Uma Adaptação do Método *Binary Relevance* Utilizando Árvores de Decisão para Problemas de Classificação Multirrótulo Aplicado à Gênomica Funcional

Erica Akemi Tanaka

SERVIÇO DE PÓS-GRADUAÇÃO DA BIOINFORMÁTICA USP

Data de Depósito: 13/06/2013

Assinatura:_____

Uma Adaptação do Método *Binary Relevance* Utilizando Árvores de Decisão para Problemas de Classificação Multirrótulo Aplicado à Gênomica Funcional

Erica Akemi Tanaka

Orientador: *Prof. Dr. José Augusto Baranaukas*

Dissertação de Mestrado apresentada ao Programa Interunidades de Pós-Graduação em Bioinformática da USP, como requisito para obtenção do título de Mestre em Bioinformática.

USP - Ribeirão Preto
Junho/2013

À minha família por estar sempre presente independente da situação. Especialmente, aos meus pais pelo amor incondicional e por todo apoio durante toda a caminhada.

Agradecimentos

Agradeço aos meus pais e meus familiares pelo carinho durante todo o período de mestrado.

Ao meu namorado e melhor amigo Oscar pelo apoio, incentivo e companheirismo.

Aos meus inesquecíveis e sempre presentes amigos, Rafael, Patricia, Thais e Silvio por toda amizade, companhia e o apoio que fizeram destes anos inesquecíveis.

Ao professor José Augusto Baranauskas do Departamento de Computação e Matemática pela orientação, atenção e incentivo durante a realização deste projeto.

À CAPES pelo apoio financeiro.

"Bom mesmo é ir à luta com determinação, abraçar a vida com paixão, perder com classe e vencer com ousadia, porque o mundo pertence a quem se atreve e a vida é muito pra ser insignificante." (Charles Chaplin)

Resumo

Muitos problemas de classificação descritos na literatura de aprendizado de máquina e mineração de dados dizem respeito à classificação em que cada exemplo pertence a um único rótulo. Porém, vários problemas de classificação, principalmente no campo de Bioinformática são associados a mais de um rótulo; esses problemas são conhecidos como problemas de classificação multirrótulo. O princípio básico da classificação multirrótulo é similar ao da classificação tradicional (que possui um único rótulo), sendo diferenciada no número de rótulos a serem preditos, na qual há dois ou mais rótulos. Na área da Bioinformática muitos problemas são compostos por uma grande quantidade de rótulos em que cada exemplo pode estar associado. Porém, algoritmos de classificação tradicionais são incapazes de lidar com um conjunto de exemplos mutirrótulo, uma vez que esses algoritmos foram projetados para predizer um único rótulo. Uma solução mais simples é utilizar o método conhecido como método *Binary Relevance*. Porém, estudos mostraram que tal abordagem não constitui uma boa solução para o problema da classificação multirrótulo, pois cada classe é tratada individualmente, ignorando as possíveis relações entre elas. Dessa maneira, o objetivo dessa pesquisa foi propor uma nova adaptação do método *Binary Relevance* que leva em consideração relações entre os rótulos para tentar minimizar sua desvantagem, além de também considerar a capacidade de interpretabilidade do modelo gerado, não só o desempenho. Os resultados experimentais mostraram que esse novo método é capaz de gerar árvores que relacionam os rótulos correlacionados e também possui um desempenho comparável ao de outros métodos, obtendo bons resultados usando a medida-F.

Palavras Chaves: Classificação Multirrótulo, Árvore de Decisão, Aprendizado de Maquina.

Abstract

Many classification problems described in the literature on Machine Learning and Data Mining relate to the classification in which each example belongs to a single class. However, many classification problems, especially in the field of Bioinformatics, are associated with more than one class; these problems are known as multi-label classification problems. The basic principle of multi-label classification is similar to the traditional classification (single label), and distinguished by the number of classes to be predicted, in this case, in which there are two or more labels. In Bioinformatics many problems are composed of a large number of labels that can be associated with each example. However, traditional classification algorithms are unable to cope with a set of multi-label examples, since these algorithms are designed to predict a single label. A simpler solution is to use the method known as *Binary Relevance*. However, studies have shown that this approach is not a good solution to the problem of multi-label classification because each class is treated individually, ignoring possible relations between them. Thus, the objective of this research was to propose a new adaptation of Binary Relevance method that took into account relations between labels trying to minimize its disadvantage, and also consider the ability of interpretability of the model generated, not just its performance. The experimental results show that this new method is capable of generating trees that relate labels and also has a performance comparable to other methods, obtaining good results using F-measure.

Key Words: Multi-Label Classification, Decision Tree, Machine Learning.

Sumário

Lista de Figuras

Lista de Tabelas

Capítulo 1

Introdução

Este capítulo está organizado da seguinte forma: na Seção 1.1 é apresentada uma contextualização sobre a tarefa de classificação de dados; na Seção 1.2 são apresentadas as motivações desse projeto; na Seção 1.3 são detalhados os objetivos desse projeto; na Seção 1.4 encontra-se a organização desse documento.

1.1 Contextualização

Com os avanços da tecnologia, a produção de dados biológicos e médicos tornou-se mais rápida, aumentando consideravelmente a quantidade de dados para analisar. Devido a isso, ocorreu um aumento de desenvolvimento de aplicações que utilizam técnicas de aprendizado de máquina para facilitar a analise. Um exemplo de aplicação altamente desenvolvida são os sistemas de suporte à decisão, no qual utilizam métodos de classificação com a finalidade de analisar um grande número de variáveis para assim predizer algum posicionamento para determinado problema [Pollettini, 2012, Tanaka et al., 2010, Dyego Carlos Sales de Morais, 2012].

A classificação é uma das tarefas de mineração de dados mais investigadas, com inúmeras aplicações industriais e comerciais [Witten and Frank, 2005]. Basicamente, a tarefa de classi-

ficação consiste em descobrir conhecimento que pode ser usado para predizer a classe de um exemplo, cuja classe é desconhecida, com base nos valores dos atributos que descrevem tal exemplo.

Neste sentido existem duas versões da tarefa de classificação, de acordo com o número de rótulos a serem preditos para cada exemplo: (a) Classificação de único rótulo (*Single-Label Classification*) e (b) Classificação multirrótulo (*Multi-Label Classification*). Classificação de único rótulo refere-se à tarefa padrão de classificação, onde há apenas um rótulo (atributo meta) a ser predito. Os princípios básicos da classificação multirrótulo são semelhantes aos da classificação de único rótulo; no entanto, na classificação multirrótulo há dois ou mais rótulos a serem preditos. No caso de modelos simbólicos expressos como regras, a conclusão (ou consequente) de uma regra de classificação contém uma ou mais conclusões, cada uma envolvendo um rótulo diferente.

1.2 Motivação

Desde o avanço do *hardware* e *software*, o sequenciamento automatizado de fragmentos de DNA tornou-se possível, aumentando consideravelmente a quantidade de dados biológicos disponível, aumentando também a necessidade do uso de ferramentas computacionais para o processamento e extração de conhecimento. Como resultado, as técnicas de aprendizado de máquina são amplamente utilizadas para predizer funções de genes; em seguida, as melhores predições podem ser testadas em laboratório para validar estes resultados [Schietgat et al., 2010]. No entanto, a predição da função gênica é complexa, considerando o fato que um único gene pode ter múltiplas funções. Neste caso, a classificação multirrótulo parece ser mais apropriada.

O aprendizado multirrótulo vem sendo estudado recentemente em diversas áreas, com propostas de novas técnicas de classificação, principalmente na área de Bioinformática. Um exemplo de problema multirrótulo é o projeto *Gene Ontology*[1], que é a principal iniciativa na Bioinformática que tem como objetivo padronizar a representação de genes e de produtos de gênicos entre as espécies e bancos de dados. Nele, genes e proteínas podem apresentar mais de uma função ou característica. Outro exemplo é o Catálogo Funcional MIPS (Fun-

[1] http://www.geneontology.org/

Cat MIPS) [Ruepp et al., 2004], em que os genes e proteínas podem pertencer a mais de uma classe funcional. Portanto, o estudo e desenvolvimento de técnicas computacionais para classificação multirrótulo de proteínas, genes e outros dados biológicos e médicos é de grande interesse científico, pois com esse conhecimento será possível desenvolver novos medicamentos, tratamentos de doenças, auxilio a diagnóstico, entre outras finalidades. Por exemplo, em [Schietgat et al., 2010] e [Barutcuoglu et al., 2006] são apresentados trabalhos utilizando classificação multirrótulo para predizer funções gênicas.

Porém, algoritmos de classificação tradicionais são incapazes de lidar com um conjunto de exemplos mutirrótulo, uma vez que esses algoritmos foram projetados para predizer um único rótulo. Uma solução simples é decompor o conjunto original em conjuntos de exemplos aproximadamente idênticos, onde cada um contém todos os atributos e seus valores para cada exemplo, mas contendo apenas um dos rótulos a ser predito, conhecido como método *Binary Relevance*. Tal abordagem resulta que o algoritmo de classificação seja treinado c vezes, sendo c o número de rótulos. Estudos mostraram que tal abordagem não constitui uma boa solução para o problema da classificação multirrótulo [Clare and King, 2001, Suzuki et al., 2001], pois cada rótulo é tratado individualmente, ignorando as possíveis relações entre eles. Intuitivamente, um algoritmo que descobre um classificador para mais de um rótulo pode capturar algumas relações entre os mesmas e descobrir um classificador mais simples (por exemplo, com menor número de regras). Devido a esse problema, é importante o desenvolvimento de técnicas que utilizam a abordagem de decomposição por ser simples, porém que consigam capturar as relações entre os rótulos.

1.3 Objetivo

O objetivo deste estudo é propor uma adaptação do método de *Binary Relevance* usando árvores de decisão para tratar problemas multirrótulos visando a melhora da compreensão do conhecimento extraído. Por esta razão, o método aqui proposto foi concebido para capturar relações entre rótulos, no qual *Binary Relevance* não leva em conta e, consequentemente, tentar melhorar a capacidade de generalização do modelo. Além disso, a proposta também leva em conta a capacidade de interpretação, não só o desempenho, no sentido que ela utiliza árvores de

decisão para gerar o modelo e reduz o número de árvores para tentar facilitar a interpretação por parte de especialistas.

1.4 Organização do Documento

Esse documento está organizado da seguinte maneira: no Capítulo 2 são apresentados os conceitos fundamentais sobre genômica funcional; no Capítulo 3 são apresentados os conceitos fundamentais de classificação de dados e de indução de árvores de decisão, além de detalhar os conceitos básicos e as abordagens para tratar os problemas de classificação multirrótulo e apresentar algumas métricas para avaliação de desempenho; no Capítulo 4 é apresentada a metodologia desse projeto; no Capítulo 5 são detalhados os conjuntos de exemplos, as configurações utilizadas no experimento realizado e os resultados e discussão; no Capítulo 6 são apresentadas as conclusões desse estudo; no Apêndice A são detalhados os experimentos preliminares; no Apêndice B são mostrados as taxa de acerto das árvores de decisão induzidas na primeira etapa do algoritmo proposto.

Capítulo

2

Genômica Funcional

A genômica é a ciência que estuda o genoma dos organismos a partir do seu sequenciamento completo para tentar entender a sua estrutura, organização e função. Essa ciência divide-se em estrutural, funcional e comparativa. O estudo das funções gênicas cabe à genômica funcional, que tenta compreender as mudanças no funcionamento do genoma em diferentes estágios do desenvolvimento e sob diferentes condições ambientais. Atualmente, a genômica funcional é uma das principais áreas de pesquisa pois a compreensão das mudanças no funcionamento do genoma é de grande importância para compreender as possíveis relações entre os genes com doença ou característica de interesse, por exemplo [Clare, 2003].

Inicialmente neste capítulo, na Seção 2.1 são apresentadas as definições dos conceitos DNA, RNA e proteína; na Seção 2.2 são apresentadas as definições dos conceitos gene e ORFs; na Seção 2.3 é apresentada uma pequena fundamentação sobre predição de função gênica; na Seção 2.5 são apresentadas as considerações finais desse capítulo.

2.1 DNA, RNA e Proteínas

O DNA é um composto orgânico cujas moléculas contêm as instruções genéticas que coordenam o desenvolvimento e funcionamento de todos os seres vivos. Do ponto de vista químico, o DNA é um longo polímero de nucleotídeos, cuja cadeia principal é formada por moléculas de açúcares e fosfato intercalados unidos por ligações fosfodiéster. Ligada à molécula de açúcar está uma de quatro bases nitrogenadas (Adenina, Timina, Citocina, Guanina); essa sequência de bases ao longo da molécula de DNA constitui a informação genética [Clare, 2003]. Na Figura 2.1, adaptada de [dog, 2013], é ilustrada o dogma central de biologia molecular descrito em 1958 por Francis Crick na tentativa de relacionar o DNA, o RNA e as proteínas. O DNA pode se replicar e dar origem a novas moléculas de DNA, pode ainda ser transcrito em RNA, e este por sua vez traduz o código genético em proteínas.

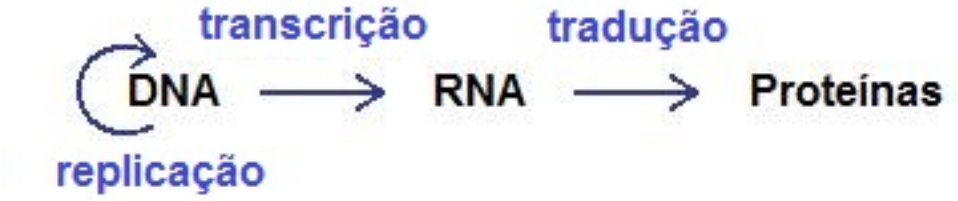


Figura 2.1: Dogma Central da Biologia Molecular

O DNA é o principal armazenador da informação genética sendo transcrita para moléculas de RNA que é um polímero de nucleotídeos sendo o responsável pela síntese de proteína da célula. As proteínas são compostos orgânicos bioquímicos de alto peso molecular constituídos por um conjunto de aminoácidos sendo consideradas as macromoléculas mais importantes das células envolvidas em vários tipos de funções como imunidade, estrutural, transporte, hormonal, metabólica, reparação e controle [Clare, 2003]. A sua estrutura e forma da molécula de proteína é altamente relevante para o trabalho da proteína, podendo ser descritas em vários níveis. Na Figura 2.2, adapatada de [Pro, 2013] (a) é ilustrada a estrutura primária que é a sequência de aminoácidos em si, já na Figura 2.2 (b) pode-se observar a estrutura secundária que é dada pelo arranjo espacial de aminoácidos próximos entre si na sequência primária; na Figura 2.2 (c) é mostrada a estrutura terciária que resulta do enrolamento das alfas-hélice ou das folhas-beta

conferindo atividade biológica às proteína e na Figura 2.2 (d) é ilustrada a estrutura quaternária que é dada pela distribuição espacial de mais de uma cadeia polipeptídica no espaço.

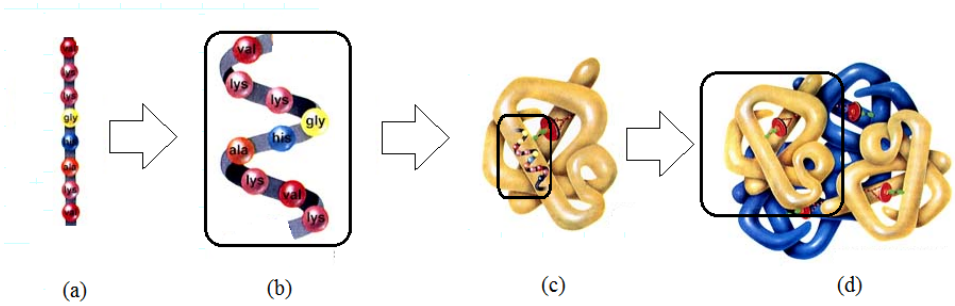


Figura 2.2: Estrutura das Proteínas. (a) Estrutura Primária, (b) Estrutura Secundária, (c) Estrutura Terciária, (d) Estrutura Quaternária

2.2 Gene e ORFs

O gene, como ilustrado na Figura 2.3 adaptada de [Gen, 2013], é um segmento de uma molécula de DNA que contém um código para a produção dos aminoácidos da cadeia polipeptídica e as sequências reguladoras para a expressão.Quando um gene se expressa, sua informação é primeiramente copiada no ácido ribonucléico (RNA), que por sua vez participa da síntese das proteínas específicas. Apesar de atualmente conhecermos como as informações contidas nos genes são codificadas em proteínas, muitas dessas proteínas/genes não possuem uma função conhecida [Carneiro et al., 2000].

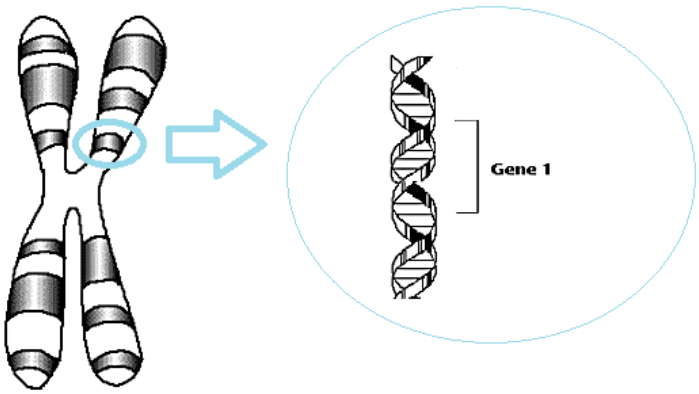


Figura 2.3: Gene

Os ORFs (Open Reading Frame) prevêem uma forte evidência para a estrutura de transcritos de RNA e são ferramentas indispensáveis para determinar a função gênica. Um ORF é uma sequência de DNA que começa com um codon de iniciação ATG (não sempre) e termina com qualquer um dos três codons de terminação (TAA, TAG, TGA).

2.3 Predição de Função Gênica

Determinar as funções dos genes e proteínas é um problema central na biologia, sendo fundamental para a compreensão dos processos moleculares e bioquímicas que sustentam a saúde ou a doença e para identificar e validar novos alvos terapêuticos e para o desenvolvimento de diagnósticos confiáveis. Na última década surgiram diversos projetos genomas visando obter o sequenciamento total ou parcial de sequências de DNA dos mais diversos organismos [Neto, 1997].

Devido a automatização do sequenciamento e com o surgimento da Bioinformática foi possível automatizar a fase de geração e da digitalização de sequências. A primeira etapa de um projeto genoma consiste no sequenciamento do DNA ou de cDNAs para gerar uma Biblioteca genômica ou de cDNAs. Após essa fase é necessário fazer a retirada sequências de adaptadores, vetores, rRNAs e cauda poli-A. Com as sequências 'limpas' prossegue-se então para a fase de montagem dos fragmentos no qual as sequências geradas pelo sequenciamento são analisados por programas que fazem cálculos de sobreposições montando as seqüências consenso (*contigs*) e fechando os espaços na sequência (*gaps*). Depois de remontar o genoma é preciso dar um significado biológico para todas as sequências geradas, isto é, identificar as regiões onde estão localizados os genes e identificar a sua função, essa fase é chamada de anotação [Carraro and Kitajima, 2002]. Com as sequências anotadas pode-se avançar para a etapa de mineração de dados (*Data Mining*) que por meio desse processo é possível selecionar, dentre todas as sequências geradas pelo projeto genoma, as possivelmente relacionadas com uma característica de interesse. Nesse contexto, existem muitos estudos na literatura que utilizam o aprendizado de máquina supervisionado para predizer funções gênicas [Marcotte EM, 1999, Chua et al., , Blockeel et al., 2006, Sharan R, 2007, Taşan et al., 2008].

2.4 Expressão Gênica

Expressão gênica é o processo pelo qual a informação hereditária contida em um gene é processada em um produto gênico funcional. Os maiores avanços nas técnicas de quantificação da expressão gênica ocorreram nos últimos tempos os quais podemos apresentar diversas técnicas como o *Microaarays* [Maskos and Southern, 1992], *Expressed Sequence Tags* (EST) [Adams M.D., 1995], *Massive Paralllel Signature Sequencing* (MPSS) [Brenner et al., 2000] e RNA-Seq [Ryan D. Morin and Marra., 2008].

A técnica *Microarray* é uma metodologia utilizada para comparar a expressão de um grande número de genes, simultaneamente, baseada na hibridização por complementaridade das moléculas de ácido nucleico, que ocorre entre a sonda depositada na lâmina e o seu RNAm correspondente extraído das amostras a serem analisadas, sendo marcadas com diferentes fluorescências (geralmente uma que emite cor vermelha e outra, verde) [Carneiro et al., 2000].

A técnica EST é baseada no sequenciamento de transcritos sendo útil para identificar transcrições genéticas e são cruciais no processo de descoberta de genes e na determinação de sequências genéticas, porém ela é extremamente trabalhosa e dispendiosa [Junior et al., 2004].

A técnica MPSS é uma ferramenta para a realização de perfis de expressão em profundidade sendo uma plataforma aberta, que analisa o nível de expressão de virtualmente todos os genes numa amostra por contagem do número de moléculas de mRNA individuais produzidos a partir de cada gene.

