*)
14: **para todo** ($Q \in S$) **faça**
15: induzirArvore(Q, P, C);
16: **fim para**
17: **fim se**
}

Algoritmo 4 Lookahead.

Pré-condição: E: um conjunto de N exemplos rotulados $\{(\vec{x}_i, y_i), i = 1, 2, \ldots, N\}$
P: a profundidade a ser considerada no *lookahead*, $P \in \mathbb{N}$
A: o conjunto de atributos ainda disponíveis
Pós-condição: criterioAtributos: vetor com o valor, para cada atributo disponível, do critério utilizado
algoritmo lookahead(E, P, A**)** {
1: a: um atributo
2: S: um particionamento (conjunto de partições criadas)
3: Q: uma partição
4: valorCriterio: um valor calculado para o critério
5: criterioParticoes: vetor com valores do critério utilizado
6: melhorValor: o valor do melhor particionamento de um atributo
7: **para todo** ($a \in$ A) **faça**
8: **para todo** (Particionamento possível de a) **faça**
9: S ← particionamento atual
10: criterioParticoes ← {}
11: **para todo** ($Q \in S$) **faça**
12: **se** ($P \neq 0$) **então**
13: criterioParticoes ← criterioParticoes + lookahead($Q, P - 1, A$)
14: **senão**
15: valorCriterio ← calcularCriterioParticao(Q)
16: criterioParticoes ← criterioParticoes + valorCriterio
17: **fim se**
18: **fim para**
19: valorCriterio ← calcularCriterio(criterioParticoes)
20: **se** (valorCriterio melhor que melhorValor) **então**
21: melhorValor ← valorCriterio
22: **fim se**
23: **fim para**
24: criterioAtributos[a] ← melhorValor
25: **fim para**
26: **retorne** criterioAtributos
}

Capítulo

5

Resultados e Discussão

5.1 Considerações Iniciais

Neste capítulo, os resultados obtidos nos experimentos realizados com as duas técnicas descritas no capítulo anterior são apresentados e discutidos.

5.2 Windowing

Para verificar as alterações propostas, foi projetado um experimento em que se usou o algoritmo J48 e diferentes versões do *windowing*. Tais versões foram resultantes da exploração de todas as combinações das seguintes opções que o algoritmo passou a oferecer: usar ou não o critério de confiança, usar ou não a ponderação do erro, usar ou não a estimação do erro de generalização no lugar do erro da janela e utilizar ou não a poda das árvores intermediárias. Todas as outras opções do algoritmo foram mantidas inalteradas. Foram produzidas, então, dezesseis configurações diferentes. Somando-as ao J48, dezessete diferentes indutores participaram do experimento. O indutor de AD base utilizado com o *windowing* foi o J48.

Na realidade, foi implementada uma versão alterada do J48, por dois motivos: para que o modelo fornecesse as medidas de desempenho propostas neste mestrado; para que houvesse a possibilidade de construção de árvores não podadas, mas que pudessem ser podadas posteriormente sem a necessidade de construí-las do início. No que diz respeito às outras características, esta implementação é idêntica ao J48 original. Em todos os casos, a configuração padrão de parâmetros deste algoritmo foi utilizada.

Foram utilizadas quarenta e uma bases reais de dados de expressão gênica, tanto oriundas de análises por *microarray* quanto por SAGE. A descrição das bases se encontra no Apêndice B.

Para cada indutor aplicado a cada base, foi utilizada validação cruzada com dez partições (Kohavi 1995). As medidas levadas em consideração foram: acurácia, AUC, altura da árvore, tamanho da árvore, tamanho da janela usada na construção do melhor classificador encontrado, coesão, compactação, relação entre coesão e compactação e tempo gasto no treinamento (esta medida foi usada por uma razão específica, a ser detalhada na Seção 5.2.4).

Para verificar se as diferenças encontradas eram significativas, foi utilizado o teste de Friedman (1940), uma técnica não-paramétrica baseada em *ranks* e largamente utilizada pela comunidade de AM. Como teste *post-hoc*, foi escolhido Benjamini & Hochberg (1995) (vide Apêndice D para maiores detalhes). Os testes executados consideraram todas as possibilidade de pares de indutores, ou seja, comparação múltipla sem a presença de um indutor controle.

As análises apresentadas neste capítulo consideram um nível de significância de 5%. Os resultados do experimento são apresentados abaixo (os valores originais das variáveis consideradas podem ser vistos no Apêndice C). Na Tabela 5.1, é apresentada a correspondência entre as diferentes configurações do *windowing* e a notação adotada para se referir a elas.

Tabela 5.1: *Notação para se referir às diferentes configurações do windowing. A presença da opção C indica a utilização do critério de confiança; a presença da opção E indica a utilização da estimação do erro da janela; a presença da opção We indica a utilização da ponderação do erro total; e a presença da opção P indica a utilização da poda das árvores intermediárias. A versão sem nenhuma das opções corresponde, na verdade, à versão original do algoritmo.*

Versão Windowing	Representação
poda, erro estimado, ponderação do erro e confiança	WPEWeC
poda, erro estimado e confiança	WPEC
erro estimado, ponderação do erro e confiança	WEWeC
erro estimado e confiança	WEC
poda, ponderação do erro e confiança	WPWeC
poda e confiança	WPC
ponderação do erro e confiança	WWeC
confiança	WC
poda, erro estimado e ponderação do erro	WPEWe
poda e erro estimado	WPE
erro estimado e ponderação do erro	WEWe
erro estimado	WE
poda e ponderação do erro	WPWe
poda	WP
ponderação do erro	WWe
original	W

5.2.1 Altura da Árvore

Considerando a interpretabilidade do modelo, quanto menor a altura da árvore, mais fácil pode-se compreendê-la sintaticamente (é claro que o fato de ser menor não faz com que ela seja necessariamente melhor, ou mesmo que esteja de alguma forma de acordo com o conceito sendo aprendido.) Na comparação realizada, o J48 teve um dos piores *ranks* e foi considerado como tendo produzido árvores significativamente mais altas que diversas versões de *windowing* testadas. O p-valor da diferença entre o J48 e o WEWeC, por exemplo, foi da ordem de 10^{-8}. Na Tabela 5.2, é mostrada a comparação entre os algoritmos.

Neste caso, seis das versões de *windowing* envolvendo a alteração relativa à confiança na classificação ocuparam as seis primeiras posições, sendo que as quatro primeiras não apresentaram diferenças significativas entre si (WEWeC, WEC, WWeC e WC, nesta ordem). A última colocada foi a WP, mas que não apresentou diferença significativa com relação ao J48 e à versão original do *windowing*. Na Tabela 5.3, é mostrado o *ranking* completo do teste.

Pode-se notar que, neste caso, as alterações propostas contribuíram para a melhora da interpretabilidade sintática do modelo, produzindo, em geral, árvores menores. Este tipo de contribuição pode ser muito importante em estudos de expressão gênica, em que se deve minerar, entre milhares de genes, somente aqueles realmente associados ao estudo em questão.

No caso da altura, foram comparados também os valores de desvio padrão dos resultados obtidos. Não foi encontrada nenhuma diferença significativa, sendo que o p-valor do teste de Friedman foi praticamente um.

Tabela 5.2: *Comparação todos contra todos por meio do teste de Friedman e post-hoc para a variável Altura da Árvore. Só é mostrado o triângulo superior direito da matriz, pois ela é simétrica. Um ○ em uma célula indica que não houve diferença alguma entre o indutor da respectiva linha e o indutor da respectiva coluna; △ (▽) indica que o indutor da linha foi melhor (pior) que o da coluna, mas não significativamente; ▲ (▼) indica que o indutor da linha foi significativamente melhor (pior) que o da coluna.*

	J48	**WPEWeC**	**WPEC**	**WEWeC**	**WEC**	**WPWeC**	**WPC**	**WWeC**	**WC**	**WPEWe**	**WPE**	**WEWe**	**WE**	**WPWe**	**WP**	**WWe**	**W**
J48	○	▼	▼	▼	▼	△	△	▼	▼	▽	▽	▽	▽	△	△	▽	▽
WPEWeC		○	△	▼	▼	▲	▲	▽	▽	△	△	△	△	▲	▲	▲	▲
WPEC			○	▼	▼	▲	▲	▼	▼	△	△	△	△	▲	▲	△	△
WEWeC				○	△	▲	▲	△	△	▲	▲	▲	▲	▲	▲	▲	▲
WEC					○	▲	▲	△	△	▲	▲	▲	▲	▲	▲	▲	▲
WPWeC						○	△	▼	▼	▽	▽	▽	▽	△	△	▽	▽
WPC							○	▼	▼	▼	▽	▼	▽	△	△	▽	▽
WWeC								○	△	▲	▲	▲	▲	▲	▲	▲	▲
WC									○	▲	▲	▲	▲	▲	▲	▲	▲
WPEWe										○	△	▽	△	▲	▲	△	△
WPE											○	▽	▽	△	△	△	△
WEWe												○	△	▲	▲	△	△
WE													○	▲	▲	△	△
WPWe														○	△	▽	▽
WP															○	▽	▽
WWe																○	○
W																	○

Tabela 5.3: *Ranking dos indutores com relação à variável Altura da Árvore. Quanto menor o rank de um indutor, mais bem colocado ele está, sendo que a lista já está ordenada de forma crescente.*

Algoritmo	Rank Médio
WEWeC	4,366
WEC	5,073
WWeC	5,659
WC	5,744
WPEWeC	7,732
WPEC	8,476
WEWe	8,683
WPEWe	8,902
WE	9,195
WPE	9,939
WWe	10,854
W	10,854
J48	11,061
WPWeC	11,085
WPC	11,573
WPWe	11,829
WP	11,976

5.2.2 Tamanho da Árvore

Os resultados obtidos neste caso foram praticamente idênticos aos obtidos para a altura da árvore, sendo que as únicas alterações foram na ordem dos *ranks* de algoritmos que não apresentaram diferenças significativas. Este resultado reitera a contribuição das alterações propostas para a melhora da interpretabilidade do modelo induzido.

Na Figura 5.1 podem ser visualizados os gráficos de caixa para a medida *Tamanho da Árvore*. O gráfico no topo da figura mostra o *boxplot* dos valores da variável propriamente dita. Pode-se perceber que as quatro caixas que ocupam as posições mais baixas do gráfico representam os quatro indutores com os menores *ranks* médios, conforme pode ser visualizado no *boxplot* na base da figura. É possível perceber também que as diferenças na variável em si são menores que as diferenças nos *ranks* médios dos indutores. Esta situação é relativamente comum em análises não-paramétricas: mesmo que não haja grandes diferenças nas observações da variável sendo analisada, se um ou mais indutores apresentarem valores sistematicamente maiores ou menores daquela variável, haverá grande chance de estes indutores serem considerados significativamente diferentes dos demais.

A comparação dos desvios padrão também não reconheceu nenhuma diferença significativa, assim como no caso da altura da árvore.

5.2.3 Tamanho da Janela

Nesta comparação, o J48 foi excluído das análises por dois motivos: 1) o algoritmo não trabalha com o conceito de janela, já que sempre utiliza todo o conjunto de exemplos no treinamento; e 2) é muito comum o *windowing* terminar o treinamento antes que tenha adicionado todos os exemplos à janela. Mesmo que fosse considerada uma "janela" para o J48 (poderia ser, por exemplo, sempre o tamanho total da base dados), ele seria considerado significativamente pior na grande maioria das vezes.

Novamente as alterações propostas surtiram efeito, tendo a versão original do *windowing* ocupado as últimas posições. As versões WEWeC, WEC, WWeC e WC, nesta ordem, foram, assim como no caso da altura e do tamanho, as quatro melhores, mas não apresentaram diferenças significativas entre si. Mais uma vez a combinação entre confiança, erro estimado e ponderação do

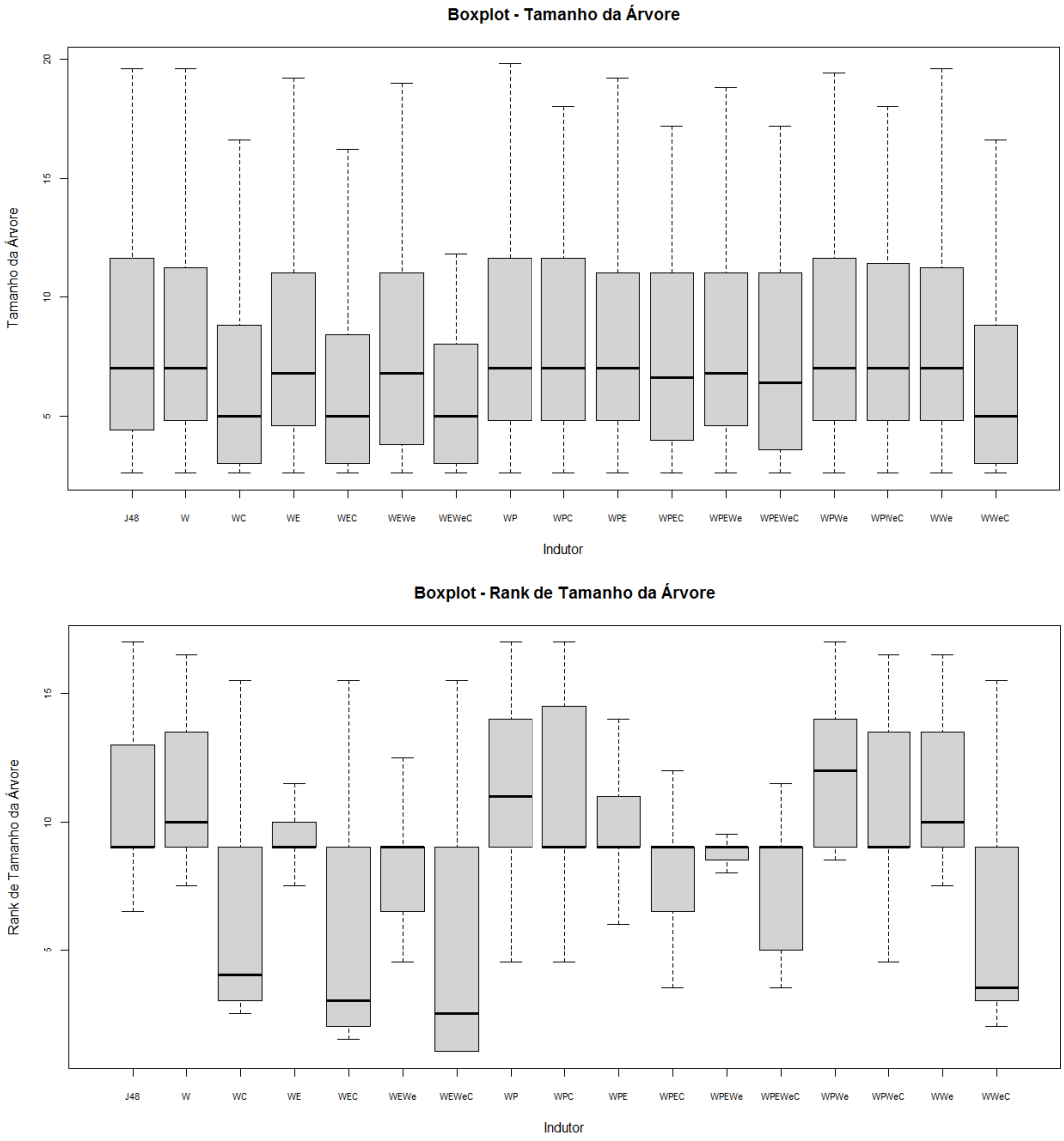


Figura 5.1: *Boxplot para a medida Tamanho da Árvore. No topo, pode ser visualizado o gráfico para os valores da variável propriamente dita. Abaixo, é mostrado o boxplot dos ranks obtidos pelos indutores para as diversas bases de dados quanto à medida em questão.*

erro produziu o melhor resultado, ou seja, produziu modelos mais interpretáveis a um preço menor, utilizando menos exemplos para treinar os classificadores. Na Tabela 5.5, é mostrado o *ranking completo do teste*.

Na Tabela 5.4, é mostrada a comparação entre todos os algoritmos. Pode-se notar que existiram muitos pares que apresentaram diferença significativa, mas o padrão dessas diferenças foi muito próximo daquele apresentado no caso da comparação da altura e tamanho do modelo.

Tabela 5.4: *Comparação todos contra todos por meio do teste de Friedman e post-hoc para a variável* Tamanho da Janela. *Só é mostrado o triângulo superior direito da matriz, pois ela é simétrica. Um ∘ em uma célula indica que não houve diferença alguma entre o indutor da respectiva linha e o indutor da respectiva coluna; △ (▽) indica que o indutor da linha foi melhor (pior) que o da coluna, mas não significativamente; ▲ (▼) indica que o indutor da linha foi significativamente melhor (pior) que o da coluna.*

	WPEWeC	**WPEC**	**WEWeC**	**WEC**	**WPWeC**	**WPC**	**WWeC**	**WC**	**WPEWe**	**WPE**	**WEWe**	**WE**	**WPWe**	**WP**	**WWe**	**W**
WPEWeC	∘	△	▼	▼	▲	▲	▼	▼	▲	▲	▲	▲	▲	▲	▲	▲
WPEC		∘	▼	▼	△	▲	▼	▼	△	▲	△	▲	▲	▲	▲	▲
WEWeC			∘	△	▲	▲	△	△	▲	▲	▲	▲	▲	▲	▲	▲
WEC				∘	▲	▲	△	△	▲	▲	▲	▲	▲	▲	▲	▲
WPWeC					∘	△	▼	▼	▽	▲	▽	△	▲	▲	△	△
WPC						∘	▼	▼	▼	△	▼	▽	△	▲	▽	▽
WWeC							∘	△	▲	▲	▲	▲	▲	▲	▲	▲
WC								∘	▲	▲	▲	▲	▲	▲	▲	▲
WPEWe									∘	▲	△	▲	▲	▲	▲	▲
WPE										∘	▼	▽	△	▲	▽	▽
WEWe											∘	△	▲	▲	△	△
WE												∘	▲	▲	△	△
WPWe													∘	△	▼	▼
WP														∘	▼	▼
WWe															∘	∘
W																∘

Tabela 5.5: *Ranking dos indutores com relação à variável Tamanho da Janela. Quanto menor o rank de um indutor, mais bem colocado ele está, sendo que a lista já está ordenada de forma crescente.*

Algoritmo	Rank Médio
WEWeC	1,902
WEC	3,037
WWeC	3,354
WC	3,549
WPEWeC	5,878
WPEC	7,695
WPEWe	8,220
WEWe	8,890
WPWeC	9,476
WE	10,634
WWe	11,037
W	11,037
WPC	11,171
WPE	11,732
WPWe	13,293
WP	15,098

5.2.4 Tempo de Treinamento

Outra variável avaliada foi o tempo de CPU[1] gasto no treinamento. Mais uma vez, o J48 foi excluído da análise. Apesar de o *windowing* diminuir o tempo de treinamento em casos específicos, no caso de bases de dados de expressão gênica, a técnica gasta sistematicamente muito mais tempo para construir o classificador. Nos testes feitos incluindo o J48, este foi significativamente certamente mais rápido que todas as versões do *windowing* (em torno de quarenta vezes mais rápido), tendo a sua média ficado abaixo dos dois segundos, enquanto as versões do *windowing* apresentaram uma média de mais de oitenta segundos.

Esta análise tinha o objetivo de verificar se alguma das versões propostas conseguiria tornar o algoritmo do *windowing* mais eficiente computacionalmente. Como se pode perceber na Tabela 5.6, o objetivo foi alcançado. A expectativa de diminuir o tempo de computação se devia principalmente ao fato de a alteração baseada na confiança na classificação fazer com que o algoritmo interrompa um determinado *trial* precocemente caso não perceba progresso ao adicionar exemplos à janela. Isto realmente ocorreu, sendo que quase todas as versões envolvendo a alteração relacionada à confiança ficaram nas primeiras posições, como pode ser visto na Tabela 5.7.

No entanto, a partir da Tabela 5.7, pode-se perceber algo que, à princípio, pode parecer incoerente: as primeiras posições são ocupadas por versões envolvendo a alteração relacionada à poda das árvores intermediárias. Porém, ainda que o processo de poda geralmente torne computacionalmente mais dispendiosa a construção das árvores, a poda acima é importante para o sucesso da atualização da janela baseada na confiança. Como citado anteriormente, a confiança de uma AD é calculada a partir da distribuição de classes do nó folha que um dado exemplo atingiu. Se a poda não for usada, a AD tende a decorar os dados, produzindo distribuições de classe com valores tendendo a 0 ou 1. Com a poda, este efeito é minimizado e a alteração relacionada à confiança se torna mais efetiva.

A versão original (W) teve um dos piores desempenhos, sendo considerada significativamente mais lenta do que muitas das versões comparadas (notavelmente, WPEWeC, WPC, WPEC e

[1] O tempo de CPU é a quantidade de tempo gasta pela CPU (*Central Processing Unit* ou Unidade Central de Processamento) para efetivamente processar instruções, ou seja, não leva em consideração, por exemplo, o tempo gasto em espera de operações de entrada e saída.

WPWeC). Algumas das versões consideradas boas nas medidas anteriores também obtiveram bom desempenho no quesito tempo, *i.e.*, produziram árvores mais interpretáveis em um tempo menor.

Tabela 5.6: *Comparação todos contra todos por meio do teste de Friedman e post-hoc para a variável Tempo de Treinamento. Só é mostrado o triângulo superior direito da matriz, pois ela é simétrica. Um ∘ em uma célula indica que não houve diferença alguma entre o indutor da respectiva linha e o indutor da respectiva coluna; △ (▽) indica que o indutor da linha foi melhor (pior) que o da coluna, mas não significativamente; ▲ (▼) indica que o indutor da linha foi significativamente melhor (pior) que o da coluna.*

	WPEWeC	**WPEC**	**WEWeC**	**WEC**	**WPWeC**	**WPC**	**WWeC**	**WC**	**WPEWe**	**WPE**	**WEWe**	**WE**	**WPWe**	**WP**	**WWe**	**W**
WPEWeC	∘	△	△	▲	△	△	▲	▲	△	▲	▲	▲	△	▲	▲	▲
WPEC		∘	△	△	△	▽	▲	▲	△	△	▲	▲	△	△	▲	▲
WEWeC			∘	△	▽	▽	△	△	▽	△	▲	▲	∘	△	▲	▲
WEC				∘	▽	▽	△	△	▽	▽	△	△	▽	▽	▲	△
WPWeC					∘	▽	△	▲	△	△	▲	▲	△	△	▲	▲
WPC						∘	▲	▲	△	△	▲	▲	△	△	▲	▲
WWeC							∘	△	▽	▽	△	△	▽	▽	▲	△
WC								∘	▽	▽	△	△	▽	▽	△	△
WPEWe									∘	△	▲	▲	△	△	▲	▲
WPE										∘	△	△	▽	▽	▲	▲
WEWe											∘	▽	▼	▼	△	△
WE												∘	▼	▼	△	△
WPWe													∘	△	▲	▲
WP														∘	▲	▲
WWe															∘	▽
W																∘

Tabela 5.7: *Ranking dos indutores com relação à variável Tempo de Treinamento. Quanto menor o rank de um indutor, mais bem colocado ele está, sendo que a lista já está ordenada de forma crescente.*

Algoritmo	Rank Médio
WPEWeC	5,524
WPC	6,439
WPEC	6,598
WPWeC	6,744
WPEWe	7,549
WEWeC	7,854
WPWe	7,854
WP	8,122
WPE	8,463
WEC	8,707
WWeC	9,146
WC	9,220
WE	10,585
WEWe	10,634
W	10,927
WWe	11,634

5.2.5 Coesão e Compactação

No caso da coesão, mais uma vez o padrão encontrado na análise do tamanho das árvores se repetiu, tanto com relação às posições do *ranking* quanto às diferenças significativas encontradas. Uma diferença interessante de se notar foi com relação ao desvio padrão da coesão, em que foram encontradas diferenças significativas. As versões que obtiveram melhor resultado nos valores de coesão propriamente ditos foram as que apresentaram maior desvio padrão, *i.e.*, maior instabilidade neste quesito (Tabela 5.8). Na verdade, ainda que o teste de Friedman tenha identificado diferenças significativas (p-valor de $3,91 \times 10^{-3}$), o teste *post-hoc* não identificou em quais pares elas ocorreram. Se o teste for afrouxado para um nível de significância de 10%, percebe-se que tais diferenças estão nos pares WEC / WPWe, WEWeC / WPWe, WEC / WP e WEWeC / WP.

Os resultados obtidos para as variáveis compactação e coesão-compactação foram coerentes com aqueles obtidos para a coesão. As versões propostas (notavelmente, WEWeC, WEC, WWeC e WC) obtiveram desempenho muito bom, sendo consideradas significativamente mais compactas que o J48 e a versão original do *windowing*. Isto pode indicar que as melhores versões conseguiram extrair o conhecimento relevante dos estudos de expressão gênica considerados, já que foram coesos (no sentido de apresentar menos nós folha por classe existente) e compactos (no sentido de terem encontrado os genes relevantes). É claro que uma análise de um especialista é essencial para realmente poder formular tal conclusão.

Na Figura 5.2 podem ser visualizados os gráficos de caixa para a medida *Compactação*. O gráfico no topo da figura mostra o *boxplot* dos valores da variável propriamente dita e o gráfico na base mostra o *boxplot* dos *ranks* obtidos pelos indutores. Neste último caso, fica evidente a superioridade dos quatro indutores considerados melhores nos testes.

Ao contrário do caso da coesão, não foram encontradas diferenças significativas no desvio padrão da compactação. Já no caso da coesão-compactação, a análise do desvio padrão se mostrou praticamente idêntica ao desvio padrão da coesão, *i.e.*, as melhores versões na variável em si se mostraram mais instáveis.

Tabela 5.8: *Ranking dos indutores com relação ao desvio padrão da variável Coesão. Quanto menor o rank de um indutor, mais bem colocado ele está, sendo que a lista já está ordenada de forma crescente.*

Algoritmo	Rank Médio
WPWe	7,329
WP	7,439
WPE	7,939
WWe	8,171
W	8,171
WPWeC	8,195
WPC	8,244
J48	8,488
WPEWe	8,793
WE	8,841
WPEC	9,305
WEWe	9,512
WPEWeC	10,000
WC	10,463
WWeC	10,549
WEWeC	10,768
WEC	10,793

5.2.6 Acurácia e AUC

No caso do *AUC*, nenhuma diferença significativa foi encontrada, tendo o teste de Friedman calculado um p-valor igual a $0,29$, bem como não houve diferenças no desvio padrão calculado para o *AUC*.

Já no caso da acurácia, diferenças significativas foram identificadas (Tabela 5.9). No entanto, as versões que haviam sido consideradas melhores nas variáveis relacionadas à interpretabilidade dos modelos induzidos obtiveram desempenho significativamente pior no caso da acurácia, inclusive com relação ao J48 e a versão original do *windowing*. No entanto, as primeiras posições do *ranking* estão ocupadas por algumas das versões propostas, ainda que não tenham superado significativamente o J48 e o *windowing* original (Tabela 5.10). A análise do desvio padrão da acurácia também mostrou diferenças significativas, apesar de elas terem ficado apenas entre os dois primeiros e os dois últimos colocados do *ranking* (Tabela 5.11). Ainda assim, novamente as versões que produziram modelos mais interpretáveis ocuparam as últimas posições.

Na Figura 5.3 podem ser visualizados os gráficos de caixa para a medida *Acurácia*. O gráfico na base da figura mostra o *boxplot* dos *ranks* obtidos pelos indutores nas diversas bases de dados. Mais uma vez é possível verificar que os indutores considerados superiores nas medidas de complexidade foram inferiores quanto à acurácia.

Na Tabela 5.12 pode ser visualizado um resumo gráfico dos resultados do experimento com *windowing*. Cada célula da tabela representa o desempenho, em termos do *rank* médio, do indutor da coluna na medida indicada na linha. Tal desempenho é definido por um valor de nível de cinza: preto, quando o *rank* médio do indutor na respectiva medida ficou abaixo de 5,2; cinza escuro, quando ficou entre 5,2 e 8,5; cinza claro, quando ficou entre 8,5 e 11,8; e cinza ainda mais claro, quando ficou acima de 11,8. Quanto mais escuro, menor o rank médio, ou seja, melhor foi o indutor naquela medida específica. Nem sempre as diferenças em nível de cinza significam diferenças estatisticamente significativas. Como se pode notar, formaram-se grupos de indutores com desempenho semelhante: os indutores das quatro primeiras colunas foram muito próximos em todas as medidas, tendo apresentado melhores resultados para acurácia e resultados mais pobres para as medidas

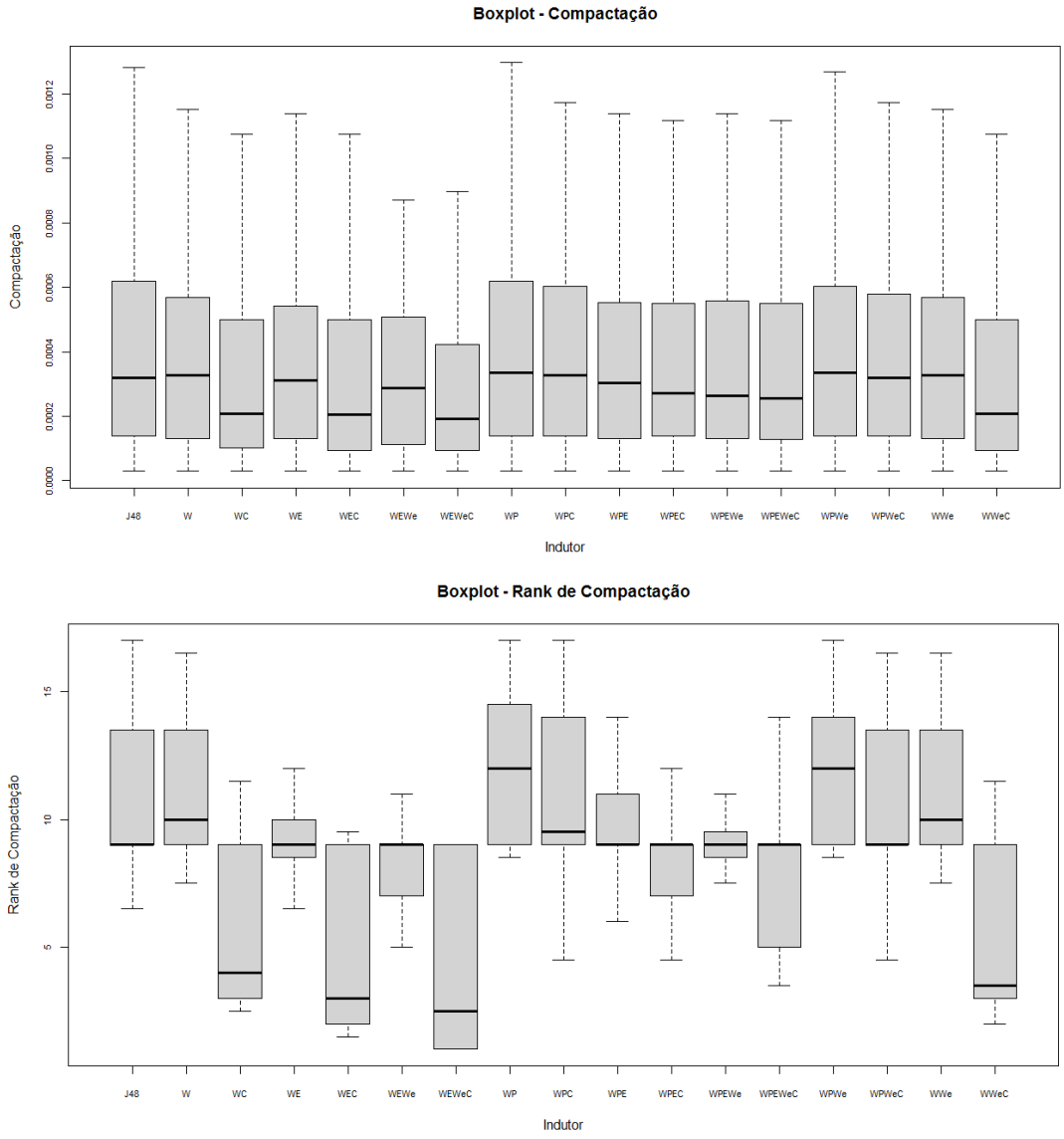


Figura 5.2: *Boxplot para a medida Compactação. No topo, pode ser visualizado o gráfico para os valores da variável propriamente dita. Abaixo, é mostrado o boxplot dos ranks obtidos pelos indutores para as diversas bases de dados quanto à medida em questão.*

relacionadas à complexidade sintática dos modelos criados; da quinta à decima primeira colunas, os indutores foram claramente melhores nas medidas relacionadas à complexidade dos modelos criados, sendo que quase todos eles apresentaram a alteração referente à confiança na classificação. O restante das colunas representa indutores com desempenho cada vez mais fraco quanto às medidas de complexidade, mas com desempenho razoável quanto à acurácia e AUC.

Tabela 5.9: *Comparação todos contra todos por meio do teste de Friedman e post-hoc para a variável Acurácia. Só é mostrado o triângulo superior direito da matriz, pois ela é simétrica. Um ∘ em uma célula indica que não houve diferença alguma entre o indutor da respectiva linha e o indutor da respectiva coluna; △ (▽) indica que o indutor da linha foi melhor (pior) que o da coluna, mas não significativamente; ▲ (▼) indica que o indutor da linha foi significativamente melhor (pior) que o da coluna.*

	J48	**WPEWeC**	**WPEC**	**WEWeC**	**WEC**	**WPWeC**	**WPC**	**WWeC**	**WC**	**WPEWe**	**WPE**	**WEWe**	**WE**	**WPWe**	**WP**	**WWe**	**W**
J48	∘	△	△	▲	△	△	△	△	△	▽	▽	△	▽	△	▽	▽	▽
WPEWeC		∘	△	△	△	△	▽	△	△	▽	▽	▽	▽	▽	▽	▽	▽
WPEC			∘	△	△	△	▽	△	△	▽	▽	▽	▽	▽	▽	▽	▽
WEWeC				∘	▽	▽	▽	▽	▽	▼	▼	▽	▼	▽	▼	▼	▼
WEC					∘	▽	▽	▽	▽	▼	▼	▽	▼	▽	▼	▽	▽
WPWeC						∘	▽	△	△	▽	▽	▽	▽	▽	▽	▽	▽
WPC							∘	△	△	▽	▽	△	▽	▽	▽	▽	▽
WWeC								∘	△	▼	▽	▽	▽	▽	▽	▽	▽
WC									∘	▼	▽	▽	▽	▽	▽	▽	▽
WPEWe										∘	△	△	△	△	△	△	△
WPE											∘	△	∘	△	▽	△	△
WEWe												∘	▽	▽	▽	▽	▽
WE													∘	△	▽	△	△
WPWe														∘	▽	▽	▽
WP															∘	△	△
WWe																∘	∘
W																	∘

Tabela 5.10: *Ranking dos indutores com relação à variável Acurácia. Quanto menor o rank de um indutor, mais bem colocado ele está, sendo que a lista já está ordenada de forma crescente.*

Algoritmo	Rank Médio
WPEWe	7,390
WP	7,890
WPE	8,073
WE	8,073
WWe	8,244
W	8,244
J48	8,317
WPWe	8,402
WPC	8,439
WEWe	8,524
WPEWeC	8,915
WPEC	9,085
WPWeC	9,268
WWeC	10,659
WC	10,671
WEC	11,293
WEWeC	11,512

Tabela 5.11: *Ranking dos indutores com relação ao desvio padrão da variável Acurácia. Quanto menor o rank de um indutor, mais bem colocado ele está, sendo que a lista já está ordenada de forma crescente.*

Algoritmo	Rank Médio
WPEWeC	7,439
WPE	7,488
WPEC	8,085
WEWe	8,354
WWe	8,378
W	8,378
WPEWe	8,415
WP	8,537
WE	8,646
WPWe	8,793
WPC	8,829
J48	8,878
WPWeC	8,976
WEWeC	10,659
WWeC	10,829
WC	11,049
WEC	11,268

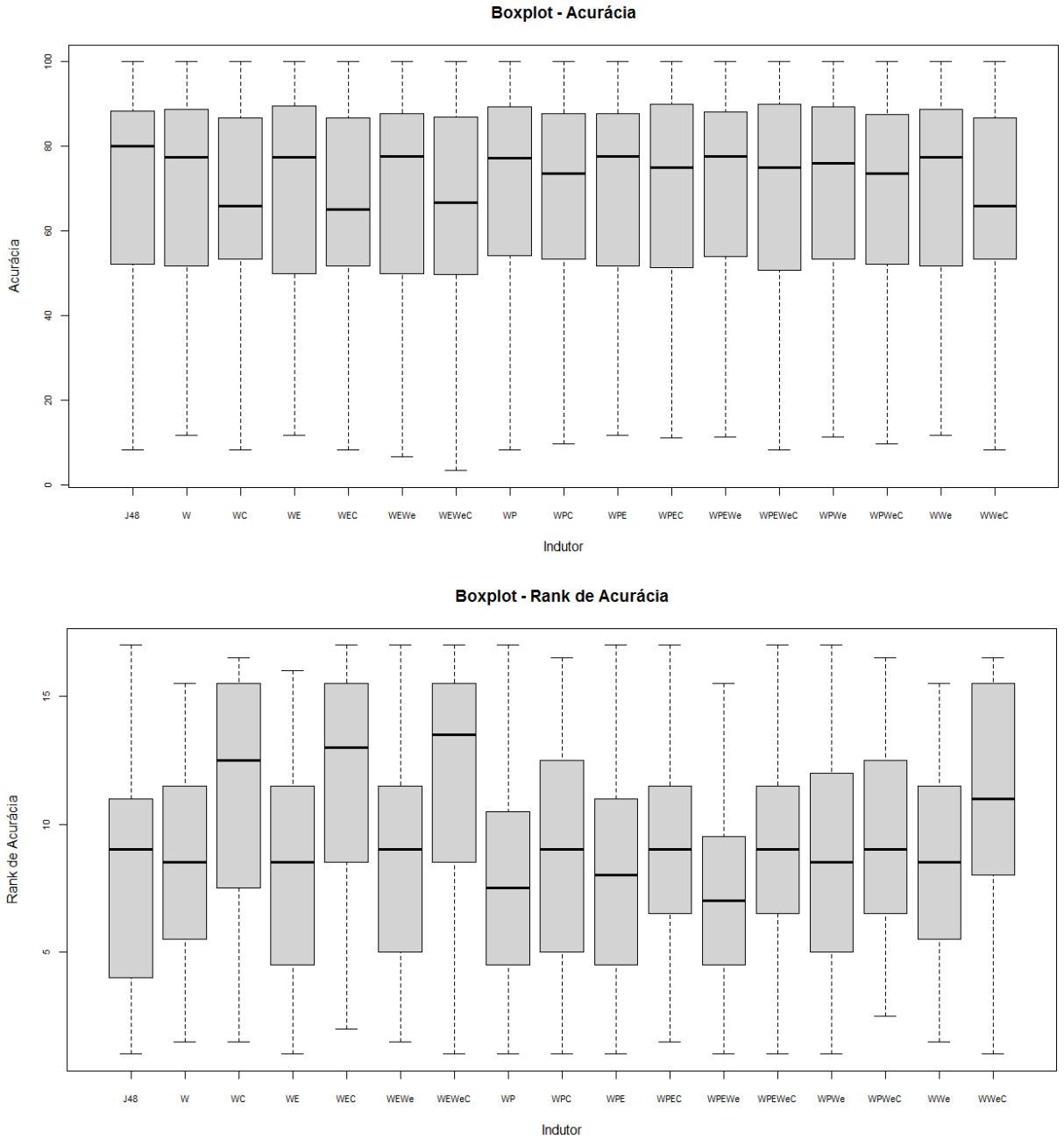
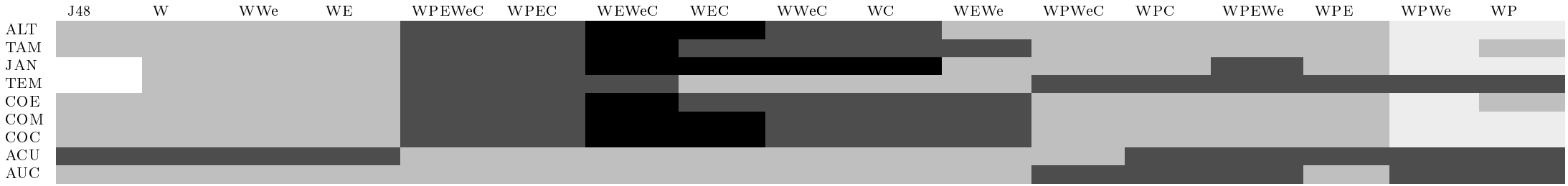


Figura 5.3: *Boxplot para a medida Acurácia. No topo, pode ser visualizado o gráfico para os valores da variável propriamente dita. Abaixo, é mostrado o boxplot dos ranks obtidos pelos indutores para as diversas bases de dados quanto à medida em questão.*

Tabela 5.12: *Resumo dos resultados do experimento com* windowing*: as colunas representam cada indutor testado e as linhas representam as medidas analisadas. Cada célula da tabela representa o desempenho, em termos do* rank *médio, do indutor da coluna na medida da linha. Tal desempenho é definido por um valor de nível de cinza: preto, quando o* rank *médio do indutor na respectiva medida ficou abaixo de 5,2; cinza escuro, quando ficou entre 5,2 e 8,5; cinza claro, quando ficou entre 8,5 e 11,8; e cinza ainda mais claro, quando ficou acima de 11,8. Quanto mais escuro, menor o rank médio, ou seja, melhor foi o indutor naquela medida específica. Quando a célula está totalmente branca, o resultado não se aplica. É o caso do* J48 *nas medidas Tempo de Treinamento e Tamanho da Janela. ALT representa a medida Altura da Árvore; TAM representa a medida Tamanho da Árvore; JAN representa a medida Tamanho da Janela; TEM representa a medida Tempo de Treinamento; COE representa a medida Coesão; COM representa a medida Compactação; COC representa a medida Coesão-Compactação; ACU representa a medida Acurácia; AUC representa a área sob a curva ROC.*



5.3 Lookahead

Experimentos também foram projetados para que fosse possível verificar o resultado da aplicação de *lookahead* de um nível em dados de expressão gênica. Vinte e uma bases de dados foram utilizadas, todas provenientes de dados reais de perfis de expressão gênica e descritas no Apêndice B. O *lookahead* foi utilizado com o indutor J48 e comparado à aplicação de J48 puro nas mesmas bases.