A técnica RNA-Seq é o primeira baseada em sequenciamento que permite que um transcriptoma completo seja pesquisado em larga escala e de maneira quantitativa, oferecendo resolução de até uma única base para anotação e níveis de expressão gênica digitais em escala genômica, normalmente a um custo bem menor, quando comparado às técnicas de *Microaarays* ou de *Expressed Sequence Tags* [Wang et al., 2009].

2.5 Considerações Finais

Nesse capítulo foram apresentados conceitos fundamentais sobre genômica funcional relatando que atualmente ela é uma das principais áreas de pesquisa. Logo após foram descritos

e conceituados os termos DNA, RNA, proteína, gene e ORFs. Finalmente, foi apresentada uma pequena fundamentação sobre predição de função gênica e de técnicas de quantificação da expressão gênica. No próximo capítulo são apresentados os conceitos fundamentais de classificação multirrótulo, apresentando as técnicas utilizadas para resolver problemas de classificação multirrótulo e algumas métricas usadas para avaliação de classificadores mutirrrótulos.

Capítulo

3

Aprendizado Multirrótulo

Inicialmente neste capítulo, na Seção 3.1 são apresentados os conceitos básicos da classificação de único rótulo detalhando como é realizado o processo de indução de uma árvore de decisão já que a metodologia aqui proposta a utiliza para gerar modelos interpretáveis pelo ser humano; na Seção 3.2 o problema de classificação multirrótulo é apresentado, assim como uma revisão bibliográfica detalhando as duas abordagens (independentes e dependentes) para tratar esse tipo de problema na Seção 3.2.2. Também são apresentadas algumas métricas de avaliação para problema multirrótulo na Seção 3.4; na Seção 4.3 são apresentadas as considerações finais desse capítulo.

3.1 Classificação Único Rótulo

A classificação de único rótulo refere-se à classificação tradicional, na qual há somente um rótulo a ser predito. Nesses problemas, um classificador é induzido usando um conjunto de exemplos que estão associados com uma única classe y de um conjunto de classes disjuntas C, sendo $|C| > 1$ [Tsoumakas and Katakis, 2007]. Se $|C| = 2$, então o problema é chamado de classificação binária e se $|C| > 2$ o problema é chamado de classificação multi-classe. Esse

classificador é obtido por meio de algoritmos de indução (indutores), que têm como objetivo gerar um classificador que seja capaz de classificar corretamente novos exemplos.

Formalmente, um exemplo z_i é um par $z_i = (x_i, y_i)$, sendo x_i a tupla de atributos de entrada e y_i o atributo de saída que possui a classe do exemplo z_i. Dado um conjunto de exemplos, a tarefa de um indutor é induzir uma função h que mapeie os valores $x_i = (x_{i1}, x_{i2}, \ldots, x_{im})$ em uma das classes y_i. Portanto, em classificação, dado um conjunto de exemplos de treinamento, um indutor gera como saída um classificador (também denominado hipótese ou modelo) de forma que, dado um novo exemplo, ele possa predizer sua classe.

Tabela 3.1: Conjunto de exemplos no formato atributo-valor para problemas único rótulo

	X_1	X_2	$\cdots$	X_m	Y
z_1	x_{11}	x_{12}	$\cdots$	x_{1m}	y_1
z_2	x_{21}	x_{22}	$\cdots$	x_{2m}	y_2
$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\vdots$
z_N	x_{N1}	x_{N2}	$\ldots$	x_{Nm}	y_N

Um dos formatos de representação do conjunto de exemplos é conhecido como *atributo-valor*, sendo amplamente utilizado pela maioria dos indutores, conforme mostrado na Tabela 3.1 [Monard and Bar
Os dados são caracterizados por N exemplos z_1, z_2, ..., z_N, cada um contendo m atributos $X_1, X_2, \ldots, X_m$ e um rótulo Y. Nessa tabela, a linha i refere-se ao i-ésimo exemplo ($i = 1, 2, \ldots, n$) e a entrada x_{ij} refere-se ao valor do j-ésimo ($j = 1, 2, \ldots, m$) atributo X_j do exemplo i.

Como pode ser notado, exemplos são tuplas $\vec{z}_i = (x_{i1}, x_{i2}, \ldots, x_{im}, y_i) = (\vec{x}_i, y_i)$ também denotados por $z_i = (x_i, y_i)$, onde fica subentendido o fato que tanto z_i como x_i são vetores. A última coluna, $y_i = f(x_i)$, é a função que tenta-se predizer a partir dos atributos. Observa-se que cada x_i é um elemento do conjunto $X_1 \times X_2 \times \ldots \times X_m$, onde X_j é o domínio do atributo e y_i pertence a uma das $\hat{k}$ classes, isto é, $y_i \in C \equiv \{C_1, C_2, \ldots, C_{\hat{k}}\}$.

Podem ser encontrados muitos algoritmos de classificação nos quais são divididos de acordo com o paradigma de classificação a que pertencem, sendo interessante para esse estudo o paradigma simbólico que é fundamentado na construção de representações simbólicas para a generalização do conhecimento podendo ser interpretadas por humanos [Michalski, 1983]. A

seguir são apresentados os conceitos básicos de indução de árvore de decisão a qual pertence ao paradigma simbólico.

Indução de Árvore de Decisão

As árvores de decisão (AD) são um dos classificadores mais populares, sendo que ADs de tamanho moderado oferecem uma fácil interpretabilidade de seus resultados para o usuário, característica muito relevante na área da Bioinformática, já que muitos problemas têm alto grau de complexidade [Rezende, 2003]. Elas são similares a regras *IF-THEN* sendo uma estrutura muito usada na implementação de sistemas especialistas e em problemas de classificação tomando como entrada uma situação descrita por um conjunto de atributos e retorna uma decisão (classe), que é o valor predito para o valor de entrada. Uma AD chega em sua decisão pela execução de uma sequência de testes (começando pela raiz da árvore), no qual cada nó interno da árvore corresponde a um teste do valor de um dos atributos, os ramos deste nó são identificados com os possíveis valores do teste e cada nó folha especifica o valor da classe.

Uma AD é constituída por dois tipos de nós:

- nós de decisão, que contêm um teste sobre o valor do atributo, que leva a uma sub-árvore;
- nós-folha, que indicam a classe correspondente.

Para melhor compreender o funcionamento de uma árvore de decisão, considerare o exemplo da Figura 3.1. O problema é distinguir a função de uma proteína entre "Estrutural" ou "Funcional". Na figura, cada elipse é um teste em um atributo para um dado conjunto de exemplos. Cada retângulo representa uma classe, ou seja, "Estrutural" ou "Funcional".

No Algoritmo 1 é mostrada a indução de uma AD: primeiramente, é feita a escolha do melhor atributo utilizando algum critério de seleção 15. Então, esse atributo é adicionado à árvore e também é adicionado um ramo para cada valor desse atributo 7. Após isso, os exemplos são subdivididos para cada ramo correspondente, considerando o valor do atributo selecionado 9. Se todos os exemplos são da mesma classe, então é feita a associação dessa classe a uma folha 2; caso contrário, é necessário repetir os passos anteriores recursivamente para cada subconjunto criado 10 [Quinlan, 1986].

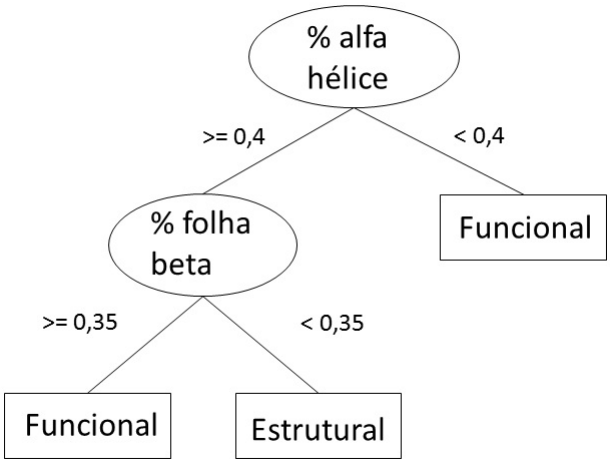


Figura 3.1: Uma AD simples para classificação de função de proteína

Algoritmo 1 buildTree

Require: conjunto de treinamento T
Ensure: ArvoreDecisão
1: **if** todos pertencem a mesma classe c **then**
2: ArvoreDecisão $\longleftarrow$ defina nó folha com classe c
3: **return** ArvoreDecisão
4: **end if**
5: Encontre o melhor atributo A em T
6: Considere $a_1, a_2, \cdots, a_r$ os possíveis valores para o atributo A
7: ArvoreDecisão $\longleftarrow$ defina nó A como raiz e considere $S(a_1), \cdots, S(a_r)$ as subárvores de A
8: **for** $i \longleftarrow 1$ to r **do**
9: Defina $T_i = \{z \in T - A = a_i\}$
10: S(a_i) = buildTree(T_i)
11: **end for**
12: **return** ArvoreDecisão

A chave para o sucesso de um algoritmo de aprendizado por AD depende do critério utilizado para escolher o atributo que particiona o conjunto de exemplos em cada iteração 15. Algumas possibilidades para escolher esse atributo são ganho máximo que seleciona o atributo que possui o maior ganho de informação esperado, isto é, seleciona o atributo que resultará no menor tamanho esperado das sub árvores, assumindo que a raíz é o nó atual; índice Gini [Breiman et al., 1984] e razão de ganho [Quinlan, 1993].

3.2 Classificação Multirrótulo

O princípio básico da classificação multirrótulo é similar ao da classificação tradicional binária, sendo diferenciado no número de rótulos a serem preditos, no qual há dois ou mais. Por exemplo, uma filme pode ser classificado como ação e aventura, um artigo no jornal pode ser classificado nas categorias música e cultura, uma proteína pode ser classificada com mais de uma função.

Define-se Y o conjunto de rótulos do problema, x_i a tupla de atributos de entrada, $y_i \in Y$ a tupla de rótulos do exemplo i e H o conjunto de classificadores para $h : x_i \rightarrow y_i$, no qual h é desconhecido. O objetivo é encontrar um classificador $h \in H$ que maximize a probabilidade de $h(x_i) = y_i$, no qual y_i é o conjunto de rótulos verdadeiros de exemplo i [Shen et al., 2004].

Na Tabela 3.2 é mostrada uma modificação do formato *atributo-valor* para tratar problemas multirrótulo. Os dados são caracterizados por N exemplos $z_1, z_2, \ldots, z_N$, cada um contendo m atributos $X_1, X_2, \ldots, X_m$ e c rótulos $Y_1, Y_2, \ldots, Y_c$. Nessa tabela, a linha i refere-se ao i-ésimo exemplo ($i = 1, 2, \ldots, N$), a entrada x_{ij} refere-se ao valor do j-ésimo ($j = 1, 2, \ldots, m$) atributo X_j do exemplo i e a saída y_{ki} refere-se ao valor do k-ésimo ($k = 1, 2, \ldots, c$) rótulo Y_k do exemplo i.

Tabela 3.2: Conjunto de exemplos no formato atributo-valor para problemas multirrótulo

	X_1	X_2	$\cdots$	X_m	Y_1	Y_2	$\cdots$	Y_c
z_1	x_{11}	x_{12}	$\cdots$	x_{1m}	y_{11}	y_{12}	$\cdots$	y_{1c}
z_2	x_{21}	x_{22}	$\cdots$	x_{2m}	y_{21}	y_{22}	$\cdots$	y_{2c}
$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$
z_N	x_{N1}	x_{N2}	$\ldots$	x_{Nm}	y_{N1}	y_{N2}	$\cdots$	y_{Nc}

Como pode ser observado, neste caso, exemplos são tuplas $\vec{z}_i$ =(x_{i1}, x_{i2}, ... ,x_{im},y_{i1}, y_{i2}, ..., y_{ic}) = ($\vec{x}_i$,$\vec{y}_i$) também denotados por $z_i = (x_i, y_i)$, onde fica subentendido o fato que o z_i, x_i e y_i são vetores. Observa-se que cada y_i é um elemento do conjunto $Y_1 \times Y_2 \times \ldots \times Y_c$, sendo que $Y_i \in \{0, 1\}$, isto é, cada rótulo tem duas classes (0 ou 1) e define-se l_i o número de rótulos j presentes no exemplo i, isto é, em que $y_{ij} = 1$.

3.2.1 Classificação Multirrótulo Binário x Classificação Multirrótulo Multi-Classe

A diferença entre a classificação multirrótulo binária e multi-classe é o número de classes distintas C de cada rótulo, isto é, o número valores distintos que cada rótulo pode possuir. Se $|C| = 2$, então o problema é chamado de classificação binária e se $|C| > 2$ o problema é chamado de classificação multi-classe.

Na Tabela 3.3 são apresentados somente os valores dos rótulos de um exemplo de problema multirrótulo binário no qual cada rótulo possui apenas 2 valores de classe, não necessariamente igual para todos os rótulos. Podemos observar o exemplo da Tabela 3.3 que as classes do rótulo Y_1 são {0,1}, do rótulo Y_2 são {No,Yes} e do rótulo Y_m são {1,2}.

Tabela 3.3: Exemplo de Problema Multirrótulo Binário

	Y_1	Y_2	$\cdots$	Y_m
z_1	0	Yes	$\cdots$	1
z_2	0	No	$\cdots$	1
z_3	1	No	$\cdots$	2
$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$
z_{N-1}	0	Yes	$\cdots$	1
z_N	1	No	$\ldots$	2

Na Tabela 3.4 são apresentados somente os valores dos rótulos de um exemplo de problema multirrótulo e multi-classe no qual cada rótulo possui mais de dois valores de classe, sendo observado que o rótulo Y_1 possui 3 valores de classes {a,b,c}, o rótulo Y_2 possui 4 valores de classes {x,y,z,w} e o rótulo Y_m possui 3 valores de classes {1,2,3}.

Tabela 3.4: Exemplo de Problema Multirrótulo e MultiClasse

	Y_1	Y_2	$\cdots$	Y_m
z_1	a	x	$\cdots$	1
z_2	b	y	$\cdots$	1
z_3	a	z	$\cdots$	3
$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$
z_{N-1}	b	w	$\cdots$	1
z_N	c	x	$\ldots$	2

Para os problemas em que há rótulos com mais de duas classes, chamados de multi-classe,

é possível realizar uma binarização das classes [Aly, 2005, Lee and Oh, 2003] para transformar em um conjunto de rótulos binário; então para cada rótulo com $\hat{k}$ classes, sendo $\hat{k} > 2$, é transformado em $\hat{k}$ rótulos binários.

3.2.2 Abordagens para Tratamento de Problemas Multirrótulo

Diferentes técnicas têm sido propostas na literatura para tratar problemas de classificação multirrótulo. Em algumas delas, classificadores único rótulo podem ser combinados para tratar problemas de classificação multirrótulo. Outras técnicas modificam classificadores único rótulo, por meio de adaptações em seus algoritmos, para permitir a utilização em problemas multirrótulo [Tsoumakas and Katakis, 2007]. Nessa subseção são apresentas algumas abordagens estudadas na literatura para tratar problemas de classificação multirrótulo.

Abordagens Independentes de Algoritmo

A abordagem independente de algoritmo lida com o problema transformando-o em um conjunto de problemas único rótulo, isto é, o conjunto de dados multirrótulo é transformado em um ou mais conjuntos de dados único-rótulo, dependendo do tipo de transformação. Após essa transformação é realizada a aplicação de algum algoritmo de classificação no(s) conjunto(s) de dados único-rótulo, para assim induzir um ou um conjunto classificadores para predizer todos rótulos de um novo exemplo. Essa transformação pode ser realizada baseando-se em rótulo ou em exemplo.

Transformação Baseada em Rótulo

Nesse tipo de transformação, o conjunto de exemplos original é dividido em conjuntos de exemplos no qual cada um contém todos os atributos e seus valores para cada exemplo, mas contendo apenas um dos rótulos a ser predito. Portanto, são utilizados c classificadores, sendo c o número de rótulos do problema e cada classificador gerado é treinado para distinguir um rótulo contra todos os demais rótulos envolvidos. Essa técnica é também chamada de técnica binária ou *One-versus-All* (OVA) ou *Binary-Relevance* (BR) [Tsoumakas et al., 2010]. Porém, essa técnica assume que os rótulos são independentes entre si, algo que nem sempre é verdade,

já que ignorar as possíveis relações entre os rótulos pode resultar em uma baixa capacidade de generalização [Cerri et al., 2009].

Um exemplo é ilustrado desta técnica na Figura 3.2, adaptada de [Cerri et al., 2009] bem como os das Figuras 3.3, 3.4, 3.5, 3.6 e 3.7, na qual é apresentado um problema multirrótulo de classificação de funções de proteínas com três rótulos "Estrutural", "Hormonal" e "Reguladora". O problema é dividido em três problemas binários, gerando três classificadores sendo que o i-ésimo classificador ($i = 1, \ldots, 3$) é treinado para considerar os exemplos pertencentes ao i-ésimo rótulo como positivos e os outros como negativos a fim de distinguir o i-ésimo rótulo dos demais. Porém, essa técnica não constitui uma boa solução, pois ignora as possíveis relações entre os rótulos, além de ser computacionalmente cara, dependendo no número de rótulos do problema.

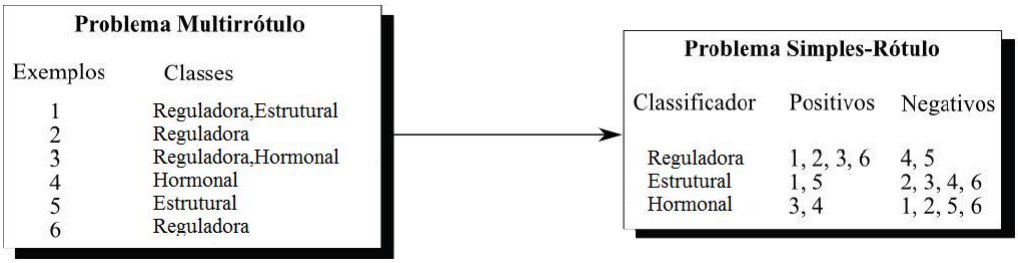


Figura 3.2: Exemplo de transformação baseada em rótulo

No trabalho de [Cherman et al., 2010] é proposta uma extensão do método BR, denominada BR+, no qual considera as relações entre os rótulos. Também são construídos c problemas de classificação binária de maneira análoga ao BR. A diferença está nos atributos descritores que além de conter os atributos X também contém todos os rótulos como descritores, exceto o próprio rótulo a ser predito. Nesse trabalho foi concluído que o BR+ mostra uma melhora na qualidade de predição em comparação com os métodos LP (descrito na subseção seguinte) e BR.

Transformação Baseada em Exemplos

Nesse tipo de transformação, o problema multirrótulo é transformado em um ou mais problemas de único rótulo, baseando-se no conjunto de rótulos associados a cada exemplo. Três diferentes estratégias são conhecidas para esse tipo de transformação, sendo uma delas baseada na

criação de rótulos. Essa estratégia pode ser chamada de *Label-Powerset* [Tsoumakas et al., 2010] e se baseia na combinação de mais de um rótulo para criar um novo rótulo; porém o número de rótulos pode aumentar consideravelmente e alguns podem terminar com poucos exemplos. Pode-se observar na Figura 3.3 que os rótulos criados "Reguladora_Estrutural" e "Reguladora_Hormonal" têm apenas um exemplo.

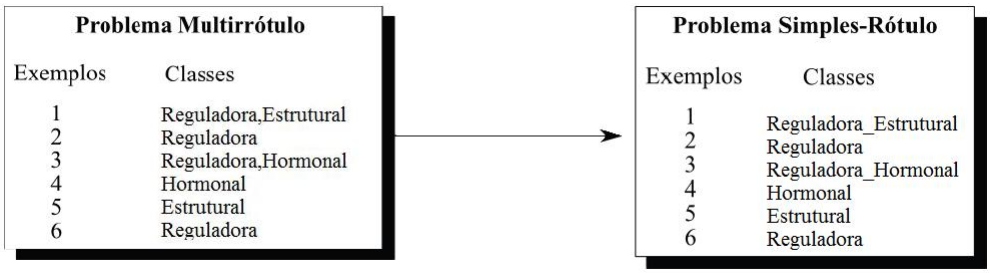


Figura 3.3: Criação de novos rótulos

No trabalho de [Tsoumakas and Vlahavas, 2007] é apresentado o algoritmo RAkEL (RAndom k-labELsets) que constrói um *ensemble* de classificadores *Label-Powerset* (LP) e cada classificador LP é treinado com um pequeno subconjunto k aleatório de rótulos. Esse método leva em conta as relações dos rótulos e, ao mesmo tempo, evita os problemas do LP citados acima. Foi concluído que o método RAkEL tem melhor desempenho em comparação aos métodos BR e LP.

Outra estratégia é a eliminação de exemplos, sendo a estratégia mais simples e menos eficaz que se baseia-se na retirada de exemplos que contenham mais de um rótulo; assim o problema multirrótulo deixa de existir, como ilustrado na Figura 3.4. Porém, ocorre uma perda de informação que pode ser relevante para o problema abordado.

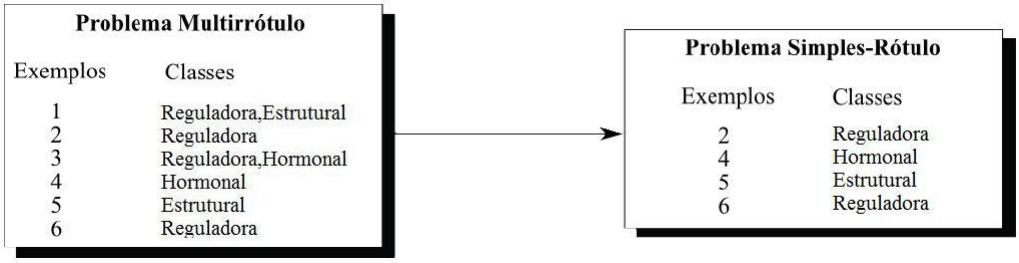


Figura 3.4: Eliminação de Exemplos com mais de uma classe

A terceira estratégia é a conversão de exemplos, sendo baseada na conversão de exemplos

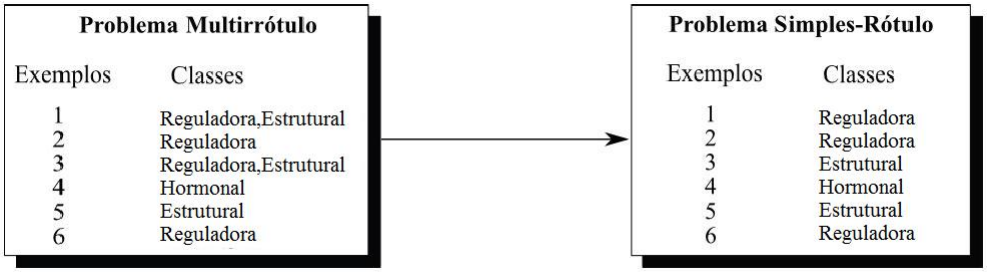


Figura 3.5: Transformação por Eliminação de rótulos

multirrótulos em exemplos único rótulo e existem duas variações dessa estratégia: eliminação e decomposição. Na primeira variação, todos os exemplos com mais de um rótulo são convertidos em exemplos único rótulo, escolhendo um dos rótulos associados ao exemplo e simplesmente eliminando os demais, gerando assim uma perda de informação, como pode ser observado na Figura 3.5. Essa escolha pode ser feita de modo determinista ou aleatória. Na segunda variação, a conversão dos exemplo com mais de um rótulo divide o problema multirrótulo com c rótulos e N exemplos em K conjuntos de problemas de único rótulo. O valor de K varia de 1 (quando nenhum exemplo possui mais de que um rótulo), a $(c-1)N$, se todos os exemplos possuem $c-1$ rótulos. Esse processo decomposição pode ser dividido em dois métodos: aditivo e multiplicativo. O método aditivo, ilustrado na Figura 3.6, considera que para cada exemplo, cada um dos possíveis rótulos será o rótulo positivo em sequência, esse método é também conhecido como *cross-training* [Shen et al., 2004]. Por exemplo, se os rótulos "Reguladora" e "Estrutural" aparecem nos exemplos multirrótulo, quando o classificador para o rótulo "Reguladora" for treinado, todos os exemplos multirrótulos que possuem o rótulo "Reguladora" se tornam exemplos único-rótulo para o rótulo "Reguladora" e o mesmo acontece para os outros rótulos.

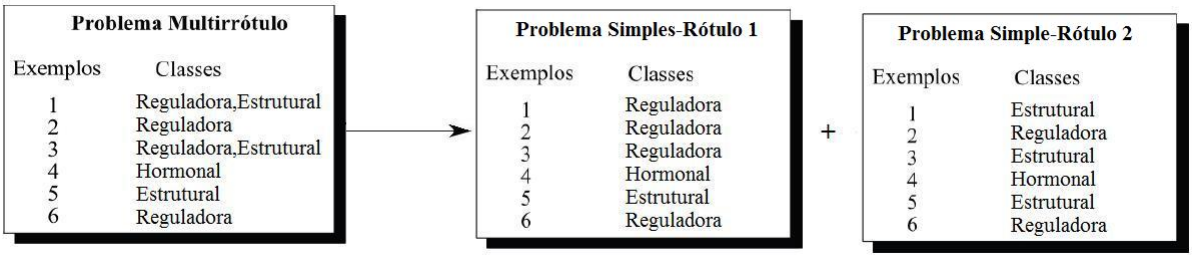


Figura 3.6: Transformação por Decomposição de rótulos - Método Aditivo

No método multiplicativo é realizada uma combinação de todos os possíveis problemas único rótulo é utilizada. O número de classificadores nesse método é igual ao $\prod l_i$, que é o produtório do número de rótulos presentes em cada exemplo i. É ilustrado na Figura 3.7 um exemplo desse método no qual o número de classificadores é dado pelo produto $2 \times 1 \times 2 \times 1 \times 1 \times 1$ no qual cada número corresponde à quantidade de rótulos que cada exemplo possui.

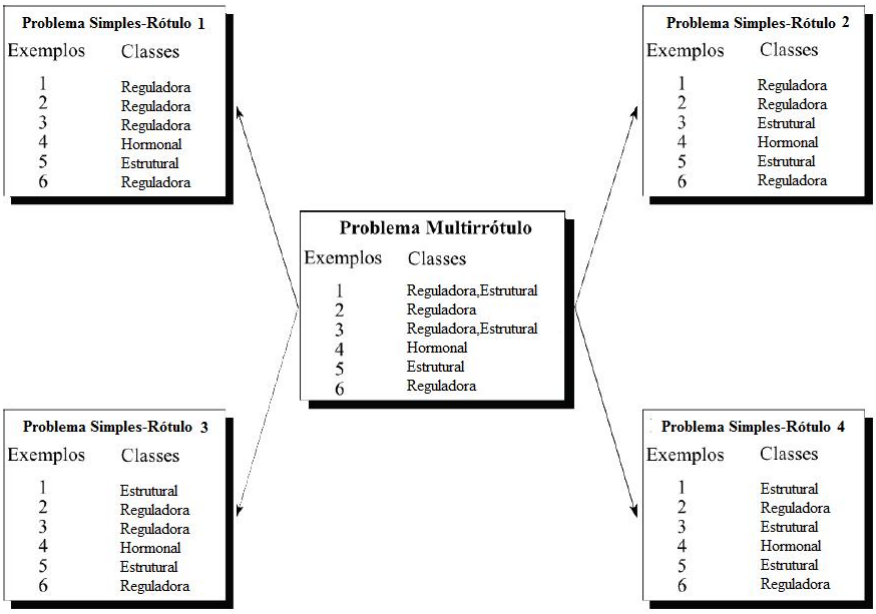


Figura 3.7: Transformação por Decomposição de rótulos - Método Multiplicativo

Abordagens Dependentes de Algoritmo

A abordagem dependente de algoritmo modifica internamente o algoritmo dos classificadores tradicionais (único rótulo) para que possam ser utilizados em problemas multirrótulos.

No trabalho de [Clare and King, 2001] é apresentado um estudo utilizando árvores de decisão para classificação hierárquica multirrótulo para analisar informações de *S. cerevisiae* e tentar predizer novas funções gênicas. Para analisar esses dados foram desenvolvidas estratégias de reamostragem e modificações no algoritmo C4.5 [Quinlan, 1993]. Foi feita uma alteração na formula da entropia, utilizando a soma das entropias de todos os rótulos.

Em [Alves et al., 2008] são propostas duas versões de um Sistema Imunológico Artificial (AIS), que é um paradigma de inteligência computacional relativamente recente para predizer funções de proteínas descritas na ontologia Gene Ontology (GO). A abordagem proposta

chamada MHCAIS (Multi-label Hierarchical Classification with an Artificial Immune System) é um algoritmo de classificação adaptado para problema multirrótulo e hierárquico. A primeira versão do MHCAIS constrói um classificador global para predizer todos os rótulos, enquanto a segunda versão constrói um classificador local para predizer cada rótulo. Em ambas as versões o classificador é expresso com um conjunto de regras *IF-THEN*, que tem a vantagem de representar o conhecimento compreensível para usuários biólogos. Os resultados dos experimentos desse trabalho mostram que a versão global obteve um pior desempenho considerando a precisão, porém obteve um bom desempenho considerando a medida revocação, com relação a versão local. A versão global tem a vantagem de construir um modelo mais simples.

No trabalho de [Comité et al., 2001] para tratar problemas multirrótulo é proposta uma extensão do ADTrees (Alternating Decision Trees) [Freund and Mason, 1999] chamado ADT-Boost.MH que combina os algoritmos ADTboost [Freund and Mason, 1999] e do AdaBoost.MH [Schapire and Singer, 1999]. Uma ADTrees consiste em nós de decisão que especificam uma condição de predição e em nós previsão que contêm um único número. O AdaBoost.MH é a versão para multirrótulo do AdaBoost no qual tem o objetivo de encontrar um forte classificador (com alta acurácia) combinando vários classificadores fracos (com baixa acurácia) e o ADTboost é a extensão do AdaBoost usando ADTrees.

No trabalho de [Tsoumakas et al., 2011] uma ferramenta chamada MuLAM foi desenvolvida baseada na biblioteca de aprendizado de maquina Weka [Witten and Frank, 2005], contendo vários métodos de modificações de algoritmos, como o ML-kNN [Zhang and Zhou, 2007] (Multi-Label k-Nearest Neighbours), BPMLL [Zhang, 2006] (Back-Propagation Multi-Label Learning), entre vários outros. O algoritmo ML-KNN é baseado no algoritmo kNN: para cada exemplo, os rótulos que são associados com k exemplos vizinhos são recuperados e é realizada uma contagem dos vizinhos associados a cada rótulo; então o principio *maximum posteriori* é utilizado para definir os rótulos de um novo exemplo. O algoritmo BPMLL é uma adaptação do popular algoritmo *back-propagation* para aprendizado multirrótulo, sendo a principal modificação desse algoritmo é a introdução de uma nova função de erro que considera múltiplos rótulos.

Em [Blockeel et al., 1998] uma ferramenta chamada Clus foi desenvolvida usando conceitos de *Predictive Clustering Trees* (PCT), no qual árvores de decisão são construídas onde cada nó

corresponde a um grupo de exemplos do conjunto de exemplos. O PCT é uma abordagem de clusterização que adapta a indução de árvores de decisão para a clusterização. Esse procedimento usado para construção da PCT é similar a outros algoritmos de indução de árvores de decisão, como C4.5 [Quinlan, 1993] ou CART [Breiman et al., 1984]. Existem várias diferenças entre o PCT e as ADs, a primeira é que na AD os nós folha contém a classe e no PCT os nós folhas simplesmente contém um conjunto de exemplos; a segunda é que o teste é localizado no próprio nó e não nos ramos como nas ADs e a última é em relação ao critério de divisão, onde o melhor teste é aquele que maximiza a distância entre dois *subclusteres* e nas ADs o melhor teste é aquele que possuí maior entropia.

No trabalho de [Blockeel et al., 2006], Clus-HMC refere-se ao uso do Clus como um sistema de classificação multirrótulo hierárquico, que aprende uma árvore para classificar todos os rótulos usando a distancia euclidiana ponderada e Clus-SC gera uma árvore de decisão para cada rótulo. Os resultados mostram que CLUS-HMC tem um melhor desempenho preditivo que CLUS-SC; o tamanho da árvore HMC é muito menor comparado com CLUS-SC; e o aprendizado de uma única árvore HMC é muito mais rápida do que aprender muitas árvores.

3.3 Balanceamento de Classes

Normalmente no mundo real as informações estão dispostas e agrupadas de maneira irregular e desbalanceada. Porém o desbalanceamento de classes é um obstáculo para algoritmos de classificação pois dificultam a construção de modelos que consigam discriminar corretamente o conjunto minoritário do majoritário [Tahir et al., 2009]. Todavia, justamente a classe com menor número de exemplos em geral, é a mais interessante e valiosa de se identificar.

Vários estudos reportaram que diversos classificadores básicos apresentaram melhor desempenho se aplicados a conjuntos de dados balanceados [Laurikkala, 2001, Tahir et al., 2009, Estabrooks et al., 2004, Orriols and Bernadó-Mansilla, 2005]. Então, uma solução é efetuar um ajuste no conjunto de dados de forma a igualar a distribuição de exemplos entre classes utilizando amostragem, seja removendo exemplos da classe majoritária, isto é, *undersampling* ou adicionando exemplos da classe minoritária, isto é, *oversampling* [Garcia V., 2007].

3.3.1 Balanceamento de Classes para Problemas Multirrótulo

Existem muitos métodos na literatura de balanceamento de classes para problemas único-rótulo, isto é, no qual o conjunto de dados só tem apenas um rótulo a ser predito. Porém, não há muitos trabalhos de balanceamento de classes para problemas multirrótulo.

Para mensurar o desbalanceamento do conjunto de exemplos foi proposta a Equação 3.1, definindo-se n0 sendo número de exemplos que tem o rótulo Y_i=0 e n1 sendo número de exemplos que tem o rótulo Y_i=1. Quanto maior o valor do Balanceamento, mais balanceadas as classes dos rótulos estão sendo que a situação ideal ocorre quando o seu valor é 1.

$$Balancealmento(D) = \frac{\sum_{Y_i=1}^{c} n0 \times n1}{N \times c} \tag{3.1}$$

Uma maneira de solucionar o problema de desbalanceamento de classes, tendo em vista a importância do problema de baixo desempenho na classificação de conjuntos de dados desbalanceados, é usando a abordagem independente de algoritmo. Nessa abordagem o problema multirrótulo é transformado em um conjunto de problemas único rótulo e assim é possível balancear um rótulo de cada vez. Então, foi proposto nesse estudo uma forma de balanceamento de classes para problemas multirrótulo baseando-se no método Binary Relevance e usando ambos os métodos de balanceamento descrito anteriormente (*oversampling* e *undersampling*).

Na Figura 3.8 é ilustrada a proposta de como é realizado o balanceamento usando o método Binary Relevance. Primeiramente, é feita a transformação do problema multirrótulo para um conjunto de problemas único-rótulo, para isso são gerados c *datasets* D_i que contêm todos os atributos X e apenas um rótulo Y_i $(i = 1 \ldots c)$.

No caso de problemas no qual o rótulo tem apenas duas classes, cada conjunto de exemplos D_i terá uma classe majoritária e uma classe minoritária. Então para cada conjunto de exemplos D_i é feita um *undersampling* dos exemplos pertencentes a classe majoritária, isto é, é selecionada uma porção dos exemplos que pertencem a classe majoritária e um *oversampling* dos exemplos pertencentes a classe minoritária, isto é, é adicionada exemplos da classe minoritária. Então são construídas as árvores com os conjuntos de exemplos balanceados.

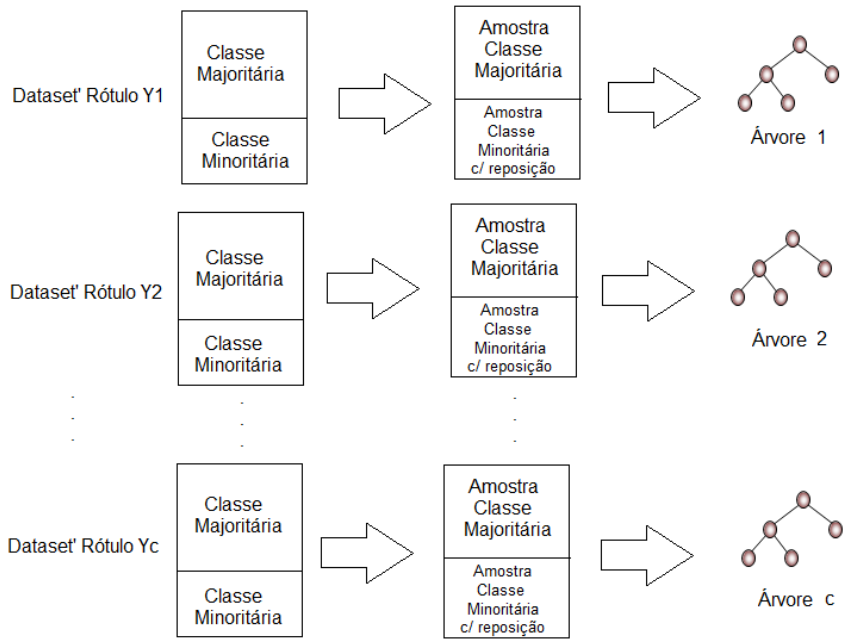


Figura 3.8: Balanceamento de Classes para problemas Multirrótulo

3.4 Métricas de Avaliação

O processo de classificação tradicional (único rótulo) geralmente envolve a divisão de dados em um conjunto de treinamento e um conjunto de teste que são disjuntos. Um algoritmo de classificação é aplicado a todos os exemplos no conjunto de treinamento, onde a classe de cada exemplo está disponível para o algoritmo. O algoritmo analisa então a relação entre os atributos e a classe para todos os exemplos de treinamento, procurando por um modelo de classificação para os dados. Em seguida, o modelo encontrado é aplicado aos exemplos no conjunto de teste, que nunca foram vistos durante o treinamento permitindo avaliar a precisão (acurácia) preditiva do modelo descoberto. Neste ponto, é crucial que os conjuntos de treinamento e teste sejam formados por conjuntos disjuntos de exemplos, ou seja, o conjunto de exemplos de teste nunca deve ter exemplos em comum com o conjunto de treinamento, a fim de caracterizar um cenário verdadeiramente preditivo, ou seja, que as medidas obtidas tenham valor estatístico. Dessa forma, é possível calcular uma medida de acurácia preditiva a partir do conjunto de teste. Mais precisamente, para cada exemplo no conjunto de teste, a classe predita pelo modelo de classificação é comparada com a classe verdadeira do exemplo, a fim de avaliar se a resposta

predita foi correta ou não. A definição padrão de acurácia (ou precisão de generalização) de um modelo de classificação é simplesmente o número de exemplos no conjunto de teste corretamente classificados por esse modelo, dividido pelo número total de exemplos do conjunto de teste [Domingos, 1997, Domingos, 1999]. Existem métricas adicionais a acurácia como a precisão, revocação, medida-F, por exemplo [Tan et al., 2006]. A precisão denota o percentual de acerto em relação a todos os exemplos tidos como positivos, a revocação denota o percentual de exemplos positivos que foram recuperados pelo classificador e a medida-F sintetiza as informações das últimas duas métricas, obtendo dessa maneira uma média harmônica entre as mesmas.

Porém a avaliação de classificadores multirrótulo não pode ser a mesma utilizada em classificadores tradicionais (único rótulo), pois na classificação multirrótulo um exemplo pode ser classificado de maneira parcialmente errada ou parcialmente correta. Isso acontece em casos em que o classificador atribui corretamente pelo menos um dos rótulos a que ele pertence, porém não associa algum rótulo que deveria ter sido associado ao exemplo, ou associa um rótulo incorreto [Vallin, 2010]. Nessa seção serão apresentadas algumas métricas propostas na literatura para avaliar classificadores multirrótulos [Tsoumakas and Katakis, 2007]. O critério de avaliação pode ser baseado na classificação realizada, utilizando os rótulos atribuídos pelo classificador para um dado exemplo ou em uma função de *ranking* que utiliza a posição em um *ranking* associado à cada rótulo pelo classificador.

No trabalho de [Schapire and Singer, 2000] é utilizada uma medida conhecida como *HammingLoss* que mensura o erro médio dos rótulos preditos. Seja D o conjunto de pares $z_i = (x_i, y_i)$ a serem classificados, N o número de exemplos em D, Y o conjunto de possíveis rótulos, h um classificador multirrótulo e $\hat{y}_i$ o conjunto de rótulos predito por h para o exemplo D_i. Nessa medida, o Δ representa a diferença simétrica entre dois conjuntos e corresponde à operação ou exclusivo (XOR) da lógica booleana. Quanto menor o valor de HammingLoss, melhor é a classificação sendo que a situação ideal ocorre quando o seu valor é zero. A equação é definida como:

$$HammingLoss(h, D) = \frac{1}{|N|} \times \sum_{i=1}^{N} \frac{|y_i \Delta \hat{y}_i|}{|Y|} \tag{3.2}$$

Algumas métricas comuns em recuperação de informação e em classificação multirrótulo incluem a precisão (3.4), a revocação (3.5) e a medida-F (3.6). Definindo tp_{Y_i}, fp_{Y_i}, tn_{Y_i} e fn_{Y_i} como o número de verdadeiros positivos, falsos positivos, verdadeiros negativos e falsos negativos avaliando o rótulo Y_i, respectivamente [Tsoumakas et al., 2010]. O cálculo dessas métricas pode ser feito usando duas operações, chamadas *macro-averaging* e *micro-averaging*. Na operação *micro-averaging* as métricas são calculadas globalmente sobre todos os rótulos, isto é, é realizada a soma de todos os tp_{Y_i}, fp_{Y_i}, tn_{Y_i} e fn_{Y_i} e então é realizado o cálculo das métricas. Na operação *macro-averaging* as métricas são calculadas localmente, isto é, é realizado o cálculo de cada métrica para cada rótulo individualmente e depois é feita uma média de todos os rótulos. A operação *micro-averaging* considera todos os exemplos com peso igual e a operação de *macro-averaging* considera todos os rótulos com peso igual independente da sua frequência [Özgür et al., 2005]. Devido a esses fatores neste estudo foi escolhida a operação de *micro-averaging*.