Para a execução dos experimentos, foi implementada a técnica de *lookahead* junto ao J48 da Weka. A análise estatística foi baseada no teste de Wilcoxon (1945) – ver Apêndice D para detalhes sobre o teste. As medidas comparadas foram a área sob a curva ROC e o tamanho das árvores finais.

Os resultados descritos a seguir consideram um nível de significância de 5%. Com relação ao *AUC*, não houve diferenças significativas, nem na medida em si nem em seu desvio padrão, ou seja, *lookahead* de um nível se mostrou estatisticamente equivalente ao J48.

Já com relação ao tamanho das árvores induzidas, o J48 puro produziu árvores significativamente menores, ou seja, *lookahead* de um nível produziu modelos supostamente mais complexos, o contrário do que normalmente se busca. O desvio padrão do tamanho da árvore não se mostrou estatisticamente diferente entre os dois indutores comparados.

Algo já esperado e que foi realmente constatado foi o custo computacional alto do *lookahead*. Tal custo é calculado sobre o número de atributos e exemplos do problema em questão. No caso de dados de expressão gênica, há poucos exemplos, mas muitíssimos atributos, fazendo com que a técnica, ainda que de apenas um nível, se torne impraticável já com um pequeno número de exemplos. Das quarenta e uma bases de dados utilizadas nos experimentos com *windowing*, apenas vinte e uma tiveram seu resultado obtido para o *lookahead*. Houve casos em que, após uma semana de processamento, não havia sido obtido o resultado de uma partição sequer da validação cruzada.

Outro aspecto que interfere na eficiência computacional do *lookahead* é o número de valores diferentes existentes para um dado atributo numérico. Quanto maior este número, mais possibilidades há para explorar. Mais uma vez, dados de expressão gênica levam desvantagem, já que praticamente só possuem atributos numéricos, cujos valores raramente se repetem.

Murthy & Salzberg (1995) descrevem um estudo do comportamento do *lookahead* com AD. Os seus experimentos foram feitos com dados artificiais e consideraram apenas dois atributos. Os autores concluíram que a técnica produziu árvores de altura levemente menor, mas de mesma acurácia e tamanho. Eles classificaram o comportamento do *lookahead* como patológico, já que, na mesma medida com que produziu árvores melhores em alguns casos, produziu árvores piores em outros.

5.4 Considerações Finais

Bons resultados foram obtidos com o *windowing*, principalmente no que se diz respeito à interpretabilidade dos modelos construídos sob o ponto de vista sintático. Já com relação ao *lookahead*, os resultados parecem menos promissores. No próximo capítulo, serão apresentadas aplicações práticas diretas deste mestrado.

Tabela 5.13: *Lookahead: Média dos valores originais da variável AUC obtidos por validação cruzada de dez partições.*

Base	J48	J48Lookahead
LYMPHOMA-ALI2000	0,90	0,81
GIST-ALL2001	0,50	0,55
COLON-ALO1999	0,80	0,74
MELANOMA-BIT2000	0,85	0,78
DLBCL-OUTCOME-SHI2002	0,51	0,48
LEUKEMIA-GOL1999	0,75	0,86
GSE11665	0,10	0,10
GSE3255	0,67	0,63
GSE360	0,57	0,66
GSE443	0,10	0,05
GSE474	0,48	0,38
GSE5473	0,48	0,63
GSE7433	0,25	0,35
GSE7898	0,10	0,05
BREAST-HED2003	0,40	0,40
LUNG-WIG2002	0,80	0,49
BREAST-MA2003	0,49	0,49
SOFT-NIE2002	0,66	0,60
OVARY-WEL2001	0,30	0,38
CNS-POM2002	0,50	0,41
PROSTATE-OUTCOME-SIN2002	0,23	0,43

Tabela 5.14: *Lookahead: Desvio padrão dos valores originais da variável AUC obtidos por validação cruzada de dez partições.*

Base	J48	J48Lookahead
LYMPHOMA-ALI2000	0,02	0,03
GIST-ALL2001	0,15	0,16
COLON-ALO1999	0,06	0,07
MELANOMA-BIT2000	0,04	0,04
DLBCL-OUTCOME-SHI2002	0,04	0,06
LEUKEMIA-GOL1999	0,07	0,03
GSE11665	0,07	0,07
GSE3255	0,07	0,06
GSE360	0,03	0,05
GSE443	0,10	0,05
GSE474	0,10	0,09
GSE5473	0,06	0,07
GSE7433	0,08	0,10
GSE7898	0,10	0,05
BREAST-HED2003	0,12	0,12
LUNG-WIG2002	0,07	0,11
BREAST-MA2003	0,05	0,03
SOFT-NIE2002	0,05	0,04
OVARY-WEL2001	0,11	0,14
CNS-POM2002	0,04	0,04
PROSTATE-OUTCOME-SIN2002	0,08	0,14

Tabela 5.15: *Lookahead: Média dos valores originais da variável Tamanho da Árvore obtidos por validação cruzada de dez partições.*

Base	J48	J48Lookahead
LYMPHOMA-ALI2000	15,00	16,60
GIST-ALL2001	3,00	3,00
COLON-ALO1999	7,00	7,20
MELANOMA-BIT2000	5,00	5,40
DLBCL-OUTCOME-SHI2002	8,20	8,60
LEUKEMIA-GOL1999	4,40	5,00
GSE11665	2,60	2,60
GSE3255	7,00	7,00
GSE360	16,80	17,00
GSE443	3,00	3,00
GSE474	5,00	5,00
GSE5473	5,00	5,00
GSE7433	5,00	5,00
GSE7898	3,00	3,00
BREAST-HED2003	3,00	3,00
LUNG-WIG2002	5,00	6,80
BREAST-MA2003	12,80	23,00
SOFT-NIE2002	10,60	11,80
OVARY-WEL2001	3,00	3,00
CNS-POM2002	8,00	7,80
PROSTATE-OUTCOME-SIN2002	4,00	5,00

Tabela 5.16: *Lookahead: Desvio padrão dos valores originais da variável Tamanho da Árvore obtidos por validação cruzada de dez partições.*

Base	J48	J48Lookahead
LYMPHOMA-ALI2000	0,42	0,27
GIST-ALL2001	0,00	0,00
COLON-ALO1999	0,00	0,36
MELANOMA-BIT2000	0,00	0,27
DLBCL-OUTCOME-SHI2002	0,33	0,27
LEUKEMIA-GOL1999	0,31	0,00
GSE11665	0,27	0,27
GSE3255	0,00	0,00
GSE360	0,20	0,30
GSE443	0,00	0,00
GSE474	0,00	0,00
GSE5473	0,00	0,00
GSE7433	0,00	0,00
GSE7898	0,00	0,00
BREAST-HED2003	0,00	0,00
LUNG-WIG2002	0,00	0,20
BREAST-MA2003	0,20	0,79
SOFT-NIE2002	0,27	0,44
OVARY-WEL2001	0,00	0,00
CNS-POM2002	0,33	0,33
PROSTATE-OUTCOME-SIN2002	0,33	0,00

Capítulo

6

Aplicação Prática

6.1 Considerações Iniciais

A seguir, são descritas duas aplicações reais dos trabalhos desenvolvidos neste mestrado.

6.2 INCT Adapta

Este projeto de mestrado encontra-se inserido no contexto do INCT[1] Adapta[2] (Instituto Nacional de Ciência e Tecnologia de Adaptações da Biota Aquática da Amazônia), que está sendo realizado na Amazônia e conta com diversos grupos de pesquisa nacionais (de todo Brasil, não só dos estados compreendidos pela Amazônia) e internacionais. Seu principal objetivo é estudar a adaptação de organismos a condições ambientais extremas, reconhecendo os genes envolvidos e desenvolvendo produtos, processos e políticas ambientais relacionados a essa adaptação. Três linhas de pesquisa são mantidas: uma que estuda as interações entre organismo e ambiente, tentando entender como alguns grupos de animais e plantas conseguem suportar variações ambientais extremas e como aqueles grupos que não o conseguem percebem as mudanças e migram; outra que busca por biomarcadores, a fim de escolher espécies que terão seu genoma e transcriptoma analisados; e uma terceira, que, a partir das outras duas, explora novos produtos e processos, a serem usados, por exemplo, como suporte à elaboração de políticas públicas (Val et al. 2008). As metodologias, técnicas e ferramentas utilizadas por AM e que foram estudadas neste trabalho para a proposta e implementação de uma metodologia específica aos dados em questão têm sido importantes e úteis ao INCT Adapta, não só, mas principalmente à segunda linha de pesquisa, ou seja, na busca por biomarcadores e na análise de dados de expressão gênica.

A diversidade biológica da Amazônia e todos os seus ecossistemas têm sido exaustivamente estudados, mas a exata noção de sua dimensão ainda está longe de ser alcançada. O rio Amazonas, por exemplo, possui uma diversidade de peixes maior que a da Europa inteira (McConnell & Lowe-McConnell 1987). Os números atuais da Amazônia são considerados subestimados (Araujo-Lima & Goulding 1997). O entendimento das rápidas mudanças ambientais que ocorrem em algumas regiões amazônicas e suas implicações é um dos desafios.

Organismos têm desenvolvido habilidades para interagir com, reduzir e até usar os efeitos dessas mudanças e de condições ambientais extremas. Acredita-se que muitos destes mecanismos sejam compartilhados entre organismos de diferentes grupos filogenéticos. O INCT Adapta considera o entendimento de tais mecanismos como importante no estabelecimento de políticas públicas, como ações de conservação ambiental e suporte a processos de intervenção ambiental (Frankham et al. 2002; Allendorf & Luikart 2007), o que resultará em produtos e processos que melhorarão a qualidade de vida dos povos amazônicos.

[1] http://www.cnpq.br/programas/inct/_apresentacao/
[2] http://adapta.inpa.gov.br/

Uma fonte de mudanças ambientais é representada pela intervenção humana, o que tem acarretado impactos imediatos no ambiente aquático. São exemplos as usinas hidrelétricas, exploração do solo, mineração, construção de rodovias, devastação florestal, impactos urbanos, entre outros. O INCT Adapta estuda as reações dos organismos aquáticos a todas as mudanças, sejam de origem natural, sejam de origem humana. Estuda também como os organismos já adaptados a estas alterações possuem tal habilidade e o que exatamente a ativa.

O Adapta tem estudado peixes, plantas aquáticas, invertebrados, anfíbios, micro-organismos e mamíferos aquáticos em ambientes naturais, ambientes impactados pelo ser humano e ambientes sob condições experimentais (microcosmos). Os resultados do estudo constituirão informações importantes para a definição de potenciais biomarcadores e de espécies que terão seu genoma analisado, o que será alcançado pela análise de expressão gênica, sequenciamento do DNA e análise por bioinformática e AM. A aplicação de ferramentas de AM ajudará a reduzir o número de variáveis a serem examinadas, auxiliando no estabelecimento de bioindicadores e biomarcadores que permitirão monitorar a qualidade ambiental.

Durante este projeto de mestrado, o autor participou ativamente do planejamento e implementação de um sistema computacional de gerenciamento de dados e informações provenientes do Adapta (Perez et al. 2011). Uma das funcionalidades que o *software* oferece é a possibilidade de construção de modelos de AM. Neste sentido, o sistema permite a leitura dos dados armazenados no banco, a definição de um conjunto de exemplos baseado naqueles dados e a construção de modelos baseados em ADs. Um exemplo de saída utilizando dados fictícios pode ser visualizado na Figura 6.1, em que a classe é o período hidrológico do ambiente da coleta quando ela foi realizada. Os atributos considerados foram: temperatura da água, concentração de CO_2 na água, concentração de O_2 na água, pH, condutividade da água e condição climática. Destes, apenas condição climática é um atributo discreto. Os outros são contínuos. O modelo construído no exemplo tenta predizer o período hidrológico em que a coleta ocorreu baseando-se nos atributos citados.

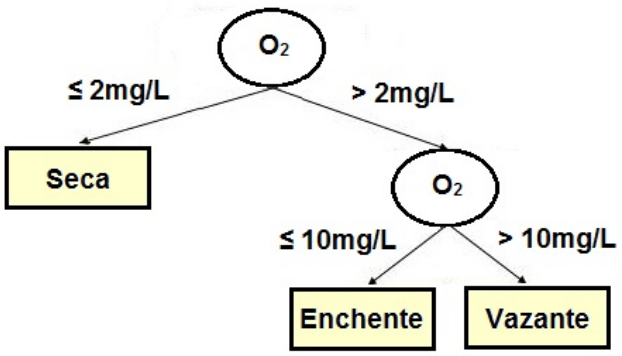


Figura 6.1: *AD induzida pela classe* J48. *O_2 representa a concentração de oxigênio na água e Seca, Enchente e Vazante representam possíveis valores para período hidrológico. Este é apenas um exemplo fictício.*

O autor contribuiu também para a escrita de um capítulo de um livro sobre poluição aquática na Amazônia, ainda a ser publicado. A contribuição se deu na escrita de tópicos referentes à construção de modelos de AM em Biologia e a testes estatísticos para comparação dos indutores desses modelos.

6.3 Atividade Microbiana de Peptídeos

A experiência no Adapta permitiu que o autor entrasse em contato com diversos pesquisadores da área biológica. Foi a partir desta experiência que surgiu uma grande oportunidade de aplicar parte das técnicas trabalhadas em um projeto científico real na biologia. Felipe Lira e Sergio Nozawa têm estudado a atividade antimicrobiana de peptídeos. Sua ideia é desenvolver peptídeos antimicrobianos mais efetivos, possivelmente até menores, a partir de peptídeos já existentes. Porém, os processos utilizados para obter tais peptídeos são caros e demorados. Ferramentas computacionais de bioinformática e AM podem auxiliar neste tipo tarefa, na tentativa de predizer a atividade antimicrobiana de um peptídeo antes que ele seja efetivamente produzido e, assim, diminuir os custos

financeiro e de tempo do projeto.

Existe grande interesse biológico e econômico em peptídeos antimicrobianos, dada a sua importância, por exemplo, no desenvolvimento de novas drogas. Normalmente, neste tipo de estudo, parte-se de um peptídeo natural e promovem-se substituições, deleções e inserções de um ou mais aminoácidos, gerando novos peptídeos, cujo potencial antimicrobiano será analisado e cujos efeitos podem ser modulados (Tossi et al. 1997). Existem extensos bancos de dados que possuem informações acumuladas sobre tais moléculas, em parte devido à dificuldade de se determinar sua atividade antimicrobiana com base apenas em sua sequência de resíduos de aminoácidos. No entanto, ainda há muito a ser feito para que se tenha uma forma definitiva de sintetizar essas moléculas artificialmente e controlar seu potencial antimicrobiano.

O peptídeo parental usado no estudo de Lira e Nozawa foi extraído de uma biblioteca de cDNA de um organismo da Amazônia. No meio das centenas de clones produzidos, uma busca utilizando BLASTX em uma base de dados do GenBank em um servidor local possibilitou a identificação de uma sequência de cDNA que codifica um peptídeo com potencial antimicrobiano. Substituições criteriosas de resíduos de aminoácidos foram realizadas no peptídeo parental e novos peptídeos foram criados, que tiveram seu potencial antimicrobiano medido por um software especializado.

A contribuição do aluno e seu orientador se deu na forma da construção de uma AD, obtida por meio da técnica de *windowing*. A base de dados utilizada no treinamento pertence ao grupo de Lira e Nozawa e era composta por sessenta peptídeos antimicrobianos, descritos por cinquenta e três atributos com informações moleculares (e.g., carga, presença de resíduos hidrofóbicos, índice de Boman, massa, contagem atômica, ponto isoelétrico). A classe considerada era composta pelos valores *nenhuma*, *baixa*, *média* e *alta* e dizia respeito à atividade antimicrobiana dos peptídeos. Neste caso, *windowing* com apenas árvores não podadas foi usado para encontrar os melhores parâmetros e os atributos mais preditivos a serem usados pelo algoritmo J48, funcionando como um otimizador de parâmetros e um filtro de atributos. Definidos os parâmetros e atributos, uma árvore final foi construída pelo J48.

Por comparação, as outras ferramentas utilizadas para realizar esta mesma tarefa apenas diziam se havia probabilidade alta ou baixa de ser um peptídeo antimicrobiano, enquanto a árvore conseguiu predizer não só o potencial antimicrobiano, como também seu espectro (nenhum, baixo, médio ou alto). Vale notar que a base utilizada serviu também para verificar a coerência nas alterações realizadas nos peptídeos.

Os peptídeos gerados pelas substituições de resíduos foram, então, passados pela árvore construída e tiveram sua atividade antimicrobiana classificada pelo modelo, ou seja, a árvore foi mais uma ferramenta de auxílio aos pesquisadores para decidirem em quais peptídeos investir. O modelo gerado pela árvore foi muito bem avaliado pelos especialistas em termos do conhecimento útil transmitido por ele e de sua efetividade em auxiliar na predição da atividade antimicrobiana e minimizar o custo envolvido. Os resultados biológicos têm sido interessantes e geraram um artigo (Lira et al. 2011), submetido a um periódico internacional relevante na área.

Selecionados os peptídeos com maior potencial antimicrobiano com a ajuda da AD, estes foram testados nas bactérias *S. aureus* e *E. coli* e o critério de análise neste caso foi o diâmetro dos discos de inibição de crescimento formados na placa de Petri. Quanto maior o disco de inibição observado, mais efetivo foi o peptídeo contra a bactéria. Foram feitos dois tipos de análise estatística: para cada peptídeo testado, comparação de sua atividade nas diferentes células bacterianas consideradas; para cada tipo de célula bacteriana considerada, comparação dos diferentes peptídeos.

Para esclarecer melhor o teste microbiológico: as bactérias são colocadas em uma placa de Petri; aplicam-se, então, discos de papel sobre a placa e aplica-se o peptídeo dissolvido sobre os discos; em seguida, as placas são submetidas a uma temperatura de 37^oC, para permitir o crescimento das bactérias. Caso tenha havido a formação de um halo no disco impedindo tal crescimento, pode-se dizer que o peptídeo funcionou. Quanto maior esse halo, mais eficaz o peptídeo pode ser considerado em termos de sua atividade antimicrobiana.

Detalhes da metodologia e resultados do trabalho encontram-se no artigo submetido à publicação. O estudo conseguiu encontrar um peptídeo menor que o parental e que apresentou atividade

antimicrobiana maior.

O estudo indica que pode ser a base para outros trabalhos que visem a entender os mecanismos de ação dos peptídeos estudados e seu uso como medicamento. Reitera também a importância das ferramentas computacionais no auxílio à descoberta de conhecimento biológico.

6.4 Considerações Finais

Neste capítulo, foram relatados resumidamente a participação do autor no INCT Adapta e as aplicações práticas dos trabalhos desenvolvidos em seu projeto de mestrado. No próximo capítulo, as principais conclusões do trabalho são fornecidas.

Capítulo

7

Conclusão

7.1 Considerações Iniciais

Neste capítulo, são apresentadas as conclusões construídas a partir dos trabalhos realizados. Será apresentada também a produção bibliográfica do autor e do grupo de pesquisa do qual faz parte, bem como os trabalhos futuros a desenvolver.

7.2 Windowing

Uma das conclusões a que se pode chegar diz respeito ao compromisso *desempenho x legibilidade* do modelo, que geralmente faz com que, quanto mais complexo um modelo, menos interpretável ele seja (Martens et al. 2011; Gamberger et al. 2004) e vice-versa. Como se pode notar nos resultados, as versões propostas consideradas mais legíveis (principalmente WEWeC, WEC, WWeC e WC) foram também as que obtiveram pior desempenho em termos de acurácia. A única versão que conseguiu balancear um pouco melhor esse compromisso foi a WPEWeC, que envolveu todas as alterações propostas, pois obteve bons resultados para as variáveis relacionadas à interpretabilidade do modelo e não foi considerada significativamente pior que nenhum outro indutor nos quesitos acurácia e AUC.

Outra constatação importante teve relação com a instabilidade dos indutores que tiveram bom desempenho nas variáveis relacionadas à interpretabilidade dos modelos. Tais indutores se apresentaram mais instáveis que os outros nas variáveis coesão, compactação, coesão-compactação e acurácia. Possivelmente, assim como existe o compromisso *desempenho x legibilidade*, pode existir um compromisso *estabilidade x legibilidade*. Modelos como SVM (*Support Vector Machines*) costumam ter melhor estabilidade, porém não são facilmente interpretáveis.

Apesar de terem sido encontradas diferenças significativas na acurácia, não ocorreu o mesmo no caso do *AUC*. Isto corrobora o fato de não haver necessariamente um compromisso (nem positivo, nem negativo) entre acurácia e *AUC*. O fato de não ter havido diferenças no *AUC* e ter havido diferenças nas variáveis de interpretabilidade faz com que se possa concluir que as alterações propostas possuem o potencial de melhorar a leitura dos modelos induzidos sem prejudicar a área sob a curva ROC.

Grande atenção tem sido dada a medidas de desempenho como *AUC* e acurácia, deixando muitas vezes a interpretabilidade do modelo em segundo plano, mesmo quando são usados classificadores simbólicos. Existem muitas técnicas de construção de AD, apesar de nenhuma delas ser universalmente a melhor, uma vez que diferentes domínios de aplicação levam a diferentes problemas, o que leva a diferentes soluções. Assim, é esperado que combinação de técnicas, como *windowing*, leve a melhores resultados em alguns casos. Foi mostrado neste trabalho que *windowing* com as alterações propostas neste mestrado tende a produzir modelos mais facilmente interpretados por humanos, sem diferenças significativas nos valores de *AUC*.

As alterações no algoritmo de *windowing* foram propostas com vistas à melhora da aplicação

da técnica em dados de expressão gênica, mas podem ser generalizadas para outros problemas de aprendizado de máquina que envolvam classificação.

A análise realizada até agora se resumiu à estatística. Ainda é importantíssimo que especialistas analisem os modelos produzidos pelos indutores propostos, para que realmente se possa concluir algo de maneira mais efetiva. Com vistas a isto, os estudos deste mestrado estão sendo estendidos e, com o auxílio de especialistas, já existe um projeto em andamento que visa aos seguintes objetivos: 1) analisar os modelos induzidos e verificar se são biologicamente plausíveis e se são melhores que aqueles induzidos pelo J48 e o *windowing* original; e 2) verificar se conhecimento novo pode ser produzido, na tentativa de encontrar genes relevantes, mas que ainda não tiverem sido associados aos problemas estudados. Esta análise está focada em dados de expressão gênica de câncer em humanos e contará inclusive com a colaboração de um especialista em transcriptoma do câncer. O estudo citado acima está em andamento e os resultados deverão ser submetidos à publicação assim que possível. Porém, algo relacionado ao objetivo de validar biologicamente os modelos produzidos foi feito e está descrito no Capítulo 6.

7.3 Lookahead

A principal conclusão a que se pode chegar neste caso é: ainda que seu custo computacional tenha sido muito alto com relação ao J48, o *lookahead* produziu árvores maiores e que não tiveram a contrapartida em *AUC*, ou seja, árvores mais complexas, mas com o mesmo desempenho.

Lookahead de um nível não foi suficiente para propiciar ganho em desempenho em dados de expressão gênica. Devido à enorme quantidade de atributos e inter-relacionamento entre eles, o algoritmo teria que explorar muito mais passos à frente para realmente poder ser eficaz. No entanto, a consideração de mais níveis é praticamente impossível, a não ser que os valores dos atributos numéricos sofram algum tipo de discretização.

A seguir, as publicações deste mestrado são descritas.

7.4 Artigos Publicados

7.4.1 Congressos Internacionais

ECML PKDD 2011 – KD-HCM

O artigo *Analysis of Decision Tree Pruning Using Windowing in Medical Datasets with Different Class Distributions* (Perez & Baranauskas 2011) foi apresentado em sessão oral na edição de 2011 da Conferência Europeia em Aprendizado de Máquina e Princípios e Práticas de Descoberta de Conhecimento em Bancos de Dados (*The European Conference on Machine Learning and Principles and Practice of Knowledge Discovery in Databases – ECML PKDD*), mais especificamente no *Workshop* em Descoberta de Conhecimento em Saúde e Medicina (*Workshop on Knowledge Discovery in Health Care and Medicine – KD-HCM*), ocorrido entre 5 e 9 de setembro em Atenas, Grécia.

ICCBBI 2011

O artigo *A Software Tool for Information Management and Data Mining of Biological Data for Studying Adaptation of Living Organisms in Amazonia* (Perez et al. 2011) foi publicado nos *proceedings* da edição 2011 da Conferência Internacional em Biociências Computacionais e Bioinformática (*International Conference on Computational Biosciences and Bio-Informatics – ICCBBI 2011*), ocorrida em julho em Bhubaneswar, Índia. O artigo se encontra no contexto das atividades do INCT Adapta.

7.4.2 Congressos Nacionais

WIM 2011

Em julho de 2011, foi realizado o WIM (XI *Workshop* de Informática Médica), em Natal, durante o XXXI Congresso da Sociedade Brasileira de Computação (CSBC), em que foi aceito o artigo intitulado *Avaliação do Algoritmo de Stacking em Dados Biomédicos* (Caffé et al. 2011). A primeira autora do artigo, Maria Izabela Ruz Caffé, é aluna de iniciação científica sob orientação do Prof.

Dr. José Augusto Baranauskas e este aluno de mestrado contribuiu para seu trabalho. O artigo explora a aplicação de variações da técnica *stacking* em dados biológicos.

7.5 Artigos Aceitos e em Fase de Publicação em Periódicos Internacionais

JHI

O artigo sobre *stacking* descrito anteriormente recebeu o prêmio de melhor trabalho em sua categoria no WIM 2011 e, por isto, foi aceito para publicação no *Journal of Health Informatics* (JHI), veículo de divulgação acadêmica oficial da Sociedade Brasileira de Informática em Saúde.

IRNet

O artigo sobre o Adapta descrito anteriormente foi aceito para publicação no periódico da Interscience Research Network (IRnet), uma instituição indiana que promove eventos e publicações científicas voltadas a profissionais, acadêmicos e pesquisadores. Esta instituição foi a responsável pelo evento ICCBBI 2011.

7.6 Artigos Submetidos

7.6.1 Congressos Internacionais

MLDM 2012

De 13 a 20 de julho de 2012, ocorrerá em Berlim, Alemanha, o MLDM 2012 (8^{th} *International Conference on Machine Learning and Data Mining*). A este evento foi submetido o artigo intitulado *How Many Trees in a Random Forest?* (Oshiro et al. 2012). A primeira autora do artigo, Thais Mayumi Oshiro, é também aluna de mestrado sob orientação do Prof. Baranauskas e este aluno de mestrado tem contribuído para seu trabalho. O artigo explora a técnica de AM conhecida como *Random Forest* (Breiman 2001).

7.6.2 Periódicos Internacionais

Aquaculture

O trabalho descrito na Seção 6.3 deu origem a um artigo (Lira et al. 2011), submetido ao periódico Aquaculture.

7.7 Artigos em Fase de Escrita

Resultados Finais do Mestrado

Um artigo contendo os resultados alcançados no mestrado está sendo escrito e pretende-se submetê-lo a um bom periódico internacional.

7.8 Capítulos de Livro

No INCT Adapta, o autor contribuiu para a escrita de um capítulo de um livro sobre poluição aquática na Amazônia, ainda a ser publicado. A contribuição se deu na escrita de tópicos referentes à construção de modelos de AM em Biologia e a testes estatísticos para comparação dos indutores desses modelos. O livro ainda não foi publicado.

7.9 Trabalhos Futuros

Como continuidade a este trabalho, pretende-se explorar mais aprofundadamente as medidas de interpretabilidade e complexidade aplicadas especificamente a árvores de decisão. Pouca literatura existe sobre o tema, que é muito importante do ponto de vista da análise do modelo por parte do especialista. Além daquelas apresentadas no texto, outras medidas já vêm sendo trabalhadas. Por exemplo, algumas delas são baseadas no balanceamento das árvores geradas. Além disto, pretende-se estudar também formas de fazer com que os indutores de árvores produzam modelos mais compreensíveis, porém sem deixar de ser informativos.

7.10 Considerações Finais

Neste mestrado, foi mostrado o potencial de algumas técnicas de AM no que se refere à tarefa de auxiliar os especialistas que utilizam a expressão gênica como base de suas análises. No entanto, ainda há muito a ser feito, principalmente com relação à instabilidade das ADs. Pontos positivos do mestrado e que poderiam ser evidenciados são: as alterações propostas no algoritmo de *windowing* produziram bons resultados quanto a aspectos de interpretabilidade dos modelos construídos; e medidas de interpretabilidade e complexidade de modelos especificamente desenvolvidas para ADs foram propostas.

Referências

Alizadeh, A. A. (2000). Distinct use large b-cell lymphoma identified by gene expression profiling. *Nature 403*(6769), 503–11.

Allander, S., Nupponen, N., Ringnér, M., Hostetter, G., Maher, G., Goldberger, N., Chen, Y., Carpten, J., Elkahloun, A. & Meltzer, P. (2001). Gastrointestinal stromal tumors with kit mutations exhibit a remarkably homogeneous gene expression profile. *Cancer Research 61*(24), 8624.

Allendorf, F. & Luikart, G. (2007). *Conservation and the genetics of populations*. Wiley-Blackwell.

Alon, U., Barkai, N., Notterman, D., Gish, K., Ybarra, S., Mack, D. & Levine, A. (1999). 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 96*(12), 6745–6750.

Amin, K. (2007). Pathway-express: A bioinformatics tool for pathway level analysis using gene expression data. *ETD Collection for Wayne State University 1*, AAI1447572.

Araujo-Lima, C. & Goulding, M. (1997). *So fruitful a fish: Ecology, conservation, and aquaculture of the Amazon's tambaqui*. Columbia Univ Pr.

Armstrong, S., Staunton, J., Silverman, L., Pieters, R., den Boer, M., Minden, M., Sallan, S., Lander, E., Golub, T., Korsmeyer, S. et al. (2002). Mll translocations specify a distinct gene expression profile that distinguishes a unique leukemia. *Nature Genetics 30*(1), 41–47.

Baldi, P. & Brunak, S. (2001). *Bioinformatics: The Machine Learning Approach, Second Edition, (Adaptive Computation and Machine Learning)*. The MIT Press.

Baranauskas, J. A. (2001). *Extração Automática de Conhecimento Utilizando Múltiplos Indutores*. Ph. D. thesis, ICMC-USP. http://www.teses.usp.br/teses/disponiveis/55/55134/tde-08102001-112806/.

Beer, D., Kardia, S., Huang, C., Giordano, T., Levin, A., Misek, D., Lin, L., Chen, G., Gharib, T., Thomas, D. et al. (2002). Gene-expression profiles predict survival of patients with lung adenocarcinoma. *Nature Medicine 8*(8), 816–824.

Ben-Dor, A., Bruhn, L., Friedman, N., Nachman, I., Schummer, M. & Yakhini, Z. (2000). Tissue classification with gene expression profiles. *Journal of Computational Biology 7*(3-4), 559–583.

Benjamini, Y. & Hochberg, Y. (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society Series B 57*, 289–300.

Bhattacharjee, A., Richards, W., Staunton, J., Li, C., Monti, S., Vasa, P., Ladd, C., Beheshti, J., Bueno, R., Gillette, M. et al. (2001). Classification of human lung carcinomas by mrna expression profiling reveals distinct adenocarcinoma subclasses. *Proceedings of the National Academy of Sciences 98*(24), 13790.

Bittner, M., Meltzer, P., Chen, Y., Jiang, Y., Seftor, E., Hendrix, M., Radmacher, M., Simon, R., Yakhini, Z., Ben-Dor, A. et al. (2000). Molecular classification of cutaneous malignant melanoma by gene expression profiling. *Nature 406*(6795), 536–540.

Bleharski, J., Li, H., Meinken, C., Graeber, T., Ochoa, M., Yamamura, M., Burdick, A., Sarno, E., Wagner, M., Röllinghoff, M. et al. (2003). Use of genetic profiling in leprosy to discriminate clinical forms of the disease. *Science 301*(5639), 1527.

Blum, A. & Mitchell, T. (1998). Combining labeled and unlabeled data with co-training. In *Proceedings of the eleventh annual conference on Computational learning theory*, pp. 92–100. ACM: Morgan Kaufmann Publishers.

Boon, K., Bailey, N., Yang, J., Steel, M., Groshong, S., Kervitsky, D., Brown, K., Schwarz, M. & Schwartz, D. (2009). Molecular phenotypes distinguish patients with relatively stable from progressive idiopathic pulmonary fibrosis (ipf). *PLoS One 4*(4), e5134.

Bradley, A. (1997). The use of the area under the roc curve in the evaluation of machine learning algorithms. *Pattern Recognition 30*(7), 1145–1159.

Breiman, L. (2001). Random forests. *Machine Learning 45*(1), 5–32.

Breiman, L., Friedman, J., Olshen, R. & Stone, C. (1984). *Classification and Regression Trees*. Pacific Grove, CA: Wadsworth & Books.

Butte, A., Tamayo, P., Slonim, D., Golub, T. & Kohane, I. (2000). Discovering functional relationships between rna expression and chemotherapeutic susceptibility using relevance networks. *Proceedings of the National Academy of Sciences 97*(22), 12182.

Caffé, M. I. R., Perez, P. S. & Baranauskas, J. A. (2011). Avaliação do algoritmo de stacking em dados biomédicos. In *Anais do XXXI Congresso da Sociedade Brasileira de Computação, XI Workshop de Informática Médica, Natal, RN*. Artigo premiado como melhor artigo completo do evento na categoria.

Caruana, R. & Niculescu-Mizil, A. (2006). An empirical comparison of supervised learning algorithms. In *Proceedings of the 23rd international conference on Machine learning*, pp. 161–168. ACM.

Catlett, J. (1991). *Megainduction*. Ph. D. thesis, Basser Department of Computer Science.

Chari, R., Lonergan, K., Ng, R., MacAulay, C., Lam, W. & Lam, S. (2007). Effect of active smoking on the human bronchial epithelium transcriptome. *BMC Genomics 8*(1), 297.

Chaussabel, D., Semnani, R., McDowell, M., Sacks, D., Sher, A. & Nutman, T. (2003). Unique gene expression profiles of human macrophages and dendritic cells to phylogenetically distinct parasites. *Blood 102*(2), 672.

Chen, Q. (2004). Inductive Learning on Partitioned Data. Master's thesis, The University of Vermont.

Cho, S. B. & Won, H.-H. (2007). Cancer classification using ensemble of neural networks with multiple significant gene subsets. *Applied Intelligence 26*(3), 243–250.

Cortes, C. & Mohri, M. (2004). Auc optimization vs. error rate minimization. In *Advances in neural information processing systems 16: proceedings of the 2003 conference*, Volume 16, pp. 313. The MIT Press.

Demšar, J. (2006). Statistical comparison of classifiers over multiple data sets. *Journal of Machine Learning Research 7*(1), 1–30.

Dettling, M. & Buhlmann, P. (2003). Boosting for tumor classification with gene expression data. *Bioinformatics 19*(9), 1061–1069.

Dietterich, T. G. (1986). Learning at the knowledge level. *Machine Learning 1*(3), 287–315. Reprinted in Shavlik and Dietterich (eds.), 1990. Readings in Machine Learning, Morgan Kaufmann Publishers, Inc.

Dietterich, T. G. (1996). Editorial. *Machine Learning 22*(1–3), 5–6.

Dietterich, T. G. (1997). Statistical tests for comparing supervised classification learning algorithms. Citeseer. ftp://ftp.cs.orst.edu/pub/tgd/papers.

Domingos, P. (1996). Efficient specific-to-general rule induction. In *Proceedings of the 2nd International Conference on Knowledge Discovery and Data Mining (KDD-96)*, pp. 319–322.

Duda, R. O., Hart, P. E. & Stork, D. G. (2001). *Pattern Classification*, Chapter Unsupervised Learning and Clustering. Pattern Classification and Scene Analysis: Pattern Classification. John Wiley & Sons, New York.

Dudoit, S., Fridlyand, J. & Speed, T. P. (2002). Comparison of discrimination methods for the classification of tumors using gene expression data. *Journal of the American Statistical Association 97*(457), 77–87.

Dunn, O. J. (1961). Multiple comparisons among means. *Journal of the American Statistical Association 56*, 52–64.

Fawcett, T. (2006). An introduction to roc analysis. *Pattern Recognition Letters 27*(8), 861–874.

Frankham, R., Ballou, D. & Briscoe, D. (2002). *Introduction to Conservation Genetics*. Cambridge UK: Cambridge University Press.

Friedman, M. (1940). A comparison of alternative tests of significance for the problem of m rankings. *The Annals of Mathematical Statistics 11*(1), 86–92.

Fürnkranz, J. (1997). More efficient windowing. In *Proceedings of the Fourteenth National Conference on Artificial Intelligence*, pp. 509–514. Citeseer.

Gamberger, D., Lavrač, N., Zelezný, F. & Tolar, J. (2004). Induction of comprehensible models for gene expression datasets by subgroup discovery methodology. *Journal of Biomedical Informatics 37*(4), 269–284.

García, S. & Herrera, F. (2008). An extension on statistical comparisons of classifiers over multiple data sets for all pairwise comparisons. *Journal of Machine Learning Research 9*, 2677–2694.

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. (1999). Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. *Science 286*, 531–537.

Gordon, G., Jensen, R., Hsiao, L., Gullans, S., Blumenstock, J., Ramaswamy, S., Richards, W., Sugarbaker, D. & Bueno, R. (2002). Translation of microarray data into clinically relevant cancer diagnostic tests using gene expression ratios in lung cancer and mesothelioma. *Cancer Research 62*(17), 4963.

Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P. & Witten, I. H. (2009). The weka data mining software: An update. *SIGKDD Explorations 11*(1), 10–18.

Hand, D. & Till, R. (2001). A simple generalisation of the area under the roc curve for multiple class classification problems. *Machine Learning 45*(2), 171–186.

Hedenfalk, I., Ringnér, M., Ben-Dor, A., Yakhini, Z., Chen, Y., Chebil, G., Ach, R., Loman, N., Olsson, H., Meltzer, P. et al. (2003). Molecular classification of familial non-brca1/brca2 breast cancer. *Proceedings of the National Academy of Sciences 100*(5), 2532.

Hochberg, Y. (1988). A sharper bonferroni procedure for multiple tests of significance. *Biometrika 75*, 800–803.

Holm, S. (1979). A simple sequentially rejective multiple test procedure. *Scandinavian Journal of Statistics 6*, 65–70.

Hommel, G. (1988). A stagewise rejective multiple test procedure based on a modified Bonferroni test. *Biometrika 75*(2), 383–386.

Hossain, M., Hassan, M. & Bailey, J. (2008). ROC-tree: A Novel Decision Tree Induction Algorithm Based on Receiver Operating Characteristics to Classify Gene Expression Data. In *Proc. of 8th SIAM Intl. Conf. on Data Mining (SDM 2008)*, pp. 455–465.

Hunt, E. B., Stone, P. J. & Marin, J. (1966). *Experiments in induction*. New York: Academic Press.

Kohavi, R. (1995). A study of cross-validation and bootstrap for accuracy estimation and model selection. In *International Joint Conference on Artificial Intelligence*, Volume 14, Montreal, Canada, pp. 1137–1145. LAWRENCE ERLBAUM ASSOCIATES LTD.

Kubat, M., Bratko, I. & Michalski, R. S. (1998). *A Review of Machine Learning Methods*, pp. 3–69. John Wiley & Sons Ltd., West Sussex, England.

Lavrač, N., Flach, P. & Zupan, R. (1999). Rule evaluation measures: A unifying view. In S. Dzeroski & P. Flach (Eds.), *Proceedings of the Ninth International Workshop on Inductive Logic Programming (ILP-99)*, Volume 1634, pp. 74–185. Springer-Verlag. Lecture Notes in Artificial Intelligence.

Lee, S., Chen, J., Zhou, G., Shi, R., Bouffard, G., Kocherginsky, M., Ge, X., Sun, M., Jayathilaka, N., Kim, Y. et al. (2006). Gene expression profiles in acute myeloid leukemia with common translocations using sage. *Proceedings of the National Academy of Sciences of the United States of America 103*(4), 1030.

Lemos, R. N. (2007). Análise de diferentes tipos de leucemia por meio de perfis de expressão gênica. Universidade de São Paulo. Trabalho de Conclusão de Curso em Informática Biomédica.

Leung, S., Chen, X., Chu, K., Yuen, S., Mathy, J., Ji, J., Chan, A., Li, R., Law, S., Troyanskaya, O. et al. (2002). Phospholipase a2 group iia expression in gastric adenocarcinoma is associated with prolonged survival and less frequent metastasis. *Proceedings of the National Academy of Sciences 99*(25), 16203.

Li, X., Rao, S., Jiang, W., Li, C., Xiao, Y., Guo, Z., Zhang, Q., Wang, L., Du, L., Li, J. et al. (2006). Discovery of time-delayed gene regulatory networks based on temporal gene expression profiling. *BMC Bioinformatics 7*(1), 26.

Lira, F., Perez, P. S., Baranauskas, J. A. & Nozawa, S. R. (2011). Development of a decision tree to predict the antimicrobial activity of synthetic peptides. Em fase de publicação.

Lonergan, K., Chari, R., Coe, B., Wilson, I., Tsao, M., Ng, R., MacAulay, C., Lam, S. & Lam, W. (2010). Transcriptome profiles of carcinoma-in-situ and invasive non-small cell lung cancer as revealed by sage. *PloS One 5*(2), e9162.

Lyons-Weiler, J., Patel, S. & Bhattacharya, S. (2003). A classification-based machine learning approach for the analysis of genome-wide expression data. *Genome Res 13(3)*, 503–512.

Ma, X., Salunga, R., Tuggle, J., Gaudet, J., Enright, E., McQuary, P., Payette, T., Pistone, M., Stecker, K., Zhang, B. et al. (2003). Gene expression profiles of human breast cancer progression. *Proceedings of the National Academy of Sciences 100*(10), 5974.

Martens, D., Vanthienen, J., Verbeke, W. & Baesens, B. (2011). Performance of classification models from a user perspective. *Decision Support Systems 51*, 782–793.

Maskos, U. & Southern, E. M. (1992). Oligonucleotide hybridizations on glass supports: a novel linker for oligonucleotide synthesis and hybridization properties of oligonucleotides synthesised in situ. *Nucleic Acids Research 20*(7), 1679–84.

Matukumalli, L., Grefenstette, J., Hyten, D., Choi, I., Cregan, P. & Van Tassell, C. (2006). Application of machine learning in SNP discovery. *BMC Bioinformatics 7*(4).

McConnell, R. & Lowe-McConnell, R. (1987). *Ecological studies in tropical fish communities*. Cambridge Univ Pr.

Michalski, R. S. (1983). A theory and methodology of inductive learning. *Artificial Intelligence 20*, 111–161.

Michie, D. (1988). Machine learning in the next five years. In *Proceedings of the Third European Working Session on Learning EWSL-88*, Glasgow, London, Pitman, pp. 107–122. Pitman.

Monard, M. C. & Baranauskas, J. A. (2003a). *Conceitos sobre Aprendizado de Máquina*, Chapter 4, pp. 89–114. In Rezende (2003).

Monard, M. C. & Baranauskas, J. A. (2003b). *Indução de Regras e Árvores de Decisão*, Chapter 5, pp. 115–139. In Rezende (2003).

Moon, D. & Marwala, T. (2008). Missing data using decision forest and computational intelligence. *Arxiv preprint arXiv:0812.1615*.

Müller, H., Neumaier, M. & Hoffmann, G. (2008). Gene expression studies with microarrays and SAGE: biological and analytical principles 1. *LaboratoriumsMedizin 32*(5).

Murthy, S. K., Kasif, S. & Salzberg, S. L. (1994). A system for induction of oblique decision trees. *Journal of Artificial Intelligence Research 2*(1), 1–32. http://www.cs.jhu.edu/~salzberg/jair94.ps.

Murthy, S. K. & Salzberg, S. L. (1995). Lookahead and pathology in decision tree induction. In R. S. Michalski, J. G. Carbonell, & T. M. Mitchell (Eds.), *Proceedings of IJCAI-95*, Montreal, pp. 1025–1031. ftp://ftp.cs.jhu.edu/pub/salzberg/lookahead.ps.

Nemenyi, P. B. (1963). *Distribution-free multiple comparisons*. Ph. D. thesis, Princeton University.

Nettleton, D., Orriols-Puig, A. & Fornells, A. (2010). A study of the effect of different types of noise on the precision of supervised learning techniques. *Artificial Intelligence Review 33*(4), 275–306.

Netto, O. P., Nozawa, S. R., Mitrowsky, R. A. R., Macedo, A. A. & Baranauskas, J. A. (2010). Applying decision trees to gene expression data from DNA microarrays: A leukemia case study. In *XXX Congress of the Brazilian Computer Society, X Workshop on Medical Informatics*, Belo Horizonte, MG, pp. 10.

Nielsen, T., West, R., Linn, S., Alter, O., Knowling, M., O'Connell, J., Zhu, S., Fero, M., Sherlock, G., Pollack, J. et al. (2002). Molecular characterisation of soft tissue tumours: a gene expression study. *The Lancet 359*(9314), 1301–1307.

Nussbaum, R. L., McInnes, R. R. & Willard, H. F. (2002). *Thompson & Thompson Genética Médica* (6 ed.). Rio de Janeiro: Guanabara Koogan.

Oshiro, T. M., Perez, P. S. & Baranauskas, J. A. (2012). How many trees in a random forest? Submitted to MLDM 2012 – 8^{th} International Conference on Machine Learning and Data Mining.

Park, J., Berggren, J., Hulver, M., Houmard, J. & Hoffman, E. (2006). Grb14, gpd1, and gdf8 as potential network collaborators in weight loss-induced improvements in insulin action in human skeletal muscle. *Physiological Genomics 27*(2), 114–121.

Perez, P. S. & Baranauskas, J. A. (2011). Analysis of decision tree pruning using windowing in medical datasets with different class distributions. In *ECML PKDD 2011 – KD-HCM (Workshop on Knowledge Discovery in Health Care and Medicine)*.

Perez, P. S., Bevilacqua, A. H., Ghelfi, A., Nozawa, S. R., Macedo, A. A. & Baranauskas, J. A. (2011). A software tool for information management and data mining of biological data for studying adaptation of living organisms in amazonia. In *ICCBBI 2011*.

Petricoin, E., Ardekani, A., Hitt, B., Levine, P., Fusaro, V., Steinberg, S., Mills, G., Simone, C., Fishman, D., Kohn, E. et al. (2002). Use of proteomic patterns in serum to identify ovarian cancer. *The lancet 359*(9306), 572–577.

Pomeroy, S., Tamayo, P., Gaasenbeek, M., Sturla, L., Angelo, M., McLaughlin, M., Kim, J., Goumnerova, L., Black, P., Lau, C. et al. (2002). Prediction of central nervous system embryonal tumour outcome based on gene expression. *Nature 415*(6870), 436–442.

Quinlan, J. (1979). Discovering rules by induction from large collections of examples. In D. Michie (Ed.), *Expert Systems in the Micro-Electronic Age*. Edinburgh University Press.

Quinlan, J. R. (1986). Induction of decision trees. *Machine Learning 1*, 81–106. Reprinted in Shavlik and Dietterich (eds.), 1990. Readings in Machine Learning, Morgan Kaufmann Publishers, Inc.

Quinlan, J. R. (1993). C4.5*: Programs for Machine Learning*. Morgan Kaufmann. San Francisco, CA.

Ramaswamy, S., Tamayo, P., Rifkin, R., Mukherjee, S., Yeang, C., Angelo, M., Ladd, C., Reich, M., Latulippe, E., Mesirov, J. et al. (2001). Multiclass cancer diagnosis using tumor gene expression signatures. *Proceedings of the National Academy of Sciences of the United States of America 98*(26), 15149.

Reinartz, T. (2002). A unifying view on instance selection. *Data Mining and Knowledge Discovery 6*(2), 191–210.

Rezende, S. O. (Ed.) (2003). *Sistemas Inteligentes - Fundamentos e Aplicações*. Manole.

Risch, N. J. (2000). Searching for genetic determinants in the new millennium. *Nature 405*, 847–856.

Rosenfeld, N., Aharonov, R., Meiri, E., Rosenwald, S., Spector, Y., Zepeniuk, M., Benjamin, H., Shabes, N., Tabak, S., Levy, A. et al. (2008). MicroRNAs accurately identify cancer tissue origin. *Nature biotechnology 26*(4), 462–469.

Rosenwald, A., Wright, G., Chan, W., Connors, J., Campo, E., Fisher, R., Gascoyne, R., Muller-Hermelink, H., Smeland, E., Giltnane, J. et al. (2002). The use of molecular profiling to predict survival after chemotherapy for diffuse large-b-cell lymphoma. *New England Journal of Medicine 346*(25), 1937–1947.

Salzberg, S. L. (1995a). Locating protein coding regions in human dna using a decision tree algorithm. *Journal of Computational Biology 2*(3), 473–485.

Salzberg, S. L. (1995b). On comparing classifiers: A critique of current research and methods. Technical Report JHU-95/06, Department of Computer Science, Johns Hopkins University. http://www.cs.jhu.edu/~salzberg/critique.ps.

Sarkar, U., Chakrabarti, P., Ghose, S. & DeSarkar, S. (1994). Improving greedy algorithms by lookahead-search. *Journal of Algorithms 16*(1), 1–23.

Schaefer, G., Nakashima, T. & Yokota, Y. (2008). Fuzzy classification for gene expression data analysis. In *Computational Intelligence in Bioinformatics*, pp. 209–218.

Schramm, A., Vandesompele, J., Schulte, J., Dreesmann, S., Kaderali, L., Brors, B., Eils, R., Speleman, F. & Eggert, A. (2007). Translating expression profiling into a clinically feasible test to predict neuroblastoma outcome. *Clinical Cancer Research 13*(5), 1459.

Shadeo, A., Chari, R., Lonergan, K., Pusic, A., Miller, D., Ehlen, T., Van Niekerk, D., Matisic, J., Richards-Kortum, R., Follen, M. et al. (2008). Up regulation in gene expression of chromatin remodelling factors in cervical intraepithelial neoplasia. *BMC Genomics 9*(1), 64.

Shipp, M., Ross, K., Tamayo, P., Weng, A., Kutok, J., Aguiar, R., Gaasenbeek, M., Angelo, M., Reich, M., Pinkus, G. et al. (2002). Diffuse large b-cell lymphoma outcome prediction by gene-expression profiling and supervised machine learning. *Nature Medicine 8*(1), 68–74.

Singh, D., Febbo, P., Ross, K., Jackson, D., Manola, J., Ladd, C., Tamayo, P., Renshaw, A., D'Amico, A., Richie, J. et al. (2002). Gene expression correlates of clinical prostate cancer behavior. *Cancer Cell 1*(2), 203–209.

Slonim, D. K. (2002). From patterns to pathways: gene expression data analysis comes of age. *Nature Genetics 32 supplement*, 502–508.

Statnikov, A., Aliferis, C., Tsamardinos, I., Hardin, D. & Levy, S. (2005). A comprehensive evaluation of multicategory classification methods for microarray gene expression cancer diagnosis. *Bioinformatics 21*(5), 631–643.

Tan, A., Naiman, D., Xu, L., Winslow, R. & Geman, D. (2005). Simple decision rules for classifying human cancers from gene expression profiles. *Bioinformatics 21*(20), 3896–3904.

Tossi, A., Tarantino, C. & Romeo, D. (1997). Design of synthetic antimicrobial peptides based on sequence analogy and amphipathicity. *European Journal of Biochemistry 250*(2), 549–558.

Turney, P. D. (1995). Technical note: Bias and the quantification of stability. *Machine Learning 20*, 23–33.

Val, A. L., Piedade, M. T. F., Zuanon, J. A. S., Gribel, R., Leone, F. A., Baldisseroto, B., Baranauskas, J. A. & Queiroz, H. L. (2008). Instituto nacional de ciência e tecnologia de adaptações da biota aquática da amazônia.

Van't Veer, L., Dai, H., Van De Vijver, M., He, Y., Hart, A., Mao, M., Peterse, H., van der Kooy, K., Marton, M., Witteveen, A. et al. (2002). Gene expression profiling predicts clinical outcome of breast cancer. *Nature 415*(6871), 530–536.

Velculescu, V. E., Zhang, L., Vogelstein, B. & Kinzler, K. W. (1995). Serial analysis of gene expression. *Science 270*(5235), 484–7.

Wang, H., Zheng, H., Simpson, D. & Azuaje, F. (2006). Machine learning approaches to supporting the identification of photoreceptor-enriched genes based on expression data. *BMC bioinformatics 7*(1), 116.

Welsh, J., Sapinoso, L., Su, A., Kern, S., Wang-Rodriguez, J., Moskaluk, C., Frierson, H. & Hampton, G. (2001). Analysis of gene expression identifies candidate markers and pharmacological targets in prostate cancer. *Cancer Research 61*(16), 5974.

Welsh, J., Zarrinkar, P., Sapinoso, L., Kern, S., Behling, C., Monk, B., Lockhart, D., Burger, R. & Hampton, G. (2001). Analysis of gene expression profiles in normal and neoplastic ovarian tissue samples identifies candidate molecular markers of epithelial ovarian cancer. *Proceedings of the National Academy of Sciences 98*(3), 1176.

Wigle, D., Jurisica, I., Radulovich, N., Pintilie, M., Rossant, J., Liu, N., Lu, C., Woodgett, J., Seiden, I., Johnston, M. et al. (2002). Molecular profiling of non-small cell lung cancer and correlation with disease-free survival. *Cancer Research 62*(11), 3005.

Wilcoxon, F. (1945). Individual comparisons by ranking methods. *Biometrics 1*, 80–83.

Witten, I. H. & Frank, E. (2005). *Data Mining: Practical Machine Learning Tools and Techniques with Java Implementations* (2 ed.). Morgan Kaufmann.

Yeoh, E., Ross, M., Shurtleff, S., Williams, W., Patel, D., Mahfouz, R., Behm, F., Raimondi, S., Relling, M., Patel, A. et al. (2002). Classification, subtype discovery, and prediction of outcome in pediatric acute lymphoblastic leukemia by gene expression profiling. *Cancer Cell 1*(2), 133–143.

Yeung, K. Y. & Bumgarner, R. E. (2003). Multiclass classification of microarray data with repeated measurements: application to cancer. *Genome Biology 4(12):R83*.

Zhou, Z.-H. & Li, M. (2010). Semi-supervised learning by disagreement. *Knowledge and Information Systems 24*, 415–439.

Apêndice

A

Ferramentas Utilizadas nas Implementações

A.1 Weka

Todos os algoritmos de aprendizado de máquina implementados neste mestrado fizeram uso de um *software* chamado Weka (Hall et al. 2009), uma poderosa coleção especializada de algoritmos de aprendizado de máquina e preprocessamento de dados, tanto clássicos quanto estado-da-arte (Witten & Frank 2005). Mais do que isto, a Weka é uma infraestrutura que permite a condução de experimentos (locais ou remotos, seriais ou paralelos), análise e visualização de bases de dados, entre outros. Segue uma licença pública (*GNU General Public License*) e permite que seu código e funcionalidade sejam utilizados, estendidos e redistribuídos gratuitamente.

Como a Weka não oferece nenhuma implementação do algoritmo de *windowing*, foi implementada uma classe Java, chamada `weka.classifiers.meta.Windowing`, como um meta-indutor, seguindo as regras e padrões definidos pela Weka, para que estivesse disponível como um de seus classificadores. Tal implementação foi baseada naquela do C4.5, incluindo a maioria das características. A única funcionalidade não implementada foi a possibilidade de se construir um único classificador baseado em regras a partir do conjunto formado pelo melhor classificador de cada *trial*, que ficou como uma funcionalidade ainda a ser implementada. Todas as alterações do algoritmo propostas neste mestrado foram também implementadas aqui. Foram também implementadas a versão modificada do J48 (descrita na Seção 5.2) e o *lookahead*.

Um detalhe interessante a ser notado é que, apesar de os experimentos com *windowing* terem sido feitos com o J48, a implementação realizada permite que qualquer indutor disponível no software Weka e que trabalhe com problemas de classificação seja utilizado como indutor base para a técnica.

A.2 R

Todos os testes estatísticos de significância foram implementados e executados no sistema computacional conhecido como R (*R Software for Statistical Computing*)[1]. O R é um sistema único, mas que traz dois grandes produtos: uma linguagem de programação completa e especializada em computação estatística; e um ambiente integrado que oferece inúmeras funcionalidades, como manipulação e análise de dados, cálculos facilitados envolvendo matrizes, ferramentas gráficas, entre outros. Também segue a *GNU General Public License* e permite que seu código e funcionalidade sejam utilizados, estendidos e redistribuídos gratuitamente.

[1] http://www.r-project.org/

Apêndice

B

Bases de Dados Utilizadas nos Experimentos

Neste apêndice, são apresentadas as bases de dados utilizadas nos experimentos realizados neste mestrado. Cada base é especificada por meio de sua descrição geral e pela tarefa que ela representou neste projeto. Na Tabela B.1, outras informações sobre as bases são apresentadas, como quantidade de exemplos, atributos, quais foram utilizadas em quais experimentos, entre outros. Nenhuma base continha exemplos cuja classe era desconhecida. Foram retirados das bases os exemplos cuja classe só aparecia uma vez, devido ao fato do *windowing* começar com janelas pequenas e as bases de dados já serem pequenas, não havendo sentido, no contexto deste trabalho, em usar uma técnica que adiciona exemplos à janela conforme precisa de mais informação para o aprendizado se só existe um exemplo representando uma dada classe. Foi o caso das bases ECML2004, GSE7898 e PROSTATE-WEL2001.

LYMPHOMA-ALI2000 (Alizadeh 2000)

- Descrição: Perfil de expressão gênica por microarray. Estudo da heterogeneidade molecular do linfoma de células B por meio da caracterização sistemática da expressão gênica de células B malignas.
- Tarefa: Classificar pacientes de acordo com a variação molecular do linfoma de células B.

GIST-ALL2001 (Allander et al. 2001)

- Descrição: Perfil de expressão gênica por microarray. Estudo de tumores do estroma gastrointestinal (GIST), comparando seu perfil de expressão gênica com o de tumores de células *spindle*.
- Tarefa: Classificar os exemplos entre GIST e tumores de células *spindle*.

COLON-ALO1999 (Alon et al. 1999)

- Descrição: Perfil de expressão gênica por microarray. Comparação do perfil de expressão gênica de células do cólon normais e malignas.
- Tarefa: Classificar células do cólon em malignas ou normais.

LEUKEMIA-GOL1999 (Golub et al. 1999)

- Descrição: Perfil de expressão gênica por microarray. Esta é uma base de dados bastante conhecida e já foi utilizada em muitos trabalhos. É composta por medidas de expressão gênica de amostras retiradas de sangue periférico e medula óssea de pacientes com AML (leucemia mieloide aguda) ou ALL (leucemia linfoblástica aguda), que são as duas possíveis classes do conjunto.
- Tarefa: Classificar pacientes como tendo AML ou ALL.

BCR-ABL-YEO2002, E2A-PBX1-YEO2002, HYPERDIP-50-YEO2002, MLL-YEO2002, T-ALL-YEO2002 e TEL-AML1-YEO2002 (Yeoh et al. 2002)

- Descrição: Perfil de expressão gênica por microarray. Estudo de classificação de subtipos de leucemia linfoblástica pediátrica. Existem seis grupos diagnósticos (BCR-ABL, E2A-PBX1, Hyperdiploid>50, MLL, T-ALL and TEL-AML1), além do grupo "Outros".
- Tarefa: Classificação de subtipos de leucemia linfoblástica pediátrica.

MELANOMA-BIT2000 (Bittner et al. 2000)

- Descrição: Perfil de expressão gênica por microarray. Descoberta de um subconjunto de melanomas por meio de análise matemática de expressão gênica.
- Tarefa: Classificação molecular de melanoma maligno cutâneo.

BREAST-VEE2002 (Van't Veer et al. 2002)

- Descrição: Perfil de expressão gênica por microarray. Estudo de pacientes com câncer de mama quanto ao ocorrido dentro do período de cinco anos após diagnóstico da doença.
- Tarefa: Classificação de pacientes com câncer de mama em: desenvolvimento de metástase dentro de cinco anos ou permanência no estado saudável por pelo menos cinco anos.

NETWORKS-BUT2000 (Butte et al. 2000)

- Descrição: Perfil de expressão gênica por microarray. Estudo para encontrar redes regulatórias e grupos de genes que afetam a susceptibilidade do câncer aos agentes anticâncer.
- Tarefa: Diferenciar entre diversos tipos de câncer: mama, leucemia, ovário, etc.

DLBCL-NIH-ROS2002 (Rosenwald et al. 2002)

- Descrição: Perfil de expressão gênica por microarray. Estudo de amostras de biópsia de linfoma de células B grandes em busca de anomalias genômicas.
- Tarefa: Classificação de pacientes com linfoma de células B grandes em duas categorias: evolução ao óbito ou sobrevivência.

DLBCL-SHI2002 e DLBCL-OUTCOME-SHI2002 (Shipp et al. 2002)

- Descrição: Perfil de expressão gênica por microarray. Estudo de linfoma de células B.
- Tarefa: DLBCL-TUMOR – classificação da morfologia das amostras em linfoma de células B e linfoma folicular; DLBCL-OUTCOME – classificação de pacientes com linfoma de células B quanto ao resultado clínico da doença, *i.e.*, curados ou fatais (entre os fatais, na verdade, há pacientes com recorrência da doença).

ECML2004

- Descrição: Perfil de expressão gênica por SAGE. Estudo global do câncer por meio de SAGE. Esta foi a base utilizada no desafio promovido pelo ECML 2004 (*European Conference on Machine Learning*).
- Tarefa: Diferenciar entre mais de vinte tipos de câncer.

GCM-RAM2001 (Ramaswamy et al. 2001)

- Descrição: Perfil de expressão gênica por microarray. Mapa global do câncer por microarray. Verificação da viabilidade da utilização de um conjunto de dados multi-classes no diagnóstico de câncer a partir da análise molecular. Foco em estágios primários da doença.
- Tarefa: Diferenciar entre quatorze tipos de câncer.

GSE360 (Chaussabel et al. 2003)

- Descrição: Perfil de expressão gênica por microarray. Macrófagos e células dendríticas humanas expostas a parasitas distintos filogeneticamente. As células foram expostas a cinco diferentes patógenos, *i.e.*, *M. tuberculosis*, *L. major*, *L. donovani*, *T. gondii* e *B. malayi*.
- Tarefa: Classificar cada célula considerando seu tipo (macrófagos ou células dendríticas) e os patógenos a que foram expostas;

GSE443 (Bleharski et al. 2003)

- Descrição: Perfil de expressão gênica por microarray. Estudo de lesão causada por hanseníase.
- Tarefa: Classificar pacientes, distinguindo entre duas formas clínicas da doença.

GSE474 (Park et al. 2006)

- Descrição: Perfil de expressão gênica por microarray. Estudo de obesidade e oxidação de gordura, na tentativa de identificar os mRNAs de proteínas envolvidas no processo.
- Tarefa: Classificação de pacientes em obesidade mórbida, obesidade e não obesidade.

GSE3255 (Lee et al. 2006)

- Descrição: Perfil de expressão gênica por SAGE. Estudo de translocações em leucemia mieloide aguda. Foram considerados pacientes com t(8;21), t(15;17), t(9;11) e inv(16).
- Tarefa: Distinguir as bibliotecas quanto às translocações consideradas no estudo.

GSE5473 (Chari et al. 2007)

- Descrição: Perfil de expressão gênica por SAGE. Estudo do efeito do ato de fumar ativo na expressão gênica do epitélio dos brônquios.
- Tarefa: Distinguir entre fumantes, ex-fumantes e nunca fumantes

GSE7433 (Shadeo et al. 2008)

- Descrição: Perfil de expressão gênica por SAGE. Estudo do tecido cervical.
- Tarefa: Distinguir entre tecido normal, displasia severa e displasia moderada.

GSE7898 (Lonergan et al. 2010)

- Descrição: Perfil de expressão gênica por SAGE. Estudo de estágio inicial de carcinoma de células escamosas de pulmão baseada em perfis globais de expressão representando estágios progressivos do desenvolvimento do câncer (precanceroso, carcinoma *in-situ* e câncer invasivo).
- Tarefa: Diferenciar os estágios de desenvolvimento do câncer (precanceroso, carcinoma *in-situ* e câncer invasivo).

GSE11665 (Boon et al. 2009)

- Descrição: Perfil de expressão gênica por SAGE. Estudo da assinatura molecular para progressão de doença em Fibrose Pulmonar Idiopática. Busca por biomarcadores relacionados à doença.
- Tarefa: Classificar paciente como tendo ou não a doença.

BREAST-HED2003 (Hedenfalk et al. 2003)

- Descrição: Perfil de expressão gênica por microarray. Descoberta de classes novas entre os tumores BRCAx.
- Tarefa: Diferenciar entre tumores BRCA1 e BRCA2.

GASTRIC-LEU2002 (Leung et al. 2002)

- Descrição: Perfil de expressão gênica por microarray. Estudo de adenocarcinoma gástrico em diferentes estágios da doença.
- Tarefa: Classificar os pacientes em: tumor gástrico primário, com metástase e mucosa normal.

LUNG-GOR2002 (Gordon et al. 2002)

- Descrição: Perfil de expressão gênica por microarray. Estudo de câncer de pulmão.
- Tarefa: Classificar entre mesotelioma pleural maligno e adenocarcinoma do pulmão.

LUNG-WIG2002 (Wigle et al. 2002)

- Descrição: Perfil de expressão gênica por microarray. Estudo de câncer de pulmão de células não pequenas (NSCLC).
- Tarefa: Pacientes que tiveram tumor devem ser classificados como tendo tido metástase local ou distante, ou como livre da doença.

LUNG-BHA2001 (Bhattacharjee et al. 2001)

- Descrição: Perfil de expressão gênica por microarray. Estudo de diferentes tumores relacionados ao pulmão.
- Tarefa: Diferenciar os pacientes entre os diferentes tipos de tumores e também pacientes normais.

LUNG-BEE2002 (Beer et al. 2002)

- Descrição: Perfil de expressão gênica por microarray. Estudo de adenocarcinoma primário de pulmão.
- Tarefa: Classificar os pacientes em tendo ou não o tumor.

BREAST-MA2003 (Ma et al. 2003)

- Descrição: Perfil de expressão gênica por microarray. Estudo de câncer de mama em seus diversos estados patológicos.
- Tarefa: Diferenciar pacientes em carcinoma ductal *in-situ*, carcinoma ductal invasivo e hiperplasia ductal atípica.

MLL-ARM2002 (Armstrong et al. 2002)

- Descrição: Perfil de expressão gênica por microarray. Estudo de leucemia em que os autores inclusive sugerem uma forma distinta de leucemia linfoblástica, que chamaram de MLL, que possui translocações do gene MLL.
- Tarefa: Diferenciar pacientes em tendo leucemia mieloide aguda, leucemia linfocítica aguda e leucemia de linhagem misturada (MLL).

SOFT-NIE2002 (Nielsen et al. 2002)

- Descrição: Perfil de expressão gênica por microarray. Estudo de tumores de tecidos moles. Busca por marcadores diagnósticos desses neoplasmas.
- Tarefa: Diferenciar entre tipos de câncer: sarcoma sinovial, tumor estromal gastrointestinal, leiomiossarcoma, lipossarcoma, histiocitoma fibroso maligno e Schwannoma.

OVARIAN-PET2002 (Petricoin et al. 2002)

- Descrição: Perfil de expressão gênica por microarray. Identificação de padrões proteômicos *in serum* no estudo de câncer de ovário. Estudo significante para mulheres com alto risco de câncer de ovário devido a histórico pessoal ou familiar de câncer.
- Tarefa: Distinguir entre pacientes com câncer e pacientes normais.

OVARY-WEL2001 (Welsh et al. 2001)

- Descrição: Perfil de expressão gênica por microarray. Estudo de câncer ovariano epitelial, buscando por marcadores moleculares da doença.
- Tarefa: Diferenciar entre pacientes com tumor e normais.

CNS-POM2002 (Pomeroy et al. 2002)

- Descrição: Perfil de expressão gênica por microarray. Estudo de tumores do sistema nervoso central.
- Tarefa: Resposta ao tratamento de pacientes com meduloblastoma, *i.e.*, pacientes que sobreviveram ao tratamento e pacientes que não sobreviveram.

PROSTATE-SIN2002 e PROSTATE-OUTCOME-SIN2002 (Singh et al. 2002)

- Descrição: Perfil de expressão gênica por microarray. Estudo de câncer de próstata.
- Tarefa: PROSTATE-TUMOR – classificação das amostras em duas classes, *i.e.*, normal e tumor; PROSTATE-OUTCOME – classificação do resultado clínico em pacientes com tumor. Os pacientes foram avaliados com relação à recorrência da doença após cirurgia em um período de quatro anos.

PROSTATE-WEL2001 (Welsh et al. 2001)

- Descrição: Perfil de expressão gênica por microarray. Estudo do comportamento e desenvolvimento de câncer de próstata, focando na caracterização do câncer de próstata primário.
- Tarefa: Classificação de amostras de tecido da próstata e linhagens celulares.

Tabela B.1: *Informações das bases de dados utilizadas nos experimentos. Na coluna N, está indicado o número de exemplos de cada base; na coluna c, está indicado o número de classes distintas existentes; ATRI, $a_{\#}$ e a_a representam o número total de atributos e o número de atributos numéricos e nominais, respectivamente; AUSE representa a porcentagem de atributos com valores ausentes, não considerando o atributo classe; na penúltima coluna (WIN), indicam-se quais bases de dados foram usadas nos experimentos de windowing; na última coluna (LOOK), indicam-se quais bases de dados foram usadas nos experimentos de lookahead.*

Base	N	c	ATRI($a_{\#}$,a_a)	AUSE	WIN	LOOK
LYMPHOMA-ALI2000	96	9	4026 (4026, 0)	5,09%	♦	♦
GIST-ALL2001	19	2	1987 (1987, 0)	0,00%	♦	♦
COLON-ALO1999	62	2	2000 (2000, 0)	0,00%	♦	♦
LEUKEMIA-GOL1999	72	2	7129 (7129, 0)	0,00%	♦	♦
BCR-ABL-YEO2002	327	2	12558 (12558, 0)	0,00%	♦	
MELANOMA-BIT2000	38	3	8067 (8067, 0)	0,00%	♦	♦
BREAST-VEE2002	97	2	24481 (24481, 0)	0,00%	♦	
NETWORKS-BUT2000	68	9	7245 (7245, 0)	0,00%	♦	
DLBCL-NIH-ROS2002	240	2	7399 (7399, 0)	10,30%	♦	
DLBCL-OUTCOME-SHI2002	58	2	7129 (7129, 0)	0,00%	♦	♦
DLBCL-SHI2002	77	2	7129 (7129, 0)	0,00%	♦	
E2A-PBX1-YEO2002	327	2	12558 (12558, 0)	0,00%	♦	
ECML2004	72	25	27679 (27679, 0)	0,00%	♦	
GCM-RAM2001	190	14	16063 (16063, 0)	0,00%	♦	
GSE11665	14	2	22721 (22721, 0)	0,00%	♦	♦
GSE3255	22	4	26670 (26670, 0)	0,00%	♦	♦
GSE360	28	12	12625 (12625, 0)	0,00%	♦	♦
GSE443	11	2	12625 (12625, 0)	0,00%	♦	♦
GSE474	24	3	22283 (22283, 0)	0,00%	♦	♦
GSE5473	24	3	32810 (32810, 0)	0,00%	♦	♦
GSE7433	16	3	44259 (44259, 0)	0,00%	♦	♦
GSE7898	11	2	34988 (34988, 0)	0,00%	♦	♦
BREAST-HED2003	16	2	4795 (4795, 0)	0,00%	♦	♦
HYPERDIP-50-YEO2002	327	2	12558 (12558, 0)	0,00%	♦	
GASTRIC-LEU2002	126	3	6688 (6688, 0)	5,87%	♦	
LUNG-BHA2001	203	5	12600 (12600, 0)	0,00%	♦	
LUNG-BEE2002	96	2	7129 (7129, 0)	0,00%	♦	
LUNG-GOR2002	181	2	12533 (12533, 0)	0,00%	♦	
LUNG-WIG2002	39	2	2880 (2880, 0)	5,96%	♦	♦
BREAST-MA2003	61	3	1946 (1941, 5)	0,30%	♦	♦
MLL-ARM2002	72	3	12582 (12582, 0)	0,00%	♦	
MLL-YEO2002	327	2	12558 (12558, 0)	0,00%	♦	
SOFT-NIE2002	46	6	5520 (5520, 0)	6,72%	♦	♦
OVARIAN-PET2002	253	2	15154 (15154, 0)	0,00%	♦	
OVARY-WEL2001	38	2	7129 (7129, 0)	0,00%	♦	♦
CNS-POM2002	60	2	7129 (7129, 0)	0,00%	♦	♦
PROSTATE-OUTCOME-SIN2002	21	2	12600 (12600, 0)	0,00%	♦	♦
PROSTATE-SIN2002	136	2	12600 (12600, 0)	0,00%	♦	
T-ALL-YEO2002	327	2	12558 (12558, 0)	0,00%	♦	
TEL-AML1-YEO2002	327	2	12558 (12558, 0)	0,00%	♦	
PROSTATE-WEL2001	48	7	12626 (12626, 0)	0,00%	♦	

Apêndice

C

Resultados Originais

Neste apêndice, são mostradas as tabelas com os resultados originais da aplicação dos algoritmos nas bases de dados utilizadas. Cada valor representa, na verdade, a média (ou desvio padrão) dos valores obtidos por meio de validação cruzada de dez partições.