As equações são definidas considerando o *micro-averaging*.

$$Acuracia(h, D) = \frac{\sum_{Y_i=1}^{c} tp_{Y_i} + \sum_{Y_i=1}^{c} tn_{Y_i}}{\sum_{Y_i=1}^{c} tp_{Y_i} + \sum_{Y_i=1}^{c} fp_{Y_i} + \sum_{Y_i=1}^{c} tn_{Y_i} + \sum_{Y_i=1}^{c} fn_{Y_i}} \tag{3.3}$$

$$Precisao(h, D) = \frac{\sum_{Y_i=1}^{c} tp_{Y_i}}{\sum_{Y_i=1}^{c} tp_{Y_i} + \sum_{Y_i=1}^{c} fp_{Y_i}} \tag{3.4}$$

$$Revocacao(h, D) = \frac{\sum_{Y_i=1}^{c} tp_{Y_i}}{\sum_{Y_i=1}^{c} tp_{Y_i} + \sum_{Y_i=1}^{c} fn_{Y_i}} \tag{3.5}$$

$$Medida - F(h, D) = \frac{2}{\frac{1}{Precisao} + \frac{1}{Revocacao}} \tag{3.6}$$

3.5 Considerações Finais

Nesse capítulo foram apresentados conceitos fundamentais de indução de árvores de decisão e de classificação de dados convencional, isto é, único rótulo. Uma vez definidos os principais conceitos e processos envolvidos em uma classificação de dados convencional, foram abordados os conceitos de classificação multirrótulo. Para isso, foram apresentadas as técnicas utilizadas para resolver problemas de classificação multirrótulo e também foram apresentadas algumas métricas usadas para avaliação de classificadores mutirrrótulos. No próximo capítulo é apresentada a proposta desse estudo, os conjuntos de dados utilizados, assim como as configurações experimentais, resultados e discussões preliminares.

Capítulo 4

Proposta de Trabalho

Neste capítulo, na Seção 4.1 é apresentada a metodologia proposta, detalhando as etapas que serão seguidas; na Seção 4.2 são mostrados os trabalhos relacionados à proposta; e na Seção 4.3 são apresentadas as considerações finais desse capítulo.

4.1 Metodologia BR-RT

A proposta de projeto é implementar uma adaptação do método Binary Relevance utilizando árvores de decisão para tratar problemas multirrótulos, visando melhorar o desempenho em relação aos métodos já existentes na literatura e bem como melhorar a compreensão de profissionais da área da Bioinformática. Para isso, esse novo método capturará as possíveis relações entre os rótulos.

Foi escolhida a utilização de árvores de decisão em vez de regras de decisão por diversas razões, entre elas, o processo de aprendizagem de regras é mais lento que de árvores, as árvores de decisão determinam quais atributos são os mais importantes (possuem seleção de atributos embarcada), a classificação de novos dados usando árvores é rápida.

Antes de introduzir a metodologia proposta na forma de algoritmo, algumas notações são

Algoritmo 2 Binary Relevance with relation Labels - BR-RLb

Require: conjunto de exemplos multi-label D contendo m atributos $X_1, \ldots, X_m$ e c rótulos $Y_1, \ldots, Y_c$
Ensure: ArvoresEstendidas

1: $G \leftarrow \emptyset$
2: $Estendido \leftarrow \emptyset$
3: **for** $i \leftarrow 1$ **to** c **do**
4: $\quad A_i \leftarrow$ InduçaoAD(D_l^i)
5: $\quad$ **for** $w \leftarrow 1$ **to** c **do**
6: $\quad\quad$ **if** $Y_w \subset A_i$ **then**
7: $\quad\quad\quad G \leftarrow G \cup \{(Y_i, Y_w)\}$
8: $\quad\quad$ **end if**
9: $\quad$ **end for**
10: **end for**
11: **for** $i \leftarrow 1$ **to** c **do**
12: $\quad T_i \leftarrow$ InduçaoAD(D_a^i)
13: $\quad S \leftarrow A_i$
14: $\quad T_i' \leftarrow T_i$
15: $\quad$ **loop**
16: $\quad\quad SR \leftarrow$ SelecionaTodasRegras(S), nas quais $R_j^{T_i'} = R_k^S$
17: $\quad\quad Rule^{T_i'} \leftarrow$ ConstroiRegras(SR), na forma $L_j^{T_i'} \rightarrow R_j^{T_i'} \wedge R_k^S$
18: $\quad\quad L(Rule^{T_i'}) \leftarrow$ calcula a precisão de Laplace de $Rule^{T_i'}$
19: $\quad\quad \Omega \leftarrow$ seleciona a regra com maior $L(Rule^{T_i'})$
20: $\quad\quad$ Estendido $\leftarrow$ Estendido $\cup \{Y_1, Y_{i-1}, Y_{i+1}, \ldots, Y_c\} \cap \Omega$
21: $\quad\quad T_i' \leftarrow T_i' \cup$ Estendido
22: $\quad\quad$ **if** Há rótulos a serem considerados em $\{Y_1, \ldots, Y_{i-1}, Y_{i+1}, \ldots, Y_c\}$ **then**
23: $\quad\quad\quad SL \leftarrow$ seleciona um rótulo y considerando a melhor acurácia de A_y de Estendido
24: $\quad\quad\quad S \leftarrow A_{SL}$
25: $\quad\quad$ **else**
26: $\quad\quad\quad$ exit loop
27: $\quad\quad$ **end if**
28: $\quad$ **end loop**
29: **end for**
30: ArvoresEstendidas $\leftarrow \emptyset$
31: **for** $j \leftarrow 1$ **to** $C(G)$ **do**
32: $\quad$ ArvoresEstendidas $\leftarrow$ ArvoresEstendidas $\cup$ {seleciona T_i' com menor HammingLoss}
33: **end for**
34: **return** ArvoresEstendidas

necessárias, a saber:

- D: o conjunto de exemplos completo com todos os atributos e rótulos $\{X_1, \ldots, X_m, Y_1, \ldots, Y_c\}$;
- D_l: o conjunto de exemplos de rótulos, definido como $D_l \equiv D \backslash \{X_1, \ldots, X_m\}$;
- D_a: o conjunto de exemplos de atributos, no qual $D_a \equiv D \backslash \{Y_1, \ldots, Y_c\}$;
- D_l^i: conjunto de exemplos especifico para um rótulo, definido como $D_l^i \equiv \{Y_1, \ldots, Y_{i-1}, Y_{i+1}, \ldots, Y_c\} \cup \{Y_i\}$, no qual $\{Y_1, \ldots, Y_{i-1}, Y_{i+1}, \ldots, Y_c\}$ representa os atributos de aprendizado e $\{Y_i\}$ o rótulo (conceito a ser aprendido);
- D_a^i: conjunto de exemplos contendo todos os atributos e o rótulo Y_i que representa o atributo-alvo, definido como $D^i \equiv D_a \cup \{Y_i\}$;
- $Rule_j^t$: j-ésima regra da árvore t, na qual $R_j^t \equiv B^t \rightarrow E^t$, ou seja a regra 'if B^t then E^t'.

A metodologia proposta para tratar problemas multirrótulo pode ser vista no Algoritmo 2, o qual é dividido em três etapas.

Na primeira etapa (Linhas 3-10), esquematizada na Figura 4.1, é realizada a indução de c árvores de decisão, utilizando o Algoritmo 1, levando em consideração somente os rótulos. Nessa situação, para cada rótulo Y_i ($i = 1, \ldots, c$) uma árvore de decisão A_i é induzida, usando como atributos os $c - 1$ remanescentes rótulos (Y_1, ..., Y_{i-1}, Y_{i+1}, ..., Y_c) e o rótulo Y_i como atributo-alvo (Linha 4). Após isso, as c árvores induzidas anteriormente são convertidas em uma estrutura de grafo G, inicialmente vazia. Seja $C(G)$ o número de componentes conexos do grafo G. Para cada A_i, uma aresta conectando os rótulos Y_i e Y_j é adicionada em G se os rótulos Y_i e Y_j são conectados em A_i (Linha 7). Na Figura 4.3(a) é ilustrado um exemplo de como o grafo é construído a partir de um conjunto de árvores $A_1, \ldots, A_4$.

Essa etapa tenta encontrar grupos de rótulos relacionados, sendo representados por um componente conexo em G. No final dessa etapa, existem três possíveis situações: (1) $C(G) = 1$, todos os rótulos são relacionados entre si e, portanto, só há um único componente conexo no grafo G, que contém todos os rótulos; (2) $C(G) = c$, nenhum rótulo tem relação com os demais,

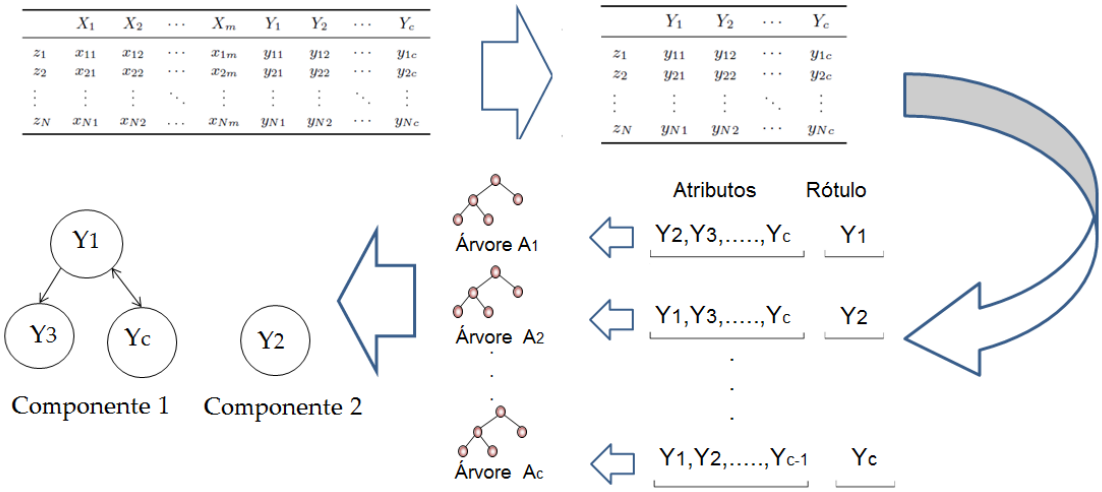


Figura 4.1: Esquema da metodologia BR-RT - Etapa 1

então G contem c componentes conexos; e (3) $1 < C(G) < c$, existem algumas relações entre alguns rótulos.

Antes de explicar a segunda etapa do algoritmo BR-RT, algumas considerações sobre critério de seleção da melhor regra são necessárias. Para que possamos considerar uma regra boa ou ruim, fazemos uso de certas métricas de avaliação de regras. Considerando cada regra no formato $R \rightarrow L$, sendo R a premissa e L a conclusão, podemos obter a matriz de contingência desta regra, para uma situação de duas classes, conforme a Tabela 4.1. Para um problema com mais de duas classes, é obtida uma matriz de contingência para cada classe.

Na Tabela 4.1, L é o conjunto de exemplos os quais a regra classifica, como positivos e $\overline{L}$ o conjunto de exemplos os quais a regra classifica como negativos. R sao os exemplos que pertencem a classe positiva e $\overline{R}$ sao os exemplos que pertencem a classe negativa. Assim, VP (verdadeiro positivo) é o conjunto dos exemplos em que a premissa e a conclusão são verdadeiras; FN é o conjunto dos exemplos em que a premissa é falsa e a conclusão é verdadeira; FP é o conjunto dos exemplos em que a premissa é verdadeira e a conclusão é falsa; por fim, VN é o conjunto dos exemplos em que a premissa e a conclusão são falsas. Para finalizar, r é o número de exemplos do conjunto R; $\overline{r}$ é o número de exemplos do conjunto $\overline{R}$; l é o número de exemplos do conjunto L; $\overline{l}$ é o número de exemplos do conjunto $\overline{L}$; n é o número total de exemplos.

Existem vários critérios para selecionar a melhor regra listados a seguir [Rezende, 2003]:

- Precisão Positiva:

Tabela 4.1: Matriz de Contingência

	L	$\overline{L}$	
R	VP	FN	r
$\overline{R}$	FP	VN	$\overline{r}$
	l	$\bar{l}$	n

$$L(R \rightarrow L) = \frac{VP}{l} \tag{4.1}$$

- Precisão Negativa:

$$L(R \rightarrow L) = \frac{VN}{\bar{l}} \tag{4.2}$$

- Precisão de Laplace Positiva:

$$L(R \rightarrow L) = \frac{VP + 1}{l + 2} \tag{4.3}$$

- Precisão de Laplace Negativa:

$$L(R \rightarrow L) = \frac{VN + 1}{\bar{l} + 2} \tag{4.4}$$

- Cobertura:

$$L(R \rightarrow L) = \frac{l}{n} \tag{4.5}$$

- Suporte:

$$L(R \rightarrow L) = \frac{VP}{n} \tag{4.6}$$

- Sensibilidade:

$$L(R \rightarrow L) = \frac{VP}{r} \tag{4.7}$$

- Especificidade:

$$L(R \rightarrow L) = \frac{VN}{\overline{r}} \tag{4.8}$$

- Precisão Total:

$$L(R \rightarrow L) = \frac{VP + VN}{n} \tag{4.9}$$

Para escolher a melhor métrica para selecionar as regras foram analisadas as vantagens e desvantagens de cada métrica citada anteriormente. As métricas precisão negativa, precisão de laplace negativa, especificidade e precisão total consideram o número de exemplos em que a premissa e a conclusão são falsas (VN) e por isso não são interessantes para selecionar a melhor regra, pois o importante é o número de exemplos em que a premissa e a conclusão são verdadeiras.

As métricas em que o divisor é o número total de exemplos também não são interessantes pois ao construir todas as possíveis regras para determinado ramo da árvore, o que diferencia uma regra da outra é a conclusão, portanto ao dividir pelo número total de exemplos acaba não se considerando a conclusão da regra.

A métrica precisão positiva e a sensibilidade possuem uma propriedade indesejada pois privilegia as regras com menos exemplos FN e FP sem considerar o número de exemplos VP, por exemplo uma regra 1 com VP =100, FN =2 e FP=2 e outra regra 2 com VP=10, FN=0 e FP=0. Então a regra 1 tem precisão positiva e sensibilidade igual a 0,98 e a regra 2 tem precisão positiva e sensibilidade igual a 1 indicando que a regra 2 é melhor que a regra 1. Uma solução para essa propriedade indesejada apresentada pelo precisão positiva é substitui-la pela precisão de laplace positiva, portanto decidiu-se usar essa métrica para selecionar as regras.

Na segunda etapa (Linhas 14-33), ilustrada na Figura 4.2, é realizada a indução das árvores T_i, porém considerando todos os atributos $X_1, \ldots, X_m$ e somente um rótulo Y_i por vez (Linha 16), utilizando o Algoritmo 1.

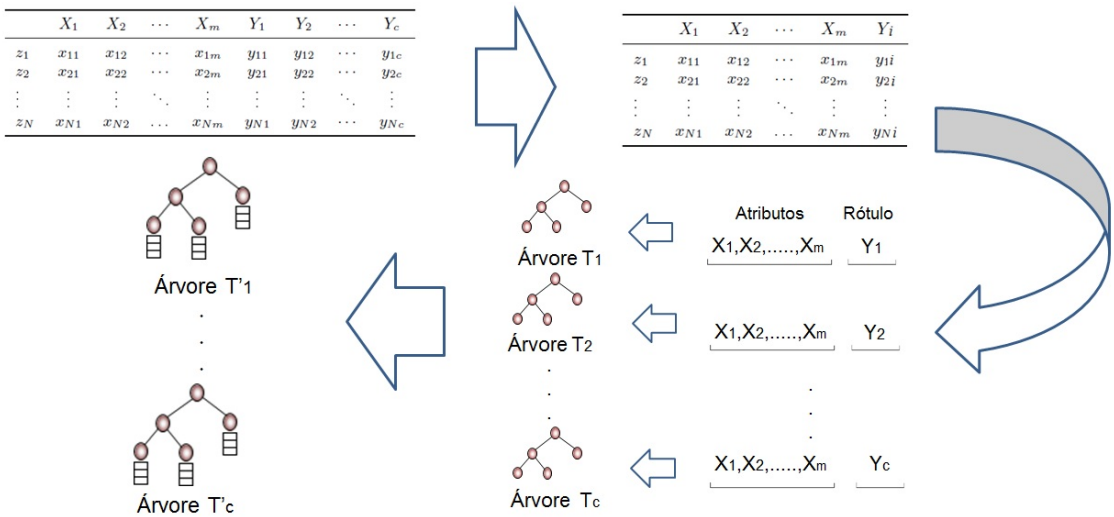
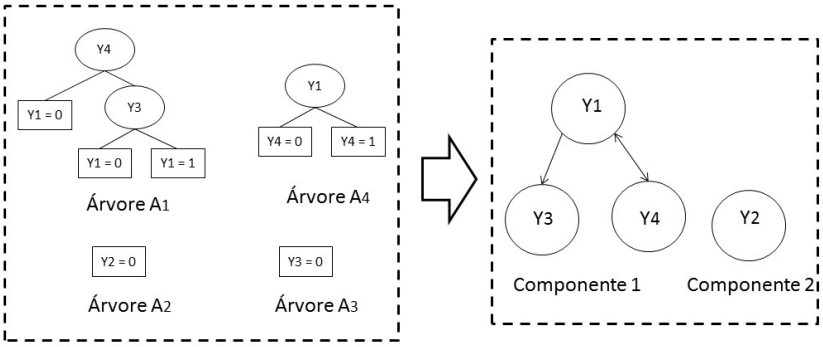


Figura 4.2: Esquema da metodologia BR-RT - Etapa 2

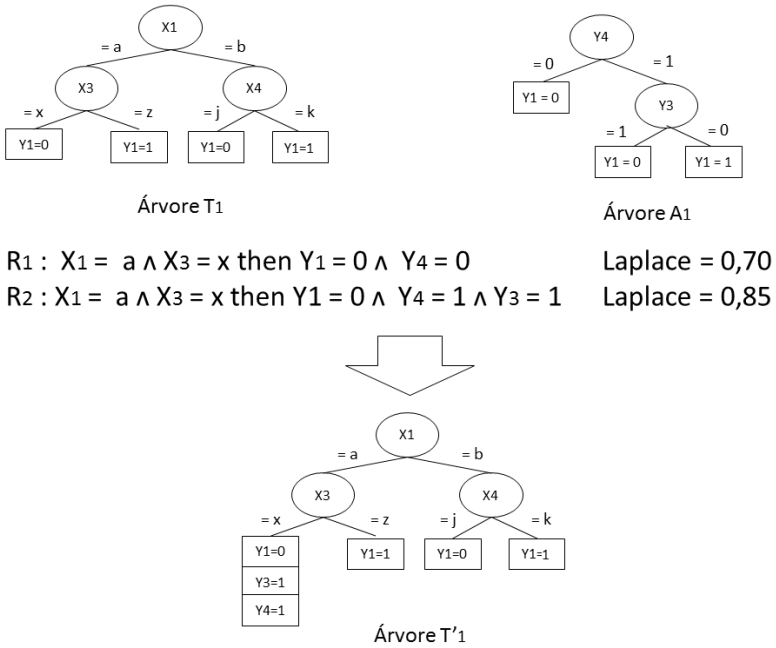
Após isso, cada árvore de decisão T_i é estendida (Linha 19), usando o componente conexo em G resultando em novas árvores T_i', que têm uma lista de rótulos em cada nó folha, isto é, se todos os rótulos são relacionados (primeiro caso acima) então a árvore T_i' será estendida para incluir todos os rótulos em suas folhas. Se houver dois ou mais componentes conexos (terceiro caso acima), a árvore T_i' estenderá apenas os rótulos que são parte do seu componente em G.

Para isso, primeiramente a árvore A_i é selecionada para começar a extensão da arvore T_i, no qual $S \leftarrow A_i$ (Linha 17).

Uma regra é criada para cada ramo de T_i. Para cada regra j de T_i são selecionadas todas as k regras de S, no qual $R_j^{T_i} = R_k^S$ (Linha 20). Depois disso, todas as k regras $Rule_k^{T_i}$ são construídas de forma lógica $Rule_k^{T_i} \equiv L^{T_i} \rightarrow R^{T_i} \wedge R^S$ (Linha 21), isto é, a premissa e a conclusão da k-ésima regra de T_i são unidas com a premissa da regra de S. Então, é calculada a precisão de Laplace (Linha 22) para todas as regras $Rule_k^{T_i}$, sendo escolhida a regra com maior precisão de Laplace (Linha 23).



(a)



(b)

Figura 4.3: A figura 4.3a ilustra a transformação das árvores A_i (esquerda) em grafo G (direita) e a figura 4.3b ilustra a extensão da árvore T_1

Na Figura 4.3(b) é ilustrada como é realizada a extensão de uma árvore, no qual o cálculo da

precisão de Laplace é feita para cada regra $Rule_k^{T_i}$, escolhendo o maior valor, como mencionado previamente. Ela ilustra, como é estendido o primeiro ramo da árvore T_1.

Se nem todos os rótulos foram estendidos do componente de T_i (Linha 26), então o processo de extensão continua selecionando uma outra árvore considerando somente as árvores pertencentes à Estendido, sendo Estendido um subconjunto de $\{A_1, \ldots, A_{i-1}, A_{i+1}, \ldots, A_c\}$, no qual somente se Y_i aparecer na regra selecionada então A_i é considerada como parte de Estendido. Do contrário, se a extensão da árvore T_i acabou (Linha 30) o laço termina.