Tabela C.1: *Windowing: Média dos valores originais da variável Altura da Árvore obtidos por validação cruzada de dez partições.*

Base	J48	WPEWeC	WPEC	WEWeC	WEC	WPWeC	WPC	WWeC	WC	WPEWe	WPE	WEWe	WE	WPWe	WP	WWe	W
LYMPHOMA-ALI2000	5,00000	5,40000	5,40000	5,00000	5,00000	5,40000	5,40000	5,00000	5,00000	5,30000	5,40000	5,40000	5,40000	5,40000	5,50000	5,40000	5,40000
GIST-ALL2001	2,00000	2,00000	2,00000	2,00000	2,00000	2,00000	2,00000	2,00000	2,00000	2,00000	2,00000	2,00000	2,00000	2,00000	2,00000	2,00000	2,00000
COLON-ALO1999	4,00000	3,70000	3,80000	3,10000	3,40000	4,00000	4,00000	3,30000	3,30000	3,90000	4,00000	3,90000	3,90000	4,00000	4,00000	4,00000	4,00000
LEUKEMIA-GOL1999	2,70000	2,30000	2,50000	2,00000	2,00000	2,90000	2,90000	2,00000	2,00000	2,80000	2,90000	2,70000	2,90000	2,90000	2,90000	2,90000	2,90000
BCR-ABL-YEO2002	4,00000	3,70000	3,70000	3,10000	3,20000	4,10000	4,20000	3,30000	3,30000	3,80000	3,80000	3,70000	3,70000	4,20000	4,20000	3,90000	3,90000
MELANOMA-BIT2000	3,00000	3,00000	3,00000	3,00000	3,00000	3,00000	3,00000	3,00000	3,00000	3,00000	3,00000	3,00000	3,00000	3,00000	3,00000	3,00000	3,00000
BREAST-VEE2002	6,10000	5,60000	5,50000	3,20000	3,20000	6,00000	6,10000	3,20000	3,30000	5,80000	5,60000	5,50000	5,60000	6,30000	6,30000	5,70000	5,70000
NETWORKS-BUT2000	8,10000	7,80000	7,80000	6,00000	6,30000	8,10000	8,10000	6,30000	6,30000	7,50000	7,60000	7,50000	7,50000	7,70000	7,70000	7,60000	7,60000
DLBCL-NIH-ROS2002	13,40000	13,10000	13,00000	12,30000	12,50000	13,80000	13,60000	13,80000	13,80000	13,40000	13,60000	13,40000	13,40000	14,40000	14,40000	13,40000	13,40000
DLBCL-OUT-SHI2002	4,40000	4,40000	4,40000	2,70000	2,80000	4,60000	4,60000	2,80000	2,90000	4,50000	4,60000	4,40000	4,40000	4,60000	4,60000	4,50000	4,50000
DLBCL-SHI2002	3,10000	3,00000	3,10000	2,00000	2,00000	3,10000	3,10000	2,10000	2,10000	3,00000	3,00000	3,00000	3,00000	3,10000	3,10000	3,00000	3,00000
E2A-PBX1-YEO2002	2,00000	2,00000	2,00000	2,00000	2,00000	2,00000	2,00000	2,00000	2,00000	2,00000	2,00000	2,00000	2,00000	2,00000	2,00000	2,00000	2,00000
ECML2004	20,80000	20,40000	20,40000	18,30000	18,60000	20,40000	20,50000	18,60000	18,60000	20,60000	20,60000	20,60000	20,60000	20,70000	20,80000	20,60000	20,60000
GCM-RAM2001	13,70000	13,40000	13,30000	11,50000	11,50000	13,40000	13,60000	11,50000	11,50000	14,00000	13,90000	13,70000	13,30000	14,20000	14,20000	14,00000	14,00000
GSE11665	1,80000	1,80000	1,80000	1,80000	1,80000	1,80000	1,80000	1,80000	1,80000	1,80000	1,80000	1,80000	1,80000	1,80000	1,80000	1,80000	1,80000
GSE3255	4,00000	4,00000	4,00000	4,00000	4,00000	4,00000	4,00000	4,00000	4,00000	4,00000	4,00000	4,00000	4,00000	4,00000	4,00000	4,00000	4,00000
GSE360	4,90000	4,90000	5,00000	4,10000	4,20000	5,00000	5,00000	4,10000	4,10000	4,70000	4,90000	4,50000	4,90000	5,00000	4,90000	5,00000	5,00000
GSE443	2,00000	2,00000	2,00000	2,00000	2,00000	2,00000	2,00000	2,00000	2,00000	2,00000	2,00000	2,00000	2,00000	2,00000	2,00000	2,00000	2,00000
GSE474	3,00000	3,00000	3,00000	3,00000	3,00000	3,00000	3,00000	3,00000	3,00000	3,00000	3,00000	3,00000	3,00000	3,00000	3,00000	3,00000	3,00000
GSE5473	3,00000	3,00000	3,00000	3,00000	3,00000	3,00000	3,00000	3,00000	3,00000	3,00000	3,00000	3,00000	3,00000	3,00000	3,00000	3,00000	3,00000
GSE7433	3,00000	3,00000	3,00000	3,00000	3,00000	3,00000	3,00000	3,00000	3,00000	3,00000	3,00000	3,00000	3,00000	3,00000	3,00000	3,00000	3,00000
GSE7898	2,00000	2,00000	2,00000	2,00000	2,00000	2,00000	2,00000	2,00000	2,00000	2,00000	2,00000	2,00000	2,00000	2,00000	2,00000	2,00000	2,00000
BREAST-HED2003	2,00000	2,00000	2,00000	2,00000	2,00000	2,00000	2,00000	2,00000	2,00000	2,00000	2,00000	2,00000	2,00000	2,00000	2,00000	2,00000	2,00000
HYPERDIP-50-YEO2002	5,90000	5,20000	5,20000	4,60000	4,90000	5,90000	5,80000	4,80000	4,80000	5,30000	5,20000	4,90000	5,10000	6,20000	6,10000	5,40000	5,40000
GASTRIC-LEU2002	3,70000	3,50000	3,60000	3,30000	3,30000	3,80000	3,80000	3,30000	3,30000	3,70000	3,70000	3,80000	3,80000	3,70000	3,70000	3,90000	3,90000
LUNG-BHA2001	5,00000	5,00000	5,00000	4,70000	4,70000	5,00000	5,00000	4,70000	4,70000	5,00000	5,00000	5,00000	5,00000	5,00000	5,00000	5,10000	5,10000
LUNG-BEE2002	2,00000	2,00000	2,00000	2,00000	2,00000	2,00000	2,00000	2,00000	2,00000	2,00000	2,00000	2,00000	2,00000	2,00000	2,00000	2,00000	2,00000
LUNG-GOR2002	3,80000	2,60000	3,00000	2,00000	2,00000	3,50000	3,50000	2,00000	2,00000	2,80000	3,00000	2,40000	2,80000	3,20000	3,20000	3,20000	3,20000
LUNG-WIG2002	3,00000	3,00000	3,00000	2,90000	2,90000	3,00000	3,10000	3,00000	3,00000	3,00000	3,00000	3,00000	3,00000	3,00000	3,00000	3,20000	3,20000
BREAST-MA2003	6,30000	5,70000	6,20000	3,90000	4,30000	6,20000	6,50000	4,60000	4,60000	6,10000	6,10000	6,20000	6,20000	6,30000	6,30000	6,10000	6,10000
MLL-ARM2002	3,00000	3,00000	3,00000	3,00000	3,00000	3,00000	3,00000	3,00000	3,00000	3,00000	3,00000	3,00000	3,00000	3,00000	3,00000	3,00000	3,00000
MLL-YEO2002	4,00000	3,90000	4,00000	3,20000	3,30000	4,40000	4,40000	3,30000	3,30000	4,00000	4,20000	4,20000	4,20000	4,30000	4,30000	4,20000	4,20000
SOFT-NIE2002	5,10000	4,60000	4,60000	4,00000	4,00000	4,60000	4,60000	4,00000	4,00000	4,90000	4,90000	4,90000	4,90000	4,90000	4,90000	4,90000	4,90000
OVARIAN-PET2002	4,00000	3,00000	3,00000	2,80000	2,80000	3,10000	3,10000	2,90000	2,90000	3,00000	3,00000	3,00000	3,00000	3,00000	3,00000	3,00000	3,00000
OVARY-WEL2001	2,00000	2,00000	2,00000	2,00000	2,00000	2,00000	2,00000	2,00000	2,00000	2,00000	2,00000	2,00000	2,00000	2,00000	2,00000	2,00000	2,00000
CNS-POM2002	4,50000	4,30000	4,40000	2,70000	2,90000	4,50000	4,70000	2,90000	2,90000	4,50000	4,50000	4,40000	4,50000	4,60000	4,80000	4,50000	4,50000
PROSTATE-OUT-SIN2002	2,50000	2,00000	2,00000	2,00000	2,00000	2,00000	2,00000	2,00000	2,00000	2,30000	2,50000	2,10000	2,50000	2,50000	2,50000	2,50000	2,50000
PROSTATE-SIN2002	5,50000	5,40000	5,50000	3,40000	3,50000	5,60000	5,60000	3,60000	3,60000	5,50000	5,20000	5,00000	5,00000	5,50000	5,50000	5,50000	5,50000
T-ALL-YEO2002	2,00000	2,00000	2,00000	2,00000	2,00000	2,00000	2,00000	2,00000	2,00000	2,00000	2,00000	2,00000	2,00000	2,00000	2,00000	2,00000	2,00000
TEL-AML1-YEO2002	4,30000	3,70000	3,90000	3,10000	3,20000	4,20000	4,20000	3,60000	3,60000	3,60000	3,90000	3,70000	3,70000	4,10000	4,10000	3,90000	3,90000
PROSTATE-WEL2001	5,00000	4,60000	4,60000	4,50000	4,70000	4,70000	4,70000	4,70000	4,70000	4,60000	4,80000	4,80000	4,80000	4,80000	4,90000	4,80000	4,80000

Tabela C.2: *Windowing: Desvio padrão dos valores originais da variável Altura da Árvore obtidos por validação cruzada de dez partições.*

Base	J48	WPEWeC	WPEC	WEWeC	WEC	WPWeC	WPC	WWeC	WC	WPEWe	WPE	WEWe	WE	WPWe	WP	WWe	W
LYMPHOMA-ALI2000	0,000	0,163	0,163	0,000	0,000	0,163	0,163	0,000	0,000	0,153	0,163	0,163	0,163	0,163	0,167	0,163	0,163
GIST-ALL2001	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000
COLON-ALO1999	0,000	0,153	0,133	0,180	0,163	0,000	0,000	0,153	0,153	0,100	0,000	0,100	0,100	0,000	0,000	0,000	0,000
LEUKEMIA-GOL1999	0,153	0,153	0,167	0,000	0,000	0,100	0,100	0,000	0,000	0,133	0,100	0,153	0,100	0,100	0,100	0,100	0,100
BCR-ABL-YEO2002	0,000	0,153	0,153	0,100	0,133	0,100	0,133	0,153	0,153	0,133	0,133	0,153	0,153	0,133	0,133	0,100	0,100
MELANOMA-BIT2000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000
BREAST-VEE2002	0,314	0,305	0,269	0,291	0,291	0,258	0,277	0,291	0,260	0,133	0,163	0,224	0,163	0,153	0,153	0,153	0,153
NETWORKS-BUT2000	0,233	0,327	0,327	0,258	0,213	0,233	0,233	0,213	0,213	0,269	0,267	0,269	0,224	0,213	0,300	0,305	0,305
DLBCL-NIH-ROS2002	0,600	0,458	0,471	0,597	0,619	0,574	0,542	0,646	0,646	0,400	0,427	0,452	0,452	0,600	0,600	0,452	0,452
DLBCL-OUTCOME-SHI2002	0,163	0,163	0,163	0,153	0,133	0,163	0,163	0,133	0,100	0,167	0,163	0,163	0,163	0,163	0,163	0,167	0,167
DLBCL-SHI2002	0,100	0,000	0,100	0,000	0,000	0,100	0,100	0,100	0,100	0,000	0,000	0,000	0,000	0,100	0,100	0,000	0,000
E2A-PBX1-YEO2002	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000
ECML2004	0,200	0,221	0,221	0,300	0,267	0,221	0,167	0,267	0,267	0,163	0,163	0,163	0,163	0,153	0,200	0,163	0,163
GCM-RAM2001	0,472	0,305	0,260	0,453	0,453	0,400	0,371	0,453	0,453	0,422	0,314	0,517	0,423	0,359	0,359	0,494	0,494
GSE11665	0,133	0,133	0,133	0,133	0,133	0,133	0,133	0,133	0,133	0,133	0,133	0,133	0,133	0,133	0,133	0,133	0,133
GSE3255	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000
GSE360	0,100	0,100	0,000	0,180	0,133	0,000	0,000	0,100	0,100	0,153	0,100	0,224	0,100	0,000	0,100	0,000	0,000
GSE443	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000
GSE474	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000
GSE5473	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000
GSE7433	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000
GSE7898	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000
BREAST-HED2003	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000
HYPERDIP-50-YEO2002	0,233	0,200	0,200	0,267	0,233	0,277	0,291	0,200	0,200	0,260	0,291	0,314	0,314	0,250	0,277	0,267	0,267
GASTRIC-LEU2002	0,153	0,167	0,163	0,153	0,153	0,200	0,200	0,153	0,153	0,153	0,153	0,133	0,133	0,153	0,153	0,100	0,100
LUNG-BHA2001	0,000	0,000	0,000	0,153	0,153	0,000	0,000	0,153	0,153	0,000	0,000	0,000	0,000	0,000	0,000	0,100	0,100
LUNG-BEE2002	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000
LUNG-GOR2002	0,200	0,163	0,258	0,000	0,000	0,224	0,224	0,000	0,000	0,133	0,149	0,163	0,200	0,133	0,133	0,133	0,133
LUNG-WIG2002	0,000	0,000	0,000	0,100	0,100	0,000	0,100	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,133	0,133
BREAST-MA2003	0,213	0,335	0,200	0,314	0,367	0,250	0,167	0,427	0,427	0,180	0,180	0,200	0,200	0,213	0,213	0,180	0,180
MLL-ARM2002	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000
MLL-YEO2002	0,000	0,180	0,149	0,133	0,153	0,163	0,163	0,153	0,153	0,211	0,133	0,133	0,133	0,153	0,153	0,133	0,133
SOFT-NIE2002	0,180	0,163	0,163	0,000	0,000	0,163	0,163	0,000	0,000	0,100	0,100	0,100	0,100	0,100	0,100	0,100	0,100
OVARIAN-PET2002	0,000	0,000	0,000	0,133	0,133	0,100	0,100	0,100	0,100	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000
OVARY-WEL2001	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000
CNS-POM2002	0,167	0,153	0,163	0,213	0,180	0,167	0,153	0,180	0,180	0,167	0,167	0,221	0,167	0,163	0,133	0,167	0,167
PROSTATE-OUTCOME-SIN2002	0,167	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,153	0,167	0,100	0,167	0,167	0,167	0,167	0,167
PROSTATE-SIN2002	0,224	0,267	0,269	0,305	0,269	0,267	0,267	0,221	0,221	0,269	0,291	0,298	0,298	0,269	0,269	0,269	0,269
T-ALL-YEO2002	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000
TEL-AML1-YEO2002	0,153	0,153	0,100	0,100	0,200	0,133	0,133	0,163	0,163	0,163	0,100	0,153	0,153	0,100	0,100	0,100	0,100
PROSTATE-WEL2001	0,149	0,163	0,163	0,167	0,153	0,153	0,153	0,153	0,153	0,163	0,200	0,200	0,200	0,200	0,180	0,200	0,200

Tabela C.3: *Windowing: Média dos valores originais da variável* Tamanho da Árvore *obtidos por validação cruzada de dez partições.*

Base	J48	WPEWeC	WPEC	WEWeC	WEC	WPWeC	WPC	WWeC	WC	WPEWe	WPE	WEWe	WE	WPWe	WP	WWe	W
LYMPHOMA-ALI2000	15,00000	16,00000	16,00000	16,20000	16,20000	16,00000	16,00000	16,20000	16,20000	15,60000	15,80000	15,80000	15,80000	16,00000	15,80000	16,00000	16,00000
GIST-ALL2001	3,00000	3,00000	3,00000	3,00000	3,00000	3,00000	3,00000	3,00000	3,00000	3,00000	3,00000	3,00000	3,00000	3,00000	3,00000	3,00000	3,00000
COLON-ALO1999	7,00000	6,40000	6,60000	5,20000	5,80000	7,00000	7,00000	5,80000	5,80000	6,80000	7,00000	6,80000	6,80000	7,00000	7,00000	7,00000	7,00000
LEUKEMIA-GOL1999	4,40000	3,60000	4,00000	3,00000	3,00000	4,80000	4,80000	3,00000	3,00000	4,60000	4,80000	4,40000	4,80000	4,80000	4,80000	4,80000	4,80000
BCR-ABL-YEO2002	9,00000	7,40000	7,60000	5,20000	5,40000	9,00000	9,40000	5,60000	5,60000	7,20000	7,20000	7,20000	7,20000	9,40000	9,40000	7,40000	7,40000
MELANOMA-BIT2000	5,00000	5,00000	5,00000	5,00000	5,00000	5,00000	5,00000	5,00000	5,00000	5,00000	5,00000	5,00000	5,00000	5,00000	5,00000	5,00000	5,00000
BREAST-VEE2002	11,60000	10,60000	10,80000	5,60000	5,60000	11,40000	11,60000	5,60000	6,00000	11,00000	11,00000	10,80000	11,00000	11,60000	11,60000	11,00000	11,00000
NETWORKS-BUT2000	19,60000	17,20000	17,20000	16,40000	16,60000	18,00000	18,00000	16,60000	16,60000	18,80000	19,20000	19,00000	19,20000	19,40000	19,80000	19,60000	19,60000
DLBCL-NIH-ROS2002	31,80000	30,80000	30,80000	29,80000	30,00000	32,80000	33,40000	31,60000	31,60000	31,20000	31,00000	31,00000	31,00000	32,40000	32,40000	31,20000	31,20000
DLBCL-OUTCOME-SHI2002	8,20000	8,40000	8,60000	4,40000	4,60000	8,80000	8,80000	4,60000	4,80000	8,40000	8,80000	8,00000	8,20000	8,80000	8,80000	8,40000	8,40000
DLBCL-SHI2002	6,20000	5,00000	5,40000	3,00000	3,00000	5,40000	5,40000	3,20000	3,20000	5,40000	5,80000	5,20000	5,60000	5,80000	5,80000	5,80000	5,80000
E2A-PBX1-YEO2002	3,00000	3,00000	3,00000	3,00000	3,00000	3,00000	3,00000	3,00000	3,00000	3,00000	3,00000	3,00000	3,00000	3,00000	3,00000	3,00000	3,00000
ECML2004	44,20000	43,80000	43,80000	37,00000	37,80000	44,00000	44,00000	37,80000	37,80000	44,40000	44,40000	44,40000	44,40000	44,40000	44,40000	44,60000	44,60000
GCM-RAM2001	37,80000	36,00000	36,40000	29,80000	29,80000	37,60000	37,60000	29,80000	29,80000	37,60000	37,60000	38,00000	37,60000	38,60000	38,60000	38,00000	38,00000
GSE11665	2,60000	2,60000	2,60000	2,60000	2,60000	2,60000	2,60000	2,60000	2,60000	2,60000	2,60000	2,60000	2,60000	2,60000	2,60000	2,60000	2,60000
GSE3255	7,00000	7,00000	7,00000	7,00000	7,00000	7,00000	7,00000	7,00000	7,00000	7,00000	7,00000	7,00000	7,00000	7,00000	7,00000	7,00000	7,00000
GSE360	16,80000	16,80000	17,20000	11,80000	14,40000	17,20000	17,20000	14,40000	14,40000	16,20000	17,00000	15,00000	17,00000	17,20000	17,00000	17,20000	17,20000
GSE443	3,00000	3,00000	3,00000	3,00000	3,00000	3,00000	3,00000	3,00000	3,00000	3,00000	3,00000	3,00000	3,00000	3,00000	3,00000	3,00000	3,00000
GSE474	5,00000	5,00000	5,00000	5,00000	5,00000	5,00000	5,00000	5,00000	5,00000	5,00000	5,00000	5,00000	5,00000	5,00000	5,00000	5,00000	5,00000
GSE5473	5,00000	5,00000	5,00000	5,00000	5,00000	5,00000	5,00000	5,00000	5,00000	5,00000	5,00000	5,00000	5,00000	5,00000	5,00000	5,00000	5,00000
GSE7433	5,00000	5,00000	5,00000	5,00000	5,00000	5,00000	5,00000	5,00000	5,00000	5,00000	5,00000	5,00000	5,00000	5,00000	5,00000	5,00000	5,00000
GSE7898	3,00000	3,00000	3,00000	3,00000	3,00000	3,00000	3,00000	3,00000	3,00000	3,00000	3,00000	3,00000	3,00000	3,00000	3,00000	3,00000	3,00000
BREAST-HED2003	3,00000	3,00000	3,00000	3,00000	3,00000	3,00000	3,00000	3,00000	3,00000	3,00000	3,00000	3,00000	3,00000	3,00000	3,00000	3,00000	3,00000
HYPERDIP-50-YEO2002	17,00000	14,80000	14,80000	11,60000	12,80000	17,40000	17,20000	13,20000	13,20000	15,00000	14,80000	13,60000	14,60000	18,20000	18,00000	15,20000	15,20000
GASTRIC-LEU2002	8,40000	7,20000	7,80000	5,60000	5,60000	8,20000	8,20000	5,60000	5,60000	8,20000	8,40000	8,00000	8,20000	8,40000	8,40000	8,60000	8,60000
LUNG-BHA2001	11,00000	11,00000	11,00000	10,40000	10,40000	11,00000	11,00000	10,40000	10,40000	11,00000	11,00000	11,00000	11,00000	11,00000	11,00000	11,20000	11,20000
LUNG-BEE2002	3,00000	3,00000	3,00000	3,00000	3,00000	3,00000	3,00000	3,00000	3,00000	3,00000	3,00000	3,00000	3,00000	3,00000	3,00000	3,00000	3,00000
LUNG-GOR2002	6,60000	4,20000	5,00000	3,00000	3,00000	6,00000	6,00000	3,00000	3,00000	4,60000	5,00000	3,80000	4,60000	5,40000	5,40000	5,40000	5,40000
LUNG-WIG2002	5,00000	5,00000	5,00000	4,80000	4,80000	5,00000	5,20000	5,00000	5,00000	5,00000	5,00000	5,00000	5,00000	5,00000	5,00000	5,40000	5,40000
BREAST-MA2003	12,80000	11,40000	12,60000	8,00000	8,40000	13,20000	13,00000	8,80000	8,80000	12,40000	12,80000	13,00000	13,00000	13,40000	13,40000	12,80000	12,80000
MLL-ARM2002	5,00000	5,00000	5,00000	5,00000	5,00000	5,00000	5,00000	5,00000	5,00000	5,00000	5,00000	5,00000	5,00000	5,00000	5,00000	5,00000	5,00000
MLL-YEO2002	9,00000	7,60000	8,20000	5,40000	6,00000	9,20000	9,20000	6,00000	6,00000	8,20000	9,20000	9,00000	9,20000	9,40000	9,40000	9,20000	9,20000
SOFT-NIE2002	10,60000	10,60000	10,60000	10,60000	10,60000	10,60000	10,60000	10,60000	10,60000	10,60000	10,60000	10,60000	10,60000	10,60000	10,60000	10,60000	10,60000
OVARIAN-PET2002	10,20000	5,20000	5,20000	4,60000	4,60000	6,00000	6,00000	4,80000	4,80000	5,00000	5,00000	5,00000	5,00000	5,60000	5,60000	5,00000	5,00000
OVARY-WEL2001	3,00000	3,00000	3,00000	3,00000	3,00000	3,00000	3,00000	3,00000	3,00000	3,00000	3,00000	3,00000	3,00000	3,00000	3,00000	3,00000	3,00000
CNS-POM2002	8,00000	7,60000	7,80000	4,40000	4,80000	8,00000	8,40000	5,00000	5,00000	8,00000	8,00000	7,80000	8,00000	8,20000	8,60000	8,00000	8,00000
PROSTATE-OUTCOME-SIN2002	4,00000	3,00000	3,00000	3,00000	3,00000	3,00000	3,00000	3,00000	3,00000	3,60000	4,00000	3,20000	4,00000	4,00000	4,00000	4,00000	4,00000
PROSTATE-SIN2002	11,80000	11,60000	11,80000	6,00000	6,20000	12,00000	12,00000	6,40000	6,40000	11,80000	11,40000	11,00000	11,00000	11,80000	11,80000	11,80000	11,80000
T-ALL-YEO2002	3,00000	3,00000	3,00000	3,00000	3,00000	3,00000	3,00000	3,00000	3,00000	3,00000	3,00000	3,00000	3,00000	3,00000	3,00000	3,00000	3,00000
TEL-AML1-YEO2002	10,00000	7,20000	7,80000	5,80000	6,20000	9,80000	9,80000	7,00000	7,00000	7,60000	8,60000	8,20000	8,80000	9,20000	9,60000	9,20000	9,20000
PROSTATE-WEL2001	12,20000	12,20000	12,20000	11,60000	11,80000	12,20000	12,20000	11,80000	11,80000	12,20000	12,20000	12,20000	12,20000	12,20000	12,20000	12,20000	12,20000

Tabela C.4: *Windowing: Desvio padrão dos valores originais da variável Tamanho da Árvore obtidos por validação cruzada de dez partições.*

Base	J48	WPEWeC	WPEC	WEWeC	WEC	WPWeC	WPC	WWeC	WC	WPEWe	WPE	WEWe	WE	WPWe	WP	WWe	W
LYMPHOMA-ALI2000	0,422	0,333	0,333	0,327	0,327	0,333	0,333	0,327	0,327	0,427	0,327	0,327	0,327	0,333	0,327	0,333	0,333
GIST-ALL2001	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000
COLON-ALO1999	0,000	0,305	0,267	0,359	0,327	0,000	0,000	0,327	0,327	0,200	0,000	0,200	0,200	0,000	0,000	0,000	0,000
LEUKEMIA-GOL1999	0,305	0,305	0,333	0,000	0,000	0,200	0,200	0,000	0,000	0,267	0,200	0,305	0,200	0,200	0,200	0,200	0,200
BCR-ABL-YEO2002	0,000	0,581	0,600	0,200	0,267	0,298	0,499	0,305	0,305	0,467	0,467	0,554	0,554	0,499	0,499	0,400	0,400
MELANOMA-BIT2000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000
BREAST-VEE2002	0,427	0,400	0,200	0,733	0,733	0,267	0,305	0,733	0,683	0,000	0,000	0,200	0,000	0,305	0,305	0,000	0,000
NETWORKS-BUT2000	0,305	0,359	0,359	0,305	0,267	0,447	0,447	0,267	0,267	0,359	0,200	0,298	0,200	0,267	0,327	0,305	0,305
DLBCL-NIH-ROS2002	1,041	0,629	0,629	0,743	0,745	0,554	0,653	0,521	0,521	0,467	0,516	0,516	0,516	0,521	0,521	0,467	0,467
DLBCL-OUTCOME-SHI2002	0,327	0,305	0,267	0,305	0,267	0,200	0,200	0,267	0,200	0,305	0,200	0,333	0,327	0,200	0,200	0,305	0,305
DLBCL-SHI2002	0,327	0,000	0,267	0,000	0,000	0,267	0,267	0,200	0,200	0,267	0,327	0,200	0,305	0,327	0,327	0,327	0,327
E2A-PBX1-YEO2002	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000
ECML2004	0,442	0,533	0,611	0,943	0,904	0,538	0,538	0,904	0,904	0,427	0,427	0,427	0,427	0,427	0,427	0,499	0,499
GCM-RAM2001	0,743	0,856	0,733	0,952	0,952	0,792	0,733	0,952	0,952	0,792	0,846	0,803	0,846	0,653	0,653	0,803	0,803
GSE11665	0,267	0,267	0,267	0,267	0,267	0,267	0,267	0,267	0,267	0,267	0,267	0,267	0,267	0,267	0,267	0,267	0,267
GSE3255	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000
GSE360	0,200	0,359	0,200	0,743	0,305	0,200	0,200	0,305	0,305	0,442	0,298	1,033	0,298	0,200	0,298	0,200	0,200
GSE443	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000
GSE474	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000
GSE5473	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000
GSE7433	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000
GSE7898	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000
BREAST-HED2003	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000
HYPERDIP-50-YEO2002	0,789	0,554	0,554	1,035	0,867	0,833	0,867	0,867	0,867	0,730	0,757	0,897	0,833	0,800	0,856	0,757	0,757
GASTRIC-LEU2002	0,305	0,554	0,442	0,305	0,305	0,327	0,327	0,305	0,305	0,327	0,305	0,447	0,442	0,305	0,305	0,267	0,267
LUNG-BHA2001	0,000	0,000	0,000	0,305	0,305	0,000	0,000	0,305	0,305	0,000	0,000	0,000	0,000	0,000	0,000	0,200	0,200
LUNG-BEE2002	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000
LUNG-GOR2002	0,400	0,327	0,516	0,000	0,000	0,447	0,447	0,000	0,000	0,267	0,298	0,327	0,400	0,267	0,267	0,267	0,267
LUNG-WIG2002	0,000	0,000	0,000	0,200	0,200	0,000	0,200	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,267	0,267
BREAST-MA2003	0,200	0,499	0,267	0,856	0,792	0,359	0,000	0,867	0,867	0,427	0,359	0,422	0,422	0,267	0,267	0,467	0,467
MLL-ARM2002	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000
MLL-YEO2002	0,000	0,427	0,327	0,267	0,538	0,467	0,467	0,538	0,538	0,442	0,200	0,298	0,200	0,267	0,267	0,200	0,200
SOFT-NIE2002	0,267	0,267	0,267	0,267	0,267	0,267	0,267	0,267	0,267	0,267	0,267	0,267	0,267	0,267	0,267	0,267	0,267
OVARIAN-PET2002	0,533	0,200	0,200	0,267	0,267	0,447	0,447	0,200	0,200	0,000	0,000	0,000	0,000	0,305	0,305	0,000	0,000
OVARY-WEL2001	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000
CNS-POM2002	0,333	0,305	0,327	0,427	0,359	0,333	0,305	0,422	0,422	0,333	0,333	0,442	0,333	0,327	0,267	0,333	0,333
PROSTATE-OUTCOME-SIN2002	0,333	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,305	0,333	0,200	0,333	0,333	0,333	0,333	0,333
PROSTATE-SIN2002	0,680	0,670	0,680	0,683	0,611	0,615	0,615	0,521	0,521	0,611	0,653	0,667	0,667	0,611	0,611	0,611	0,611
T-ALL-YEO2002	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000
TEL-AML1-YEO2002	0,615	0,554	0,442	0,442	0,611	0,442	0,442	0,516	0,516	0,427	0,400	0,743	0,467	0,359	0,427	0,359	0,359
PROSTATE-WEL2001	0,327	0,327	0,327	0,305	0,327	0,327	0,327	0,327	0,327	0,327	0,327	0,327	0,327	0,327	0,327	0,327	0,327

Tabela C.5: *Windowing: Média dos valores originais da variável Tamanho da Janela obtidos por validação cruzada de dez partições.*

Base	WPEWeC	WPEC	WEWeC	WEC	WPWeC	WPC	WWeC	WC	WPEWe	WPE	WEWe	WE	WPWe	WP	WWe	W
LYMPHOMA-ALI2000	51,70000	52,90000	38,40000	40,30000	52,90000	53,70000	39,10000	40,30000	60,60000	62,70000	62,90000	62,90000	61,50000	66,20000	62,80000	62,80000
GIST-ALL2001	8,00000	8,00000	8,00000	8,00000	8,30000	8,30000	8,00000	8,00000	8,00000	8,00000	8,00000	8,00000	8,20000	8,20000	8,00000	8,00000
COLON-ALO1999	33,10000	35,90000	24,90000	27,90000	37,10000	40,90000	28,00000	28,00000	35,20000	36,30000	36,00000	36,00000	35,60000	38,40000	36,30000	36,30000
LEUKEMIA-GOL1999	24,90000	28,50000	18,20000	18,30000	34,80000	34,90000	18,30000	18,30000	34,10000	36,40000	32,10000	35,90000	39,30000	41,30000	35,40000	35,40000
BCR-ABL-YEO2002	117,10000	118,30000	80,60000	84,30000	117,50000	118,10000	83,50000	83,50000	121,00000	121,00000	124,70000	124,70000	130,10000	134,40000	122,40000	122,40000
MELANOMA-BIT2000	17,90000	18,00000	13,10000	13,30000	17,90000	17,90000	13,30000	13,30000	22,00000	22,40000	22,40000	22,40000	23,20000	27,00000	22,40000	22,40000
BREAST-VEE2002	75,40000	78,80000	34,00000	34,60000	79,10000	79,60000	34,60000	37,30000	78,60000	80,50000	78,10000	79,90000	81,20000	82,00000	79,60000	79,60000
NETWORKS-BUT2000	48,20000	48,90000	36,50000	37,80000	50,50000	50,50000	37,40000	37,80000	56,30000	56,90000	56,20000	56,60000	56,60000	58,20000	56,60000	56,60000
DLBCL-NIH-ROS2002	207,40000	208,00000	205,70000	205,80000	208,60000	210,50000	207,50000	207,50000	205,70000	208,40000	205,10000	205,10000	207,20000	208,10000	208,10000	208,10000
DLBCL-OUTCOME-SHI2002	46,50000	47,30000	21,80000	22,60000	47,20000	47,80000	21,80000	22,60000	47,10000	48,10000	45,20000	46,50000	48,50000	49,20000	47,10000	47,10000
DLBCL-SHI2002	42,40000	42,90000	17,30000	17,90000	42,90000	44,90000	18,50000	18,50000	42,50000	44,90000	43,30000	43,90000	47,80000	51,00000	44,90000	44,90000
E2A-PBX1-YEO2002	58,30000	58,30000	58,30000	58,30000	58,30000	58,30000	58,30000	58,30000	58,30000	58,30000	58,30000	58,30000	58,30000	58,30000	58,30000	58,30000
ECML2004	59,80000	60,30000	51,30000	52,50000	59,90000	60,10000	52,50000	52,50000	62,80000	62,90000	62,90000	62,90000	62,80000	64,20000	62,90000	62,90000
GCM-RAM2001	131,30000	134,90000	92,50000	92,50000	133,30000	134,50000	92,50000	92,50000	137,90000	144,10000	141,20000	143,50000	142,00000	146,50000	142,10000	142,10000
GSE11665	7,00000	7,00000	7,00000	7,00000	8,10000	8,10000	7,00000	7,00000	7,00000	7,00000	7,00000	7,00000	8,10000	8,10000	7,00000	7,00000
GSE3255	10,60000	10,60000	8,80000	9,00000	10,60000	10,80000	9,00000	9,00000	13,10000	13,30000	13,30000	13,30000	15,50000	15,50000	13,30000	13,30000
GSE360	21,70000	22,30000	15,00000	18,20000	22,20000	23,30000	18,20000	18,20000	21,90000	24,00000	20,10000	23,80000	23,60000	24,40000	23,60000	23,60000
GSE443	6,70000	6,70000	6,20000	6,20000	8,30000	8,30000	6,20000	6,20000	6,60000	6,60000	6,60000	6,60000	8,60000	8,60000	6,60000	6,60000
GSE474	11,70000	12,10000	9,60000	9,70000	12,10000	12,60000	9,70000	9,70000	18,60000	18,60000	17,20000	18,60000	18,90000	19,90000	18,60000	18,60000
GSE5473	11,80000	11,80000	9,40000	9,80000	11,80000	12,00000	9,80000	9,80000	18,10000	18,10000	18,10000	18,10000	19,60000	19,60000	18,10000	18,10000
GSE7433	9,10000	9,10000	7,40000	7,50000	9,10000	9,60000	7,50000	7,50000	11,90000	12,10000	12,10000	12,10000	13,30000	13,30000	12,10000	12,10000
GSE7898	6,80000	6,80000	6,60000	6,60000	7,80000	7,80000	6,60000	6,60000	6,90000	6,90000	6,90000	6,90000	8,10000	8,10000	6,90000	6,90000
BREAST-HED2003	8,60000	8,60000	7,40000	7,40000	8,80000	9,10000	7,40000	7,40000	9,80000	9,90000	9,90000	9,90000	11,70000	11,70000	9,90000	9,90000
HYPERDIP-50-YEO2002	187,50000	191,30000	147,60000	156,00000	192,20000	194,30000	162,80000	162,80000	193,80000	198,90000	181,40000	199,90000	203,50000	204,70000	198,00000	198,00000
GASTRIC-LEU2002	50,90000	54,60000	34,60000	35,50000	58,10000	58,10000	35,50000	35,50000	54,90000	58,40000	55,80000	57,40000	62,00000	65,20000	59,70000	59,70000
LUNG-BHA2001	83,90000	84,60000	66,90000	66,90000	83,90000	93,70000	66,90000	66,90000	90,90000	95,30000	94,70000	95,30000	92,40000	103,30000	95,50000	95,50000
LUNG-BEE2002	22,80000	22,80000	19,20000	19,20000	24,00000	25,00000	19,20000	19,20000	24,20000	24,20000	24,20000	24,20000	32,90000	32,90000	24,20000	24,20000
LUNG-GOR2002	53,90000	65,20000	32,90000	32,90000	81,50000	81,50000	33,20000	33,20000	61,10000	65,10000	54,40000	62,80000	69,20000	69,20000	68,00000	68,00000
LUNG-WIG2002	21,90000	22,20000	20,10000	20,50000	22,20000	25,90000	21,20000	21,20000	23,20000	23,70000	23,60000	23,60000	23,70000	25,70000	23,80000	23,80000
BREAST-MA2003	48,10000	51,50000	32,20000	33,30000	50,30000	52,20000	34,10000	34,10000	49,60000	51,10000	50,00000	50,00000	50,40000	51,20000	49,50000	49,50000
MLL-ARM2002	27,00000	27,00000	20,40000	21,60000	27,00000	27,30000	21,60000	21,60000	37,30000	38,60000	38,60000	38,60000	39,10000	44,50000	38,60000	38,60000
MLL-YEO2002	106,70000	114,90000	76,20000	78,00000	113,80000	116,90000	79,00000	79,00000	107,50000	118,40000	111,30000	118,30000	114,90000	117,80000	118,30000	118,30000
SOFT-NIE2002	32,70000	32,90000	20,80000	23,90000	32,50000	32,70000	22,80000	23,90000	34,70000	35,00000	34,50000	35,00000	36,10000	37,70000	34,90000	34,90000
OVARIAN-PET2002	68,10000	68,10000	54,00000	54,00000	72,40000	72,40000	54,40000	54,40000	66,30000	69,40000	69,40000	69,40000	71,20000	71,20000	69,40000	69,40000
OVARY-WEL2001	11,90000	11,90000	11,50000	11,50000	13,30000	13,30000	11,50000	11,50000	12,90000	12,90000	12,90000	12,90000	16,00000	16,00000	12,90000	12,90000
CNS-POM2002	44,30000	45,10000	23,70000	26,20000	46,00000	48,40000	26,60000	26,60000	44,50000	45,10000	41,50000	44,60000	45,60000	48,10000	44,60000	44,60000
PROSTATE-OUTCOME-SIN2002	9,40000	9,40000	8,30000	8,30000	9,40000	9,50000	8,30000	8,30000	11,70000	15,20000	9,90000	15,20000	15,90000	17,20000	15,20000	15,20000
PROSTATE-SIN2002	90,80000	94,00000	43,00000	44,20000	92,10000	94,70000	44,70000	44,70000	87,90000	91,50000	89,30000	89,80000	90,90000	91,60000	90,30000	90,30000
T-ALL-YEO2002	59,00000	59,00000	59,00000	59,00000	63,70000	63,70000	59,00000	59,00000	59,30000	59,30000	59,30000	59,30000	69,60000	69,60000	59,30000	59,30000
TEL-AML1-YEO2002	109,50000	115,20000	92,20000	95,40000	132,60000	133,30000	102,70000	102,70000	119,50000	129,30000	114,80000	125,60000	131,20000	139,70000	129,00000	129,00000
PROSTATE-WEL2001	26,00000	26,00000	19,90000	22,30000	26,00000	26,00000	22,30000	22,30000	32,00000	34,00000	34,00000	34,00000	35,20000	35,60000	34,00000	34,00000