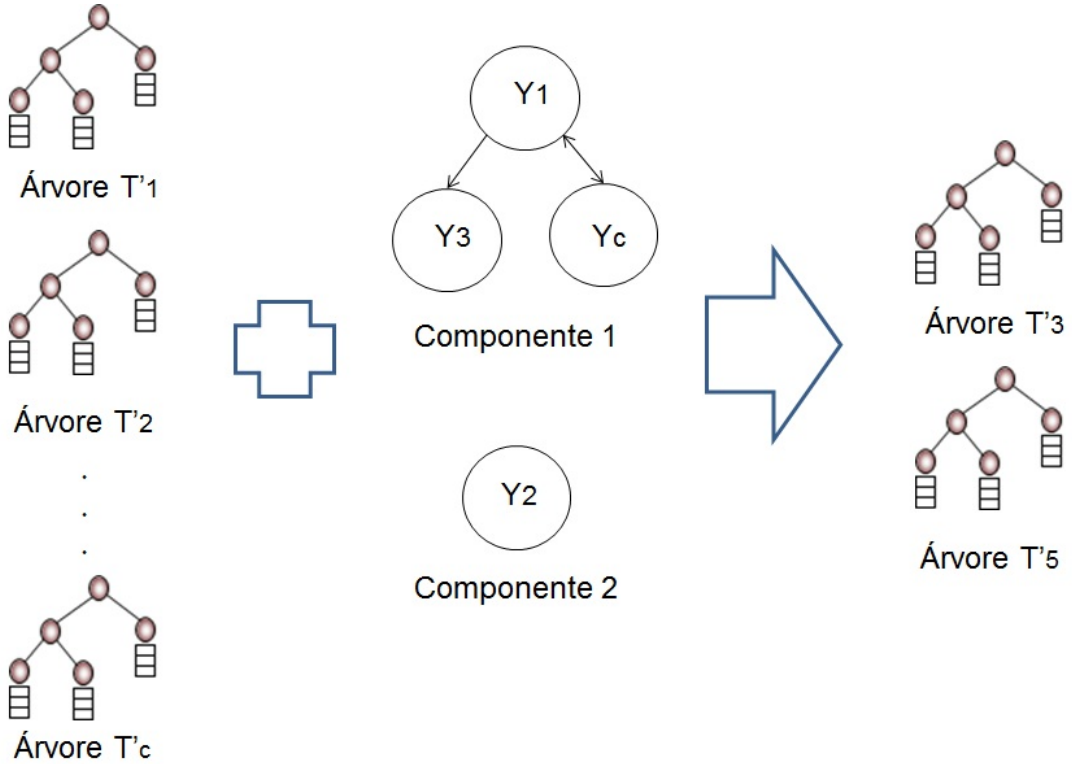


Figura 4.4: Esquema da metodologia BR-RT - Etapa 3

Na terceira etapa (Linhas 30-13), esquematizada na Figura 4.4, ocorre a seleção de uma árvore por componente conexo, sendo selecionada a com menor HammingLoss (Linha 12).

Na próxima seção são apresentados os trabalhos relacionados com essa proposta.

4.2 Trabalhos Relacionados

Nessa seção é descrita a relação entre alguns trabalhos mencionados da Seção 3.2.2 e a proposta aqui apresentada. Como já mencionado o método *Binary Relevance* apresenta uma

desvantagem por não considerar as relações entre os rótulos ao induzir o conjunto de classificadores.

O método aqui proposto se relaciona com o trabalho de [Cherman et al., 2010] já que ambos exploram a relação entre os rótulos e também são construídos c problemas de classificação binária, sendo diferenciada nos atributos usados como descritores que contém os rótulos como atributo descritor exceto o rótulo a ser predito. Já na proposta deste mestrado são construídos c problemas de classificação binária de maneira análoga ao BR, que posteriormente, serão estendidas e selecionadas diminuindo, possivelmente, o número de árvores a serem analisadas.

No trabalho de [Alves et al., 2008] é considerada a indução de um classificador global para predizer todos os rótulos assim como o método aqui proposto no qual, na melhor hipótese, todos os rótulos têm relação entre si, gerando um único classificador para todos os rótulos. Outra relação entre a proposta e este trabalho é que o classificador é expresso com um conjunto de regras *IF-THEN* e a(s) árvore(s) estendidas podem também serem vistas como um conjunto de regras *IF-THEN*, que têm a vantagem de representar o conhecimento compreensível para especialistas.

No trabalho de [Clare and King, 2001] é apresentado um estudo utilizando árvores de decisão, assim como o método aqui proposto. Assim como no trabalho de [Blockeel et al., 1998] no qual a ferramenta Clus no qual árvores de decisão são construídas onde cada nó corresponde a um grupo de exemplos do conjunto de exemplos. Outro trabalho de [Blockeel et al., 2006] no qual é apresentada o Clus-HMC que refere-se ao uso do Clus como um sistema de classificação multirrótulo hierárquico, que aprende uma árvore para classificar todos os rótulos, assim como o método aqui proposto no melhor caso (há somente um componente conexo).

4.3 Considerações Finais

Nesse capítulo foi apresentada a metodologia proposta, detalhando as suas etapas e foram apresentados os trabalhos relacionados à proposta. No próximo capítulo são apresentados os experimentos realizados com o Algoritmo 2 em comparação a alguns outros métodos citados na Seção 3.2.2. Os experimentos foram realizados utilizando dez conjuntos de dados biológicos.

Capítulo

5

Experimentos

Neste capítulo, na Seção 5.1.1 são detalhados os conjuntos de exemplos usado no experimento; na Seção 5.1.2 é apresentada a metodologia experimental utilizada, isto é, como o experimento foi conduzido; na Seção 5.1.4 são apresentados os resultados e as discussões; na Seção 5.2 são apresentadas as considerações finais desse capítulo.

5.1 Base Função de Proteína

5.1.1 Conjuntos de Exemplos

Os conjuntos de exemplos usados nos experimentos desse trabalho são na área da genômica funcional, relacionado ao organismo *Saccharomyces cerevisiae*, no qual os rótulos são estruturados hierarquicamente de acordo com o catálogo FunCat [Mewes et al., 2004] desenvolvido pelo MIPS disponível em 24/04/2002[1]. Esse catálogo provê descrições funcionais de proteínas, sendo estruturada como uma árvore com 4 níveis de profundidade. Na Figura 5.1 é mostrado um exemplo de como é a hieraquia, tirado do catálogo FunCat, na qual os níveis estão diferenciados

[1]http://www.aber.ac.uk/ dcswww/Research/bio/dss/yeastpreds/yeast/classes.txt

```
[1,0,0,0],"METABOLISM"
 [1,1,0,0],"amino acid metabolism"
  [1,1,1,0],"amino acid biosynthesis"
    [1,1,1,11],"biosynthesis of serine"
    [1,1,1,11],"biosynthesis of the cysteine-aromatic
    [1,1,1,15],"biosynthesis of the pyruvate family"
    [1,1,1,7],"biosynthesis of lysine"
    [1,1,1,7],"biosynthesis of the aspartate family"
```

Figura 5.1: Exemplo do catálogo FunCat

pela cor.

O *S. cerevisiae* ou levedura é um organismo eucarioto unicelular, usado na produção de pão e cerveja, sendo usado na produção de álcool também. É um dos organismos mais estudados na biologia e tem sido usado em vários estudos ao longo dos anos [Vens et al., 2008].

Os conjuntos de dados foram disponíbilizados pelo Universidade Católica de Leuven[2] e cada um descreve diferentes aspectos sobre os genes da levedura, incluindo cinco tipos de dados: estatísticas sobre a sequência, fenótipo, estrutura secundária, homologia e expressão. Nos experimentos foram usados os seguintes dez conjuntos de exemplos:

- **Seq**: contém estatísticas sobre a sequência que depende da sequência de aminoácidos da proteína codificada por um gene, sendo coletado a partir de várias fontes incluindo o ProtParam [Gasteiger et al., 2005] e MIPS;
- **Pheno**: contém dados sobre o fenótipo do organismo, sendo coletado a partir de várias fontes incluindo TRIPLES [Kumar et al., 2000], EUROFAN [Oliver et al., 1996] e MIPS;
- **CellCycle, Church, Derisi, Expr, Eisen, Gasch1, Gasch2, SPO**: dados de *microarray* de [Spellman et al., 1998], [Roth1JT et al., 1998], [Roth1JT et al., 1998],[Clare, 2003], [Eisen et al., 1998], [Gasch et al., 2000], [Gasch et al., 2001], [Chu et al., 1998], respectivamente;

Como esses conjuntos de dados são hierárquicos, primeiramente, foi realizado um pré-processamento para transformá-los em dados não hierárquicos. Nesse caso, um vetor binário

[2]http://dtai.cs.kuleuven.be/clus/hmc-ens/

Tabela 5.1: Características dos Conjuntos de Exemplos

Conjunto de exemplos	Número de Exemplos	Número de Atributos
derisi	3733	63
seq	3932	478
pheno	1592	69
gasch2	3788	52
expr	3788	551
church	3764	27
gasch1	3733	173
cellcycle	3766	77
spo	3711	80
eisen	2425	79

foi criado, sendo que cada posição do vetor corresponde a uma categoria contida nos rótulos do conjunto de exemplos hierárquico. Portanto, a k-ésima posição do vetor corresponde ao k-ésimo rótulo e recebe o valor 1 se o exemplo pertencer a esse rótulo, senão recebe o valor 0;

Então cada exemplo foi transformado de hierárquico para não hierárquico considerando apenas o primeiro nível da hierarquia. Por exemplo, no caso da anotação FunCat se um exemplo tem os rótulos 5/1/0/0@9/1/1/0 indica que ele pertence aos rótulos: 5/1/0/0 e 9/1/1/0, isto é, os rótulos são separados por '@' e o nível da hierarquia por '/'. Considerando apenas o primeiro nível o exemplo anterior é considerado pertencente aos rótulos 5 e 9 (o valor 1 é atribuído apenas para esses rótulos). Portanto, só recebem valor 1 os rótulos que pertencem ao exemplo e os demais recebem o valor 0.

Na Tabela 5.1 são apresentadas as características gerais do conjunto de exemplos descritos anteriormente, mostrando o número de exemplos, número de atributos, a densidade e a cardinalidade de cada conjunto de exemplos.

5.1.2 Metodologia Experimental

Os experimentos foram realizados usando a biblioteca Weka. No método proposto, as árvores de decisão foram baseadas no algoritmo J48 [Quinlan, 1993] com configuração *default*, isto é, árvores não binárias, com poda e peso mínimo igual a 2 (número mínimo de exemplos por folha). Avaliou-se o método proposto comparando-o com cinco outros métodos da biblioteca MuLAM: BR, LP, RAkEL e MLkNN. Para os três primeiros métodos foi utilizado o algoritmo

J48 também com configurações *default* para construção dos classificadores e nos dois últimos métodos foram usadas as suas configurações *default*. Além disso, a biblioteca Clus também foi utilizada para comparação, usando como configuração: redução de variância como heurística, árvores não binárias, nenhum método de poda e peso mínimo igual a 2 (número mínimo de exemplos por folha).

Para analisar o desempenho foi utilizada validação cruzada com 10 partições para cada método e cada conjunto de exemplos, computando a métrica Medida-F descritas pela Equação 3.6.

5.1.3 Analise Estatística

A análise estatística dos resultados obtidos é uma importante ferramenta para validação e para a adequada extrapolação dos resultados obtidos para a população estudada. Nesse contexto, existem dois tipos de testes estatísticos: teste paramétricos e teste não paramétrico.

Os teste paramétricos baseiam-se em medidas intervalares da variável dependente (um parâmetro ou característica) e a utilização deste tipo de testes exige que a amostra tenha uma distribuição normal, uma variancia homogênea e os intervalos contínuos e iguais [Rice, 2001]. Já os testes não paramétricos [Graczyk et al., 2010] podem ser utilizados quando testes paramétricos não se aplicam, ou seja quando a distribuição da amostra não é normal ou a variância não é homogênea. Segundo [Demšar, 2006] em aprendizado de máquina a melhor opção é se fazer uso de testes não paramétricos.

Então, para analisar a significância dos resultados foi usado o teste não paramétrico de Friedman [Friedman, 1940], considerando um nível de significância de 5% e Benjamini-Hochberg [Benjamini and Hochberg, 1995] como teste *post-hoc*.

5.1.4 Resultados e Discussão

Nesta seção serão mostrados os resultados da métrica Medida-F para os quatro níveis da hierarquia separadamente, sendo mostrado também o *rank* médio obtido por meio do teste de Friedman com nível de significância de 5%. Os melhores resultados para cada conjunto de dados são mostrados em negrito e o melhor desempenho geral é visto analisando o *rank* médio sendo considerada a melhor versão aquela que obteve menor *rank* médio. Além disso, são mostrados

também os resultados do teste *post-hoc*, no qual o símbolo $\triangle$ ($\blacktriangle$) significa que a variação de uma especifica linha é melhor (significativamente) que a variação de uma especifica coluna, enquanto o símbolo $\triangledown$ ($\blacktriangledown$) significa que a variação de uma especifica linha é pior (significativamente) que a variação de uma especifica coluna.

Resultados no Primeiro Nível da Hierarquia

Considerando o conjunto de dados com 16 rótulos, isto é, só o primeiro nível da hierarquia foi considerado podemos observar que nas Tabelas B.1 e B.2, localizadas no Apêndice B, são mostradas as taxas de acerto das árvores de cada rótulos geradas na primeira etapa das abordagens BR-RTb Pru e BR-RTb Unpr-Pr(1), respectivamente. Podemos observar nas tabelas que as taxas de acerto da maioria das árvores são maiores que 80% para BR-RTb Pru e 70% para BR-RTb Unpr-Pr(1), exceto a árvore do rótulo Localização celular em ambas as tabelas. Portanto, como as árvores criadas para capturar as relações entre os rótulos em média obtiveram uma boa taxa de acerto podemos concluir que as relações obtidas a partir dessas árvores são significativas.

Analisando os resultados da métrica Medida-F que são mostrados na Tabelas 5.2 podemos observar que o método BR-RTb Pru obteve o segundo melhor *rank* médio sendo que não há grande diferença entre o *rank* médio do método RAkEL, que obteve o melhor desempenho. Além disso, podemos ver pelo test *post-hoc* que não há diferença significativa no desempenho entre BR-RTb Pru e RAkEL. Pode-se ser visualizado também no test *post-hoc* que o método BR-RTb Pru obteve desempenho significativamente melhor que os métodos BR-CTb Unpr-Pr(1), LP, Clus Pru e Clus Unpr. Considerando o método BR-RTb Unpr-Pr(1) podemos ver que ele obteve o pior desempenho sendo pior que todos os outros método comparados e sendo significativamente pior que os métodos BR, RAkEL, MLkNN e Clus Pru.

Resultados no Segundo Nível da Hierarquia

Considerando o conjunto de dados com 102 rótulos, isto é, o segundo nível da hierarquia foi considerado podemos observar que nas Tabelas B.3 e B.4, localizadas no Apêndice B, são mostradas as taxas de acerto das árvores de cada rótulos geradas na primeira etapa das aborda-

Tabela 5.2: Medida-F obtidos no experimento - nível 1

Conjunto de exemplos	BR-RTb Pru	BR-RTb Unpr-Pr(1)	BR	LP	RAkEL	MLkNN	Clus pru	Clus unpr
Medida-F								
pheno	0,393	0,368	0,393	0,377	**0,397**	0,377	0,379	0,374
seq	0,457	0,234	0,457	0,413	**0,490**	0,435	0,405	0,408
church	**0,459**	0,318	0,390	0,397	0,390	0,386	0,387	0,374
cellcycle	**0,461**	0,101	0,426	0,371	0,426	0,403	0,388	0,376
derisi	0,400	0,390	**0,417**	0,370	0,385	0,405	0,399	0,375
eisen	0,501	0,313	0,498	0,468	**0,535**	0,515	0,474	0,479
gasch2	**0,461**	0,258	0,424	0,386	0,425	0,429	0,394	0,375
spo	**0,461**	0,362	0,379	0,369	0,396	0,407	0,386	0,375
gasch1	0,427	0,114	0,422	0,407	**0,457**	0,421	0,396	0,397
expr	0,380	0,122	0,429	0,389	**0,467**	0,422	0,397	0,400
Resultado do Teste *post-hoc*								
BR.CT.pru	o	▲	△	▲	▽	△	▲	▲
BR.CTb.unpr.pru1	x	o	▼	▽	▼	▼	▼	▽
BR	x	x	o	▲	▽	△	△	▲
LP	x	x	x	o	▼	▽	▽	△
RAkEL	x	x	x	x	o	△	▲	▲
MLkNN	x	x	x	x	x	o	△	▲
Clus pru	x	x	x	x	x	x	o	△
Clus unpr	x	x	x	x	x	x	x	o
Rank Médio	2,400	7,700	3,000	5,850	**2,300**	3,450	5,200	6,100

gens BR-RTb Pru e BR-RTb Unpr-Pr(1), respectivamente. Observando as tabelas verifica-se que as taxas de acerto da maioria das árvores são maiores que 80% para ambas abordagens, BR-RTb Pru e BR-RTb Unpr-Pr(1). Consequentemente, como as árvores criadas para capturar as relações entre os rótulos em média obtiveram uma boa taxa de acerto podemos concluir que as relações obtidas a partir dessas árvores são significativas.

Examinando os resultados da métrica Medida-F que são mostrados na Tabelas 5.3 podemos observar que os métodos BR-RTb Pru e o BR-RTb Unpr-Pr(1) obtiveram o quinto e o sexto melhor *rank* médio, respectivamente. Porém, podemos observar pelo teste *post-hoc* que o método BR-RTb Pru é melhor que o Clus Pru e considerando que apenas os métodos BR-RTb e Clus produzem modelos que podem ser interpretados pelo homem para compreensão do problema. Neste nível ocorreu do método Clus Pru só obter como resultado árvores folhas classificando todos os rótulos como verdadeiro negativo (tn_{Y_i}) por isso que a medida-F de todas as bases foi igual a zero, exceto o conjunto de dados 'Seq', já que o cálculo dela leva em consideração a precisão (número de tp_{Y_i} por $tp_{Y_i} + fp_{Y_i}$) e a revocação (número de tp_{Y_i} por $tp_{Y_i} + fn_{Y_i}$). Então como as árvores folhas classificam todos os rótulos como verdadeiro negativo significa que qualquer exemplo que for classificado por elas concluirá que não pertence a nenhum rótulo.

Tabela 5.3: Medida-F obtidos no experimento - nível 2

Conjunto de exemplos	BR-RT Pru	BR-RTb Unpr-Pr(1)	BR	LP	RAkEL	MLkNN	Clus pru	Clus unpr
Medida-F								
pheno	0,054	0,021	0,058	**0,094**	0,055	0,044	0,000	0,037
seq	0,070	0,056	0,221	0,160	**0,227**	0,085	0,002	0,157
church	0,049	0,032	0,095	0,098	**0,101**	0,010	0,000	0,108
cellcycle	0,060	0,020	**0,164**	0,131	0,160	0,077	0,000	0,103
derisi	0,052	0,021	0,102	**0,126**	0,114	0,070	0,000	0,109
eisen	0,068	0,034	0,267	0,210	**0,275**	0,179	0,000	0,200
gasch2	0,058	0,026	0,150	**0,155**	0,154	0,095	0,000	0,138
spo	0,056	0,047	**0,116**	0,112	0,115	0,076	0,000	0,102
gasch1	0,058	0,043	0,223	0,171	**0,230**	0,139	0,000	0,151
expr	0,071	0,045	0,221	0,164	**0,231**	0,112	0,000	0,152
Resultado do Teste *post-hoc*								
BR.CT.pru	o	△	▼	▼	▼	▽	△	▽
BR.CTb.unpr.pru1	x	o	▼	▼	▼	▽	△	▼
BR	x	x	o	△	▽	▲	▲	△
LP	x	x	x	o	▽	▲	▲	△
RAkEL	x	x	x	x	o	▲	▲	△
MLkNN	x	x	x	x	x	o	▲	▽
Clus.pru	x	x	x	x	x	x	o	▼
Clus.unpr	x	x	x	x	x	x	x	o
Rank Médio	5,700	6,900	2,300	2,400	**1,700**	5,200	8,000	3,800

Resultados no Terceiro Nível da Hierarquia

Levando em conta o terceiro nível da hierarquia do MIPS, os conjuntos de dados possuem 89 rótulos. Podemos observar nas Tabelas B.5 e B.6, localizadas no Apêndice B, que são mostradas as taxas de acerto das árvores de cada rótulos geradas na primeira etapa das abordagens BR-RTb pru e BR-RTb unpr-pru1, respectivamente, no qual as linhas correspondem aos rótulos e as colunas corresponde as bases de dados. Nas tabelas as taxas de acerto da maioria das árvores são maiores que 95% para BR-RTb Pru e 70% para BR-RTb Unpr-Pr(1). Então podemos concluir que as relações obtidas a partir dessas árvores são significativas, pois as árvores criadas para capturar as relações entre os rótulos em média obtiveram uma boa taxa de acerto.