Tabela C.6: *Windowing: Desvio padrão dos valores originais da variável Tamanho da Janela obtidos por validação cruzada de dez partições.*

Base	WPEWeC	WPEC	WEWeC	WEC	WPWeC	WPC	WWeC	WC	WPEWe	WPE	WEWe	WE	WPWe	WP	WWe	W
LYMPHOMA-ALI2000	2,011	1,353	1,267	0,761	1,353	1,367	1,251	0,761	1,166	1,230	1,140	1,140	1,078	1,332	1,104	1,104
GIST-ALL2001	0,000	0,000	0,000	0,000	0,213	0,213	0,000	0,000	0,000	0,000	0,000	0,000	0,133	0,133	0,000	0,000
COLON-ALO1999	1,378	1,858	2,292	2,479	1,016	1,149	2,503	2,503	0,512	0,633	0,558	0,558	0,636	1,231	0,633	0,633
LEUKEMIA-GOL1999	2,610	2,676	0,574	0,597	1,638	1,643	0,597	0,597	2,424	1,869	2,900	1,767	2,196	1,257	1,772	1,772
BCR-ABL-YEO2002	3,598	4,028	6,176	6,566	4,390	4,408	6,202	6,202	4,061	4,061	4,088	4,088	4,985	4,949	3,789	3,789
MELANOMA-BIT2000	0,233	0,258	0,623	0,578	0,233	0,233	0,578	0,578	0,632	0,872	0,872	0,872	0,841	1,095	0,872	0,872
BREAST-VEE2002	1,887	1,228	4,442	4,362	1,329	1,514	4,362	4,017	1,077	1,108	1,320	1,059	0,629	0,447	0,933	0,933
NETWORKS-BUT2000	1,041	0,875	0,860	0,574	0,833	0,833	0,805	0,574	0,539	0,657	0,629	0,819	0,562	0,533	0,733	0,733
DLBCL-NIH-ROS2002	1,352	1,461	1,211	1,113	1,416	0,792	1,138	1,138	1,383	0,909	1,804	1,804	1,365	1,100	1,320	1,320
DLBCL-OUTCOME-SHI2002	1,587	1,499	1,533	1,267	1,548	1,124	1,533	1,267	0,767	0,504	1,263	0,958	0,601	0,554	0,971	0,971
DLBCL-SHI2002	2,104	2,253	0,895	0,994	2,253	1,858	1,327	1,327	1,772	2,052	1,932	2,126	2,736	1,585	2,052	2,052
E2A-PBX1-YEO2002	0,153	0,153	0,153	0,153	0,153	0,153	0,153	0,153	0,153	0,153	0,153	0,153	0,153	0,153	0,153	0,153
ECML2004	0,574	0,496	1,126	1,025	0,586	0,547	1,025	1,025	0,250	0,277	0,277	0,277	0,250	0,327	0,277	0,277
GCM-RAM2001	1,535	1,804	3,181	3,181	1,430	1,881	3,181	3,181	0,737	2,208	2,289	2,713	1,626	1,655	1,792	1,792
GSE11665	0,298	0,298	0,298	0,298	0,504	0,504	0,298	0,298	0,298	0,298	0,298	0,298	0,482	0,482	0,298	0,298
GSE3255	0,163	0,163	0,250	0,211	0,163	0,200	0,211	0,211	0,407	0,472	0,472	0,472	0,401	0,401	0,472	0,472
GSE360	0,396	0,260	0,667	0,490	0,250	0,300	0,490	0,490	0,862	0,211	1,501	0,291	0,340	0,221	0,340	0,340
GSE443	0,260	0,260	0,133	0,133	0,260	0,260	0,133	0,133	0,221	0,221	0,221	0,221	0,305	0,305	0,221	0,221
GSE474	0,423	0,314	0,267	0,260	0,314	0,221	0,260	0,260	0,427	0,427	0,879	0,427	0,504	0,348	0,427	0,427
GSE5473	0,442	0,442	0,163	0,200	0,442	0,394	0,200	0,200	0,605	0,605	0,605	0,605	0,521	0,521	0,605	0,605
GSE7433	0,233	0,233	0,221	0,224	0,233	0,221	0,224	0,224	0,314	0,314	0,314	0,314	0,335	0,335	0,314	0,314
GSE7898	0,200	0,200	0,163	0,163	0,250	0,250	0,163	0,163	0,233	0,233	0,233	0,233	0,314	0,314	0,233	0,233
BREAST-HED2003	0,371	0,371	0,163	0,163	0,389	0,348	0,163	0,163	0,327	0,379	0,379	0,379	0,517	0,517	0,379	0,379
HYPERDIP-50-YEO2002	2,888	2,856	13,708	11,900	1,489	1,506	12,080	12,080	3,918	4,373	8,565	4,046	4,425	4,077	4,232	4,232
GASTRIC-LEU2002	3,146	3,152	2,655	2,311	1,952	1,952	2,311	2,311	2,443	1,950	2,260	2,423	2,654	2,032	1,700	1,700
LUNG-BHA2001	2,892	3,212	3,153	3,153	2,892	3,077	3,153	3,153	2,278	2,092	2,216	2,092	1,565	3,180	2,242	2,242
LUNG-BEE2002	0,800	0,800	0,512	0,512	0,760	0,954	0,512	0,512	0,800	0,800	0,800	0,800	1,991	1,991	0,800	0,800
LUNG-GOR2002	5,425	7,653	0,547	0,547	5,766	5,766	0,574	0,574	2,953	3,683	4,665	4,482	3,518	3,518	3,266	3,266
LUNG-WIG2002	0,567	0,696	1,260	1,249	0,743	1,187	0,757	0,757	0,467	0,597	0,763	0,763	0,668	0,831	0,727	0,727
BREAST-MA2003	1,402	0,543	3,540	3,317	0,517	0,696	3,557	3,557	0,792	0,458	0,774	0,774	0,476	0,490	0,806	0,806
MLL-ARM2002	1,229	1,229	0,884	0,542	1,229	1,126	0,542	0,542	1,300	1,772	1,772	1,772	1,567	1,195	1,772	1,772
MLL-YEO2002	4,374	2,862	4,454	4,955	2,594	2,131	4,619	4,619	3,071	3,899	3,590	3,810	2,722	2,760	3,810	3,810
SOFT-NIE2002	0,857	0,836	0,892	0,752	0,820	0,804	1,153	0,752	0,517	0,650	1,003	0,650	0,722	0,367	0,623	0,623
OVARIAN-PET2002	2,137	2,137	1,856	1,856	3,862	3,862	1,675	1,675	2,679	2,762	2,762	2,762	3,183	3,183	2,762	2,762
OVARY-WEL2001	0,314	0,314	0,224	0,224	0,539	0,539	0,224	0,224	0,623	0,623	0,623	0,623	1,229	1,229	0,623	0,623
CNS-POM2002	0,804	0,836	2,399	1,896	0,683	0,452	2,130	2,130	1,003	1,080	1,934	0,980	1,118	0,458	0,980	0,980
PROSTATE-OUTCOME-SIN2002	0,400	0,400	0,213	0,213	0,400	0,401	0,213	0,213	1,325	0,964	0,888	0,964	1,140	0,593	0,964	0,964
PROSTATE-SIN2002	3,428	4,058	3,935	3,444	3,863	3,454	3,141	3,141	2,514	2,272	3,044	2,736	2,742	2,513	2,490	2,490
T-ALL-YEO2002	0,394	0,394	0,394	0,394	0,746	0,746	0,394	0,394	0,539	0,539	0,539	0,539	1,565	1,565	0,539	0,539
TEL-AML1-YEO2002	4,847	4,237	7,431	8,601	3,531	3,688	6,578	6,578	5,640	4,372	7,217	3,748	3,720	5,471	3,899	3,899
PROSTATE-WEL2001	0,650	0,650	1,329	0,559	0,650	0,650	0,559	0,559	0,715	0,931	0,931	0,931	1,020	0,980	0,931	0,931

Tabela C.7: *Windowing: Média dos valores originais da variável Tempo de Treinamento obtidos por validação cruzada de dez partições.*

Base	WPEWeC	WPEC	WEWeC	WEC	WPWeC	WPC	WWeC	WC	WPEWe	WPE	WEWe	WE	WPWe	WP	WWe	W
LYMPHOMA-ALI2000	21,98300	22,16900	22,19400	23,62600	23,07200	22,27900	22,25600	22,33500	31,06500	31,74400	31,70300	32,10800	31,05800	33,15500	31,67300	32,04200
GIST-ALL2001	0,23600	0,23000	0,24300	0,27900	0,23400	0,23200	0,24500	0,23900	0,22900	0,23600	0,25200	0,24900	0,23400	0,23400	0,25100	0,28400
COLON-ALO1999	5,89100	5,97200	5,85200	5,94500	5,81100	5,80600	6,09100	5,95800	4,69400	4,88900	5,37600	4,89100	4,73300	4,74500	4,97700	5,67200
LEUKEMIA-GOL1999	5,42600	6,18600	5,84200	5,68700	5,45600	6,31600	5,70900	5,78400	15,30300	13,59200	16,20100	14,15700	13,60500	13,61800	14,12300	15,68200
BCR-ABL-YEO2002	69,69200	67,58800	71,47400	70,97400	70,45500	67,34100	75,90000	72,93600	79,89100	77,58800	91,64700	82,03000	79,28800	77,44600	88,71300	82,07300
MELANOMA-BIT2000	4,48100	4,31200	4,61900	4,53400	4,66700	4,36900	4,60000	4,63500	9,86400	10,62300	11,06200	10,58800	10,30900	9,48100	11,04700	10,60300
BREAST-VEE2002	246,18500	252,48500	247,27600	250,78100	251,42400	245,47100	259,14000	247,11200	236,49200	242,83400	250,47700	239,87700	239,30300	239,23600	251,80600	239,69700
NETWORKS-BUT2000	35,98000	40,35900	36,63900	36,39600	37,54400	38,36500	39,41000	39,50100	64,37400	73,14900	67,83200	68,50300	65,77200	63,90200	70,99300	66,08400
DLBCL-NIH-ROS2002	679,17400	665,23100	695,93600	698,30400	662,70500	683,10200	733,91000	706,97500	469,24700	488,34400	494,71900	494,53100	478,11000	471,11200	530,34900	496,80200
DLBCL-OUTCOME-SHI2002	22,07700	22,89000	22,21100	23,93800	21,85000	24,67000	22,50100	24,09000	28,78400	29,56500	29,33100	29,91200	28,87900	28,90900	30,80800	29,17300
DLBCL-SHI2002	13,79500	14,82500	13,83000	14,70900	13,58100	13,61900	14,00800	14,47500	19,14200	22,37300	20,57900	22,59800	19,26100	19,30500	20,50800	21,39500
E2A-PBX1-YEO2002	2,17500	2,13000	2,47900	2,54300	2,09800	2,12100	2,49800	2,64400	2,11300	2,44200	2,63900	2,47100	2,09400	2,09700	2,56000	2,53500
ECML2004	342,48000	350,55400	340,66300	342,09800	347,73700	345,81000	355,81200	358,10000	313,60400	325,76600	321,99700	318,14200	322,32400	316,73200	319,69200	329,20500
GCM-RAM2001	796,64600	799,91300	800,13200	818,10600	802,12400	805,35900	810,15800	807,78400	850,62600	851,38500	855,93600	879,43700	866,64700	859,46600	864,65800	892,52400
GSE11665	2,93900	3,25100	3,19300	3,17100	3,37400	3,28200	3,33900	3,41400	3,19100	3,20400	3,19900	3,14700	3,29400	2,96900	3,45400	3,15100
GSE3255	9,16600	9,69800	9,42900	9,97600	10,20600	8,94600	9,80100	9,85300	17,93900	16,10100	17,00800	17,40100	16,48700	17,33000	18,00700	16,88500
GSE360	35,88300	38,94000	36,07300	37,26400	38,21100	36,82600	36,47900	36,88400	29,91000	29,98600	29,63200	30,86700	28,95400	28,42800	31,73400	30,73100
GSE443	2,39200	2,66300	2,46800	2,46700	2,72500	2,39900	2,83700	2,44300	2,51700	2,65800	2,61100	2,76300	2,50800	2,51600	2,60300	2,62600
GSE474	10,96900	9,64900	10,37800	10,01900	10,42400	10,90600	10,26500	9,99500	27,31300	26,34100	26,96400	28,22800	27,30800	28,95400	28,84100	28,42500
GSE5473	11,20200	10,65600	10,69700	10,97700	11,48600	10,94000	11,09800	11,33100	24,28400	23,75600	25,21600	24,90000	25,32500	23,85200	25,04800	25,86400
GSE7433	12,03500	12,25800	12,10700	12,10400	12,69200	12,01300	12,47600	12,70900	19,64400	19,06100	19,89300	19,91800	19,92200	20,20500	20,21800	19,61200
GSE7898	6,25200	5,87100	6,29100	6,64400	6,06700	6,30100	6,54400	6,64000	7,29000	7,13800	7,16100	7,21800	6,97800	7,26100	7,21100	7,23000
BREAST-HED2003	1,14800	1,15000	1,19200	1,30900	1,14900	1,14600	1,19400	1,37000	1,78000	1,78100	1,60900	1,61000	1,54900	1,78200	1,61600	1,61300
HYPERDIP-50-YEO2002	335,68700	329,16100	351,32600	351,24900	339,65400	335,06500	350,30400	353,74000	276,82800	280,30900	276,82300	281,46400	298,88600	303,43600	286,36400	284,98800
GASTRIC-LEU2002	32,96900	33,15900	33,82200	34,39900	33,21100	34,54400	34,59100	35,80000	29,81400	30,75900	32,24500	31,77000	29,91900	30,00500	31,08400	33,84800
LUNG-BHA2001	138,86200	133,65900	143,40900	142,63200	139,20000	138,60000	142,39300	141,47500	146,57700	143,26100	148,29300	148,03100	137,55500	147,03400	148,48000	144,80200
LUNG-BEE2002	3,37500	3,37800	4,15800	3,60800	3,37800	3,44900	3,61300	4,16600	4,72500	4,73300	5,11600	5,12100	4,73600	4,72300	5,83900	5,62600
LUNG-GOR2002	40,00200	39,97600	41,77700	42,90000	41,09700	40,72500	41,73300	41,37900	53,22400	52,53900	54,51700	57,18300	53,58500	51,98100	56,21900	55,67800
LUNG-WIG2002	3,87000	3,98900	4,03300	4,24300	3,97900	3,98300	4,14400	4,11400	2,76900	2,82100	2,87500	3,40900	2,81300	2,82000	2,92600	3,40400
BREAST-MA2003	14,61800	14,40200	14,27400	14,53000	13,91300	14,79700	14,43000	14,51400	11,71800	11,80000	11,89700	12,83100	12,27200	12,21200	12,00700	12,73800
MLL-ARM2002	10,83400	10,26300	11,56500	11,23800	11,40700	11,23600	11,43800	11,07000	30,50800	29,48800	30,23400	30,00100	29,58000	27,87600	30,03400	29,60000
MLL-YEO2002	67,33400	68,35600	74,06600	73,22300	68,36400	70,13500	72,38700	74,87100	55,22000	53,90700	57,67200	58,06400	58,44300	62,23000	57,16400	56,96100
SOFT-NIE2002	16,05800	16,53600	15,20900	15,07900	14,64500	14,67800	15,01300	15,27800	19,40100	19,48500	22,19900	21,35900	19,97400	19,40300	21,28100	19,97600
OVARIAN-PET2002	44,86300	48,02600	49,28700	47,45000	47,54100	44,90200	48,87500	47,84700	41,91400	41,22800	43,83200	43,23200	41,62400	43,32900	44,62200	43,36500
OVARY-WEL2001	2,25400	2,16200	2,11600	2,30000	1,98700	1,98400	2,12300	2,11800	2,32800	2,32800	2,55200	2,48500	2,33100	2,33200	2,49300	2,48700
CNS-POM2002	23,36700	23,62000	25,08800	24,62900	23,56600	26,43400	25,70500	25,04100	27,57000	27,75900	28,47300	31,83200	28,25600	28,32100	29,85800	30,29800
PROSTATE-OUTCOME-SIN2002	3,28700	3,10700	3,59500	3,25100	3,17500	3,31100	3,40300	3,30900	8,22600	8,64100	9,21000	9,16400	8,45700	8,46800	9,15100	9,47500
PROSTATE-SIN2002	134,20200	131,39700	133,27600	137,69600	127,41800	132,87200	133,09600	136,63800	108,53900	109,57300	114,42300	112,06200	112,39500	108,91200	114,77200	112,77400
T-ALL-YEO2002	6,47500	6,64000	7,70400	8,22100	6,41500	6,46800	7,58600	7,87300	6,43900	6,53800	7,32700	7,15400	6,29400	5,98900	7,22800	6,99400
TEL-AML1-YEO2002	105,97200	107,25900	109,87700	113,00100	108,47500	114,37400	111,74200	111,27100	87,85600	79,22300	80,37700	81,65600	80,66600	81,52400	83,60500	81,09800
PROSTATE-WEL2001	18,93200	18,58800	20,60500	20,76500	19,71100	18,52900	19,85500	20,04400	38,93600	40,02500	38,96400	38,27300	38,88700	37,07900	40,93800	41,06000

Tabela C.8: *Windowing: Desvio padrão dos valores originais da variável Tempo de Treinamento obtidos por validação cruzada de dez partições.*

Base	WPEWeC	WPEC	WEWeC	WEC	WPWeC	WPC	WWeC	WC	WPEWe	WPE	WEWe	WE	WPWe	WP	WWe	W
LYMPHOMA-ALI2000	2,749	3,278	2,618	3,204	3,532	3,316	2,606	2,606	2,784	3,003	2,814	2,957	2,783	4,078	2,799	2,612
GIST-ALL2001	0,047	0,050	0,052	0,059	0,047	0,053	0,049	0,049	0,057	0,057	0,061	0,060	0,057	0,058	0,060	0,070
COLON-ALO1999	0,877	0,892	0,873	0,893	0,875	0,880	0,902	0,891	0,795	0,783	0,796	0,790	0,787	0,788	0,767	0,913
LEUKEMIA-GOL1999	1,539	1,739	1,531	1,578	1,546	1,791	1,580	1,477	3,407	2,868	3,377	2,933	2,886	2,871	2,946	3,340
BCR-ABL-YEO2002	11,513	11,569	13,536	11,233	12,286	12,200	15,216	13,261	18,038	16,427	21,953	16,340	15,432	18,250	19,193	17,355
MELANOMA-BIT2000	0,289	0,104	0,330	0,113	0,294	0,129	0,263	0,322	1,765	1,715	1,714	1,838	1,736	1,729	1,706	1,718
BREAST-VEE2002	54,098	52,784	53,081	49,492	43,523	49,427	49,163	48,616	22,705	30,703	26,769	24,811	21,687	27,428	34,195	24,459
NETWORKS-BUT2000	5,405	5,693	6,439	5,704	6,701	5,961	6,640	6,041	4,395	6,800	5,781	5,567	5,609	4,507	3,979	5,448
DLBCL-NIH-ROS2002	75,130	71,361	61,884	58,924	59,152	68,105	62,721	62,573	53,617	69,696	42,587	47,570	46,772	58,490	52,147	38,262
DLBCL-OUTCOME-SHI2002	6,249	7,463	6,681	7,974	6,613	7,051	6,754	7,486	2,160	3,055	2,029	2,403	2,526	3,002	2,875	2,099
DLBCL-SHI2002	2,802	3,336	2,694	3,530	2,684	2,731	2,699	2,907	4,146	4,902	4,142	5,164	4,194	4,209	4,536	4,870
E2A-PBX1-YEO2002	0,127	0,096	0,011	0,121	0,012	0,056	0,022	0,186	0,049	0,016	0,165	0,014	0,020	0,013	0,138	0,122
ECML2004	35,317	43,043	38,459	34,448	38,139	38,343	42,358	41,413	14,647	17,393	13,184	11,487	14,006	7,125	13,844	11,635
GCM-RAM2001	172,900	176,716	157,089	161,346	146,897	168,930	167,174	160,600	70,547	63,039	63,261	69,869	86,038	85,655	49,141	59,703
GSE11665	0,982	1,213	0,998	1,076	1,138	1,112	1,226	1,254	1,174	1,158	1,053	1,062	1,202	1,003	1,232	1,063
GSE3255	0,628	0,796	0,368	0,745	0,318	0,283	0,707	0,548	2,197	2,496	2,587	1,981	2,513	2,895	2,152	2,038
GSE360	4,219	4,351	4,198	3,475	5,139	4,045	4,417	4,085	2,968	3,184	2,619	3,107	2,254	1,785	2,486	2,062
GSE443	0,237	0,272	0,280	0,264	0,292	0,259	0,307	0,258	0,410	0,333	0,423	0,591	0,403	0,408	0,421	0,425
GSE474	0,395	0,424	0,644	0,396	0,634	0,427	0,474	0,392	3,725	3,210	3,102	3,313	5,127	2,755	4,699	4,009
GSE5473	0,367	0,596	0,763	0,847	0,339	0,605	0,753	0,842	4,122	4,011	3,704	3,511	3,681	3,627	3,877	3,867
GSE7433	1,195	0,720	0,976	0,734	0,576	0,374	0,781	0,876	1,849	2,836	2,069	2,601	3,056	2,837	3,055	2,414
GSE7898	0,664	0,421	0,504	0,653	0,530	0,757	0,634	0,407	0,879	0,866	0,970	1,147	0,890	0,736	0,972	0,767
BREAST-HED2003	0,071	0,071	0,080	0,124	0,069	0,069	0,075	0,088	0,301	0,298	0,268	0,271	0,258	0,299	0,268	0,267
HYPERDIP-50-YEO2002	42,704	35,613	47,400	53,306	39,307	37,395	50,966	53,871	42,826	43,964	36,072	41,868	42,314	41,061	40,944	38,494
GASTRIC-LEU2002	7,361	7,431	7,714	8,064	7,444	8,712	8,592	9,312	5,320	5,874	4,400	5,586	5,346	5,357	5,326	6,635
LUNG-BHA2001	29,134	28,092	29,137	29,815	28,529	27,318	29,294	28,565	31,967	31,086	32,696	32,013	28,543	31,100	32,232	29,805
LUNG-BEE2002	0,196	0,203	0,283	0,240	0,202	0,232	0,242	0,279	0,785	0,781	0,822	0,818	0,785	0,785	0,958	1,059
LUNG-GOR2002	17,027	16,661	16,586	16,823	17,798	17,113	17,348	17,033	10,761	9,258	10,701	10,376	10,645	10,124	12,309	12,650
LUNG-WIG2002	0,448	0,465	0,382	0,403	0,464	0,462	0,379	0,416	0,305	0,301	0,280	0,327	0,310	0,306	0,286	0,325
BREAST-MA2003	1,721	1,494	1,577	1,593	1,413	1,725	1,575	1,554	0,649	0,661	0,650	1,040	0,723	0,689	0,647	0,938
MLL-ARM2002	1,234	0,745	1,280	0,907	1,297	1,223	1,339	1,145	4,321	5,444	4,765	4,327	4,596	4,129	4,728	4,755
MLL-YEO2002	6,869	6,973	8,517	7,081	7,192	8,259	7,010	8,438	8,753	9,173	9,122	9,506	10,857	11,250	9,339	8,556
SOFT-NIE2002	1,832	2,349	1,785	1,766	1,709	1,709	1,738	1,726	2,055	2,048	3,095	0,983	2,865	2,049	3,596	2,061
OVARIAN-PET2002	8,231	9,631	9,652	7,683	9,196	8,264	10,113	9,624	7,157	7,356	7,038	7,951	8,257	9,301	9,510	8,294
OVARY-WEL2001	0,332	0,381	0,270	0,261	0,254	0,254	0,271	0,269	0,362	0,354	0,473	0,377	0,352	0,352	0,379	0,375
CNS-POM2002	5,389	5,476	5,952	5,864	5,430	6,596	5,558	6,117	4,727	4,751	4,780	4,990	5,883	4,519	5,540	5,904
PROSTATE-OUTCOME-SIN2002	0,398	0,205	0,395	0,228	0,159	0,238	0,379	0,258	0,879	1,128	1,194	1,158	0,981	0,985	1,180	1,277
PROSTATE-SIN2002	29,407	31,863	29,248	32,521	29,932	26,203	30,371	30,239	16,529	19,459	18,700	17,779	18,697	16,561	19,518	18,881
T-ALL-YEO2002	0,621	0,878	0,806	0,690	0,626	0,624	0,848	0,859	0,825	0,788	0,957	0,990	0,735	0,563	0,752	0,654
TEL-AML1-YEO2002	19,026	17,355	18,720	17,802	22,073	20,868	18,049	18,590	19,581	15,984	16,392	18,030	19,334	17,152	19,283	17,236
PROSTATE-WEL2001	3,932	3,600	3,209	3,962	4,061	3,311	3,249	3,934	2,899	2,840	3,760	3,307	3,452	3,239	3,969	4,058

Tabela C.9: *Windowing: Média dos valores originais da variável Coesão obtidos por validação cruzada de dez partições.*

Base	J48	WPEWeC	WPEC	WEWeC	WEC	WPWeC	WPC	WWeC	WC	WPEWe	WPE	WEWe	WE	WPWe	WP	WWe	W
LYMPHOMA-ALI2000	0,56338	0,54596	0,54596	0,54265	0,54265	0,54596	0,54596	0,54265	0,54265	0,55301	0,54926	0,54926	0,54926	0,54596	0,54926	0,54596	0,54596
GIST-ALL2001	0,66667	0,66667	0,66667	0,66667	0,66667	0,66667	0,66667	0,66667	0,66667	0,66667	0,66667	0,66667	0,66667	0,66667	0,66667	0,66667	0,66667
COLON-ALO1999	0,40000	0,43000	0,42000	0,49667	0,46000	0,40000	0,40000	0,46000	0,46000	0,41000	0,40000	0,41000	0,41000	0,40000	0,40000	0,40000	0,40000
LEUKEMIA-GOL1999	0,55000	0,61667	0,58333	0,66667	0,66667	0,51667	0,51667	0,66667	0,66667	0,53333	0,51667	0,55000	0,51667	0,51667	0,51667	0,51667	0,51667
BCR-ABL-YEO2002	0,33333	0,39667	0,39000	0,49000	0,48000	0,33524	0,32690	0,47000	0,47000	0,40000	0,40000	0,40333	0,40333	0,32690	0,32690	0,39000	0,39000
MELANOMA-BIT2000	0,60000	0,60000	0,60000	0,60000	0,60000	0,60000	0,60000	0,60000	0,60000	0,60000	0,60000	0,60000	0,60000	0,60000	0,60000	0,60000	0,60000
BREAST-VEE2002	0,27619	0,29643	0,29048	0,49190	0,49190	0,27857	0,27500	0,49190	0,46524	0,28571	0,28571	0,29048	0,28571	0,27500	0,27500	0,28571	0,28571
NETWORKS-BUT2000	0,49211	0,52684	0,52684	0,53934	0,53603	0,51507	0,51507	0,53603	0,53603	0,50325	0,49737	0,50031	0,49737	0,49474	0,48947	0,49211	0,49211
DLBCL-NIH-ROS2002	0,11580	0,11873	0,11873	0,12255	0,12182	0,11198	0,11022	0,11584	0,11584	0,11716	0,11789	0,11789	0,11789	0,11322	0,11322	0,11716	0,11716
DLBCL-OUTCOME-SHI2002	0,36000	0,35333	0,34667	0,55000	0,53333	0,34000	0,34000	0,53333	0,51667	0,35333	0,34000	0,36667	0,36000	0,34000	0,34000	0,35333	0,35333
DLBCL-SHI2002	0,44000	0,50000	0,48000	0,66667	0,66667	0,48000	0,48000	0,65000	0,65000	0,48000	0,46000	0,49000	0,47000	0,46000	0,46000	0,46000	0,46000
E2A-PBX1-YEO2002	0,66667	0,66667	0,66667	0,66667	0,66667	0,66667	0,66667	0,66667	0,66667	0,66667	0,66667	0,66667	0,66667	0,66667	0,66667	0,66667	0,66667
ECML2004	0,53659	0,53895	0,53900	0,58203	0,57661	0,53780	0,53780	0,57661	0,57661	0,53543	0,53543	0,53543	0,53543	0,53543	0,53543	0,53432	0,53432
GCM-RAM2001	0,43261	0,44520	0,44220	0,49413	0,49413	0,43402	0,43394	0,49413	0,49413	0,43404	0,43411	0,43138	0,43411	0,42721	0,42721	0,43138	0,43138
GSE11665	0,73333	0,73333	0,73333	0,73333	0,73333	0,73333	0,73333	0,73333	0,73333	0,73333	0,73333	0,73333	0,73333	0,73333	0,73333	0,73333	0,73333
GSE3255	0,57143	0,57143	0,57143	0,57143	0,57143	0,57143	0,57143	0,57143	0,57143	0,57143	0,57143	0,57143	0,57143	0,57143	0,57143	0,57143	0,57143
GSE360	0,60316	0,60346	0,59714	0,69267	0,64211	0,59714	0,59714	0,64211	0,64211	0,61298	0,60030	0,63649	0,60030	0,59714	0,60030	0,59714	0,59714
GSE443	0,66667	0,66667	0,66667	0,66667	0,66667	0,66667	0,66667	0,66667	0,66667	0,66667	0,66667	0,66667	0,66667	0,66667	0,66667	0,66667	0,66667
GSE474	0,60000	0,60000	0,60000	0,60000	0,60000	0,60000	0,60000	0,60000	0,60000	0,60000	0,60000	0,60000	0,60000	0,60000	0,60000	0,60000	0,60000
GSE5473	0,60000	0,60000	0,60000	0,60000	0,60000	0,60000	0,60000	0,60000	0,60000	0,60000	0,60000	0,60000	0,60000	0,60000	0,60000	0,60000	0,60000
GSE7433	0,60000	0,60000	0,60000	0,60000	0,60000	0,60000	0,60000	0,60000	0,60000	0,60000	0,60000	0,60000	0,60000	0,60000	0,60000	0,60000	0,60000
GSE7898	0,66667	0,66667	0,66667	0,66667	0,66667	0,66667	0,66667	0,66667	0,66667	0,66667	0,66667	0,66667	0,66667	0,66667	0,66667	0,66667	0,66667
BREAST-HED2003	0,66667	0,66667	0,66667	0,66667	0,66667	0,66667	0,66667	0,66667	0,66667	0,66667	0,66667	0,66667	0,66667	0,66667	0,66667	0,66667	0,66667
HYPERDIP-50-YEO2002	0,20288	0,22667	0,22667	0,29159	0,26159	0,19914	0,20136	0,25524	0,25524	0,22564	0,22842	0,24691	0,23199	0,19146	0,19369	0,22342	0,22342
GASTRIC-LEU2002	0,45000	0,50143	0,47429	0,57000	0,57000	0,45714	0,45714	0,57000	0,57000	0,45714	0,45000	0,46714	0,46000	0,45000	0,45000	0,44286	0,44286
LUNG-BHA2001	0,50000	0,50000	0,50000	0,51667	0,51667	0,50000	0,50000	0,51667	0,51667	0,50000	0,50000	0,50000	0,50000	0,50000	0,50000	0,49545	0,49545
LUNG-BEE2002	0,66667	0,66667	0,66667	0,66667	0,66667	0,66667	0,66667	0,66667	0,66667	0,66667	0,66667	0,66667	0,66667	0,66667	0,66667	0,66667	0,66667
LUNG-GOR2002	0,42667	0,56667	0,52000	0,66667	0,66667	0,45667	0,45667	0,66667	0,66667	0,53333	0,50667	0,60000	0,54000	0,48000	0,48000	0,48000	0,48000
LUNG-WIG2002	0,50000	0,50000	0,50000	0,51667	0,51667	0,50000	0,49000	0,50000	0,50000	0,50000	0,50000	0,50000	0,50000	0,50000	0,50000	0,48000	0,48000
BREAST-MA2003	0,33750	0,36905	0,34167	0,47940	0,46226	0,33083	0,33333	0,45155	0,45155	0,34667	0,33833	0,33500	0,33500	0,32667	0,32667	0,33917	0,33917
MLL-ARM2002	0,60000	0,60000	0,60000	0,60000	0,60000	0,60000	0,60000	0,60000	0,60000	0,60000	0,60000	0,60000	0,60000	0,60000	0,60000	0,60000	0,60000
MLL-YEO2002	0,33333	0,38333	0,36000	0,48000	0,45667	0,33238	0,33238	0,45667	0,45667	0,36190	0,32857	0,33524	0,32857	0,32381	0,32381	0,32857	0,32857
SOFT-NIE2002	0,55636	0,55636	0,55636	0,55636	0,55636	0,55636	0,55636	0,55636	0,55636	0,55636	0,55636	0,55636	0,55636	0,55636	0,55636	0,55636	0,55636
OVARIAN-PET2002	0,30857	0,49000	0,49000	0,53333	0,53333	0,45333	0,45333	0,51667	0,51667	0,50000	0,50000	0,50000	0,50000	0,47000	0,47000	0,50000	0,50000
OVARY-WEL2001	0,66667	0,66667	0,66667	0,66667	0,66667	0,66667	0,66667	0,66667	0,66667	0,66667	0,66667	0,66667	0,66667	0,66667	0,66667	0,66667	0,66667
CNS-POM2002	0,36667	0,38000	0,37333	0,55667	0,52333	0,36667	0,35333	0,51333	0,51333	0,36667	0,36667	0,37667	0,36667	0,36000	0,34667	0,36667	0,36667
PROSTATE-OUTCOME-SIN2002	0,58333	0,66667	0,66667	0,66667	0,66667	0,66667	0,66667	0,66667	0,66667	0,61667	0,58333	0,65000	0,58333	0,58333	0,58333	0,58333	0,58333
PROSTATE-SIN2002	0,27651	0,28048	0,27690	0,47000	0,45333	0,27214	0,27214	0,43667	0,43667	0,27571	0,28405	0,29238	0,29238	0,27571	0,27571	0,27571	0,27571
T-ALL-YEO2002	0,66667	0,66667	0,66667	0,66667	0,66667	0,66667	0,66667	0,66667	0,66667	0,66667	0,66667	0,66667	0,66667	0,66667	0,66667	0,66667	0,66667
TEL-AML1-YEO2002	0,31270	0,40333	0,37667	0,46333	0,45333	0,31548	0,31548	0,41000	0,41000	0,38333	0,34857	0,37381	0,34381	0,33048	0,32095	0,33048	0,33048
PROSTATE-WEL2001	0,55641	0,55641	0,55641	0,56987	0,56538	0,55641	0,55641	0,56538	0,56538	0,55641	0,55641	0,55641	0,55641	0,55641	0,55641	0,55641	0,55641

Tabela C.10: *Windowing: Desvio padrão dos valores originais da variável Coesão obtidos por validação cruzada de dez partições.*

Base	J48	WPEWeC	WPEC	WEWeC	WEC	WPWeC	WPC	WWeC	WC	WPEWe	WPE	WEWe	WE	WPWe	WP	WWe	W
LYMPHOMA-ALI2000	0,007	0,006	0,006	0,005	0,005	0,006	0,006	0,005	0,005	0,007	0,005	0,005	0,005	0,006	0,005	0,006	0,006
GIST-ALL2001	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000
COLON-ALO1999	0,000	0,015	0,013	0,023	0,016	0,000	0,000	0,016	0,016	0,010	0,000	0,010	0,010	0,000	0,000	0,000	0,000
LEUKEMIA-GOL1999	0,025	0,025	0,028	0,000	0,000	0,017	0,017	0,000	0,000	0,022	0,017	0,025	0,017	0,017	0,017	0,017	0,017
BCR-ABL-YEO2002	0,000	0,024	0,025	0,010	0,013	0,009	0,012	0,015	0,015	0,019	0,019	0,023	0,023	0,012	0,012	0,016	0,016
MELANOMA-BIT2000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000
BREAST-VEE2002	0,008	0,009	0,005	0,037	0,037	0,005	0,005	0,037	0,032	0,000	0,000	0,005	0,000	0,005	0,005	0,000	0,000
NETWORKS-BUT2000	0,004	0,006	0,006	0,005	0,004	0,007	0,007	0,004	0,004	0,005	0,003	0,004	0,003	0,004	0,004	0,004	0,004
DLBCL-NIH-ROS2002	0,003	0,002	0,002	0,003	0,003	0,002	0,002	0,002	0,002	0,002	0,002	0,002	0,002	0,002	0,002	0,002	0,002
DLBCL-OUTCOME-SHI2002	0,011	0,010	0,009	0,025	0,022	0,007	0,007	0,022	0,017	0,010	0,007	0,011	0,011	0,007	0,007	0,010	0,010
DLBCL-SHI2002	0,016	0,000	0,013	0,000	0,000	0,013	0,013	0,017	0,017	0,013	0,016	0,010	0,015	0,016	0,016	0,016	0,016
E2A-PBX1-YEO2002	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000
ECML2004	0,003	0,003	0,004	0,006	0,006	0,003	0,003	0,006	0,006	0,002	0,002	0,002	0,002	0,002	0,002	0,003	0,003
GCM-RAM2001	0,005	0,006	0,005	0,008	0,008	0,005	0,005	0,008	0,008	0,005	0,006	0,005	0,006	0,004	0,004	0,005	0,005
GSE11665	0,044	0,044	0,044	0,044	0,044	0,044	0,044	0,044	0,044	0,044	0,044	0,044	0,044	0,044	0,044	0,044	0,044
GSE3255	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000
GSE360	0,003	0,005	0,003	0,016	0,005	0,003	0,003	0,005	0,005	0,007	0,004	0,020	0,004	0,003	0,004	0,003	0,003
GSE443	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000
GSE474	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000
GSE5473	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000
GSE7433	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000
GSE7898	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000
BREAST-HED2003	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000
HYPERDIP-50-YEO2002	0,008	0,007	0,007	0,028	0,018	0,008	0,009	0,018	0,018	0,009	0,010	0,012	0,011	0,008	0,009	0,010	0,010
GASTRIC-LEU2002	0,011	0,024	0,018	0,015	0,015	0,012	0,012	0,015	0,015	0,012	0,011	0,018	0,018	0,011	0,011	0,010	0,010
LUNG-BHA2001	0,000	0,000	0,000	0,008	0,008	0,000	0,000	0,008	0,008	0,000	0,000	0,000	0,000	0,000	0,000	0,005	0,005
LUNG-BEE2002	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000
LUNG-GOR2002	0,027	0,027	0,035	0,000	0,000	0,028	0,028	0,000	0,000	0,022	0,020	0,027	0,029	0,013	0,013	0,013	0,013
LUNG-WIG2002	0,000	0,000	0,000	0,017	0,017	0,000	0,010	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,013	0,013
BREAST-MA2003	0,004	0,012	0,006	0,031	0,028	0,007	0,000	0,030	0,030	0,008	0,007	0,008	0,008	0,004	0,004	0,009	0,009
MLL-ARM2002	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000
MLL-YEO2002	0,000	0,017	0,011	0,013	0,023	0,013	0,013	0,023	0,023	0,013	0,005	0,009	0,005	0,006	0,006	0,005	0,005
SOFT-NIE2002	0,007	0,007	0,007	0,007	0,007	0,007	0,007	0,007	0,007	0,007	0,007	0,007	0,007	0,007	0,007	0,007	0,007
OVARIAN-PET2002	0,015	0,010	0,010	0,022	0,022	0,020	0,020	0,017	0,017	0,000	0,000	0,000	0,000	0,015	0,015	0,000	0,000
OVARY-WEL2001	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000
CNS-POM2002	0,011	0,010	0,011	0,031	0,026	0,011	0,010	0,029	0,029	0,011	0,011	0,017	0,011	0,011	0,009	0,011	0,011
PROSTATE-OUTCOME-SIN2002	0,028	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,025	0,028	0,017	0,028	0,028	0,028	0,028	0,028
PROSTATE-SIN2002	0,015	0,016	0,016	0,038	0,032	0,015	0,015	0,022	0,022	0,015	0,016	0,016	0,016	0,015	0,015	0,015	0,015
T-ALL-YEO2002	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000
TEL-AML1-YEO2002	0,012	0,023	0,017	0,020	0,032	0,010	0,010	0,022	0,022	0,017	0,012	0,028	0,014	0,010	0,012	0,010	0,010
PROSTATE-WEL2001	0,007	0,007	0,007	0,007	0,007	0,007	0,007	0,007	0,007	0,007	0,007	0,007	0,007	0,007	0,007	0,007	0,007

Tabela C.11: *Windowing: Média dos valores originais da variável Compactação obtidos por validação cruzada de dez partições.*

Base	J48	WPEWeC	WPEC	WEWeC	WEC	WPWeC	WPC	WWeC	WC	WPEWe	WPE	WEWe	WE	WPWe	WP	WWe	W
LYMPHOMA-ALI2000	0,00174	0,00186	0,00186	0,00184	0,00184	0,00186	0,00186	0,00184	0,00184	0,00181	0,00184	0,00184	0,00184	0,00186	0,00184	0,00186	0,00186
GIST-ALL2001	0,00050	0,00050	0,00050	0,00050	0,00050	0,00050	0,00050	0,00050	0,00050	0,00050	0,00050	0,00050	0,00050	0,00050	0,00050	0,00050	0,00050
COLON-ALO1999	0,00150	0,00135	0,00140	0,00105	0,00120	0,00150	0,00150	0,00120	0,00120	0,00145	0,00150	0,00145	0,00145	0,00150	0,00150	0,00150	0,00150
LEUKEMIA-GOL1999	0,00024	0,00018	0,00021	0,00014	0,00014	0,00027	0,00027	0,00014	0,00014	0,00025	0,00027	0,00024	0,00027	0,00027	0,00027	0,00027	0,00027
BCR-ABL-YEO2002	0,00032	0,00026	0,00026	0,00017	0,00018	0,00032	0,00033	0,00018	0,00018	0,00025	0,00025	0,00025	0,00025	0,00033	0,00033	0,00026	0,00026
MELANOMA-BIT2000	0,00025	0,00025	0,00025	0,00025	0,00025	0,00025	0,00025	0,00025	0,00025	0,00025	0,00025	0,00025	0,00025	0,00025	0,00025	0,00025	0,00025
BREAST-VEE2002	0,00022	0,00020	0,00020	0,00009	0,00009	0,00021	0,00022	0,00009	0,00010	0,00020	0,00020	0,00020	0,00020	0,00022	0,00022	0,00020	0,00020
NETWORKS-BUT2000	0,00128	0,00112	0,00112	0,00106	0,00108	0,00117	0,00117	0,00108	0,00108	0,00123	0,00126	0,00124	0,00126	0,00127	0,00130	0,00128	0,00128
DLBCL-NIH-ROS2002	0,00207	0,00201	0,00201	0,00195	0,00196	0,00215	0,00219	0,00207	0,00207	0,00204	0,00203	0,00203	0,00203	0,00212	0,00212	0,00204	0,00204
DLBCL-OUTCOME-SHI2002	0,00051	0,00052	0,00053	0,00024	0,00025	0,00055	0,00055	0,00025	0,00027	0,00052	0,00055	0,00049	0,00051	0,00055	0,00055	0,00052	0,00052
DLBCL-SHI2002	0,00037	0,00028	0,00031	0,00014	0,00014	0,00031	0,00031	0,00015	0,00015	0,00031	0,00034	0,00030	0,00032	0,00034	0,00034	0,00034	0,00034
E2A-PBX1-YEO2002	0,00008	0,00008	0,00008	0,00008	0,00008	0,00008	0,00008	0,00008	0,00008	0,00008	0,00008	0,00008	0,00008	0,00008	0,00008	0,00008	0,00008
ECML2004	0,00078	0,00077	0,00077	0,00065	0,00066	0,00077	0,00077	0,00066	0,00066	0,00078	0,00078	0,00078	0,00078	0,00078	0,00078	0,00078	0,00078
GCM-RAM2001	0,00115	0,00109	0,00110	0,00090	0,00090	0,00114	0,00114	0,00090	0,00090	0,00114	0,00114	0,00115	0,00114	0,00116	0,00116	0,00115	0,00115
GSE11665	0,00004	0,00004	0,00004	0,00004	0,00004	0,00004	0,00004	0,00004	0,00004	0,00004	0,00004	0,00004	0,00004	0,00004	0,00004	0,00004	0,00004
GSE3255	0,00011	0,00011	0,00011	0,00011	0,00011	0,00011	0,00011	0,00011	0,00011	0,00011	0,00011	0,00011	0,00011	0,00011	0,00011	0,00011	0,00011
GSE360	0,00062	0,00056	0,00057	0,00040	0,00050	0,00058	0,00060	0,00050	0,00050	0,00057	0,00062	0,00051	0,00060	0,00060	0,00062	0,00060	0,00060
GSE443	0,00008	0,00008	0,00008	0,00008	0,00008	0,00008	0,00008	0,00008	0,00008	0,00008	0,00008	0,00008	0,00008	0,00008	0,00008	0,00008	0,00008
GSE474	0,00009	0,00009	0,00009	0,00009	0,00009	0,00009	0,00009	0,00009	0,00009	0,00009	0,00009	0,00009	0,00009	0,00009	0,00009	0,00009	0,00009
GSE5473	0,00006	0,00006	0,00006	0,00006	0,00006	0,00006	0,00006	0,00006	0,00006	0,00006	0,00006	0,00006	0,00006	0,00006	0,00006	0,00006	0,00006
GSE7433	0,00005	0,00005	0,00005	0,00005	0,00005	0,00005	0,00005	0,00005	0,00005	0,00005	0,00005	0,00005	0,00005	0,00005	0,00005	0,00005	0,00005
GSE7898	0,00003	0,00003	0,00003	0,00003	0,00003	0,00003	0,00003	0,00003	0,00003	0,00003	0,00003	0,00003	0,00003	0,00003	0,00003	0,00003	0,00003
BREAST-HED2003	0,00021	0,00021	0,00021	0,00021	0,00021	0,00021	0,00021	0,00021	0,00021	0,00021	0,00021	0,00021	0,00021	0,00021	0,00021	0,00021	0,00021
HYPERDIP-50-YEO2002	0,00064	0,00055	0,00055	0,00042	0,00047	0,00065	0,00064	0,00049	0,00049	0,00056	0,00055	0,00050	0,00054	0,00068	0,00068	0,00057	0,00057
GASTRIC-LEU2002	0,00055	0,00046	0,00051	0,00034	0,00034	0,00054	0,00054	0,00034	0,00034	0,00054	0,00055	0,00052	0,00054	0,00055	0,00055	0,00057	0,00057
LUNG-BHA2001	0,00040	0,00040	0,00040	0,00037	0,00037	0,00040	0,00040	0,00037	0,00037	0,00040	0,00040	0,00040	0,00040	0,00040	0,00040	0,00040	0,00040
LUNG-BEE2002	0,00014	0,00014	0,00014	0,00014	0,00014	0,00014	0,00014	0,00014	0,00014	0,00014	0,00014	0,00014	0,00014	0,00014	0,00014	0,00014	0,00014
LUNG-GOR2002	0,00022	0,00013	0,00016	0,00008	0,00008	0,00020	0,00020	0,00008	0,00008	0,00014	0,00016	0,00011	0,00014	0,00018	0,00018	0,00018	0,00018
LUNG-WIG2002	0,00069	0,00069	0,00069	0,00066	0,00066	0,00069	0,00073	0,00069	0,00069	0,00069	0,00069	0,00069	0,00069	0,00069	0,00069	0,00076	0,00076
BREAST-MA2003	0,00303	0,00267	0,00298	0,00180	0,00190	0,00313	0,00308	0,00200	0,00200	0,00293	0,00303	0,00308	0,00308	0,00319	0,00319	0,00303	0,00303
MLL-ARM2002	0,00016	0,00016	0,00016	0,00016	0,00016	0,00016	0,00016	0,00016	0,00016	0,00016	0,00016	0,00016	0,00016	0,00016	0,00016	0,00016	0,00016
MLL-YEO2002	0,00032	0,00026	0,00029	0,00018	0,00020	0,00033	0,00033	0,00020	0,00020	0,00029	0,00033	0,00032	0,00033	0,00033	0,00033	0,00033	0,00033
SOFT-NIE2002	0,00087	0,00087	0,00087	0,00085	0,00087	0,00087	0,00087	0,00087	0,00087	0,00087	0,00087	0,00087	0,00087	0,00087	0,00087	0,00087	0,00087
OVARIAN-PET2002	0,00030	0,00014	0,00014	0,00012	0,00012	0,00017	0,00017	0,00013	0,00013	0,00013	0,00013	0,00013	0,00013	0,00015	0,00015	0,00013	0,00013
OVARY-WEL2001	0,00014	0,00014	0,00014	0,00014	0,00014	0,00014	0,00014	0,00014	0,00014	0,00014	0,00014	0,00014	0,00014	0,00014	0,00014	0,00014	0,00014
CNS-POM2002	0,00049	0,00046	0,00048	0,00024	0,00027	0,00049	0,00052	0,00028	0,00028	0,00049	0,00049	0,00048	0,00049	0,00051	0,00053	0,00049	0,00049
PROSTATE-OUTCOME-SIN2002	0,00012	0,00008	0,00008	0,00008	0,00008	0,00008	0,00008	0,00008	0,00008	0,00010	0,00012	0,00009	0,00012	0,00012	0,00012	0,00012	0,00012
PROSTATE-SIN2002	0,00043	0,00042	0,00043	0,00020	0,00021	0,00044	0,00044	0,00021	0,00021	0,00043	0,00041	0,00040	0,00040	0,00043	0,00043	0,00043	0,00043
T-ALL-YEO2002	0,00008	0,00008	0,00008	0,00008	0,00008	0,00008	0,00008	0,00008	0,00008	0,00008	0,00008	0,00008	0,00008	0,00008	0,00008	0,00008	0,00008
TEL-AML1-YEO2002	0,00036	0,00025	0,00027	0,00019	0,00021	0,00035	0,00035	0,00024	0,00024	0,00026	0,00030	0,00029	0,00031	0,00033	0,00034	0,00033	0,00033
PROSTATE-WEL2001	0,00044	0,00044	0,00044	0,00041	0,00042	0,00044	0,00044	0,00042	0,00042	0,00044	0,00044	0,00044	0,00044	0,00044	0,00044	0,00044	0,00044

Tabela C.12: *Windowing: Desvio padrão dos valores originais da variável Compactação obtidos por validação cruzada de dez partições. Obs.: os valores mostrados na tabela correspondem aos valores originais multiplicados por* 10^2.

Base	J48	WPEWeC	WPEC	WEWeC	WEC	WPWeC	WPC	WWeC	WC	WPEWe	WPE	WEWe	WE	WPWe	WP	WWe	W
LYMPHOMA-ALI2000	0,00524	0,00413	0,00413	0,00662	0,00662	0,00413	0,00413	0,00662	0,00662	0,00530	0,00405	0,00405	0,00405	0,00413	0,00405	0,00413	0,00413
GIST-ALL2001	0,00000	0,00000	0,00000	0,00000	0,00000	0,00000	0,00000	0,00000	0,00000	0,00000	0,00000	0,00000	0,00000	0,00000	0,00000	0,00000	0,00000
COLON-ALO1999	0,00000	0,00764	0,00667	0,00897	0,00817	0,00000	0,00000	0,00817	0,00817	0,00500	0,00000	0,00500	0,00500	0,00000	0,00000	0,00000	0,00000
LEUKEMIA-GOL1999	0,00215	0,00215	0,00235	0,00000	0,00000	0,00141	0,00141	0,00000	0,00000	0,00188	0,00141	0,00215	0,00141	0,00141	0,00141	0,00141	0,00141
BCR-ABL-YEO2002	0,00000	0,00232	0,00240	0,00080	0,00107	0,00119	0,00198	0,00122	0,00122	0,00187	0,00187	0,00221	0,00221	0,00198	0,00198	0,00160	0,00160
MELANOMA-BIT2000	0,00000	0,00000	0,00000	0,00000	0,00000	0,00000	0,00000	0,00000	0,00000	0,00000	0,00000	0,00000	0,00000	0,00000	0,00000	0,00000	0,00000
BREAST-VEE2002	0,00088	0,00082	0,00041	0,00150	0,00150	0,00055	0,00063	0,00150	0,00139	0,00000	0,00000	0,00041	0,00000	0,00063	0,00063	0,00000	0,00000
NETWORKS-BUT2000	0,00211	0,00248	0,00248	0,00211	0,00184	0,00309	0,00309	0,00184	0,00184	0,00248	0,00138	0,00206	0,00138	0,00184	0,00225	0,00211	0,00211
DLBCL-NIH-ROS2002	0,00607	0,00424	0,00424	0,00501	0,00503	0,00375	0,00442	0,00352	0,00352	0,00315	0,00348	0,00348	0,00348	0,00352	0,00352	0,00315	0,00315
DLBCL-OUTCOME-SHI2002	0,00229	0,00214	0,00187	0,00215	0,00188	0,00140	0,00140	0,00188	0,00141	0,00214	0,00140	0,00233	0,00229	0,00140	0,00140	0,00214	0,00214
DLBCL-SHI2002	0,00229	0,00000	0,00187	0,00000	0,00000	0,00187	0,00187	0,00141	0,00141	0,00187	0,00229	0,00140	0,00214	0,00229	0,00229	0,00229	0,00229
E2A-PBX1-YEO2002	0,00000	0,00000	0,00000	0,00000	0,00000	0,00000	0,00000	0,00000	0,00000	0,00000	0,00000	0,00000	0,00000	0,00000	0,00000	0,00000	0,00000
ECML2004	0,00097	0,00108	0,00120	0,00171	0,00164	0,00110	0,00110	0,00164	0,00164	0,00096	0,00096	0,00096	0,00096	0,00077	0,00096	0,00096	0,00096
GCM-RAM2001	0,00231	0,00267	0,00229	0,00297	0,00297	0,00247	0,00229	0,00297	0,00297	0,00247	0,00263	0,00250	0,00263	0,00187	0,00187	0,00250	0,00250
GSE11665	0,00059	0,00059	0,00059	0,00059	0,00059	0,00059	0,00059	0,00059	0,00059	0,00059	0,00059	0,00059	0,00059	0,00059	0,00059	0,00059	0,00059
GSE3255	0,00000	0,00000	0,00000	0,00000	0,00000	0,00000	0,00000	0,00000	0,00000	0,00000	0,00000	0,00000	0,00000	0,00000	0,00000	0,00000	0,00000
GSE360	0,00107	0,00220	0,00231	0,00236	0,00169	0,00207	0,00212	0,00169	0,00169	0,00308	0,00107	0,00490	0,00131	0,00131	0,00107	0,00131	0,00131
GSE443	0,00000	0,00000	0,00000	0,00000	0,00000	0,00000	0,00000	0,00000	0,00000	0,00000	0,00000	0,00000	0,00000	0,00000	0,00000	0,00000	0,00000
GSE474	0,00000	0,00000	0,00000	0,00000	0,00000	0,00000	0,00000	0,00000	0,00000	0,00000	0,00000	0,00000	0,00000	0,00000	0,00000	0,00000	0,00000
GSE5473	0,00000	0,00000	0,00000	0,00000	0,00000	0,00000	0,00000	0,00000	0,00000	0,00000	0,00000	0,00000	0,00000	0,00000	0,00000	0,00000	0,00000
GSE7433	0,00000	0,00000	0,00000	0,00000	0,00000	0,00000	0,00000	0,00000	0,00000	0,00000	0,00000	0,00000	0,00000	0,00000	0,00000	0,00000	0,00000
GSE7898	0,00000	0,00000	0,00000	0,00000	0,00000	0,00000	0,00000	0,00000	0,00000	0,00000	0,00000	0,00000	0,00000	0,00000	0,00000	0,00000	0,00000
BREAST-HED2003	0,00000	0,00000	0,00000	0,00000	0,00000	0,00000	0,00000	0,00000	0,00000	0,00000	0,00000	0,00000	0,00000	0,00000	0,00000	0,00000	0,00000
HYPERDIP-50-YEO2002	0,00314	0,00220	0,00220	0,00412	0,00345	0,00331	0,00345	0,00345	0,00345	0,00291	0,00302	0,00357	0,00332	0,00318	0,00341	0,00302	0,00302
GASTRIC-LEU2002	0,00228	0,00414	0,00330	0,00229	0,00229	0,00243	0,00243	0,00229	0,00229	0,00243	0,00228	0,00334	0,00330	0,00228	0,00228	0,00199	0,00199
LUNG-BHA2001	0,00000	0,00000	0,00000	0,00122	0,00122	0,00000	0,00000	0,00122	0,00122	0,00000	0,00000	0,00000	0,00000	0,00000	0,00000	0,00079	0,00079
LUNG-BEE2002	0,00000	0,00000	0,00000	0,00000	0,00000	0,00000	0,00000	0,00000	0,00000	0,00000	0,00000	0,00000	0,00000	0,00000	0,00000	0,00000	0,00000
LUNG-GOR2002	0,00159	0,00131	0,00205	0,00000	0,00000	0,00177	0,00177	0,00000	0,00000	0,00107	0,00119	0,00131	0,00159	0,00105	0,00105	0,00105	0,00105
LUNG-WIG2002	0,00000	0,00000	0,00000	0,00347	0,00347	0,00000	0,00348	0,00000	0,00000	0,00000	0,00000	0,00000	0,00000	0,00000	0,00000	0,00464	0,00464
BREAST-MA2003	0,00514	0,01282	0,00685	0,02199	0,02033	0,00923	0,00000	0,02226	0,02226	0,01097	0,00923	0,01084	0,01084	0,00685	0,00685	0,01199	0,01199
MLL-ARM2002	0,00000	0,00000	0,00000	0,00000	0,00000	0,00000	0,00000	0,00000	0,00000	0,00000	0,00000	0,00000	0,00000	0,00000	0,00000	0,00000	0,00000
MLL-YEO2002	0,00000	0,00171	0,00131	0,00107	0,00215	0,00185	0,00185	0,00215	0,00215	0,00176	0,00079	0,00119	0,00079	0,00105	0,00105	0,00079	0,00079
SOFT-NIE2002	0,00241	0,00241	0,00241	0,00276	0,00241	0,00241	0,00241	0,00241	0,00241	0,00241	0,00241	0,00241	0,00241	0,00241	0,00241	0,00241	0,00241
OVARIAN-PET2002	0,00176	0,00066	0,00066	0,00088	0,00088	0,00148	0,00148	0,00066	0,00066	0,00000	0,00000	0,00000	0,00000	0,00101	0,00101	0,00000	0,00000
OVARY-WEL2001	0,00000	0,00000	0,00000	0,00000	0,00000	0,00000	0,00000	0,00000	0,00000	0,00000	0,00000	0,00000	0,00000	0,00000	0,00000	0,00000	0,00000
CNS-POM2002	0,00233	0,00214	0,00229	0,00300	0,00252	0,00233	0,00214	0,00296	0,00296	0,00233	0,00233	0,00310	0,00233	0,00229	0,00187	0,00233	0,00233
PROSTATE-OUTCOME-SIN2002	0,00133	0,00000	0,00000	0,00000	0,00000	0,00000	0,00000	0,00000	0,00000	0,00122	0,00133	0,00080	0,00133	0,00133	0,00133	0,00133	0,00133
PROSTATE-SIN2002	0,00270	0,00266	0,00270	0,00271	0,00242	0,00244	0,00244	0,00206	0,00206	0,00242	0,00259	0,00265	0,00265	0,00242	0,00242	0,00242	0,00242
T-ALL-YEO2002	0,00000	0,00000	0,00000	0,00000	0,00000	0,00000	0,00000	0,00000	0,00000	0,00000	0,00000	0,00000	0,00000	0,00000	0,00000	0,00000	0,00000
TEL-AML1-YEO2002	0,00244	0,00221	0,00177	0,00177	0,00244	0,00176	0,00176	0,00206	0,00206	0,00171	0,00159	0,00296	0,00186	0,00143	0,00169	0,00143	0,00143
PROSTATE-WEL2001	0,00129	0,00129	0,00129	0,00158	0,00169	0,00129	0,00129	0,00169	0,00169	0,00129	0,00129	0,00129	0,00129	0,00129	0,00129	0,00129	0,00129

Tabela C.13: *Windowing: Média dos valores originais da variável Coesão-Compactação obtidos por validação cruzada de dez partições.*

Base	J48	WPEWeC	WPEC	WEWeC	WEC	WPWeC	WPC	WWeC	WC	WPEWe	WPE	WEWe	WE	WPWe	WP	WWe	W
LYMPHOMA-ALI2000	0,02755	0,02908	0,02908	0,02897	0,02897	0,02908	0,02908	0,02897	0,02897	0,02846	0,02878	0,02878	0,02878	0,02908	0,02878	0,02908	0,02908
GIST-ALL2001	0,01295	0,01295	0,01295	0,01295	0,01295	0,01295	0,01295	0,01295	0,01295	0,01295	0,01295	0,01295	0,01295	0,01295	0,01295	0,01295	0,01295
COLON-ALO1999	0,03000	0,02771	0,02847	0,02294	0,02542	0,03000	0,03000	0,02542	0,02542	0,02924	0,03000	0,02924	0,02924	0,03000	0,03000	0,03000	0,03000
LEUKEMIA-GOL1999	0,01034	0,00834	0,00934	0,00684	0,00684	0,01134	0,01134	0,00684	0,00684	0,01084	0,01134	0,01034	0,01134	0,01134	0,01134	0,01134	0,01134
BCR-ABL-YEO2002	0,01457	0,01236	0,01262	0,00923	0,00953	0,01454	0,01498	0,00984	0,00984	0,01214	0,01214	0,01210	0,01210	0,01498	0,01498	0,01245	0,01245
MELANOMA-BIT2000	0,00996	0,00996	0,00996	0,00996	0,00996	0,00996	0,00996	0,00996	0,00996	0,00996	0,00996	0,00996	0,00996	0,00996	0,00996	0,00996	0,00996
BREAST-VEE2002	0,01251	0,01173	0,01191	0,00686	0,00686	0,01237	0,01252	0,00686	0,00734	0,01208	0,01208	0,01191	0,01208	0,01252	0,01252	0,01208	0,01208
NETWORKS-BUT2000	0,02553	0,02300	0,02300	0,02212	0,02235	0,02385	0,02385	0,02235	0,02235	0,02470	0,02513	0,02491	0,02513	0,02533	0,02573	0,02553	0,02553
DLBCL-NIH-ROS2002	0,04273	0,04211	0,04211	0,04130	0,04146	0,04367	0,04412	0,04275	0,04275	0,04244	0,04228	0,04228	0,04228	0,04337	0,04337	0,04244	0,04244
DLBCL-OUTCOME-SHI2002	0,01796	0,01831	0,01865	0,01034	0,01084	0,01900	0,01900	0,01084	0,01134	0,01831	0,01900	0,01762	0,01796	0,01900	0,01900	0,01831	0,01831
DLBCL-SHI2002	0,01427	0,01184	0,01265	0,00684	0,00684	0,01265	0,01265	0,00734	0,00734	0,01265	0,01346	0,01225	0,01306	0,01346	0,01346	0,01346	0,01346
E2A-PBX1-YEO2002	0,00515	0,00515	0,00515	0,00515	0,00515	0,00515	0,00515	0,00515	0,00515	0,00515	0,00515	0,00515	0,00515	0,00515	0,00515	0,00515	0,00515
ECML2004	0,01897	0,01883	0,01883	0,01648	0,01677	0,01890	0,01890	0,01677	0,01677	0,01904	0,01904	0,01904	0,01904	0,01908	0,01904	0,01906	0,01906
GCM-RAM2001	0,02549	0,02458	0,02479	0,02129	0,02129	0,02539	0,02539	0,02129	0,02129	0,02539	0,02539	0,02559	0,02539	0,02582	0,02582	0,02559	0,02559
GSE11665	0,00306	0,00306	0,00306	0,00306	0,00306	0,00306	0,00306	0,00306	0,00306	0,00306	0,00306	0,00306	0,00306	0,00306	0,00306	0,00306	0,00306
GSE3255	0,00694	0,00694	0,00694	0,00694	0,00694	0,00694	0,00694	0,00694	0,00694	0,00694	0,00694	0,00694	0,00694	0,00694	0,00694	0,00694	0,00694
GSE360	0,01566	0,01491	0,01513	0,01103	0,01335	0,01524	0,01555	0,01335	0,01335	0,01483	0,01571	0,01352	0,01551	0,01557	0,01571	0,01557	0,01557
GSE443	0,00514	0,00514	0,00514	0,00514	0,00514	0,00514	0,00514	0,00514	0,00514	0,00514	0,00514	0,00514	0,00514	0,00514	0,00514	0,00514	0,00514
GSE474	0,00599	0,00599	0,00599	0,00599	0,00599	0,00599	0,00599	0,00599	0,00599	0,00599	0,00599	0,00599	0,00599	0,00599	0,00599	0,00599	0,00599
GSE5473	0,00494	0,00494	0,00494	0,00494	0,00494	0,00494	0,00494	0,00494	0,00494	0,00494	0,00494	0,00494	0,00494	0,00494	0,00494	0,00494	0,00494
GSE7433	0,00425	0,00425	0,00425	0,00425	0,00425	0,00425	0,00425	0,00425	0,00425	0,00425	0,00425	0,00425	0,00425	0,00425	0,00425	0,00425	0,00425
GSE7898	0,00309	0,00309	0,00309	0,00309	0,00309	0,00309	0,00309	0,00309	0,00309	0,00309	0,00309	0,00309	0,00309	0,00309	0,00309	0,00309	0,00309
BREAST-HED2003	0,00834	0,00834	0,00834	0,00834	0,00834	0,00834	0,00834	0,00834	0,00834	0,00834	0,00834	0,00834	0,00834	0,00834	0,00834	0,00834	0,00834
HYPERDIP-50-YEO2002	0,02250	0,02059	0,02059	0,01720	0,01856	0,02282	0,02265	0,01896	0,01896	0,02074	0,02055	0,01938	0,02034	0,02349	0,02332	0,02091	0,02091
GASTRIC-LEU2002	0,01743	0,01517	0,01633	0,01215	0,01215	0,01708	0,01708	0,01215	0,01215	0,01708	0,01743	0,01668	0,01703	0,01743	0,01743	0,01778	0,01778
LUNG-BHA2001	0,01409	0,01409	0,01409	0,01342	0,01342	0,01409	0,01409	0,01342	0,01342	0,01409	0,01409	0,01409	0,01409	0,01409	0,01409	0,01429	0,01429
LUNG-BEE2002	0,00684	0,00684	0,00684	0,00684	0,00684	0,00684	0,00684	0,00684	0,00684	0,00684	0,00684	0,00684	0,00684	0,00684	0,00684	0,00684	0,00684
LUNG-GOR2002	0,01130	0,00742	0,00872	0,00516	0,00516	0,01039	0,01039	0,00516	0,00516	0,00818	0,00886	0,00667	0,00810	0,00954	0,00954	0,00954	0,00954
LUNG-WIG2002	0,01863	0,01863	0,01863	0,01785	0,01785	0,01863	0,01927	0,01863	0,01863	0,01863	0,01863	0,01863	0,01863	0,01863	0,01863	0,01991	0,01991
BREAST-MA2003	0,04481	0,04102	0,04429	0,03046	0,03186	0,04578	0,04534	0,03302	0,03302	0,04372	0,04477	0,04525	0,04525	0,04631	0,04631	0,04473	0,04473
MLL-ARM2002	0,00797	0,00797	0,00797	0,00797	0,00797	0,00797	0,00797	0,00797	0,00797	0,00797	0,00797	0,00797	0,00797	0,00797	0,00797	0,00797	0,00797
MLL-YEO2002	0,01457	0,01271	0,01353	0,00953	0,01036	0,01474	0,01474	0,01036	0,01036	0,01350	0,01480	0,01454	0,01480	0,01503	0,01503	0,01480	0,01480
SOFT-NIE2002	0,01964	0,01964	0,01964	0,01942	0,01964	0,01964	0,01964	0,01964	0,01964	0,01964	0,01964	0,01964	0,01964	0,01964	0,01964	0,01964	0,01964
OVARIAN-PET2002	0,01446	0,00840	0,00840	0,00744	0,00744	0,00947	0,00947	0,00778	0,00778	0,00812	0,00812	0,00812	0,00812	0,00896	0,00896	0,00812	0,00812
OVARY-WEL2001	0,00684	0,00684	0,00684	0,00684	0,00684	0,00684	0,00684	0,00684	0,00684	0,00684	0,00684	0,00684	0,00684	0,00684	0,00684	0,00684	0,00684
CNS-POM2002	0,01762	0,01693	0,01727	0,01025	0,01125	0,01762	0,01831	0,01165	0,01165	0,01762	0,01762	0,01721	0,01762	0,01796	0,01865	0,01762	0,01762
PROSTATE-OUTCOME-SIN2002	0,00703	0,00514	0,00514	0,00514	0,00514	0,00514	0,00514	0,00514	0,00514	0,00627	0,00703	0,00552	0,00703	0,00703	0,00703	0,00703	0,00703
PROSTATE-SIN2002	0,01757	0,01736	0,01756	0,01020	0,01057	0,01779	0,01779	0,01095	0,01095	0,01758	0,01715	0,01671	0,01671	0,01758	0,01758	0,01758	0,01758
T-ALL-YEO2002	0,00515	0,00515	0,00515	0,00515	0,00515	0,00515	0,00515	0,00515	0,00515	0,00515	0,00515	0,00515	0,00515	0,00515	0,00515	0,00515	0,00515
TEL-AML1-YEO2002	0,01566	0,01210	0,01297	0,01010	0,01059	0,01547	0,01547	0,01184	0,01184	0,01271	0,01402	0,01334	0,01425	0,01477	0,01523	0,01477	0,01477
PROSTATE-WEL2001	0,01402	0,01402	0,01402	0,01330	0,01350	0,01402	0,01402	0,01350	0,01350	0,01402	0,01402	0,01402	0,01402	0,01402	0,01402	0,01402	0,01402

Tabela C.14: *Windowing: Desvio padrão dos valores originais da variável Coesão-Compactação obtidos por validação cruzada de dez partições. Obs.: os valores mostrados na tabela correspondem aos valores originais multiplicados por* 10^2.