Examinando os resultados da métrica Medida-F que são mostrados na Tabelas 5.4 podemos observar que os métodos BR-RTb Pru e o BR-RTb Unpr-Pr(1) obtiveram o quarto e o sétimo melhores *rank* médio, respectivamente. Pode-se ser visualizado também no test *post-hoc* que o método BR-RT Pru obteve desempenho significativamente melhor que os métodos BR-RTb Unpr-Pr(1) e Clus Pru. Considerando o método BR-RTb Unpr-Pr(1) podemos ver que ele obteve o segundo pior desempenho comparando com os outros método exceto comparando com Clus Pru no qual foi melhor. Porém, podemos observar pelo teste *post-hoc* que o método BR-RTb Pru é significativamente melhor que o Clus Pru e não é significativamente pior que os outros métodos. Neste nível também ocorreu do método Clus Pru só obter como resultado

Tabela 5.4: Medida-F obtidos no experimento - nível 3

Conjunto de exemplos	BR-RT Pru	BR-RTb Unpr-Pr(1)	BR	LP	RAkEL	MLkNN	Clus pru	Clus unpr
Medida-F								
pheno	**0,044**	0,002	0,013	0,002	0,012	0,004	0,000	0,000
seq	0,032	0,000	**0,136**	0,080	0,108	0,025	0,000	0,081
church	0,027	0,000	0,006	0,013	0,005	0,002	0,000	**0,033**
cellcycle	0,040	0,000	0,080	0,058	**0,073**	0,007	0,000	0,042
derisi	0,030	0,002	0,016	0,049	0,014	0,002	0,000	**0,045**
eisen	0,055	0,000	**0,132**	0,087	0,113	0,040	0,000	0,066
gasch2	0,042	0,000	0,067	**0,070**	0,052	0,064	0,000	0,042
spo	0,035	0,000	0,058	**0,064**	0,048	0,010	0,000	0,062
gasch1	0,036	0,000	**0,120**	0,090	0,092	0,037	0,000	0,054
expr	0,048	0,013	**0,133**	0,086	0,109	0,021	0,000	0,058
Resultado do Teste *post-hoc*								
BR.CT.pru	o	▲	▽	▽	▽	△	▲	▽
BR.CTb.unpr.pru1	x	o	▼	▼	▼	▽	△	▼
BR	x	x	o	△	△	▲	▲	△
LP	x	x	x	o	△	▲	▲	△
RAkEL	x	x	x	x	o	△	▲	△
MLkNN	x	x	x	x	x	o	△	▽
Clus.pru	x	x	x	x	x	x	o	▼
Clus.unpr	x	x	x	x	x	x	x	o
Rank Médio	4.250	7.150	**2.000**	2.750	3.100	5.450	7.600	3.700

árvores folhas classificando todos os rótulos como verdadeiro negativo por isso que a medida-F de todas as bases foi igual a zero.

Resultados no Quarto Nível da Hierarquia

Considerando o quarto nível da hierarquia do MIPS, os conjuntos de dados possuem 42 rótulos. Podemos observar nas Tabelas B.7 e B.8, localizadas no Apêndice B, que são mostrados as taxas de acerto das árvores de cada rótulos geradas na primeira etapa das abordagens BR-RTb Pru e BR-RTb Unpr-Pr(1), respectivamente, no qual as linhas correspondem aos rótulos e as colunas corresponde as bases de dados. Podemos observar nas tabelas que as taxas de acerto de todas das árvores são maiores que 95% para BR-RTb Pru e 70% para BR-RTb Unpr-Pr(1). Logo, como as árvores criadas para capturar as relações entre os rótulos em média obtiveram uma boa taxa de acerto podemos concluir que as relações obtidas a partir dessas árvores são significativas.

Examinando os resultados da métrica Medida-F que são mostrados na Tabelas 5.5 podemos observar que os métodos BR-RTb Pru e o BR-RTb Unpr-Pr(1) obtiveram o terceiro e o quinto melhor *rank* médio, respectivamente. Pode-se ser visualizado também no test *post-hoc* que o método BR-RTb Pru obteve desempenho melhor que os métodos BR-RTb Unpr-Pr(1), RAkEL, MLkNN e Clus Pru sendo significativamente melhor nos últimos. Considerando o método BR-

Tabela 5.5: Medida-F obtidos no experimento - nível 4

Conjunto de exemplos	BR-RT Pru	BR-RTb Unpr-Pr(1)	BR	LP	RAkEL	MLkNN	Clus pru	Clus unpr
Medida-F								
pheno	**0,037**	0,002	0,000	0,000	0,000	0,000	0,000	0,000
seq	0,029	0,000	0,007	0,082	0,083	0,000	0,000	**0,084**
church	0,020	0,000	0,002	0,000	0,000	0,000	0,000	**0,039**
cellcycle	0,044	0,000	**0,075**	0,044	0,044	0,002	0,000	0,051
derisi	0,026	0,002	0,000	0,045	0,000	0,000	0,000	**0,062**
eisen	0,049	0,000	0,098	**0,091**	0,002	0,002	0,000	0,069
gasch2	0,018	0,000	0,000	**0,056**	0,002	0,000	0,000	0,023
spo	0,027	0,024	0,002	0,054	0,002	0,002	0,000	**0,067**
gasch1	0,037	0,004	**0,077**	0,060	0,049	0,000	0,000	0,042
expr	0,037	0,000	0,098	**0,069**	0,065	0,000	0,000	0,056
Resultado do Teste *post-hoc*								
BR.CT.pru	o	△	△	▽	△	▲	▲	▽
BR.CTb.unpr.pru1	x	o	▽	▼	▽	△	△	▼
BR	x	x	o	▽	△	▲	▲	▽
LP	x	x	x	o	△	▲	▲	▽
RAkEL	x	x	x	x	o	△	△	▽
MLkNN	x	x	x	x	x	o	△	▼
Clus.pru	x	x	x	x	x	x	o	▼
Clus.unpr	x	x	x	x	x	x	x	o
Rank Médio	3.400	5.750	3.650	2.950	4.550	6.350	6.900	**2.450**

RTb Unpr-Pr(1) podemos ver que ele obteve desempenho melhor que os métodos MLkNN e Clus Pru. Neste nível ocorreu do método Clus Pru só obter como resultado árvores folhas classificando todos os rótulos como verdadeiro negativo (tn_{Y_i}) por isso que a medida-F de todas as bases foi igual a zero.

5.2 Considerações Finais

Nesse capítulo foram detalhados os conjuntos de dados utilizados nos experimentos preliminares, descreveu-se os experimentos realizados, os resultados obtidos e as discussões. No próximo capítulo é apresentada a conclusão desse trabalho.

Capítulo 6

Conclusão

Neste trabalho um estudo do problema de classificação multirrótulo foi conduzido. Problemas com essas características são comuns em Bioinformática, no qual há mais do que um rótulo para ser predito, isto é, um exemplo pode ser relacionado com mais de um rótulo fazendo a tarefa de classificação mais difícil.

Os métodos de classificação multirrótulo podem ser divididos em dois tipos de abordagens: abordagem independente de algoritmo e abordagem dependente de algoritmo, sendo escolhida a abordagem independente de algoritmo para investigação e desenvolvimento da proposta. A abordagem independente de algoritmo lida com o problema multirrótulo transformando-o em um conjunto de problemas único rótulo. Após essa transformação é realizado a aplicação de algum algoritmo de classificação para assim induzir um ou um conjunto classificadores para predizer todos rótulos de um novo exemplo.

A fim de melhorar desempenho e compreensão do modelo obtido, neste estudo propôs-se uma adaptação para o método Binary Relevance para superar sua desvantagem: considerando as relações entre os rótulos e, portanto, esta pode melhorar a generalização de o modelo induzido e, possivelmente, pode diminuir o número de classificadores a serem analisados por um especialista humano. Quando todos os rótulos estão relacionados, a proposta encontra um único

classificador (árvore de decisão) que pode classificar todos os rótulos.

Com intuito de comparar os resultados do método proposto com outros métodos, experimentos foram realizados com os métodos BR, LP, RAkEL, MLkNN e Clus. Para que os resultados pudessem ser comparados com os obtidos pelo método BR-RTb foram usadas configuração semelhantes para todos os métodos. Os experimentos realizados envolveram dez conjuntos de dados na área da genômica funcional relacionado ao organismo *Saccharomyces cerevisiae*, no qual os rótulos são estruturados hierarquicamente de acordo com o catálogo FunCat desenvolvido pelo MIPS. Os resultados foram avaliados usando a métrica medida-F e todas as comparações entre os desempenhos dos métodos foram analisadas pelo teste estatístico de *Friedam* para avaliar a significância estatísticas dos resultados com nível de significância de 5%. É importante salientar que segundo [Gamberger et al., 2004] classicadores mais simples, como o aqui proposto, podem apresentar um desempenho mais baixo do que classificadores mais complexos (por exemplo, RAkEL e MLkNN, que não é simbólico).

Este documento está organizado da seguinte maneira: na Seção 6.1 é apresentada um resumo dos principais resultados obtidos; na Seção 6.2 são apresentadas as contribuições deste estudo e as publicações geradas; na Seção 6.3 são apresentadas possíveis trabalhos futuros.

6.1 Principais Resultados

Considerando os resultados do primeiro nível da hierarquia, os métodos RAkEL e BR-RTb Pru obtiveram os melhores desempenhos quando analisadas utilizando como métrica a medida-F. Porém, as analises estatísticas mostraram não haver diferença significativa na comparação entre esses dois métodos.

No segundo nível, as variações do método BR-RTb acabaram não obtendo resultados favoráveis em comparação com os métodos BR, LP, RAkEL, MLkNN e Clus Unpr, porém foi melhor que o Clus Pru. Como já mencionado anteriormente dentre os métodos comparados o único que produz como saída um modelo que possa ser interpretado é o método Clus e como as variações do BR-RTb avaliadas são podadas é mais adequado comparar com o Clus Pru (que é podado também). Como a proposta do BR-RTb, além de tentar obter bons desempenhos, é também construir modelos mais facilmente interpretáveis para o homem, então apesar das

variações BR-RTb não apresentarem bons resultados em relação a todos os métodos comparados, ele foi melhor que o Clus que nesse nível, para todas as bases de dados exceto a 'Seq' no qual foi gerada como modelos árvores folha classificando todos os rótulos como 0, isto é, gerando árvores que classificam os exemplos como não sendo pertencentes a nenhum rótulo.

Analisando os resultados do terceiro nível, os resultados foram parecidos com as do segundo nível no qual as variações do método BR-RTb também acabaram não obtendo resultados favoráveis em comparação com os métodos BR, LP, RAkEL, MLkNN e Clus Unpr, porém foi melhor que o Clus Pru sendo que este obteve para todas as bases de dados árvores folhas classificando todos os rótulos como 0.

Finalmente, no quarto nível as variações do método BR-RTb obtiveram melhores resultados comparados ao dos segundo e terceiro níveis obtendo o terceiro e quinto *rank* médio. A variação BR-RTb Pru foi melhor que os métodos BR-RTb Unpr-Pru(1), BR, RAkEL, MLkNN e Clus Pru sendo significativamente melhor que os dois últimos. Já a variação BR-RTb Unpr-Pru(1) só foi melhor que os métodos MLkNN e Clus Pru. Neste nível também o método Clus Pru gerou como modelo árvores folha que classificam os exemplos como não sendo pertencentes a nenhum rótulo.

Com base nos resultados apresentados podemos observar que a variação BR-RTb Pru do método proposto em comparação com o método Clus Pru, em todos os níveis teve melhor resultado apesar de não ter obtido bons resultados comparados com os outros métodos comparados em relação ao segundo e terceiro nível. Uma explicação para esse resultado pode ser o alto número de rótulos no segundo e terceiro níveis e como cada rótulo tem um alto grau de desbalanceamento que pode acabar prejudicando as fases de indução das árvores consequentemente prejudicando o modelo final. Outra observação é que o método Clus Pru do segundo ao quarto nível gera somente árvores folha isto mostra que o Clus podado não conseguiu lidar com o problema do desbalanceamento de classes.

6.2 Contribuições e Publicações

A partir de um trabalho [Tanaka et al., 2010] sobre reprodução assistida, no qual foi desenvolvido um sistema de suporte à decisão foi evidenciado a importância de estudo e desen-

volvimento de novas técnicas para analisar problemas com vários rótulos. Devido a isso, foi realizada um estudo e desenvolvimento de uma nova abordagem para problemas multirrótulos.

Uma das principais contribuições deste estudo foi a proposta da abordagem BR-RT que é uma adaptação do método de *Binary Relevance* usando árvores de decisão para tratar problemas multirrótulos visando a melhora da compreensão do conhecimento extraído, para isso o método BR-RT captura as relações entre rótulos, no qual *Binary Relevance* não leva em conta, e, consequentemente, tentar melhorar a capacidade de generalização do modelo. Além disso, os resultados apresentados pelo método proposto foram promissores.

Outra importante contribuição é o estudo sobre o desbalanceamento de classes em problemas multirrótulos no qual foi proposto uma equação para mensurar o desbalanceamento do conjunto de dados multirrótulo e foi proposta uma forma de tentar amenizar o desbalanceamento que foi analisada e testada utilizando o conceito da abordagem independente de algoritmo, isto é, foram criadas conjuntos de dados com todos os atributos e somente um rótulo e a partir deles feito o balanceamento gerando conjuntos de dados balanceados para cada rótulo que serão usados posteriormente para indução das árvores de decisão.

Além disso, foram feitas contribuições na área de Bioinformática, as quais foram relatadas na publicação [Tanaka and Baranauskas, 2012]. Neste artigo foi relatada a proposta deste estudo e foram realizados experimentos iniciais com o conjunto de dados utilizado neste estudo utilizando apenas o primeiro nível da hierarquia.

6.3 Trabalhos Futuros

Muito pode ser realizado em pesquisas futuras considerando o BR-RT. Em sua versão original, o método só é capaz de trabalhar com conjunto de dados binários. Então, um primeiro trabalho futuro seria a adaptação do algoritmo para aceitar rótulos multi-classe, isto é, aceitar rótulos com com mais de duas classes. Um segundo trabalho futuro é a na fase de extensão das árvores para classificar todos os rótulos, no algoritmo original uma vez descoberto a classe de um rótulo para um determinado ramo ele não é revisto nas demais iterações de descoberta das classes dos demais rótulos possibilitando uma propagação de erros. Por exemplo, se o conjunto de dados tem cinco rótulo todos relacionados entre si, então a saída do método será apenas

uma árvore que classificará todos os cinco rótulos. Então, se na fase de descoberta das classes dos rótulos de determinado ramo for descoberto que o rótulo $Y_2 = 1$ e $Y_5 = 0$, esses rótulos não são mais revistos. Por fim, outro trabalho futuro é a analise sobre o tempo de execução que na versão original é computacionalmente custosa dependendo do conjunto de dados.

Referências Bibliográficas

[dog, 2013] (2013). dogma. http://www.algoritmosgeneticos.com.br/Figura02.02.jpg. acessado em 06/05/2013. Citado na página 6.

[Pro, 2013] (2013). estruturas proteica. http://www.enq.ufsc.br/labs/probio/disc_eng_bioq/trabalhos_pos2003/const_microorg/proteinas.htm. acessado em 06/05/2013. Citado na página 6.

[Gen, 2013] (2013). gene. http://www.accessexcellence.org/RC/VL/GG/genes.php. acessado em 06/05/2013. Citado na página 7.

[Adams M.D., 1995] Adams M.D., Kerlavage A.R., F. R. F. R. B. C. L. N. K. E. W. K. G. J. W. O. e. a. (1995). Initial assessment of human gene diversity and expression patterns based upon 83 million nucleotides of cdna sequence. *Nature*, pages 3–174. Citado na página 9.

[Alves et al., 2008] Alves, R. T., Delgado, M. R., and Freitas, A. A. (2008). Multi-label hierarchical classification of protein functions with artificial immune systems. *Advances in Bioinformatics and Computational Biology*, pages 1–12. Citado nas páginas 21 and 38.

[Aly, 2005] Aly, M. (2005). Survey on multiclass classification methods. Citado na página 17.

[Barutcuoglu et al., 2006] Barutcuoglu, Z., Schapire, R. E., and Troyanskaya, O. G. (2006). Hierarchical multi-label prediction of gene function. *Bioinformatics*. Citado na página 3.

[Benjamini and Hochberg, 1995] Benjamini, Y. and 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. Citado na página 42.

[Blockeel et al., 1998] Blockeel, H., Raedt, L. D., and Ramon, J. (1998). Top-down induction of clustering trees. In *Proceedings of the 15th International Conference on Machine Learning*, ICML '98, pages 55–63. Citado nas páginas 22 and 38.

[Blockeel et al., 2006] Blockeel, H., Schietgat, L., Struyf, J., Clare, A., and Dzeroski, S. (2006). Hierarchical multilabel classification trees for gene function prediction. Citado nas páginas 9, 23, and 38.

[Breiman et al., 1984] Breiman, L., Friedman, J., Stone, C. J., and Olshen, R. (1984). *Classification and Regression Trees*. Wadsworth & Books, Pacific Grove, CA. Citado nas páginas 14 and 23.

[Brenner et al., 2000] Brenner, S., Johnson, M., Bridgham, J., Golda, G., Lloyd, D. H., Johnson, D., Luo, S., McCurdy, S., Foy, M., Ewan, M., Roth, R., George, D., Eletr, S., Albrecht, G., Vermaas, E., Williams, S. R., Moon, K., Burcham, T., Pallas, M., DuBridge, R. B., Kirchner, J., Fearon, K., Mao, J., and Corcoran, K. (2000). Gene expression analysis by massively parallel signature sequencing (MPSS) on microbead arrays. *Nature biotechnology*, 18(6):630–634. Citado na página 9.

[Carneiro et al., 2000] Carneiro, N. P., Carneiro, A. A., Guimarães, C. T., and Paiva, E. (2000). Desvendando o código genético. *Biotecnologia Ciência & Desenvolvimento*, (17):50–58. Citado nas páginas 7 and 9.

[Carraro and Kitajima, 2002] Carraro, D. M. and Kitajima, J. P. (2002). Sequenciamento e bioinformática de genomas bacterianos. *Biotecnologia Ciência e Desenvolvimento*, (28):16–20. Citado na página 8.

[Cerri et al., 2009] Cerri, R., da Silva, R., and de Carvalho, A. (2009). Comparing methods for multilabel classification of proteins using machine learning techniques. pages 109–120. Springer. Citado na página 18.

[Cherman et al., 2010] Cherman, E. A., Metz, J., and Monard, M. C. (2010). Métodos multirrótulo independentes de algoritmo: um estudo de caso. In *Anais da XXXVI Conferencia Latinoamericana de Informática (CLEI)*, pages 1–14, Asuncion, Paraguay. Publicado em CD-ROM. Citado nas páginas 18 and 38.

[Chu et al., 1998] Chu, S., DeRisi, J., Eisen, M., Mulholland, J., Botstein, D., Brown, P., and Herskowitz, I. (1998). The transcriptional program of sporulation in budding yeast. *Science*, 282(5389):699. Citado na página 40.

[Chua et al.,] Chua, H. N., Sung, W.-K., and Wong, L. Exploiting indirect neighbours and topological weight to predict protein function from protein–protein interactions. *Bioinformatics*, 22(13):1623–1630. Citado na página 9.

[Clare, 2003] Clare, A. (2003). *Machine learning and data mining for yeast functional genomics*. PhD thesis, The University of Wales. Citado nas páginas 5, 6, and 40.

[Clare and King, 2001] Clare, A. and King, R. D. (2001). Knowledge discovery in multi-label phenotype data. *Lecture Notes in Computer Science*, pages 42–53. Citado nas páginas 3, 21, and 38.

[Comité et al., 2001] Comité, F. D., Gilleron, R., and Tommasi, M. (2001). Learning Multi-label Alternating Decision Trees and Applications. In *Proceedings of CAP'01 : Conférence en Apprentissage Automatique*, pages 195–210. Citado na página 22.

[Demšar, 2006] Demšar, J. (2006). Statistical comparison of classifiers over multiple data sets. *Journal of Machine Learning Research*, 7(1):1–30. Citado na página 42.

[Domingos, 1997] Domingos, P. (1997). Knowledge acquisition from examples via multiple models. In *In Proceedings of the Fourteenth International Conference on Machine Learning*, pages 98–106. Morgan Kaufmann. Citado na página 26.

[Domingos, 1999] Domingos, P. (1999). The role of occam's razor in knowledge discovery. *Data Mining and Knowledge Discovery*, 3(4):409–425. Citado na página 26.

[Dyego Carlos Sales de Morais, 2012] Dyego Carlos Sales de Morais, Bruno Carlos Sales de Morais, J. V. d. M. J. C. M. G. d. G. (2012). Sistema móvel de apoio a decisão médica aplicado ao diagnóstico de asma intelimed. *VIII Simpósio Brasileiro de Sistemas de Informação*, pages 528–539. Citado na página 1.

[Eisen et al., 1998] Eisen, M., Spellman, P., Brown, P., and Botstein, D. (1998). Cluster analysis and display of genome-wide expression patterns. *Proceedings of the National Academy of Sciences*, 95(25):14863. Citado na página 40.

[Estabrooks et al., 2004] Estabrooks, A., Jo, T., and Japkowicz, N. (2004). A multiple resampling method for learning from imbalanced data sets. *Computational Intelligence*, 20(1):18–36. Citado na página 23.