Base	J48	WPEWeC	WPEC	WEWeC	WEC	WPWeC	WPC	WWeC	WC	WPEWe	WPE	WEWe	WE	WPWe	WP	WWe	W
LYMPHOMA-ALI2000	0,06498	0,04998	0,04998	0,06912	0,06912	0,04998	0,04998	0,06912	0,06912	0,06514	0,04897	0,04897	0,04897	0,04998	0,04897	0,04998	0,04998
GIST-ALL2001	0,00000	0,00000	0,00000	0,00000	0,00000	0,00000	0,00000	0,00000	0,00000	0,00000	0,00000	0,00000	0,00000	0,00000	0,00000	0,00000	0,00000
COLON-ALO1999	0,00000	0,11669	0,10185	0,15005	0,12475	0,00000	0,00000	0,12475	0,12475	0,07639	0,00000	0,07639	0,07639	0,00000	0,00000	0,00000	0,00000
LEUKEMIA-GOL1999	0,07647	0,07647	0,08343	0,00000	0,00000	0,05006	0,05006	0,00000	0,00000	0,06675	0,05006	0,07647	0,05006	0,05006	0,05006	0,05006	0,05006
BCR-ABL-YEO2002	0,00000	0,08177	0,08450	0,03048	0,04064	0,03652	0,05712	0,04656	0,04656	0,06547	0,06547	0,07800	0,07800	0,05712	0,05712	0,05509	0,05509
MELANOMA-BIT2000	0,00000	0,00000	0,00000	0,00000	0,00000	0,00000	0,00000	0,00000	0,00000	0,00000	0,00000	0,00000	0,00000	0,00000	0,00000	0,00000	0,00000
BREAST-VEE2002	0,03270	0,03177	0,01641	0,07768	0,07768	0,01973	0,02261	0,07768	0,07059	0,00000	0,00000	0,01641	0,00000	0,02261	0,02261	0,00000	0,00000
NETWORKS-BUT2000	0,03103	0,03892	0,03892	0,03414	0,02980	0,04837	0,04837	0,02980	0,02980	0,03753	0,02031	0,03100	0,02031	0,02708	0,03317	0,03103	0,03103
DLBCL-NIH-ROS2002	0,06980	0,05024	0,05024	0,06098	0,06130	0,04286	0,04983	0,04027	0,04027	0,03673	0,04072	0,04072	0,04072	0,04045	0,04045	0,03673	0,03673
DLBCL-OUTCOME-SHI2002	0,05635	0,05272	0,04601	0,07647	0,06675	0,03451	0,03451	0,06675	0,05006	0,05272	0,03451	0,05752	0,05635	0,03451	0,03451	0,05272	0,05272
DLBCL-SHI2002	0,06607	0,00000	0,05395	0,00000	0,00000	0,05395	0,05395	0,05006	0,05006	0,05395	0,06607	0,04046	0,06180	0,06607	0,06607	0,06607	0,06607
E2A-PBX1-YEO2002	0,00000	0,00000	0,00000	0,00000	0,00000	0,00000	0,00000	0,00000	0,00000	0,00000	0,00000	0,00000	0,00000	0,00000	0,00000	0,00000	0,00000
ECML2004	0,01681	0,01932	0,02187	0,03419	0,03265	0,01964	0,01964	0,03265	0,03265	0,01656	0,01656	0,01656	0,01656	0,01438	0,01656	0,01709	0,01709
GCM-RAM2001	0,03685	0,04383	0,03771	0,05132	0,05132	0,03934	0,03637	0,05132	0,05132	0,03977	0,04231	0,04019	0,04231	0,03009	0,03009	0,04019	0,04019
GSE11665	0,05107	0,05107	0,05107	0,05107	0,05107	0,05107	0,05107	0,05107	0,05107	0,05107	0,05107	0,05107	0,05107	0,05107	0,05107	0,05107	0,05107
GSE3255	0,00000	0,00000	0,00000	0,00000	0,00000	0,00000	0,00000	0,00000	0,00000	0,00000	0,00000	0,00000	0,00000	0,00000	0,00000	0,00000	0,00000
GSE360	0,01827	0,03534	0,03306	0,05970	0,03022	0,02976	0,03129	0,03022	0,03022	0,05504	0,01995	0,10130	0,02214	0,01909	0,01995	0,01909	0,01909
GSE443	0,00000	0,00000	0,00000	0,00000	0,00000	0,00000	0,00000	0,00000	0,00000	0,00000	0,00000	0,00000	0,00000	0,00000	0,00000	0,00000	0,00000
GSE474	0,00000	0,00000	0,00000	0,00000	0,00000	0,00000	0,00000	0,00000	0,00000	0,00000	0,00000	0,00000	0,00000	0,00000	0,00000	0,00000	0,00000
GSE5473	0,00000	0,00000	0,00000	0,00000	0,00000	0,00000	0,00000	0,00000	0,00000	0,00000	0,00000	0,00000	0,00000	0,00000	0,00000	0,00000	0,00000
GSE7433	0,00000	0,00000	0,00000	0,00000	0,00000	0,00000	0,00000	0,00000	0,00000	0,00000	0,00000	0,00000	0,00000	0,00000	0,00000	0,00000	0,00000
GSE7898	0,00000	0,00000	0,00000	0,00000	0,00000	0,00000	0,00000	0,00000	0,00000	0,00000	0,00000	0,00000	0,00000	0,00000	0,00000	0,00000	0,00000
BREAST-HED2003	0,00000	0,00000	0,00000	0,00000	0,00000	0,00000	0,00000	0,00000	0,00000	0,00000	0,00000	0,00000	0,00000	0,00000	0,00000	0,00000	0,00000
HYPERDIP-50-YEO2002	0,06721	0,05056	0,05056	0,12214	0,09193	0,07030	0,07312	0,09232	0,09232	0,06701	0,06937	0,08398	0,07735	0,06806	0,07277	0,06950	0,06950
GASTRIC-LEU2002	0,05363	0,10429	0,08143	0,06170	0,06170	0,05734	0,05734	0,06170	0,06170	0,05734	0,05363	0,08252	0,08194	0,05363	0,05363	0,04681	0,04681
LUNG-BHA2001	0,00000	0,00000	0,00000	0,03373	0,03373	0,00000	0,00000	0,03373	0,03373	0,00000	0,00000	0,00000	0,00000	0,00000	0,00000	0,02030	0,02030
LUNG-BEE2002	0,00000	0,00000	0,00000	0,00000	0,00000	0,00000	0,00000	0,00000	0,00000	0,00000	0,00000	0,00000	0,00000	0,00000	0,00000	0,00000	0,00000
LUNG-GOR2002	0,06827	0,06165	0,08834	0,00000	0,00000	0,07379	0,07379	0,00000	0,00000	0,05033	0,05111	0,06165	0,07089	0,04069	0,04069	0,04069	0,04069
LUNG-WIG2002	0,00000	0,00000	0,00000	0,07876	0,07876	0,00000	0,06366	0,00000	0,00000	0,00000	0,00000	0,00000	0,00000	0,00000	0,00000	0,08488	0,08488
BREAST-MA2003	0,05265	0,13713	0,07020	0,27549	0,25262	0,08985	0,00000	0,27572	0,27572	0,10956	0,09168	0,10659	0,10659	0,06456	0,06456	0,11847	0,11847
MLL-ARM2002	0,00000	0,00000	0,00000	0,00000	0,00000	0,00000	0,00000	0,00000	0,00000	0,00000	0,00000	0,00000	0,00000	0,00000	0,00000	0,00000	0,00000
MLL-YEO2002	0,00000	0,05862	0,04246	0,04064	0,07639	0,05676	0,05676	0,07639	0,07639	0,05537	0,02292	0,03652	0,02292	0,03056	0,03056	0,02292	0,02292
SOFT-NIE2002	0,04355	0,04355	0,04355	0,04521	0,04355	0,04355	0,04355	0,04355	0,04355	0,04355	0,04355	0,04355	0,04355	0,04355	0,04355	0,04355	0,04355
OVARIAN-PET2002	0,05937	0,02776	0,02776	0,04577	0,04577	0,05908	0,05908	0,03433	0,03433	0,00000	0,00000	0,00000	0,00000	0,04240	0,04240	0,00000	0,00000
OVARY-WEL2001	0,00000	0,00000	0,00000	0,00000	0,00000	0,00000	0,00000	0,00000	0,00000	0,00000	0,00000	0,00000	0,00000	0,00000	0,00000	0,00000	0,00000
CNS-POM2002	0,05752	0,05272	0,05635	0,10060	0,08362	0,05752	0,05272	0,09574	0,09574	0,05752	0,05752	0,08060	0,05752	0,05635	0,04601	0,05752	0,05752
PROSTATE-OUTCOME-SIN2002	0,06277	0,00000	0,00000	0,00000	0,00000	0,00000	0,00000	0,00000	0,00000	0,05753	0,06277	0,03766	0,06277	0,06277	0,06277	0,06277	0,06277
PROSTATE-SIN2002	0,07537	0,07594	0,07716	0,10707	0,09302	0,07031	0,07031	0,07434	0,07434	0,06973	0,07406	0,07541	0,07541	0,06973	0,06973	0,06973	0,06973
T-ALL-YEO2002	0,00000	0,00000	0,00000	0,00000	0,00000	0,00000	0,00000	0,00000	0,00000	0,00000	0,00000	0,00000	0,00000	0,00000	0,00000	0,00000	0,00000
TEL-AML1-YEO2002	0,06483	0,07800	0,06072	0,06385	0,09316	0,04888	0,04888	0,07301	0,07301	0,05862	0,05000	0,10052	0,05750	0,04330	0,05105	0,04330	0,04330
PROSTATE-WEL2001	0,03209	0,03209	0,03209	0,03553	0,03808	0,03209	0,03209	0,03808	0,03808	0,03209	0,03209	0,03209	0,03209	0,03209	0,03209	0,03209	0,03209

Tabela C.15: *Windowing: Média dos valores originais da variável Acurácia obtidos por validação cruzada de dez partições.*

Base	J48	WPEWeC	WPEC	WEWeC	WEC	WPWeC	WPC	WWeC	WC	WPEWe	WPE	WEWe	WE	WPWe	WP	WWe	W
LYMPHOMA-ALI2000	81,44444	76,11111	76,11111	66,66667	63,33333	76,11111	77,22222	64,44444	63,33333	77,55556	77,55556	72,55556	72,55556	73,44444	82,55556	70,33333	70,33333
GIST-ALL2001	80,00000	95,00000	95,00000	95,00000	95,00000	95,00000	95,00000	95,00000	95,00000	95,00000	95,00000	95,00000	95,00000	95,00000	95,00000	95,00000	95,00000
COLON-ALO1999	82,14286	79,28571	77,14286	72,85714	74,52381	80,47619	82,14286	74,52381	74,52381	75,95238	74,04762	75,71429	75,71429	74,28571	77,14286	77,38095	77,38095
LEUKEMIA-GOL1999	78,92857	83,21429	86,07143	92,85714	95,71429	79,10714	81,96429	95,71429	95,71429	76,07143	76,07143	78,92857	77,50000	76,07143	76,07143	77,50000	77,50000
BCR-ABL-YEO2002	93,87311	96,01326	95,70076	92,34848	92,95455	93,57955	93,57955	92,95455	92,95455	95,11364	95,11364	94,48864	94,48864	94,49811	94,49811	94,48864	94,48864
MELANOMA-BIT2000	84,16667	69,16667	65,83333	59,16667	59,16667	69,16667	69,16667	59,16667	59,16667	76,66667	76,66667	76,66667	76,66667	81,66667	79,16667	76,66667	76,66667
BREAST-VEE2002	63,11111	59,77778	58,66667	61,88889	59,88889	58,66667	61,66667	59,88889	63,00000	59,66667	64,00000	57,88889	62,00000	58,55556	59,55556	61,88889	61,88889
NETWORKS-BUT2000	30,95238	20,47619	19,04762	26,42857	22,14286	20,71429	20,71429	23,57143	22,14286	32,14286	32,14286	32,61905	32,38095	36,90476	25,00000	25,00000	25,00000
DLBCL-NIH-ROS2002	52,08333	50,83333	49,16667	56,66667	56,25000	52,08333	54,58333	59,16667	59,16667	50,41667	52,50000	50,00000	50,00000	54,16667	54,16667	51,66667	51,66667
DLBCL-OUT-SHI2002	53,33333	49,66667	51,33333	52,00000	57,00000	53,33333	53,33333	62,66667	61,00000	54,66667	53,00000	53,00000	51,00000	54,66667	54,66667	54,66667	54,66667
DLBCL-SHI2002	72,50000	80,35714	80,35714	84,28571	84,28571	80,35714	80,53571	85,53571	85,53571	77,67857	80,35714	77,67857	78,92857	76,07143	76,07143	80,35714	80,35714
E2A-PBX1-YEO2002	100,00000	100,00000	100,00000	100,00000	100,00000	100,00000	100,00000	100,00000	100,00000	100,00000	100,00000	100,00000	100,00000	100,00000	100,00000	100,00000	100,00000
ECML2004	14,10714	11,07143	11,07143	14,82143	8,21429	9,64286	9,64286	8,21429	8,21429	11,25000	14,10714	14,10714	14,10714	11,25000	12,67857	12,67857	12,67857
GCM-RAM2001	58,94737	57,89474	56,31579	47,36842	47,36842	56,31579	54,73684	47,36842	47,36842	60,52632	60,00000	64,73684	64,73684	60,52632	60,00000	64,21053	64,21053
GSE11665	80,00000	80,00000	80,00000	80,00000	80,00000	80,00000	80,00000	80,00000	80,00000	80,00000	80,00000	80,00000	80,00000	80,00000	80,00000	80,00000	80,00000
GSE3255	46,66667	58,33333	58,33333	48,33333	53,33333	58,33333	68,33333	53,33333	53,33333	58,33333	51,66667	51,66667	51,66667	53,33333	53,33333	51,66667	51,66667
GSE360	8,33333	8,33333	11,66667	3,33333	16,66667	11,66667	15,00000	16,66667	16,66667	11,66667	11,66667	6,66667	11,66667	11,66667	8,33333	11,66667	11,66667
GSE443	80,00000	75,00000	75,00000	65,00000	65,00000	45,00000	45,00000	65,00000	65,00000	85,00000	85,00000	85,00000	85,00000	45,00000	45,00000	85,00000	85,00000
GSE474	26,66667	36,66667	41,66667	51,66667	53,33333	38,33333	40,00000	53,33333	53,33333	21,66667	21,66667	26,66667	21,66667	21,66667	18,33333	21,66667	21,66667
GSE5473	36,66667	65,00000	65,00000	58,33333	53,33333	65,00000	71,66667	53,33333	53,33333	36,66667	36,66667	36,66667	36,66667	36,66667	36,66667	36,66667	36,66667
GSE7433	55,00000	45,00000	45,00000	35,00000	30,00000	45,00000	35,00000	30,00000	30,00000	60,00000	60,00000	60,00000	60,00000	60,00000	60,00000	60,00000	60,00000
GSE7898	40,00000	40,00000	40,00000	40,00000	40,00000	50,00000	50,00000	40,00000	40,00000	40,00000	40,00000	40,00000	40,00000	60,00000	60,00000	40,00000	40,00000
BREAST-HED2003	80,00000	65,00000	65,00000	50,00000	50,00000	65,00000	60,00000	50,00000	50,00000	80,00000	80,00000	80,00000	80,00000	80,00000	80,00000	80,00000	80,00000
HYPERDIP-YEO2002	88,36174	87,20644	87,80303	86,83712	86,53409	86,88447	87,79356	85,62500	85,62500	88,12500	87,82197	87,78409	89,64015	89,36553	89,36553	88,75947	88,75947
GASTRIC-LEU2002	91,21795	84,10256	89,03846	84,10256	83,26923	87,43590	87,43590	84,03846	84,03846	84,93590	85,70513	85,70513	88,07692	87,37179	86,53846	88,14103	88,14103
LUNG-BHA2001	93,11905	92,16667	92,16667	86,71429	86,71429	92,16667	91,16667	86,71429	86,71429	92,66667	92,66667	91,16667	92,66667	91,16667	91,66667	92,66667	92,66667
LUNG-BEE2002	99,00000	94,88889	94,88889	89,55556	89,55556	96,00000	96,00000	89,55556	89,55556	97,88889	97,88889	97,88889	97,88889	99,00000	99,00000	97,88889	97,88889
LUNG-GOR2002	95,05848	93,91813	93,36257	94,53216	94,53216	94,47368	94,47368	95,58480	95,58480	97,25146	96,69591	96,14035	96,14035	96,69591	96,69591	96,14035	96,14035
LUNG-WIG2002	84,16667	86,66667	84,16667	66,66667	63,33333	86,66667	81,66667	65,83333	65,83333	84,16667	84,16667	86,66667	86,66667	86,66667	86,66667	76,66667	76,66667
BREAST-MA2003	40,71429	41,19048	39,52381	49,76190	49,76190	39,52381	39,76190	46,42857	46,42857	54,04762	50,71429	40,47619	40,47619	52,85714	55,71429	39,52381	39,52381
MLL-ARM2002	84,82143	73,57143	73,57143	68,39286	68,39286	73,57143	73,57143	68,39286	68,39286	80,89286	82,14286	82,14286	82,14286	80,89286	86,25000	82,14286	82,14286
MLL-YEO2002	96,32576	96,02273	95,10417	93,83523	95,06629	95,71023	96,02273	96,00379	96,00379	96,63826	95,70076	96,30682	95,70076	96,61932	96,61932	95,70076	95,70076
SOFT-NIE2002	41,00000	44,00000	44,00000	37,50000	37,50000	44,00000	44,00000	38,00000	37,50000	42,00000	42,00000	42,00000	42,00000	33,00000	36,50000	38,00000	38,00000
OVARIAN-PET2002	95,60000	98,80000	98,80000	96,00000	96,00000	98,80000	98,80000	96,00000	96,00000	98,80000	98,80000	98,80000	98,80000	98,80000	98,80000	98,80000	98,80000
OVARY-WEL2001	85,00000	90,00000	90,00000	77,50000	77,50000	75,83333	75,83333	77,50000	77,50000	87,50000	87,50000	87,50000	87,50000	80,00000	80,00000	87,50000	87,50000
CNS-POM2002	58,33333	56,66667	56,66667	48,33333	51,66667	56,66667	60,00000	53,33333	53,33333	46,66667	43,33333	46,66667	45,00000	45,00000	51,66667	45,00000	45,00000
PROST-OUT-SIN2002	33,33333	33,33333	33,33333	66,66667	66,66667	33,33333	38,33333	66,66667	66,66667	36,66667	33,33333	46,66667	33,33333	33,33333	38,33333	33,33333	33,33333
PROSTATE-SIN2002	79,45055	80,21978	81,59341	75,54945	74,12088	80,98901	82,41758	74,12088	74,12088	81,59341	78,68132	77,91209	78,68132	80,16484	80,16484	80,16484	80,16484
T-ALL-YEO2002	99,68750	99,68750	99,68750	99,68750	99,68750	99,68750	99,68750	99,68750	99,68750	99,68750	99,68750	99,68750	99,68750	99,68750	99,68750	99,68750	99,68750
TEL-AML1-YEO2002	94,49811	95,72917	96,03220	94,46970	95,37879	96,65720	96,96023	95,69129	95,69129	96,02273	96,62879	97,23485	97,24432	96,31629	95,70076	97,54735	97,54735
PROSTATE-WEL2001	60,00000	60,00000	60,00000	45,00000	47,00000	58,00000	58,00000	47,00000	47,00000	65,00000	64,00000	64,00000	64,00000	59,50000	59,50000	64,00000	64,00000

Tabela C.16: *Windowing: Desvio padrão dos valores originais da variável Acurácia obtidos por validação cruzada de dez partições.*

Base	J48	WPEWeC	WPEC	WEWeC	WEC	WPWeC	WPC	WWeC	WC	WPEWe	WPE	WEWe	WE	WPWe	WP	WWe	W
LYMPHOMA-ALI2000	3,280	4,031	4,031	4,563	3,934	4,031	3,892	3,845	3,934	5,372	5,575	6,327	6,327	5,120	4,711	5,605	5,605
GIST-ALL2001	11,055	5,000	5,000	5,000	5,000	5,000	5,000	5,000	5,000	5,000	5,000	5,000	5,000	5,000	5,000	5,000	5,000
COLON-ALO1999	5,110	4,789	7,056	3,940	4,709	5,333	5,110	4,709	4,709	5,423	4,454	4,458	4,458	5,427	5,009	5,048	5,048
LEUKEMIA-GOL1999	4,943	2,906	3,021	3,195	2,182	2,352	2,970	2,182	2,182	5,729	5,729	6,167	5,778	5,729	5,729	5,778	5,778
BCR-ABL-YEO2002	0,795	0,659	0,687	1,243	1,145	1,459	1,459	1,145	1,145	0,669	0,669	0,770	0,770	0,756	0,756	0,770	0,770
MELANOMA-BIT2000	4,383	4,977	3,611	8,375	8,375	4,977	4,977	8,375	8,375	5,932	5,932	5,932	5,932	4,083	7,375	5,932	5,932
BREAST-VEE2002	6,285	4,865	3,784	5,444	5,058	4,817	4,445	5,058	5,154	3,659	4,101	3,738	4,514	4,539	4,682	5,019	5,019
NETWORKS-BUT2000	4,682	2,222	2,100	3,557	4,462	3,921	3,921	4,411	4,462	4,563	4,563	5,336	4,644	4,507	3,079	3,079	3,079
DLBCL-NIH-ROS2002	3,121	3,391	3,819	2,576	2,795	3,688	3,750	3,716	3,716	2,739	2,992	4,648	4,648	2,240	2,240	2,992	2,992
DLBCL-OUTCOME-SHI2002	3,651	3,386	3,791	6,907	5,222	3,651	3,651	5,183	5,307	4,949	3,238	4,081	5,141	3,485	3,485	5,778	5,778
DLBCL-SHI2002	5,108	3,501	2,305	4,686	4,686	2,305	3,005	4,575	4,575	4,327	3,501	2,882	3,550	5,052	5,052	3,501	3,501
E2A-PBX1-YEO2002	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000
ECML2004	4,263	4,000	4,000	4,038	3,613	4,126	4,126	3,613	3,613	3,552	4,765	4,765	4,765	3,552	4,489	4,489	4,489
GCM-RAM2001	3,310	3,138	3,333	4,368	4,368	2,724	3,059	4,368	4,368	3,617	3,697	2,083	2,083	3,259	3,347	2,456	2,456
GSE11665	8,165	8,165	8,165	8,165	8,165	8,165	8,165	8,165	8,165	8,165	8,165	8,165	8,165	8,165	8,165	8,165	8,165
GSE3255	11,863	11,719	11,719	10,672	11,863	11,719	7,222	11,863	11,863	11,719	12,031	12,031	12,031	11,863	11,863	12,031	12,031
GSE360	5,693	5,693	6,111	3,333	7,027	6,111	6,310	7,027	7,027	6,111	6,111	4,445	6,111	6,111	5,693	6,111	6,111
GSE443	13,333	13,437	13,437	15,000	15,000	15,723	15,723	15,000	15,000	10,672	10,672	10,672	10,672	15,723	15,723	10,672	10,672
GSE474	7,935	6,939	9,379	10,957	8,165	8,975	11,706	8,165	8,165	7,876	7,876	7,935	7,876	7,876	6,310	7,876	7,876
GSE5473	8,535	8,407	8,407	10,614	11,055	8,407	8,259	11,055	11,055	8,535	8,535	8,535	8,535	8,535	8,535	8,535	8,535
GSE7433	13,844	13,844	13,844	13,017	13,333	13,844	13,017	13,333	13,333	12,472	12,472	12,472	12,472	12,472	12,472	12,472	12,472
GSE7898	16,330	16,330	16,330	16,330	16,330	16,667	16,667	16,330	16,330	16,330	16,330	16,330	16,330	16,330	16,330	16,330	16,330
BREAST-HED2003	8,165	10,672	10,672	10,541	10,541	10,672	12,472	10,541	10,541	8,165	8,165	8,165	8,165	8,165	8,165	8,165	8,165
HYPERDIP-50-YEO2002	1,112	2,041	1,900	1,779	1,667	1,963	1,912	1,380	1,380	2,028	1,953	2,068	1,499	1,973	1,973	2,297	2,297
GASTRIC-LEU2002	2,503	3,421	2,874	2,447	2,298	2,045	2,045	2,188	2,188	2,420	2,253	2,352	2,207	2,085	2,406	1,776	1,776
LUNG-BHA2001	2,370	2,205	2,205	2,653	2,653	2,205	2,730	2,653	2,653	2,077	2,077	2,730	2,077	2,730	2,431	2,077	2,077
LUNG-BEE2002	1,000	2,266	2,266	2,236	2,236	2,211	2,211	2,236	2,236	1,410	1,410	1,410	1,410	1,000	1,000	1,410	1,410
LUNG-GOR2002	1,266	1,935	1,817	1,976	1,976	1,657	1,657	1,612	1,612	1,235	1,223	1,181	1,664	1,477	1,477	1,443	1,443
LUNG-WIG2002	4,383	4,513	4,383	5,271	6,236	4,513	5,528	6,143	6,143	4,383	4,383	4,513	4,513	4,513	4,513	7,935	7,935
BREAST-MA2003	5,848	5,789	5,747	6,233	7,156	3,810	5,778	7,056	7,056	7,777	5,601	6,428	6,428	7,527	7,029	5,747	5,747
MLL-ARM2002	3,851	6,104	6,104	5,881	6,255	6,104	6,104	6,255	6,255	4,533	4,525	4,525	4,525	4,533	3,552	4,525	4,525
MLL-YEO2002	1,192	0,801	1,226	1,400	1,251	1,319	1,125	1,232	1,232	0,715	1,322	1,213	1,322	1,341	1,341	1,322	1,322
SOFT-NIE2002	5,044	4,521	4,521	7,967	8,275	4,521	4,521	8,172	8,275	7,000	7,930	7,930	7,930	7,789	5,480	8,982	8,982
OVARIAN-PET2002	1,514	0,611	0,611	1,461	1,461	0,611	0,611	1,461	1,461	0,611	0,611	0,611	0,611	0,611	0,611	0,611	0,611
OVARY-WEL2001	6,667	5,528	5,528	8,700	8,700	9,416	9,416	8,700	8,700	5,590	5,590	5,590	5,590	8,165	8,165	5,590	5,590
CNS-POM2002	4,479	6,667	6,667	3,889	5,241	5,666	3,685	5,443	5,443	6,479	5,666	5,443	4,339	6,111	7,638	4,339	4,339
PROSTATE-OUTCOME-SIN2002	10,541	7,453	7,453	7,453	7,453	7,453	6,597	7,453	7,453	11,055	10,541	9,230	10,541	10,541	9,954	10,541	10,541
PROSTATE-SIN2002	2,755	2,380	3,111	5,284	5,592	2,647	3,247	5,592	5,592	2,688	2,483	2,396	2,483	2,386	2,386	2,386	2,386
T-ALL-YEO2002	0,312	0,312	0,312	0,312	0,312	0,312	0,312	0,312	0,312	0,312	0,312	0,312	0,312	0,312	0,312	0,312	0,312
TEL-AML1-YEO2002	1,189	0,804	0,786	1,932	1,679	1,054	1,010	1,617	1,617	1,361	0,958	0,961	1,062	1,000	1,048	0,998	0,998
PROSTATE-WEL2001	6,498	5,773	5,773	8,333	8,439	6,110	6,110	8,439	8,439	8,596	5,617	5,617	5,617	8,248	7,690	5,617	5,617

Tabela C.17: *Windowing: Média dos valores originais da variável AUC obtidos por validação cruzada de dez partições.*

Base	J48	WPEWeC	WPEC	WEWeC	WEC	WPWeC	WPC	WWeC	WC	WPEWe	WPE	WEWe	WE	WPWe	WP	WWe	W
LYMPHOMA-ALI2000	0,89972	0,86153	0,86153	0,80472	0,78139	0,86153	0,86778	0,78833	0,78139	0,87750	0,87583	0,83147	0,83147	0,84439	0,87842	0,81967	0,81967
GIST-ALL2001	0,50000	0,55000	0,55000	0,55000	0,55000	0,55000	0,55000	0,55000	0,55000	0,55000	0,55000	0,55000	0,55000	0,55000	0,55000	0,55000	0,55000
COLON-ALO1999	0,79792	0,75417	0,74167	0,67917	0,70417	0,75417	0,79167	0,70417	0,70417	0,70417	0,67292	0,69167	0,69167	0,68542	0,72292	0,70417	0,70417
LEUKEMIA-GOL1999	0,75000	0,80583	0,83500	0,93000	0,95250	0,76917	0,79833	0,95250	0,95250	0,70250	0,70250	0,75250	0,74250	0,70250	0,70250	0,74250	0,74250
BCR-ABL-YEO2002	0,57188	0,62016	0,71694	0,79294	0,79617	0,65731	0,65731	0,79617	0,79617	0,68715	0,68715	0,66215	0,66215	0,68715	0,63876	0,58876	0,58876
MELANOMA-BIT2000	0,85000	0,75417	0,73750	0,68333	0,68333	0,76250	0,76250	0,68333	0,68333	0,80000	0,80833	0,80833	0,80833	0,83333	0,80833	0,80833	0,80833
BREAST-VEE2002	0,62467	0,61583	0,61067	0,62250	0,60250	0,58667	0,61667	0,60250	0,63500	0,59617	0,63617	0,56983	0,60817	0,57167	0,58967	0,61667	0,61667
NETWORKS-BUT2000	0,60929	0,54595	0,54000	0,58738	0,56000	0,55524	0,55524	0,57190	0,56000	0,60952	0,60857	0,60619	0,61071	0,64833	0,58095	0,57381	0,57381
DLBCL-NIH-ROS2002	0,50509	0,52571	0,50571	0,56532	0,54781	0,50200	0,54825	0,59511	0,59511	0,50455	0,51848	0,46828	0,46828	0,53812	0,53812	0,51036	0,51036
DLBCL-OUTCOME-SHI2002	0,51111	0,50000	0,51667	0,52083	0,56250	0,54167	0,54167	0,61250	0,59583	0,56111	0,53333	0,53333	0,50833	0,54583	0,54583	0,53333	0,53333
DLBCL-SHI2002	0,65083	0,75667	0,75667	0,76500	0,79500	0,75667	0,75000	0,80333	0,80333	0,74000	0,75667	0,74000	0,73167	0,69667	0,69667	0,75667	0,75667
E2A-PBX1-YEO2002	1,00000	1,00000	1,00000	1,00000	1,00000	1,00000	1,00000	1,00000	1,00000	1,00000	1,00000	1,00000	1,00000	1,00000	1,00000	1,00000	1,00000
ECML2004	0,58333	0,58631	0,58631	0,57381	0,54435	0,57917	0,57917	0,54435	0,54435	0,55774	0,57560	0,57560	0,57560	0,55595	0,56845	0,57679	0,57679
GCM-RAM2001	0,78321	0,77255	0,76372	0,71704	0,71704	0,76525	0,75341	0,71704	0,71704	0,78612	0,78225	0,81216	0,80879	0,78633	0,78356	0,80881	0,80881
GSE11665	0,10000	0,10000	0,10000	0,10000	0,10000	0,10000	0,10000	0,10000	0,10000	0,10000	0,10000	0,10000	0,10000	0,10000	0,10000	0,10000	0,10000
GSE3255	0,66667	0,70833	0,70833	0,63333	0,68333	0,70833	0,78333	0,68333	0,68333	0,70833	0,66667	0,66667	0,66667	0,70833	0,70833	0,66667	0,66667
GSE360	0,56667	0,56667	0,55000	0,50833	0,59167	0,54167	0,56667	0,56667	0,56667	0,59167	0,60000	0,60000	0,60833	0,60000	0,57500	0,60000	0,60000
GSE443	0,10000	0,05000	0,05000	0,05000	0,05000	0,05000	0,05000	0,05000	0,05000	0,05000	0,05000	0,05000	0,05000	0,05000	0,05000	0,05000	0,05000
GSE474	0,47500	0,47500	0,52500	0,60000	0,62500	0,50000	0,52500	0,62500	0,62500	0,45000	0,45000	0,42500	0,45000	0,42500	0,40000	0,45000	0,45000
GSE5473	0,47500	0,70000	0,70000	0,60833	0,57500	0,70000	0,77500	0,57500	0,57500	0,45000	0,45000	0,45000	0,45000	0,45000	0,45000	0,45000	0,45000
GSE7433	0,25000	0,32500	0,32500	0,22500	0,17500	0,35000	0,40000	0,17500	0,17500	0,27500	0,27500	0,27500	0,27500	0,22500	0,22500	0,27500	0,27500
GSE7898	0,10000	0,00000	0,00000	0,00000	0,00000	0,10000	0,10000	0,00000	0,00000	0,00000	0,00000	0,00000	0,00000	0,10000	0,10000	0,00000	0,00000
BREAST-HED2003	0,40000	0,35000	0,35000	0,30000	0,30000	0,35000	0,40000	0,30000	0,30000	0,40000	0,40000	0,40000	0,40000	0,40000	0,40000	0,40000	0,40000
HYPERDIP-50-YEO2002	0,83297	0,79620	0,80694	0,78626	0,78324	0,80451	0,81953	0,75962	0,75962	0,76097	0,74338	0,75122	0,76454	0,83885	0,83335	0,77736	0,77736
GASTRIC-LEU2002	0,92428	0,82555	0,86363	0,85182	0,85182	0,89979	0,89979	0,85598	0,85598	0,85416	0,85454	0,85308	0,87871	0,89066	0,88073	0,88755	0,88755
LUNG-BHA2001	0,93791	0,92671	0,92814	0,89665	0,89665	0,92671	0,91782	0,89665	0,89665	0,93288	0,93145	0,90839	0,93145	0,91496	0,92329	0,93145	0,93145
LUNG-BEE2002	0,99444	0,92708	0,92708	0,85347	0,85347	0,93333	0,93333	0,85347	0,85347	0,98819	0,98819	0,98819	0,98819	0,99444	0,99444	0,98819	0,98819
LUNG-GOR2002	0,92083	0,88750	0,88417	0,90417	0,90417	0,90417	0,90417	0,91083	0,91083	0,94750	0,94417	0,90083	0,92750	0,94417	0,94417	0,92750	0,92750
LUNG-WIG2002	0,80000	0,86667	0,81667	0,63333	0,60833	0,86667	0,78333	0,63333	0,63333	0,80000	0,80000	0,86667	0,86667	0,81667	0,81667	0,77500	0,77500
BREAST-MA2003	0,48722	0,51431	0,47889	0,57917	0,59333	0,49556	0,48038	0,58417	0,58417	0,58903	0,57111	0,48833	0,48833	0,59101	0,60708	0,47153	0,47153
MLL-ARM2002	0,89131	0,80119	0,80119	0,76750	0,76750	0,80119	0,80369	0,76750	0,76750	0,85595	0,86167	0,86167	0,86167	0,85595	0,89952	0,86167	0,86167
MLL-YEO2002	0,79349	0,76766	0,74102	0,82866	0,85777	0,83683	0,79183	0,88610	0,88610	0,88777	0,88371	0,86355	0,88371	0,91183	0,91183	0,88371	0,88371
SOFT-NIE2002	0,65583	0,67000	0,67000	0,64250	0,63750	0,66583	0,69083	0,63583	0,63750	0,67083	0,66667	0,66167	0,66667	0,60833	0,62750	0,64667	0,64667
OVARIAN-PET2002	0,95104	0,98819	0,98819	0,96389	0,96389	0,98819	0,98819	0,96389	0,96389	0,99063	0,99063	0,99063	0,99063	0,99063	0,99063	0,99063	0,99063
OVARY-WEL2001	0,30000	0,33333	0,33333	0,36667	0,36667	0,28333	0,28333	0,36667	0,36667	0,33333	0,33333	0,33333	0,33333	0,26667	0,26667	0,33333	0,33333
CNS-POM2002	0,50208	0,50833	0,52083	0,49583	0,50833	0,52500	0,56875	0,50833	0,50833	0,40833	0,37083	0,42083	0,39583	0,40208	0,50833	0,39583	0,39583
PROSTATE-OUTCOME-SIN2002	0,22500	0,27500	0,27500	0,55000	0,55000	0,27500	0,32500	0,55000	0,55000	0,27500	0,22500	0,32500	0,22500	0,22500	0,27500	0,22500	0,22500
PROSTATE-SIN2002	0,78592	0,79048	0,80298	0,73851	0,72185	0,80357	0,81711	0,72423	0,72423	0,80551	0,77307	0,76682	0,76932	0,78884	0,78884	0,78884	0,78884
T-ALL-YEO2002	0,99821	0,99821	0,99821	0,99821	0,99821	0,99821	0,99821	0,99821	0,99821	0,99821	0,99821	0,99821	0,99821	0,99821	0,99821	0,99821	0,99821
TEL-AML1-YEO2002	0,92250	0,93002	0,93752	0,91740	0,93315	0,95092	0,95442	0,93940	0,93940	0,93769	0,94594	0,94713	0,95459	0,94125	0,93891	0,95659	0,95659
PROSTATE-WEL2001	0,74750	0,75083	0,75083	0,67417	0,70083	0,73833	0,73833	0,70083	0,70083	0,77667	0,77917	0,77917	0,77917	0,76833	0,77500	0,77917	0,77917

Tabela C.18: *Windowing: Desvio padrão dos valores originais da variável AUC obtidos por validação cruzada de dez partições.*

Base	J48	WPEWeC	WPEC	WEWeC	WEC	WPWeC	WPC	WWeC	WC	WPEWe	WPE	WEWe	WE	WPWe	WP	WWe	W
LYMPHOMA-ALI2000	0,018	0,025	0,025	0,028	0,026	0,025	0,023	0,026	0,026	0,031	0,034	0,038	0,038	0,034	0,034	0,034	0,034
GIST-ALL2001	0,149	0,157	0,157	0,157	0,157	0,157	0,157	0,157	0,157	0,157	0,157	0,157	0,157	0,157	0,157	0,157	0,157
COLON-ALO1999	0,065	0,058	0,075	0,042	0,053	0,064	0,064	0,053	0,053	0,071	0,059	0,062	0,062	0,069	0,064	0,067	0,067
LEUKEMIA-GOL1999	0,068	0,032	0,040	0,031	0,027	0,033	0,040	0,027	0,027	0,075	0,075	0,076	0,073	0,075	0,075	0,073	0,073
BCR-ABL-YEO2002	0,087	0,093	0,068	0,062	0,061	0,075	0,075	0,061	0,061	0,094	0,094	0,062	0,062	0,071	0,065	0,110	0,110
MELANOMA-BIT2000	0,041	0,038	0,028	0,074	0,074	0,034	0,034	0,074	0,074	0,047	0,046	0,046	0,046	0,037	0,066	0,046	0,046
BREAST-VEE2002	0,059	0,047	0,040	0,054	0,051	0,046	0,042	0,051	0,051	0,042	0,043	0,038	0,045	0,054	0,058	0,049	0,049
NETWORKS-BUT2000	0,024	0,013	0,012	0,022	0,027	0,028	0,028	0,028	0,027	0,023	0,024	0,029	0,026	0,026	0,021	0,019	0,019
DLBCL-NIH-ROS2002	0,042	0,038	0,045	0,042	0,033	0,047	0,039	0,046	0,046	0,036	0,034	0,052	0,052	0,030	0,030	0,041	0,041
DLBCL-OUTCOME-SHI2002	0,039	0,033	0,037	0,068	0,057	0,038	0,038	0,056	0,057	0,054	0,031	0,040	0,053	0,032	0,032	0,056	0,056
DLBCL-SHI2002	0,073	0,050	0,043	0,075	0,069	0,043	0,045	0,070	0,070	0,046	0,050	0,047	0,048	0,065	0,065	0,050	0,050
E2A-PBX1-YEO2002	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000
ECML2004	0,028	0,021	0,021	0,022	0,017	0,023	0,023	0,017	0,017	0,017	0,021	0,021	0,021	0,018	0,029	0,022	0,022
GCM-RAM2001	0,018	0,017	0,018	0,024	0,024	0,015	0,017	0,024	0,024	0,020	0,021	0,013	0,011	0,018	0,018	0,015	0,015
GSE11665	0,067	0,067	0,067	0,067	0,067	0,067	0,067	0,067	0,067	0,067	0,067	0,067	0,067	0,067	0,067	0,067	0,067
GSE3255	0,073	0,091	0,091	0,084	0,082	0,091	0,059	0,082	0,082	0,091	0,090	0,090	0,090	0,080	0,080	0,090	0,090
GSE360	0,032	0,037	0,048	0,026	0,032	0,052	0,045	0,035	0,035	0,040	0,039	0,037	0,040	0,043	0,029	0,043	0,043
GSE443	0,100	0,050	0,050	0,050	0,050	0,050	0,050	0,050	0,050	0,050	0,050	0,050	0,050	0,050	0,050	0,050	0,050
GSE474	0,102	0,045	0,069	0,100	0,067	0,065	0,087	0,067	0,067	0,073	0,073	0,075	0,073	0,084	0,077	0,073	0,073
GSE5473	0,058	0,082	0,082	0,102	0,099	0,082	0,069	0,099	0,099	0,050	0,050	0,050	0,050	0,050	0,050	0,050	0,050
GSE7433	0,083	0,118	0,118	0,087	0,065	0,125	0,119	0,065	0,065	0,102	0,102	0,102	0,102	0,095	0,095	0,102	0,102
GSE7898	0,100	0,000	0,000	0,000	0,000	0,100	0,100	0,000	0,000	0,000	0,000	0,000	0,000	0,100	0,100	0,000	0,000
BREAST-HED2003	0,125	0,107	0,107	0,082	0,082	0,107	0,125	0,082	0,082	0,125	0,125	0,125	0,125	0,125	0,125	0,125	0,125
HYPERDIP-50-YEO2002	0,026	0,029	0,027	0,026	0,024	0,030	0,026	0,025	0,025	0,044	0,044	0,039	0,044	0,028	0,029	0,050	0,050
GASTRIC-LEU2002	0,035	0,042	0,034	0,022	0,022	0,027	0,027	0,021	0,021	0,037	0,036	0,032	0,031	0,036	0,034	0,034	0,034
LUNG-BHA2001	0,020	0,022	0,022	0,020	0,020	0,022	0,026	0,020	0,020	0,023	0,022	0,029	0,022	0,027	0,023	0,022	0,022
LUNG-BEE2002	0,006	0,054	0,054	0,065	0,065	0,055	0,055	0,065	0,065	0,008	0,008	0,008	0,008	0,006	0,006	0,008	0,008
LUNG-GOR2002	0,026	0,038	0,037	0,038	0,038	0,037	0,037	0,036	0,036	0,025	0,025	0,035	0,038	0,027	0,027	0,036	0,036
LUNG-WIG2002	0,073	0,053	0,062	0,064	0,074	0,053	0,074	0,074	0,074	0,073	0,073	0,053	0,053	0,075	0,075	0,080	0,080
BREAST-MA2003	0,049	0,053	0,048	0,047	0,061	0,035	0,053	0,061	0,061	0,070	0,048	0,056	0,056	0,063	0,061	0,055	0,055
MLL-ARM2002	0,028	0,045	0,045	0,043	0,044	0,045	0,046	0,044	0,044	0,035	0,035	0,035	0,035	0,035	0,027	0,035	0,035
MLL-YEO2002	0,064	0,058	0,065	0,058	0,044	0,066	0,073	0,043	0,043	0,040	0,056	0,058	0,056	0,053	0,053	0,056	0,056
SOFT-NIE2002	0,048	0,024	0,024	0,052	0,051	0,023	0,022	0,052	0,051	0,046	0,059	0,060	0,059	0,055	0,052	0,065	0,065
OVARIAN-PET2002	0,017	0,006	0,006	0,013	0,013	0,006	0,006	0,013	0,013	0,005	0,005	0,005	0,005	0,005	0,005	0,005	0,005
OVARY-WEL2001	0,113	0,129	0,129	0,138	0,138	0,114	0,114	0,138	0,138	0,129	0,129	0,129	0,129	0,103	0,103	0,129	0,129
CNS-POM2002	0,039	0,061	0,065	0,043	0,049	0,058	0,042	0,049	0,049	0,059	0,053	0,054	0,045	0,057	0,085	0,045	0,045
PROSTATE-OUTCOME-SIN2002	0,079	0,079	0,079	0,117	0,117	0,079	0,075	0,117	0,117	0,095	0,079	0,092	0,079	0,079	0,079	0,079	0,079
PROSTATE-SIN2002	0,036	0,026	0,034	0,060	0,063	0,030	0,037	0,063	0,063	0,031	0,027	0,025	0,026	0,027	0,027	0,027	0,027
T-ALL-YEO2002	0,002	0,002	0,002	0,002	0,002	0,002	0,002	0,002	0,002	0,002	0,002	0,002	0,002	0,002	0,002	0,002	0,002
TEL-AML1-YEO2002	0,018	0,017	0,012	0,030	0,026	0,017	0,017	0,026	0,026	0,023	0,015	0,018	0,017	0,015	0,015	0,016	0,016
PROSTATE-WEL2001	0,040	0,037	0,037	0,054	0,058	0,038	0,038	0,058	0,058	0,055	0,040	0,040	0,040	0,052	0,052	0,040	0,040

Apêndice

D

Estratégia de Avaliação Experimental de Algoritmos

D.1 Considerações Iniciais

A obtenção de um classificador permite que a tarefa de classificação em questão seja realizada, mas ainda se faz necessário medir o quão bem o modelo a realiza. Para isto, existem inúmeras medidas, sendo a taxa de acertos (ou de erros) provavelmente a mais popular. Porém, existem diferentes aspectos nessa análise e cada medida permite quantificar um ou alguns desses aspectos. Por exemplo, o desbalanceamento de classes faz com que a taxa de acertos pura não seja a medida mais adequada, já que é assumido que ela considera os diferentes rótulos de classe como igualmente distribuídos. Se houver pesos diferentes para os erros em cada um dos rótulos, a taxa de acertos pode levar a uma conclusão equivocada. Como é apresentado a seguir, a área sob a curva ROC corrige aquela deficiência. Existe a necessidade também de se medir a eficácia de um algoritmo. Para isto, faz-se necessário medir o desempenho dos modelos criados por ele em diversos conjuntos de exemplos, ou pelo menos em diferentes conjuntos do mesmo problema. Finalmente, devem existir formas de comparar o desempenho de diferentes algoritmos, a fim de verificar possíveis diferenças.

Neste capítulo são apresentados alguns dos conceitos e procedimentos utilizados neste projeto para avaliar e comparar o desempenho de classificadores. Detalhes adicionais sobre comparação de algoritmos podem ser encontrados em (Salzberg 1995b; Dietterich 1997; Demšar 2006; García & Herrera 2008).

Para se medir o desempenho de um classificador em um dado problema, será descrito o conceito de curva ROC e sua área. Adicionalmente, há situações em que se deve comparar o desempenho de dois ou mais algoritmos neste mesmo problema. Supondo que a medida de desempenho utilizada seja a acurácia, pode-se realizar validação cruzada (Kohavi 1995), possivelmente várias vezes, calcular a média dos valores de acurácia encontrados para cada indutor e escolher, como melhor, aquele algoritmo que obteve a maior acurácia média. Para verificar se a diferença encontrada foi significativa, será descrita uma forma de comparar o desempenho médio de diversos algoritmos em diversos problemas.

D.2 Curvas ROC e AUC

Nesta seção serão tratadas as curvas ROC e sua aplicação em avaliação de classificadores. Apesar de terem sido concebidas na área de teoria de detecção de sinais, as curvas ROC têm encontrado aplicações em diversas outras áreas e vêm sendo cada vez mais utilizadas em AM.

Considere um problema de classificação em que os exemplos representam pacientes e há duas classes possíveis: S, indicando um paciente saudável e D, indicando um paciente doente. Se um classificador for construído a partir de uma base de dados de treinamento e um exemplo for apresentado

a esse classificador, existem basicamente quatro possíveis situações, que podem ser visualizadas na Tabela D.1. Em análise ROC, existem duas possibilidades: *positivo* e *negativo*. O que cada um deles significa depende do problema tratado. No caso do exemplo acima, a classe S será considerada positiva (sendo, então, representada por P) e a classe D, negativa (sendo representada por N). As quatro situações de um modelo ao classificar um exemplo são:

(i) a classe predita foi a P e a classe verdadeira era realmente a P, configurando um *verdadeiro positivo (VP)*;

(ii) a classe predita foi a P, mas a classe verdadeira era a N, configurando um *falso positivo (FP)*;

(iii) a classe predita foi a N, mas a classe verdadeira era a P, configurando um *falso negativo (FN)*;

(iv) a classe predita foi a N e a classe verdadeira era realmente a N, configurando um *verdadeiro negativo (VN)*.

Tabela D.1: *Matriz de confusão indicando as quatro possíveis situações em que um classificador pode se encontrar ao predizer a classe de um determinado exemplo. As colunas mostram as classes preditas (P_p e N_p); as linhas mostram as classes verdadeiras (P_v e N_v). Nas células da matriz, V significa verdadeiro, F significa falso, P significa positivo e N significa negativo.*

	P_p	N_p
P_v	VP	FN
N_v	FP	VN

Na matriz de confusão da Tabela D.1, cada célula mostra a frequência de cada uma das quatro situações, considerando que o classificador tenha realizado a predição de vários exemplos. Portanto, na matriz, VP representa a frequência de exemplos que apresentam a classe positiva e foram classificados como positivos; FP representa a frequência de exemplos que apresentam a classe negativa, mas que foram classificados como positivos; e assim por diante. Portanto, VP, FP, FN e VN são frequências. Entre outras medidas que podem ser calculadas com estas frequências, quatro podem ser destacadas, cujas fórmulas se encontram em D.1: a taxa de verdadeiros positivos (TVP), que é a taxa de acertos considerando somente os exemplos cuja classe é verdadeiramente positiva (primeira linha da tabela); a taxa de falsos negativos (TFN), que é a taxa de erro considerando somente os exemplos cuja classe é verdadeiramente positiva, sendo, portanto, complementar a TVP; a taxa de verdadeiros negativos (TVN), que é a taxa de acertos considerando somente os exemplos cuja classe é verdadeiramente negativa (segunda linha da tabela); e a taxa de falsos positivos (TFP), que é a taxa de erro considerando somente os exemplos cuja classe é verdadeiramente negativa, sendo, portanto, complementar a TVN. A TVP é conhecida como *sensibilidade* e a TVN, como *especificidade*.

$$TVP = \frac{VP}{VP+FN} \qquad TFN = \frac{FN}{VP+FN} = 1 - TVP$$
$$TVN = \frac{VN}{VN+FP} \qquad TFP = \frac{FP}{VN+FP} = 1 - TVN \tag{D.1}$$

Gráficos ROC são aqueles em que a ordenada representa a TVP e a abscissa, a TFP, ficando ambas no intervalo $[0, 1]$, cujo exemplo é mostrado na Figura D.1. Cada ponto neste gráfico representa o par (TVP, TFP) de um dado classificador. Um modelo que sempre classifica os exemplos como pertencendo à classe negativa, independentemente dos seus atributos, terá o seu ponto em $C = (0, 0)$. Já aquele que sempre classifica os exemplos como pertencendo à classe positiva terá o

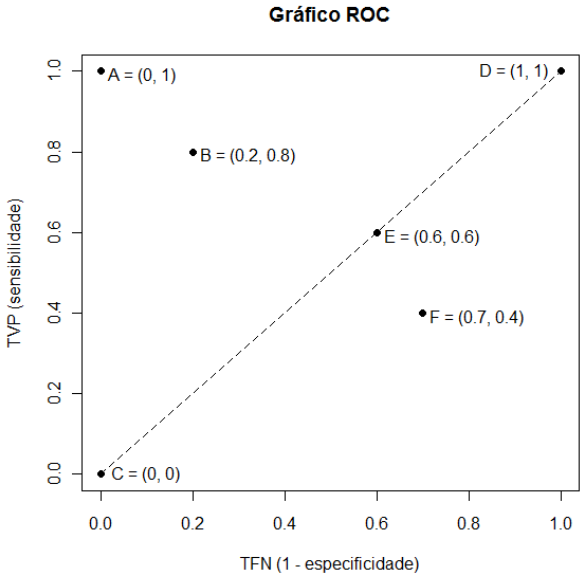


Figura D.1: *Gráfico ROC mostrando pontos calculados para diferentes classificadores: o ponto A representa o classificador ideal; o ponto B representa um classificador melhor que o aleatório, mas não perfeito; o ponto C pertence a um classificador que sempre prediz a classe negativa, enquanto o ponto D pertence a um classificador que sempre prediz a classe positiva; o ponto E representa um classificador aleatório que "chuta" a classe positiva com $p = 0,6$; já o ponto F representa um classificador pior que o aleatório.*

seu ponto em $D = (1, 1)$. O ponto $A = (0, 1)$ representa o classificador ideal, ou seja, aquele que sempre acerta quando o exemplo é positivo e nunca erra quando o exemplo é negativo. Desta forma, quanto mais alto e mais à esquerda o ponto estiver, melhor é considerado o classificador. A reta que liga os pontos $C = (0, 0)$ e $D = (1, 1)$ representa a classificação aleatória, ou seja, classificadores cujos pontos caírem nela predizem a classe de um determinado exemplo segundo uma probabilidade p de ser positiva e $q = 1 - p$ de ser negativa. O valor de p dirá em que lugar da reta o classificador cairá. Por exemplo, se $p = 0,5$, o ponto do modelo cairá em $(0,5; 0,5)$. Qualquer classificador que caia abaixo dessa reta é considerado pior que um classificador aleatório.

Dados um classificador e um conjunto de exemplos de teste, a matriz de confusão pode ser construída, o que permite desenhar um único ponto no gráfico ROC. Muitos modelos, ao classificarem um exemplo, fornecem, além da classe predita, um *score*, que indica quanto o exemplo classificado pertence à classe predita. Assim, valores maiores indicam que o exemplo está mais próximo da classe positiva e valores menores indicam que ele está mais próximo da classe negativa. É claro que, para tomar a decisão final, o classificador tem que escolher um ponto de corte nesse *score*, ou seja, toda vez que um exemplo for classificado nesse ponto ou acima, o classificador dirá que o exemplo é da classe positiva; caso contrário, da classe negativa. Normalmente esse ponto de corte é exatamente no meio do intervalo em que se encontra o *score*. Por exemplo, se esse intervalo é $[0, 1]$, o corte será em $0,5$ (o que seria esperado em distribuições balanceadas). De qualquer maneira, o ponto de corte pode estar em qualquer lugar do intervalo $(-\infty, \infty)$. Mesmo que o classificador não produza um *score*, existem técnicas que tornam isto possível, *e.g.*, baseadas em *ensembles* (Fawcett 2006).

Se os exemplos da base de dados utilizada na análise forem colocados em ordem decrescente do *score* que o classificador produziu para cada um deles e se o ponto de corte for sendo variado de ∞ a $-\infty$, cada ponto escolhido gerará uma matriz de confusão e um ponto no gráfico ROC. Veja que não é necessário tratar o corte como contínuo. Os candidatos a corte podem ser os próprios valores de *score*, já que estão ordenados e, portanto, não há na base de dados um *score* entre dois valores consecutivos (o máximo que pode ocorrer é que eles sejam iguais). Veja também que o classificador continua sempre o mesmo, o único aspecto sofrendo variação é o ponto de corte escolhido. O primeiro corte escolhido é ∞, o que fará com que todos os exemplos sejam classificados como negativos, gerando o ponto $(0, 0)$. O próximo corte será o maior *score* observado: se o exemplo

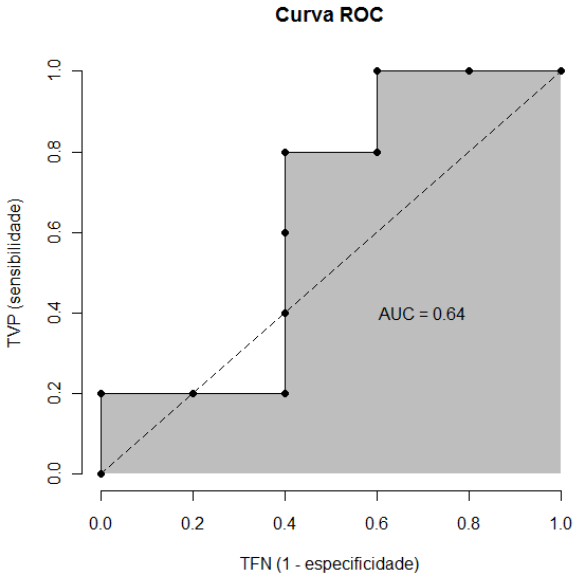


Figura D.2: *Curva ROC de um classificador considerando uma base de teste de dez exemplos. Cada ponto indicado representa um ponto de corte diferente utilizado. A linha cheia é a função escada criada a partir da ligação de tais pontos. A área cinza abaixo da função indica a medida AUC, sendo seu valor mostrado no gráfico.*

que gerou esse *score* for da classe positiva, a TVP irá aumentar e, se o exemplo for da classe negativa, a TFN irá aumentar, ou seja, os pontos gerados no gráfico ROC são monotonicamente crescentes. O último ponto a considerar é o menor *score* observado, o que fará com que todos os exemplos sejam classificados como positivos, gerando o ponto $(1, 1)$.

Se todos os pontos consecutivos do gráfico ROC calculados anteriormente forem ligados por segmentos de reta, tem-se a conhecida curva ROC (veja Figura D.2). Se fosse possível considerar todos os infinitos pontos de corte, seria obtida uma curva real, mas, como isto em geral não é possível, obtém-se, na verdade, uma função escada. A curva ROC permite medir quão bem um classificador consegue, na média, distinguir a classe positiva relativamente à negativa. O classificador ideal é aquele em que a curva ROC caminha acima na reta entre os pontos $(0, 0)$ e $(0, 1)$. Atingindo o ponto $(0, 1)$, ela caminha para a direita na reta entre os pontos $(0, 1)$ e $(1, 1)$. Pode-se notar que, mesmo para o classificador ideal, dependendo do corte escolhido, ele pode ter acurácia menor que 1. Isto ocorre porque esse *score* calculado não é calibrado, já que geralmente ele não representa uma probabilidade verdadeira (Fawcett 2006). O melhor corte será aquele que produziu o ponto $(0, 1)$ no gráfico ROC.

Uma característica importante das curvas ROC é que elas não se alteram com a modificação na distribuição de classes. Isto ocorre porque a TVP se baseia somente na primeira linha da matriz de confusão e a TFN, apenas na segunda linha, ou seja, simplesmente mudar a proporção de exemplos positivos e negativos (relação entre as duas linhas), não alterará as taxas. A relação entre os exemplos classificados corretamente e incorretamente permanecerá a mesma em cada linha, visto que, apesar da alteração na distribuição de classes, o conceito sendo aprendido é o mesmo e o classificador construído permanece fundamentalmente o mesmo. O mesmo se aplica a alterações nos custos de erros (Fawcett 2006).

Para expressar a curva ROC com apenas um número, é comum que se utilize sua área. É a conhecida medida AUC, que se encontra no intervalo $[0, 1]$, pois a curva se encontra dentro de um quadrado de lado um. Classificadores aleatórios terão $AUC = 0,5$, sendo que um modelo considerável deve ter $AUC > 0,5$. Assim como na análise feita anteriormente, o fato de um classificador ter AUC maior que o de outro não significa que ele terá um desempenho melhor, já que isto dependerá do ponto de corte usado por cada um. Além disso, sem um valor de variância para AUC, ou seja,

sem utilizar várias amostras de dados, produzindo várias curvas e valores de AUC, fica difícil concluir se a diferença entre os valores observados é significativa ou se ocorreu ao acaso. Isto pode ser resolvido, por exemplo, com a aplicação de validação cruzada e testes de hipóteses (Bradley 1997).

Quando o problema contém mais de duas classes, a complexidade aumenta consideravelmente e não é possível mais visualizar a curva ROC. Existem algumas estratégias para contornar o problema, sendo a mais utilizada aquela que constrói c gráficos ROC (c é o número de classes). Para cada rótulo de classe, é construído um gráfico, sendo considerada como classe positiva o rótulo em questão e, como classe negativa, qualquer outro rótulo, dividindo o problema multiclasse inicial em c problemas binários. Mas isto vem com um preço: a curva ROC perde a característica de ser insensível à distribuição de classes.

Usando a solução anterior, serão produzidos c valores de AUC para a análise de um classificador. Uma das formas mais usadas para combinar esses valores é realizar sua soma, ponderando cada valor pela prevalência da classe a que ele se refere. Isto pode ser visualizado em D.2, em que AUC_i e p_i representam o valor de AUC e a proporção da classe i, respectivamente. Esta formulação é rápida de ser calculada, mas não é insensível a mudanças na distribuição de classes nem nos custos de erros de classificação. Uma outra formulação, apresentada por Hand & Till (2001), apresenta essa insensibilidade, mas sua complexidade computacional aumenta, apesar de valer a pena, principalmente em casos em que tal característica é importante.

$$AUC = \sum_{i=1}^{c} AUC_i * p_i \tag{D.2}$$

Uma vez obtida a medida AUC ou outra medida qualquer, é possível comparar vários algoritmos entre si por meio de um teste estatístico não paramétrico, o que será descrito na próxima seção.

D.3 Teste de Friedman

Ao se compararem, por exemplo, cinco diferentes indutores, A, B, C, D e E, a medida de desempenho utilizada pode ser a acurácia, a taxa de erro, o tamanho do classificador final, a taxa de falsos negativos, ou qualquer outra. Independentemente da medida utilizada, o que se pode fazer é escolher uma série de diferentes conjuntos de dados, nos quais os algoritmos a serem comparados serão aplicados (o critério para esta escolha depende do estudo sendo realizado). Para cada algoritmo aplicado a cada conjunto, aplica-se uma das formas de estimação da medida de desempenho. Pode-se usar, por exemplo, realizar várias repetições de validação cruzada e tomar a média dessas repetições como a estimativa da medida. Tem-se, basicamente, uma tabela em que as linhas representam os conjuntos de dados, as colunas representam os indutores testados e cada cruzamento linha-coluna traz a estimativa da medida obtida pela aplicação de um dado algoritmo (coluna) em um dado conjunto (linha). A Tabela D.2 é um exemplo (por enquanto, os valores entre parênteses e a última linha da tabela devem ser desconsiderados). Nela são mostrados os valores de acurácia dos cinco indutores citados acima em vinte bases de dados diferentes (numeradas de 1 a 20). Considere que o valor de cada célula tenha sido obtido pela média de dez repetições de validação cruzada com dez partições do algoritmo da respectiva coluna na base de dados da respectiva linha.

Se a média da medida em cada coluna for calculada, tem-se o desempenho médio de cada indutor. Se os experimentos acima tiverem sido aplicados em um número suficientemente grande de conjuntos de dados, é comum que se considere a distribuição dessas médias como sendo Gaussiana (normal), independentemente da distribuição verdadeira das medidas propriamente ditas. Neste caso, pode-se usar o teste *t de Student*, para a comparação de dois algoritmos, e a ANOVA, para a comparação de mais de dois algoritmos. Porém, na grande maioria dos casos, o número de bases de dados utilizadas na comparação não é suficiente para justificar uma boa aproximação pela normal. Uma alternativa é a utilização de métodos não paramétricos, que vêm sendo muito utilizados em AM, ou seja, métodos que não assumem nenhuma distribuição *a priori* dos dados. Esses métodos deixam de lado as medidas propriamente ditas e se baseiam na ordenação (*ranking*) dos algoritmos.

Para a comparação de mais de dois algoritmos, tem sido usado o teste de Friedman (1940), uma espécie de versão não-paramétrica da ANOVA. Sua hipótese nula diz que a diferença encontrada

Tabela D.2: *Valores de acurácia para a aplicação dos indutores A, B, C, D e E nas bases de dados de 1 a 20. Cada valor entre parênteses indica o rank do indutor especificado na coluna com relação à base de dados especificada na linha. A última linha da tabela traz o rank médio para cada indutor.*

Bases	A	B	C	D	E
1	78,33 (4,00)	89,17 (3,00)	65,28 (5,00)	92,50 (2,00)	97,50 (1,00)
2	95,67 (5,00)	98,72 (1,00)	97,30 (3,00)	97,30 (3,00)	97,30 (3,00)
3	95,30 (2,00)	99,30 (1,00)	92,29 (4,00)	90,85 (5,00)	93,50 (3,00)
4	95,50 (2,00)	99,50 (1,00)	92,58 (4,00)	91,98 (5,00)	93,63 (3,00)
5	61,11 (3,00)	64,29 (2,00)	54,20 (4,00)	54,12 (5,00)	69,07 (1,00)
6	71,82 (1,00)	69,58 (3,00)	70,28 (2,00)	66,01 (5,00)	68,95 (4,00)
7	97,50 (1,50)	88,89 (3,00)	65,28 (5,00)	85,00 (4,00)	97,50 (1,50)
8	83,50 (1,00)	76,70 (4,00)	54,45 (5,00)	77,96 (3,00)	82,05 (2,00)
9	49,20 (1,50)	49,20 (1,50)	42,70 (5,00)	43,22 (4,00)	47,79 (3,00)
10	62,58 (5,00)	74,84 (3,00)	64,52 (4,00)	76,13 (2,00)	84,52 (1,00)
11	48,00 (3,00)	46,00 (4,00)	40,00 (5,00)	74,00 (1,00)	66,00 (2,00)
12	60,67 (3,00)	52,67 (5,00)	65,00 (2,00)	58,67 (4,00)	67,33 (1,00)
13	96,90 (2,00)	92,90 (4,00)	30,60 (5,00)	93,50 (3,00)	96,94 (1,00)
14	87,36 (2,00)	73,96 (4,00)	49,06 (5,00)	79,25 (3,00)	88,30 (1,00)
15	83,57 (1,00)	80,48 (4,00)	42,56 (5,00)	81,01 (2,00)	80,89 (3,00)
16	53,96 (4,00)	61,88 (3,00)	47,92 (5,00)	75,63 (2,00)	93,13 (1,00)
17	58,95 (2,00)	44,32 (4,00)	15,79 (5,00)	53,79 (3,00)	64,84 (1,00)
18	74,05 (2,00)	74,05 (2,00)	74,05 (2,00)	65,69 (5,00)	71,76 (4,00)
19	83,56 (1,00)	78,67 (3,00)	55,56 (5,00)	76,96 (4,00)	82,96 (2,00)
20	84,26 (1,00)	80,13 (3,00)	79,35 (4,00)	78,97 (5,00)	82,45 (2,00)
Rank médio	2,35	2,93	4,20	3,50	2,03

nos dados é devida ao acaso. No teste de *Friedman*, para cada conjunto de dados utilizado, os algoritmos são ordenados pela medida obtida por ele. Empates são resolvidos por meio do cálculo de médias. Por exemplo, para a base 2, os algoritmos C, D e E empataram em segundo lugar. Como são três algoritmos, ocupariam as posições 2, 3 e 4, cuja média é 3, que é, então, o *rank* de cada um deles. No entanto, há outras formas de se lidar com empates. Após a ordenação em todas as bases, calcula-se, para cada coluna, a média dos *ranks*, fornecendo, assim, as médias dos *ranks* para cada classificador, a partir das quais verifica-se se as diferenças obtidas são significativas. Na Tabela D.2, são mostrados os *ranks* de cada algoritmo em cada base de dados entre parênteses, mostrando, na última linha, o *rank* médio de cada algoritmo.

Primeiramente é necessário calcular o valor da função definida em D.3 (Demšar 2006), onde B é o número de bases de dados utilizadas no estudo; K é o número de indutores considerados; e R_j é o *rank* médio do indutor j.

$$F_F = \frac{(B-1)\,\chi_F^2}{B\,(K-1)-\chi_F^2}, \text{ onde } \chi_F^2 = \frac{12B}{K\,(K+1)}\left[\sum_j R_j^2 - \frac{K\,(K+1)^2}{4}\right] \tag{D.3}$$

F_F segue uma distribuição F com $(K-1)$ e $(K-1)(B-1)$ graus de liberdade (ver Figura D.3 para um exemplo de distribuição F). Com o valor calculado de F_F e os graus de liberdade envolvidos, pode-se obter o p-valor a partir da função distribuição da distribuição F. No exemplo, os graus de liberdade são $K-1=4$ e $(K-1)(B-1)=76$ e o cálculo do valor da função foi $8,41$ (note que, apesar de serem mostradas apenas duas casas decimais, consideram-se todas as casas obtidas nos cálculos).

O p-valor encontrado foi de $1,13 \times 10^{-5}$, que permite rejeitar a hipótese nula do teste para a maioria dos valores de nível de significância comumente utilizados. No entanto, o teste não diz entre quais pares de indutores as diferenças são significativas. Se o teste não tivesse rejeitado a

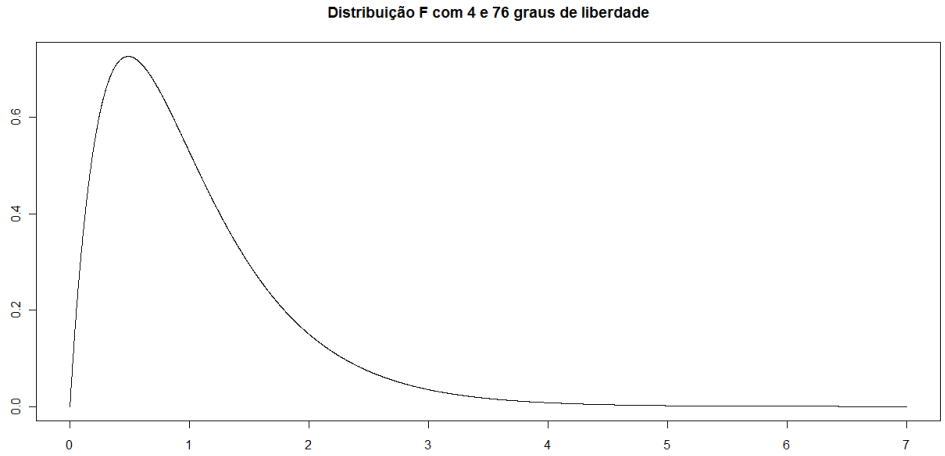


Figura D.3: *Exemplo de distribuição F com 4 e 76 graus de liberdade.*

hipótese nula, ele pararia por aí. Mas, como rejeitou, testes *post-hoc* são necessários, caso se queira identificar em quais pares de indutores se encontram as diferenças significativas apontadas, visto que nem todos os pares são necessariamente significativamente diferentes.

D.4 Testes *Post-hoc*

Neste ponto, existem basicamente duas situações, dependendo do objetivo do estudo: (i) comparação de todos os indutores com todos os outros; ou (ii) definição de um deles como controle e comparação de todos os outros com ele. Independentemente da situação, para cada par (i, j) de indutores sendo comparados, calcula-se o valor da função dada em D.5 (García & Herrera 2008), que segue uma distribuição normal padrão. O cálculo do valor da função para todos os pares, juntamente com os respectivos p-valores calculados a partir da função distribuição da distribuição normal, é mostrado na Tabela D.3. A comparação aqui é par a par, mas deve levar em consideração o fato de a comparação, na verdade, ser múltipla. Assim, um ajuste dos p-valores é necessário.

Na comparação de todos contra todos, um teste muito utilizado é o de Nemenyi (1963), que ajusta cada p-valor simplesmente multiplicando-o pelo número total de comparações, que é:

$$\frac{K(K-1)}{2} \tag{D.4}$$

No caso do exemplo, são dez comparações. O p-valor ajustado por Nemenyi é mostrado na quarta coluna da Tabela D.3. Se for usado um nível de significância de 5%, as diferenças significativas encontradas são os algoritmos A e C, C e E e D e E. O sentido da diferença pode ser definido pelo *rank* médio. Por exemplo, C e E são significativamente diferentes. Como E tem um *rank* médio menor do que C, aquele pode ser considerado significativamente melhor que este.

Na comparação contra um controle, um dos testes utilizados é o de Bonferroni (Dunn 1961), que também ajusta cada p-valor simplesmente multiplicando-o pelo número total de comparações, que, neste caso, é $(K-1)$. No caso do exemplo, são quatro comparações. Considerando como controle o algoritmo A, os p-valores ajustados por Bonferroni são mostrados na quinta coluna da Tabela D.3, mas apenas nas linhas envolvendo comparações com o indutor A. Novamente, se for usado um nível de significância de 5%, há diferenças significativas entre os algoritmos A e C, sendo A considerado significativamente melhor que C, já que seu *rank* médio é menor.

$$g = \frac{R_i - R_j}{\sqrt{\frac{K\,(K+1)}{6B}}} \tag{D.5}$$

Tabela D.3: *Valores da função g e respectivo p-valor para cada par possível de indutores. A quarta coluna traz os p-valores da comparação de todos contra todos, ajustados por Nemenyi; a quinta coluna traz os p-valores da comparação do indutor A com todos os outros, ajustados por Bonferroni.*

Pares	g	*p-valor*	*p-valor ajustado Nemenyi*	*p-valor ajustado Bonferroni*
A x B	1,15	$2,50 \times 10^{-1}$	$1,00 \times 10^{0}$	$1,00 \times 10^{0}$
A x C	3,70	$2,16 \times 10^{-4}$	$2,16 \times 10^{-3}$	$8,62 \times 10^{-4}$
A x D	2,30	$2,14 \times 10^{-2}$	$2,14 \times 10^{-1}$	$8,58 \times 10^{-2}$
A x E	0,65	$5,16 \times 10^{-1}$	$1,00 \times 10^{0}$	$1,00 \times 10^{0}$
B x C	2,55	$1,08 \times 10^{-2}$	$1,08 \times 10^{-1}$	-
B x D	1,15	$2,50 \times 10^{-1}$	$1,00 \times 10^{0}$	-
B x E	1,80	$7,19 \times 10^{-2}$	$7,19 \times 10^{-1}$	-
C x D	1,40	$1,62 \times 10^{-1}$	$1,00 \times 10^{0}$	-
C x E	4,35	$1,36 \times 10^{-5}$	$1,36 \times 10^{-4}$	-
D x E	2,95	$3,18 \times 10^{-3}$	$3,18 \times 10^{-2}$	-

D.5 Comparação de Dois Algoritmos

No caso da comparação de apenas dois algoritmos, A e B, um teste não paramétrico mais adequado é o de Wilcoxon (1945). Nas colunas 1, 2 e 3 da Tabela D.4 encontram-se, respectivamente, a identificação do conjunto de exemplos e as medidas de acurácia obtidas pelos algoritmos A e B em cada um dos conjuntos. Os passos do teste são:

1. Para cada conjunto de exemplos, calcular a diferença entre a medida obtida pelo algoritmo B e a medida obtida pelo algoritmo A, ou seja, B-A (quarta coluna da Tabela D.4);

2. Considerando apenas o valor absoluto das diferenças encontradas anteriormente (coluna 5 da Tabela D.4), calcular o *rank* de cada diferença. Pode-se notar que não estão sendo calculados os *ranks* de cada indutor em cada conjunto, como no teste de Friedman, mas sim o *rank* de cada diferença absoluta entre todas as calculadas. A menor diferença recebe *rank* 1, a segunda menor recebe *rank* 2 e assim por diante. Havendo empates, o valor médio é utilizado. Por exemplo, se a quinta e sexta menores diferenças forem iguais, dá-se o valor 5,5 como *rank* para cada uma e o próximo valor de *rank* a ser considerado será 7. O *ranking* final para este caso é mostrado na coluna 6 da Tabela D.4;

3. Calcular duas somas, sendo a primeira denominada R^+, em que são somados os valores de *rank* para as linhas em que a diferença foi positiva (penúltima linha à esquerda da Tabela D.4), e a segunda denominada R^-, em que são somados os valores de *rank* para as linhas em que a diferença foi negativa (última linha à esquerda da tabela). Os *ranks* das linhas em que a diferença é zero são tratados diferentemente. Se houver um número ímpar de diferenças nulas, uma delas é descartada; caso contrário, nenhuma é descartada. A soma dos *ranks* das diferenças nulas restantes é dividida igualmente entre R^+ e R^-. No exemplo, há duas diferenças iguais a zero, nos conjuntos 9 e 18, não sendo, portanto, necessário descartar nenhuma. O *rank* de cada uma é $1,50$, somando 3, sendo $1,50$ somado a R^+ e $1,50$ somado a R^-;

4. Calcular o valor da função R^M do teste (penúltima linha à direita da tabela), sendo este igual ao menor valor entre R^+ e R^-;

5. Com o valor de R^M e considerando o número de conjuntos de exemplos utilizados, pode-se obter o p-valor (última linha à direita da tabela) para o teste. Alternativamente, definindo-se um nível de significância, pode-se consultar a tabela do teste. Geralmente, tais tabelas trazem valores para um número de conjuntos de até 25. Para mais conjuntos, pode-se calcular o valor

da função dada em D.6, que segue uma distribuição aproximadamente normal padrão. Se a tabela for usada, deve-se comparar o valor de R^M com o valor obtido da tabela: se R^M for menor, rejeita-se a hipótese nula.

$$w = \frac{R^M - \frac{1}{4}B(B+1)}{\sqrt{\frac{1}{24}B(B+1)(2B+1)}} \tag{D.6}$$

Como se pode notar, o p-valor encontrado é grande com relação a qualquer nível de significância praticável, fazendo com que não seja possível rejeitar a hipótese nula do teste, que diz que as diferenças foram encontradas ao acaso. Caso a hipótese nula fosse rejeitada, haveria diferença significativa entre os dois indutores. Se este fosse o caso, para decidir qual dos dois é o melhor, seria necessário verificar qual soma foi menor: se R^+ fosse menor, o melhor indutor seria o B, pois a diferença foi calculada como $B - A$; caso contrário, A seria o melhor.

Tabela D.4: *Desenvolvimento do teste de Wilcoxon. Na segunda e terceira colunas, são mostrados os valores de acurácia obtidos pelos algoritmos; na quarta coluna, pode ser visualizado o resultado da acurácia de A subtraída da acurácia de B; o valor absoluto da diferença anterior é mostrado na coluna 5; o rank de cada diferença absoluta é mostrado na última coluna. Nas duas últimas linhas à esquerda, são mostradas a soma dos ranks em que B foi melhor que A (R^+) e a soma dos ranks em que A foi melhor que B (R^-), compartilhando igualmente os ranks em que A e B empataram. Nas duas últimas linhas à direita, são mostrados o valor R^M e o p-valor do teste.*

Bases	**A**	**B**	**d**	**d**	***rank***
1	78,33	89,17	10,84	10,84	17,00
2	95,67	98,72	3,05	3,05	5,00
3	95,30	99,30	4,00	4,00	9,00
4	95,50	99,50	4,00	4,00	9,00
5	61,11	64,29	3,18	3,18	7,00
6	71,82	69,58	-2,24	2,24	4,00
7	97,50	88,89	-8,61	8,61	16,00
8	83,50	76,70	-6,80	6,80	13,00
9	49,20	49,20	0,00	0,00	1,50
10	62,58	74,84	12,26	12,26	18,00
11	48,00	46,00	-2,00	2,00	3,00
12	60,67	52,67	-8,00	8,00	15,00
13	96,90	92,90	-4,00	4,00	9,00
14	87,36	73,96	-13,40	13,40	19,00
15	83,57	80,48	-3,09	3,09	6,00
16	53,96	61,88	7,92	7,92	14,00
17	58,95	44,32	-14,63	14,63	20,00
18	74,05	74,05	0,00	0,00	1,50
19	83,56	78,67	-4,89	4,89	12,00
20	84,26	80,13	-4,13	4,13	11,00
R^+	80,50			R^M	80,50
R^-	129,50			**p-valor**	0,36

D.6 Considerações Finais

Muitas medidas de desempenho utilizadas para analisar classificadores, *e.g.*, acurácia, dependem da definição de um ponto de corte específico para os valores de *score* calculados e geralmente tentam

minimizar a probabilidade de erro. Há casos em que o que se quer minimizar são os custos de erros de classificação, mas geralmente estes não são conhecidos. Nestes casos, AUC é a medida mais indicada, justamente por ser insensível aos custos, pelo menos nos problemas binários (Bradley 1997). Apesar da medida AUC ser largamente usada, a maioria dos algoritmos de aprendizado baseia seu comportamento na minimização da taxa de erro. Alguns autores defendem que a otimização de AUC seja usada ao invés da taxa de erro na construção de classificadores (Cortes & Mohri 2004).

Um ponto interessante é com relação aos classificadores cuja curva ROC apresenta pontos abaixo da reta de modelos aleatórios, ou àqueles que possuem $AUC < 0,5$: caso o classificador seja negado, ou seja, todas as suas classificações positivas sejam consideradas negativas e vice-versa, os pontos abaixo passam para cima da referida reta. É como se o indutor tivesse encontrado nos dados de treinamento informação suficiente para produzir modelos melhores que os aleatórios, mas estivesse utilizando tal informação de maneira errada (Fawcett 2006).

Os testes de Nemenyi e Bonferroni são considerados conservativos e com pouco poder para verificar as diferenças significativas apontadas pelo teste de Friedman. Por isto, seu uso tem sido mais restrito. Foram apresentados aqui por serem mais adequados para explicar os conceitos envolvidos de maneira mais compreensível. Há testes mais poderosos, como os que controlam a taxa de erro da família de comparações. São exemplos Holm (1979), Hochberg (1988) e Hommel (1988). Estes testes não ajustam todos os p-valores da mesma forma, sendo que tal ajuste depende, por exemplo, do valor absoluto das diferenças encontradas. Há ainda outros testes, como o de Benjamini & Hochberg (1995), cuja estratégia é controlar a taxa de falsos positivos, ao invés da taxa de erro da família de comparações. Como será visto adiante, este mestrado tem usado alguns destes testes considerados mais poderosos.

Como última observação, tanto em Friedman quanto em Wilcoxon, há outras formas de se lidar com empates. Além disso, há outras formas também de se lidar com diferenças iguais a zero no teste de Wilcoxon, sendo possível, por exemplo, simplesmente eliminar da análise os conjuntos em que os dois indutores foram exatamente iguais.