[Freund and Mason, 1999] Freund, Y. and Mason, L. (1999). The alternating decision tree learning algorithm,. In *Proc. 16th International Conf. on Machine Learning*, pages 124–133. Morgan Kaufmann, San Francisco, CA. Citado na página 22.

[Friedman, 1940] 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. Citado nas páginas 42 and 65.

[Gamberger et al., 2004] Gamberger, D., Lavrač, N., Zelezny, F., and Tolar, J. (2004). Induction of comprehensible models for gene expression datasets by subgroup discovery methodology. *Journal of Biomedical Informatics*, 37:269–284. Citado na página 49.

[Garcia V., 2007] Garcia V., Sánchez J.S., M. R. A. R. S. J. (2007). The class imbalance problem in pattern classification and learning. *Pattern Analysis and Learning Group*, pages 283–291. Citado na página 23.

[Gasch et al., 2001] Gasch, A., Huang, M., Metzner, S., Botstein, D., Elledge, S., and Brown, P. (2001). Genomic expression responses to dna-damaging agents and the regulatory role of the yeast atr homolog mec1p. *Molecular biology of the cell*, 12(10):2987–3003. Citado na página 40.

[Gasch et al., 2000] Gasch, A., Spellman, P., Kao, C., Carmel-Harel, O., Eisen, M., Storz, G., Botstein, D., and Brown, P. (2000). Genomic expression programs in the response of yeast cells to environmental changes. *Molecular biology of the cell*, 11(12):4241–4257. Citado na página 40.

[Gasteiger et al., 2005] Gasteiger, E., Hoogland, C., Gattiker, A., Duvaud, S., Wilkins, M., Appel, R., and Bairoch, A. (2005). Protein identification and analysis tools on the expasy server. In *The Proteomics Protocols Handbook, Humana Press*, pages 571–607. Citado na página 40.

[Graczyk et al., 2010] Graczyk, M., Lasota, T., Telec, Z., and Trawiński, B. (2010). Nonparametric statistical analysis of machine learning algorithms for regression problems. In *Proceedings of the 14th international conference on Knowledge-based and intelligent information and engineering systems: Part I*, pages 111–120. Citado na página 42.

[Junior et al., 2004] Junior, T. C., augusto Benedito, V., and de Oliveira Figueira, A. V. (2004). Análise serial da expressão gênica. *Biotecnologia Ciência & Desenvolvimento*, (33):88–100. Citado na página 9.

[Kumar et al., 2000] Kumar, A., Cheung, K., Ross-Macdonald, P., Coelho, P., Miller, P., and Snyder, M. (2000). Triples: a database of gene function in saccharomyces cerevisiae. *Nucleic Acids Research*, 28(1):81–84. Citado na página 40.

[Laurikkala, 2001] Laurikkala, J. (2001). Improving identification of difficult small classes by balancing class distribution. In *Proceedings of the 8th Conference on AI in Medicine in Europe: Artificial Intelligence Medicine*, AIME '01, pages 63–66. Citado na página 23.

[Lee and Oh, 2003] Lee, J.-S. and Oh, I.-S. (2003). Binary classification trees for multi-class classification problems. In *Proceedings of the Seventh International Conference on Document Analysis and Recognition - Volume 2*, pages 770–. IEEE Computer Society. Citado na página 17.

[Marcotte EM, 1999] Marcotte EM, Pellegrini M, T. M. Y. T. E. D. (1999). A combined algorithm for genome-wide prediction of protein function. *Nature*, (402):83–86. Citado na página 9.

[Maskos and Southern, 1992] Maskos, U. and 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 Res*, 20(7):1679–84.+. Citado na página 9.

[Mewes et al., 2004] Mewes, H. W., Frishman, D., Mayer, K. F. X., Münsterkötter, M., Noubibou, O., Rattei, T., Oesterheld, M., and Stümpflen, V. (2004). Mips: Analysis and annotation of proteins from whole genomes. *Nucleic Acids Res*, 32:41–44. Citado na página 39.

[Michalski, 1983] Michalski, R. S. (1983). A theory and methodology of inductive learning. *Artificial Intelligence*, 20:111–161. Citado na página 12.

[Monard and Baranauskas, 2003] Monard, M. C. and Baranauskas, J. A. (2003). *Conceitos sobre Aprendizado de Máquina*, chapter 4, pages 89–114. In [Rezende, 2003]. Citado na página 12.

[Neto, 1997] Neto, E. D. (1997). Projeto genoma de parasitas. *Biotecnologia Ciência e Desenvolvimento*, (2):18–21. Citado na página 8.

[Oliver et al., 1996] Oliver, S. et al. (1996). A network approach to the systematic analysis of yeast gene function. *Trends in genetics: TIG*, 12(7):241. Citado na página 40.

[Orriols and Bernadó-Mansilla, 2005] Orriols, A. and Bernadó-Mansilla, E. (2005). The class imbalance problem in learning classifier systems: a preliminary study. In *Proceedings of the 2005 workshops on Genetic and evolutionary computation*, GECCO '05, pages 74–78. ACM. Citado na página 23.

[Özgür et al., 2005] Özgür, A., Özgür, L., and Güngör, T. (2005). Text categorization with class-based and corpus-based keyword selection. In *Proceedings of the 20th international conference on Computer and Information Sciences*, ISCIS'05, pages 606–615. Citado na página 27.

[Pollettini, 2012] Pollettini, Juliana T., P. S. R. G. D. J. C. T. R. B. J. A. M. A. A. (2012). Using machine learning classifiers to assist healthcare-related decisions: Classification of electronic patient records. *Journal of Medical Systems*, pages 3861–3874. Citado na página 1.

[Quinlan, 1986] 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. Citado na página 13.

[Quinlan, 1993] Quinlan, J. R. (1993). *C4.5: Programs for Machine Learning*. Morgan Kaufmann. San Francisco, CA. Citado nas páginas 14, 21, 23, 41, 63, and 65.

[Rezende, 2003] Rezende, S. O., editor (2003). *Sistemas Inteligentes: Fundamentos e Aplicações*. Manole. Citado nas páginas 13, 32, and 57.

[Rice, 2001] Rice, J. A. (2001). *Mathematical Statistics and Data Analysis*. Duxbury Press, 3 edition. Citado na página 42.

[Roth1JT et al., 1998] Roth1JT, F., Hughes, J., Estep, P., and Church, G. (1998). Finding dna regulatory motifs within unaligned noncoding sequences clustered by whole-genome mrna quantitation. *Nature biotechnology*, 16:939. Citado na página 40.

[Ruepp et al., 2004] Ruepp, A., Zollner, A., Maier, D., Albermann, K., Hani, J., Mokrejs, M., Tetko, I., Güldener, U., Mannhaupt, G., Münsterkötter, M., and Mewes, H. W. (2004). The funcat, a functional annotation scheme for systematic classification of proteins from whole genomes. *Nucleic Acids Research*, 32(18):5539–5545. Citado na página 3.

[Ryan D. Morin and Marra., 2008] Ryan D. Morin, Matthew Bainbridge, A. F. M. H. M. K. T. J. P. H. M. R. V. S. J. J. and Marra., M. A. (2008). Profiling the hela s3 transciptome using randomly primed cdna and massively parallel short-read sequencing. *BioTechniques*, pages 81–94. Citado na página 9.

[Schapire and Singer, 1999] Schapire, R. E. and Singer, Y. (1999). Improved boosting algorithms using confidence-rated predictions. *Mach. Learn.*, 37(3):297–336. Citado na página 22.

[Schapire and Singer, 2000] Schapire, R. E. and Singer, Y. (2000). Boostexter: A boosting-based system for text categorization. In *Machine Learning*, pages 135–168. Citado na página 26.

[Schietgat et al., 2010] Schietgat, L., Vens, C., Struyf, J., Blockeel, H., Kocev, D., and Dzeroski, S. (2010). Predicting gene function using hierarchical multi-label decision tree ensembles. *BMC Bioinformatics*, 11(1):2+. Citado nas páginas 2 and 3.

[Sharan R, 2007] Sharan R, Ulitsky I, S. R. (2007). Network-based prediction of protein function. *Mol Syst Biol*, (3):88. Citado na página 9.

[Shen et al., 2004] Shen, X., Boutell, M., Luo, J., and Brown, C. (2004). Multi-label machine learning and its application to semantic scene cassification. Citado nas páginas 15 and 20.

[Spellman et al., 1998] Spellman, P., Sherlock, G., Zhang, M., Iyer, V., Anders, K., Eisen, M., Brown, P., Botstein, D., and Futcher, B. (1998). Comprehensive identification of cell cycle–regulated genes of the yeast saccharomyces cerevisiae by microarray hybridization. *Molecular biology of the cell*, 9(12):3273–3297. Citado na página 40.

[Suzuki et al., 2001] Suzuki, E., Gotoh, M., and Choki, Y. (2001). Bloomy decision tree for multi-objective classification. pages 436–447. Citado na página 3.

[Tahir et al., 2009] Tahir, M. A., Kittler, J., Mikolajczyk, K., and Yan, F. (2009). A multiple expert approach to the class imbalance problem using inverse random under sampling. In *Proceedings of the 8th International Workshop on Multiple Classifier Systems*, MCS '09, pages 82–91. Citado na página 23.

[Tan et al., 2006] Tan, P., Steinbach, M., Kumar, V., et al. (2006). *Introduction to data mining*. Pearson Addison Wesley Boston. Citado na página 26.

[Tanaka and Baranauskas, 2012] Tanaka, E. A. and Baranauskas, J. A. (2012). An adaptation of binary relevance for multi-label classification applied to functional genomics. In *Proceedings of the XXXII Congress of the Brazilian Computer Society*, page 10p. Citado na página 51.

[Tanaka et al., 2010] Tanaka, E. A., Junta, C. M., Vagnini, L., Baranauskas, J. A., and Giuliatti, S. (2010). Um sistema computacional integrando suporte à decisão na Área de reprodução humana. *Anuais do XII Congresso Brasileiro de Informática em Saúde*, page 6p. Citado nas páginas 1 and 50.

[Taşan et al., 2008] Taşan, M., Tian, W., Hill, D., Gibbons, F., Blake, J., Roth, F., et al. (2008). An en masse phenotype and function prediction system for mus musculus. *Genome biology*, 9(Suppl 1):S8. Citado na página 9.

[Tsoumakas and Katakis, 2007] Tsoumakas, G. and Katakis, I. (2007). Multi-label classification: An overview. *Int J Data Warehousing and Mining*, pages 1–13. Citado nas páginas 11, 17, and 26.

[Tsoumakas et al., 2010] Tsoumakas, G., Katakis, I., and Vlahavas, I. (2010). Mining multi-label data. pages 667–685. Springer. Citado nas páginas 17, 19, and 27.

[Tsoumakas et al., 2011] Tsoumakas, G., Spyromitros-Xioufis, E., Vilcek, J., and Vlahavas, I. (2011). Mulan: A java library for multi-label learning. *Journal of Machine Learning Research*, 12:2411–2414. Citado na página 22.

[Tsoumakas and Vlahavas, 2007] Tsoumakas, G. and Vlahavas, I. (2007). Random k-labelsets: An ensemble method for multilabel classification. *Machine Learning: ECML 2007*, 4701:406–417. Citado na página 19.

[Vallin, 2010] Vallin, R. M. M. (2010). Sistemas classificadores evolutivos para problemas multirrótulo. Tese de mestrado, Univerdade de São Paulo, Instituto de Ciências Matemáticas e de Computação. Citado na página 26.

[Vens et al., 2008] Vens, C., Struyf, J., Schietgat, L., Džeroski, S., and Blockeel, H. (2008). Decision trees for hierarchical multi-label classification. *Machine Learning*, 73(2):185–214. Citado na página 40.

[Wang et al., 2009] Wang, Z., Gerstein, M., and Snyder, M. (2009). RNA-Seq: a revolutionary tool for transcriptomics. *Nature reviews. Genetics*, 10(1):57–63. Citado na página 9.

[Witten and Frank, 2005] Witten, I. H. and Frank, E. (2005). *Data Mining: Practical Machine Learning Tools and Techniques*. Morgan Kaufmann Series in Data Management Systems. Morgan Kaufmann, second edition. Citado nas páginas 1 and 22.

[Zhang, 2006] Zhang, M. L. (2006). Multilabel Neural Networks with Applications to Functional Genomics and Text Categorization. *IEEE Transactions on Knowledge and Data Engineering*, 18(10):1338–1351. Citado na página 22.

[Zhang and Zhou, 2007] Zhang, M. L. and Zhou, Z. H. (2007). Ml-knn: A lazy learning approach to multi-label learning. *Pattern Recogn.*, 40(7):2038–2048. Citado na página 22.

Apêndice

A

Testes Preliminares

Neste apêndice apresentam-se os testes preliminares para verificação e seleção das melhores abordagens utilizada pelo algoritmo proposto.

A.1 Teste BR-RTa x BR-RTb

Nessa seção apresenta-se a ideia original da metodologia proposta na qual há uma pequena alteração na ordem das etapas do Algoritmo 2. No decorrer da pesquisa e do desenvolvimento da metodologia primeiramente decidiu-se selecionar uma árvore por componente, sendo aquela com melhor acurácia e depois estender somente aquelas selecionadas; porém um teste preliminar realizado com um conjunto de exemplos simples mostrou que não necessariamente a árvore selecionada obterá o melhor HammingLoss se todas as árvores fossem estendidas antes de selecionar. Portanto, para analisar essas duas possibilidades estatisticamente foi realizado um experimento para comparar o desempenho da versão BR-RTa no qual é feita a seleção primeiro e somente são estendidas as árvores selecionadas, visando melhorar o tempo de processamento sem comprometer o desempenho e da versão BR-RTb (igual ao Algoritmo 2) para o qual a seleção é feita após a extensão. As alterações da metodologia para a versão BR-RTa pode ser vista no Algoritmo 3, no qual na segunda etapa é realizada a seleção da melhor árvore por componente (Linha 12). Na terceira etapa é realizada a extensão das árvores previamente selecionadas, para isso foram realizados três alterações no algoritmo. A primeira mudança foi o laço (Linha 11) que na versão BT-CTb é de 1 a c alterando-o para de 1 a $C(G)$ e a segunda mudança foi a adição de uma linha no algoritmo que pegasse a árvore selecionada do componente conexo (Linha 15) e a ultima alteração foi na indução das árvores T que considera todos os atributos X e apenas um rótulo Y de cada vez. Na versão BR-RTb eram induzidas

c árvores T, porém na versão BR-RTa serão induzidas apenas $C(G)$ árvores, isto é, somente serão induzidas as árvores T_h, na qual h é a árvore A_h selecionada na etapa anterior.

Algoritmo 3 Binary Relevance with Relation Labels - BR-CLa

Require: conjunto de exemplos multi-label D contendo m atributos $X_1, \ldots, X_m$ e c rótulos $Y_1, \ldots, Y_c$
Ensure: ArvoresEstendidas

1: $G \leftarrow \emptyset$
2: $Estendido \leftarrow \emptyset$
3: **for** $i \leftarrow 1$ **to** c **do**
4: $\quad A_i \leftarrow$ InduçaoAD(D_l^i)
5: $\quad$ **for** $w \leftarrow 1$ **to** c **do**
6: $\quad\quad$ **if** $Y_w \subset A_i$ **then**
7: $\quad\quad\quad G \leftarrow G \cup \{(Y_i, Y_w)\}$
8: $\quad\quad$ **end if**
9: $\quad$ **end for**
10: **end for**
11: **for** $j \leftarrow 1$ **to** $C(G)$ **do**
12: $\quad$ ArvoresSelecionada $\leftarrow$ SelecionaMelhorArvore(j)
13: **end for**
14: **for** $i \leftarrow 1$ **to** $C(G)$ **do**
15: $\quad h \leftarrow$ ArvoresSelecionada(i)
16: $\quad T_h \leftarrow$ InduçaoAD(D_a^h)
17: $\quad S \leftarrow A_h$
18: $\quad T_i' \leftarrow T_h$
19: $\quad$ **loop**
20: $\quad\quad SR \leftarrow$ SelecionaTodasRegras(S), nas quais $E_j^{T_h'} = E_k^S$
21: $\quad\quad R^{T_h'} \leftarrow$ ConstroiRegras(SR), na forma $B_j^{T_h'} \wedge B_k^S \rightarrow E_j^{T_h'}$
22: $\quad\quad L(R^{T_h'}) \leftarrow$ calcula laplace de $R^{T_h'}$
23: $\quad\quad \Omega \leftarrow$ seleciona a regra com maior $L(R^{T_h'})$
24: $\quad\quad$ Estendido $\leftarrow$ Estendido $\cup\{Y_1, Y_{i-1}, Y_{i+1}, \ldots, Y_c\} \cap \Omega$
25: $\quad\quad T_h' \leftarrow T_h' \cup$ Estendido
26: $\quad\quad$ **if** Há rótulos a serem considerados em $\{Y_1, \ldots, Y_{i-1}, Y_{i+1}, \ldots, Y_c\}$ **then**
27: $\quad\quad\quad SL \leftarrow$ seleciona um rótulo y considerando a melhor acurácia de A_y de Estendido
28: $\quad\quad\quad S \leftarrow A_{SL}$
29: $\quad\quad$ **else**
30: $\quad\quad\quad$ exit loop
31: $\quad\quad$ **end if**
32: $\quad$ **end loop**
33: **end for**
34: **return** ArvoresEstendidas

Para a realização desse teste, utilizou-se 10 *datasets* hierárquicos relacionados ao organismo *Saccharomyces cerevisiae* disponíbilizados pelo Universidade Católica de Leuven[1]. Nesse experimento foram considerados 18 rótulos do primeiro nível da hierarquia. Devido ao conjunto de dados ser hierárquico foi necessário realizar um pré-processamento para transformá-los em dados não hierárquicos usando um vetor binário no qual cada posição do vetor corresponde ao primeiro nível da hierarquia contida nos rótulos do conjunto de exemplos hierárquico.

[1] http://dtai.cs.kuleuven.be/clus/hmc-ens/

Nas duas versões da metodologia BR-RT as árvores de decisão foram baseadas no algoritmo J48 [Quinlan, 1993] com configuração *default*, isto é, número mínimo de objetos na folha igual a 2, taxa de poda igual a 0,25 e árvores não binárias. Para analisar o desempenho foi utilizada validação cruzada com 10 partições e para analisar a significância dos resultados foi usado o teste de Friedman, considerando um nível de significância de 5% com hipótese nula de que não existe diferença entre os resultados obtidos pelas versões comparadas. Se a hipótese nula for rejeitada, isto é, há diferença entre os resultados obtidos pelas versões e então é aplicado o teste *post-hoc* de Benjamini-Hochberg.

Os resultados da métrica HammingLoss, Acurácia, Número de Nós, Precisão, Revocação e da Medida-F são mostrados nas Tabelas A.1 e A.2, também é mostrado o *rank* médio obtidos a partir do teste de Friedman que resultou em rejeição da hipótese nula para ambas as métricas demonstrando que há diferença estatística entre as duas versões. Os melhores resultados para cada conjunto de dados são mostrados em negrito, porém não indica que a versão obteve o melhor desempenho geral. O melhor desempenho geral é visto analisando o *rank* médio sendo considerada a melhor versão aquela que obteve menor *rank* médio. Então, podemos observar que em todas as métricas a versão BR-RTb obteve melhor *rank* médio.

Tabela A.1: Resultados das medidas HammingLoss, Acurácia e Número de Nós obtidos no teste BR-RTa x BR-RTb

	HammingLoss		Acurácia		Número de Nós	
Conjunto de exemplos	BR-RTb	BR-RTa	BR-RTb	BR-RTa	BR-RTb	BR-RTa
pheno	0,101	0,159	0,899	0,81	4	4
seq	0,107	0,127	0,893	0,887	16	135
church	0,112	0,112	0,888	0,868	1	26
cellcycle	0,115	0,174	0,885	0,851	1	141
derisi	0,098	0,104	0,902	0,883	1	10
eisen	0,119	0,187	0,881	0,815	1	60
gasch2	0,113	0,144	0,887	0,853	1	109
spo	0,112	0,118	0,887	0,871	117	131
gasch1	0,113	0,138	0,887	0,855	1	143
expr	0,112	0,129	0,889	0,844	4	103
Rank Médio	**1.050**	1.950	**1.000**	2.000	**1.050**	1.950

O resultado do teste *post-hoc* é mostrado na Tabela A.3, considerando as métricas HammingLoss, Acurácia, Precisão, Revocação e Medida-F, em que cada símbolo $\triangle$ ($\blacktriangle$) significa que BR-RTa é melhor (significativamente) em relação ao BR-RTb enquanto que o símbolo $\triangledown$ ($\blacktriangledown$) significa que BR-RTa é pior (significativamente) em relação ao BR-RTb.

Como pode ser observado, em ambas as métricas o BR-RTb foi melhor que o BR-RTa, sendo que para a métrica HammingLoss, Acurácia e Precisão a versão BR-RTa foi significativamente pior que BR-RTb.

Então, devido a esses resultados preliminares foi decidido usar a versão BR-RTb nos exper-

Tabela A.2: Resultados das medidas Precisão, Revocação e Medida-F obtidos no teste BR-RTa x BR-RTb

	Precisão		Revocação		Medida-F	
Conjunto de exemplos	BR-RTb	BR-RTa	BR-RTb	BR-RTa	BR-RTb	BR-RTa
pheno	0,576	0,063	0,3	0,322	0,394	0,106
seq	0,533	0,466	0,502	0,469	0,517	0,467
church	0,547	0,639	0,308	0,289	0,394	0,398
cellcycle	0,481	0,196	0,45	0,129	0,465	0,155
derisi	0,585	0,503	0,306	0,259	0,401	0,342
eisen	0,482	0,314	0,425	0,329	0,452	0,321
gasch2	0,464	0,159	0,458	0,447	0,461	0,234
spo	0,533	0,457	0,31	0,43	0,392	0,443
gasch1	0,465	0,149	0,458	0,101	0,462	0,120
expr	0,461	0,215	0,453	0,139	0,457	0,169
Rank Médio	**1.100**	1.900	**1.200**	1.800	**1.200**	1.800

Tabela A.3: Benjamini-Hochberg *post-hoc* Test BR-RTa x BR-RTb

	HammingLoss	Acurácia	Número de Nós	Precisão	Revocação	Medida-F
BR-RTb	▼	▼	▼	▼	▽	▽

imentos.

A.2 Teste de Balanceamento

Este teste foi realizado para selecionar as 2 melhores variações da metodologia BR-RT considerando a poda e o balanceamento das classes definidos como:

- Pru: variação na qual as ADs são induzidas sem balanceamento e são podadas nas etapas 1 e 3;
- Pru(13): variação na qual as ADs são induzidas são são podadas nas etapas 1 e 3 e balanceadas em ambas as etapas;
- Pru(1): variação na qual as ADs são induzidas são são podadas nas etapas 1 e 3 e somente balanceadas na etapa 1;
- Pru(3): variação na qual as ADs são induzidas são são podadas nas etapas 1 e 3 e somente balanceadas na etapa 3;
- Unpr: variação na qual as ADs são induzidas sem balanceamento e não são podadas nas etapas 1 e 3;

- Unpr(13): variação na qual as ADs são induzidas não são podadas nas etapas 1 e 3 e balanceadas em ambas as etapas;
- Unpr(1): variação na qual as ADs são induzidas não são podadas nas etapas 1 e 3 e somente balanceadas na etapa 1;
- Unpr(3): variação na qual as ADs são induzidas não são podadas nas etapas 1 e 3 e somente balanceadas na etapa 3;
- Unpr-Pru: variação na qual as ADs são induzidas sem balanceamento e não são podadas na etapa 1 e são podadas na etapa 3;
- Unpr-Pru(13): variação na qual as ADs são induzidas não são podadas na etapa 1 e são podadas na etapa 3 sendo balanceadas em ambas as etapas;
- Unpr-Prur(1): variação na qual as ADs são induzidas não são podadas na etapa 1 e são podadas na etapa 3 sendo balanceadas na etapa 1;
- Unpr-Pru(3): variação na qual as ADs são induzidas não são podadas na etapa 1 e são podadas na etapa 3 sendo balanceadas na etapa 3;

Para a realização desse teste, utilizou-se 20 *datasets* sendo que 16 foram obtidos do Repositório UCI[2] e 4 do Mulan[3]. Primeiramente, foi necessário realizar uma preprocessamento nos *datasets* obtidos do UCI para transforma-los em problemas multirrótulo, pois eles são problemas de único-rótulo multi-classe, isto é, só há um rótulo a ser predito com K classes ($K > 2$). Nesse caso, foi criado um vetor binário no qual cada posição K do vetor representa a classe K.

Na Tabela A.4 são apresentadas as características gerais do conjunto de exemplos, mostrando o fator de balanceamento, o número de exemplos, número de rótulos, número de atributos, número de instâncias e a fonte de onde foi obtida de cada conjunto de exemplos.

Os experimentos reportados foram realizados usando a biblioteca Weka no qual as árvores de decisão induzidas foram baseadas no algoritmo J48 [Quinlan, 1993] com número mínimo de objetos igual a 2, árvores não binárias, a poda e balanceamento de classes foram variadas.

Para analisar o desempenho foi utilizada validação cruzada com 10 partições para cada método e cada conjunto de exemplos, computando as métricas *HammingLoss*, Acurácia, Pecisão e Revocação descritas previamente. Para analisar a significância dos resultados foi também utilizado o teste de Friedman [Friedman, 1940], considerando um nível de significância de 5% com hipótese nula de que não existe diferença entre os resultados obtidos pelas versões comparadas. Se a hipótese nula for rejeitada, isto é, há diferença entre os resultados obtidos pelas versões e então é aplicado o teste *post-hoc* de Benjamini-Hochberg.

Os resultados das medidas HammingLoss, Acurácia, Precisão e Revocação são mostrados na Tabelas A.9, A.6, A.7 e A.10, respectivamente. É mostrado também o *rank* médio obtido

[2]http://archive.ics.uci.edu/ml/datasets.html
[3]http://mulan.sourceforge.net/datasets.html

Tabela A.4: Características gerais do conjunto de exemplos

Datasets	Balanceamento	Número de Rótulos	Número de Atributos	Número de Instâncias	Fonte
allhyper	0.042	5	29	3772	UCI
medical	0.097	43	1451	978	Mulan
allhypo	0.145	4	29	3772	UCI
genbase	0.164	27	1186	662	Mulan
ann-thyroid	0.186	3	9	7200	UCI
lympoma	0.322	9	4026	96	UCI
ecoli	0.362	8	7	336	UCI
glass	0.419	7	10	214	UCI
lymph	0.530	4	18	148	UCI
dermatology	0.531	6	36	336	UCI
postoperative	0.556	3	8	90	UCI
breast	0.548	6	9	106	UCI
scene	0.585	6	294	2407	Mulan
splice	0.820	3	60	3190	UCI
emotions	0.841	6	72	593	Mulan
lung cancer	0.860	3	56	32	UCI
contraceptive	0.861	3	9	1473	UCI
wine	0.873	3	13	177	UCI
iris	0.885	3	4	150	UCI

através do teste de Friedman com nível de significância de 5%. Os melhores resultados para cada conjunto de dados também são mostrados em negrito e o melhor desempenho geral é visto analisando o *rank* médio sendo considerada a melhor versão aquela que obteve menor *rank* médio. Além disso, é mostrado também os resultados do teste *post-hoc*, no qual o símbolo $\triangle$ ($\blacktriangle$) significa que a variação de uma especifica linha é melhor (significativamente) que a variação de uma especifica coluna, enquanto o símbolo $\triangledown$ ($\blacktriangledown$) significa que a variação de uma especifica linha é pior (significativamente) que a variação de uma especifica coluna.

Analisando os resultados em relação à métrica HammingLoss podemos observar que o Unpr e Unpr-Pr foram as duas versões que obtiveram melhor desempenho, porém não sendo significativamente melhor que as outras variações. Comparando as variações não balanceadas com as balanceadas podemos ver que as variações Pru, Unpr e Unpr-Pru são melhores que suas variações balanceadas, porém comparando somente as variações balanceadas podemos ver que as variações com balanceamento somente na primeira etapa foram melhor que as outras versões de balanceamento. Esse resultado pode ser explicado pois na métrica *HammingLoss* um erro em um único rótulo é punido como quase como um erro em todos os rótulos, não discrimina bem entre "quase correta"e completamente errado. Então,

Em relação as métricas Acurácia as duas versões que obtiveram melhor desempenho foram Pru e Unpr, porém não sendo significativamente melhor que as outras variações. Comparando as variações não balanceadas com as balanceadas podemos ver que as variações Pru e Unpr são melhores que suas variações balanceadas e que a variação Unpr-Pru obteve o mesmo *rank* médio que a variação Unpr-Pru(1). Porém, analisando somente as variações com balanceamento

Tabela A.5: HammingLoss e Teste *post-hoc* obtidos nos experimentos

Datasets	Pru	Pru(13)	Pru(1)	Pru(3)	Unpr	Unpr(13)	Unpr(1)	Unpr(3)	Unpr-Pr	Unpr-Pr(13)	Unpr-Pr(1)	Unpr-Pr(3)
HammingLoss												
allhyper	**0.010**	0.016	**0.010**	0.049	0.034	0.014	0.011	0.065	**0.010**	0.013	**0.010**	0.058
medical	0.031	0.129	0.103	0.03	**0.029**	0.133	0.134	0.03	0.031	0.142	0.126	0.037
allhypo	0.065	0.064	**0.038**	0.075	0.041	0.061	**0.038**	0.079	0.044	0.062	**0.038**	0.062
genbase	0.140	0.305	0.316	0.141	0.059	0.173	0.206	**0.058**	0.06	0.168	0.168	0.061
ann-thyroid	0.004	**0.003**	0.004	0.007	0.007	0.004	**0.003**	0.007	0.004	0.006	0.004	0.004
lympoma	0.130	0.106	0.123	0.111	0.115	0.114	0.112	0.110	0.126	**0.099**	0.119	0.102
ecoli	0.084	**0.075**	0.099	0.109	0.083	0.096	0.094	0.113	0.091	0.103	0.085	0.089
glass	0.092	0.092	0.092	**0.091**	0.092	0.093	0.092	0.092	0.092	0.092	**0.091**	0.093
lymph	0.114	0.148	0.135	**0.113**	0.127	0.165	0.164	0.125	0.124	0.178	0.179	0.114
dermatology	0.052	0.044	0.051	0.050	0.051	0.053	0.051	0.055	0.051	0.043	0.052	**0.041**
postoperative	**0.192**	0.307	**0.192**	0.422	0.200	0.281	0.218	0.318	**0.192**	0.296	**0.192**	0.340
breast	0.056	0.047	0.040	0.034	0.044	0.053	0.040	0.053	**0.031**	0.055	0.037	0.047
scene	0.071	0.099	0.071	0.098	0.074	0.088	0.068	0.089	**0.064**	0.085	0.072	0.093
splice	0.067	0.068	0.068	0.076	**0.063**	0.071	0.066	0.075	0.068	0.076	0.067	0.067
emotions	0.291	**0.237**	0.304	0.289	0.293	0.265	0.312	0.292	0.297	0.266	0.294	0.303
lung cancer	0.461	0.461	0.455	0.400	0.461	0.472	0.516	**0.383**	0.483	0.416	0.466	0.494
contraceptive	0.357	0.445	**0.356**	0.437	0.411	0.459	0.416	0.465	0.361	0.451	0.357	0.422
wine	0.169	0.176	0.181	**0.162**	0.180	0.176	0.168	0.181	0.180	0.169	0.180	0.237
iris	0.280	**0.200**	0.266	0.324	0.248	0.244	0.271	0.271	0.284	0.262	0.293	0.293
Resultado do Teste *post-hoc*												
Pru	o	△	△	△	▽	△	△	△	▽	△	△	△
Pru13	x	o	▽	△	▽	△	▽	△	▽	△	▽	△
Pru1	x	x	o	△	▽	△	△	△	▽	△	▽	△
Pru3	x	x	x	o	▽	△	▽	△	▽	△	▽	△
Unpr	x	x	x	x	o	△	△	△	△	△	△	△
Unpr13	x	x	x	x	x	o	▽	▽	▽	▽	▽	▽
Unpr1	x	x	x	x	x	x	o	△	▽	△	▽	△
Unpr3	x	x	x	x	x	x	x	o	▽	▽	▽	▽
Unpr-Pr	x	x	x	x	x	x	x	x	o	△	△	△
Unpr-Pr13	x	x	x	x	x	x	x	x	x	o	▽	▽
Unpr-Pr1	x	x	x	x	x	x	x	x	x	x	o	△
Unpr-Pr3	x	x	x	x	x	x	x	x	x	x	x	o
Rank Médio	5.684	6.211	6.158	6.658	**5.421**	7.789	6.184	8.026	**5.579**	7.026	5.947	7.316

podemos perceber que as variações Pru(13), Unpr(3) e Unpr-Pru(1) foram as melhores de cada tipo de variação considerando a poda.

Levando em consideração a métrica precisão, as as duas versões que obtiveram melhor desempenho também foram Pru e Unpr, porém não sendo significativamente melhor que as outras variações. Comparando as variações não balanceadas com as balanceadas podemos ver que as variações Pru e Unpr são melhores que suas variações balanceadas e que a variação Unpr-Pru(1) obteve melhor *rank* médio que a variação Unpr-Pru. Agora, analisando somente as variações com balanceamento podemos visualizar que as variações Pru(3), Unpr(3) e Unpr-Pru(1) foram as melhores.

Levando em consideração a métrica Revocação, as versões Pru e Pru(13) foram as melhores, porém não sendo significativamente melhor que as outras variações. Comparando as variações não balanceadas com as balanceadas podemos ver que as variações Pru e Unpr são melhores que suas variações balanceadas e que a variação Unpr-Pru(1) obteve melhor *rank* médio que a variação Unpr-Pru. Analisando somente as variações balanceadas podemos ver que as variações Pru(13), Unpr(3) e Unpr-Pru(1) foram as melhores.

Levando em consideração a métrica Medida-F, as versões Pru e Unp-Pru(1) foram as melhores, porém não sendo significativamente melhor que as outras variações. Comparando as variações não balanceadas com as balanceadas podemos ver que as variações Pru e Unpr são melhores que suas variações balanceadas e que a variação Unpr-Pru(1) obteve melhor *rank* médio que a variação Unpr-Pru. Analisando somente as variações balanceadas podemos ver que as variações Pru(3), Unpr(3) e Unpr-Pru(1) foram as melhores.

Finalmente, considerando o número de nós da árvore, as versões Pru(1) e Unpr-Pru(1) foram as melhores, sendo que a variação Pru(1) e Unpr-Pru(1) foram significativamente melhores que as variações Unpr e Unpr(3). Comparando as variações não balanceadas com as balanceadas podemos ver que todas variações de poda sem balanceamento foram piores que as variações com balanceamento, sendo Pru(1), Unpr(3) e Unpr-Pru(1) as melhores.

Como mencionado anteriormente a partir desses resultados foram escolhidas as duas melhores versões do algoritmo BR-RT e devido a esta grande variação entre os resultados, isto é, os resultados de cada métrica são diferentes entre si. Então, foi utilizado para cada versão do algoritmo BR-RT a média geométrica das seis métricas foi selecionado as versões Pru e Unpr-Pru(1).

Tabela A.6: Acurácia e Teste *post-hoc* obtidos nos experimentos

Datasets	Pru	Pru(13)	Pru(1)	Pru(3)	Unpr	Unpr(13)	Unpr(1)	Unpr(3)	Unpr-Pr	Unpr-Pr(13)	Unpr-Pr(1)	Unpr-Pr(3)
Acurácia												
allhyper	**0.989**	0.988	**0.989**	0.986	0.983	0.988	0.988	0.988	**0.989**	0.988	**0.989**	0.987
medical	0.966	0.87	0.894	0.966	**0.968**	0.847	0.851	0.963	**0.968**	0.857	0.873	0.962
allhypo	**0.984**	0.960	0.961	0.974	0.968	0.963	0.961	0.952	0.964	0.965	0.961	0.964
genbase	0.859	0.694	0.683	0.858	0.940	0.826	0.793	**0.941**	0.939	0.831	0.831	0.938
ann-thyroid	0.995	**0.996**	0.995	0.995	0.995	**0.996**	0.996	0.995	0.995	0.993	0.995	0.995
lympoma	0.869	0.894	0.876	0.888	0.884	0.887	0.888	0.89	0.873	**0.900**	0.882	0.897
ecoli	0.920	**0.927**	0.902	0.897	0.921	0.911	0.909	0.896	0.912	0.899	0.915	0.922
glass	0.907	0.907	0.907	**0.908**	0.907	0.906	0.907	0.907	0.907	0.907	**0.908**	0.906
lymph	0.885	0.851	0.864	**0.886**	0.879	0.834	0.838	0.874	0.875	0.821	0.820	0.885
dermatology	0.784	0.870	0.790	0.889	0.787	0.868	0.794	**0.906**	0.802	0.871	0.789	0.902
postoperative	**0.807**	0.733	**0.807**	0.614	0.729	0.777	0.803	0.688	**0.807**	0.729	**0.807**	0.755
breast	0.803	0.846	0.846	0.863	0.830	0.816	0.844	0.813	0.800	0.851	**0.870**	0.866
scene	0.774	0.765	0.759	0.743	0.769	0.768	**0.784**	0.761	0.78	0.76	0.776	0.752
splice	0.932	0.931	0.931	0.923	**0.936**	0.928	0.933	0.924	0.931	0.923	0.932	0.932
emotions	0.690	0.689	0.688	0.682	0.700	0.679	0.683	0.684	0.675	0.699	**0.704**	0.673
lung cancer	0.538	0.538	0.544	0.600	0.538	0.527	0.483	**0.616**	0.516	0.583	0.533	0.505
contraceptive	0.655	0.633	0.652	0.644	0.645	0.642	0.631	0.638	0.652	0.633	**0.659**	0.656
wine	0.830	0.823	0.818	**0.837**	0.819	0.823	0.831	0.818	0.819	0.830	0.819	0.770
iris	0.724	**0.799**	0.733	0.675	0.751	0.755	0.728	0.764	0.720	0.737	0.706	0.751
Resultado do Teste *post-hoc*												
Pru	o	△	△	△	△	△	△	△	△	△	△	△
Pru13	x	o	△	△	▽	△	△	△	▽	△	▽	▽
Pru1	x	x	o	▽	▽	△	▽	▽	▽	△	▽	▽
Pru3	x	x	x	o	▽	△	▽	▽	▽	△	▽	▽
Unpr	x	x	x	x	o	△	△	△	△	△	△	△
Unpr13	x	x	x	x	x	o	▽	▽	▽	▽	▽	▽
Unpr1	x	x	x	x	x	x	o	▽	▽	△	▽	▽
Unpr3	x	x	x	x	x	x	x	o	▽	△	▽	▽
Unpr-Pr	x	x	x	x	x	x	x	x	o	△	o	△
Unpr-Pr13	x	x	x	x	x	x	x	x	x	o	▽	▽
Unpr-Pr1	x	x	x	x	x	x	x	x	x	x	o	△
Unpr-Pr3	x	x	x	x	x	x	x	x	x	x	x	o
Rank Médio	**5.139**	6.306	7.278	6.583	**5.583**	7.500	7.000	6.694	6.139	7.167	6.139	6.472

Tabela A.7: Precisão e Teste *post-hoc* obtidos nos experimentos

Datasets	Pru	Pru(13)	Pru(1)	Pru(3)	Unpr	Unpr(13)	Unpr(1)	Unpr(3)	Unpr-Pr	Unpr-Pr(13)	Unpr-Pr(1)	Unpr-Pr(3)
Precisão												
allhyper	**0.972**	0.970	**0.972**	0.965	0.959	0.970	0.970	0.970	0.972	0.971	**0.972**	0.967
medical	0.333	0.063	0.074	0.339	0.436	0.048	0.052	0.254	**0.569**	0.057	0.069	0.274
allhypo	**0.968**	0.920	0.922	0.949	0.936	0.927	0.922	0.901	0.928	0.931	0.922	0.927
genbase	0.168	0.052	0.047	0.169	**0.329**	0.071	0.069	0.345	0.313	0.089	0.095	0.306
ann-thyroid	0.993	**0.994**	0.993	0.993	0.993	**0.994**	**0.994**	0.993	0.993	0.990	0.993	0.993
lympoma	0.411	0.528	0.414	0.536	0.478	0.493	0.500	0.503	0.43	**0.553**	0.472	0.552
ecoli	0.681	**0.711**	0.610	0.441	0.684	0.647	0.636	0.585	0.648	0.597	0.660	0.687
glass	0.677	0.677	0.677	**0.678**	0.677	0.673	0.676	0.677	0.676	0.677	**0.678**	0.673
lymph	**0.779**	0.703	0.725	0.773	0.754	0.663	0.678	0.747	0.751	0.639	0.651	0.762
dermatology	0.443	0.612	0.372	0.668	0.362	0.606	0.384	**0.720**	0.406	0.613	0.413	0.706
postoperative	**0.711**	0.555	**0.711**	0.127	0.644	0.666	0.700	0.533	**0.711**	0.544	**0.711**	0.633
breast	0.41	0.538	0.54	0.591	0.491	0.448	0.534	0.441	0.573	0.553	**0.610**	0.598
scene	0.353	0.329	0.315	0.266	0.335	0.334	**0.379**	0.330	0.377	0.313	0.363	0.298
splice	0.898	0.897	0.897	0.885	**0.904**	0.892	0.900	0.886	0.897	0.884	0.898	0.898
emotions	0.513	0.521	0.505	0.493	0.526	0.498	0.491	0.499	0.491	0.516	**0.527**	0.479
lung cancer	0.308	0.308	0.316	0.400	0.308	0.291	0.225	**0.425**	0.275	0.374	0.299	0.258
contraceptive	0.482	0.450	0.479	0.466	0.468	0.463	0.447	0.458	0.478	0.450	**0.488**	0.484
wine	0.746	0.734	0.727	**0.756**	0.729	0.734	0.746	0.728	0.729	0.745	0.729	0.655
iris	0.586	**0.700**	0.599	0.513	0.626	0.633	0.593	0.646	0.58	0.606	0.56	0.626
Resultado do Teste *post-hoc*												
Pru	o	△	△	△	△	△	△	△	△	△	△	△
Pru13	x	o	△	▽	▽	△	△	▽	▽	△	▽	▽
Pru1	x	x	o	▽	▽	△	△	▽	▽	△	▽	▽
Pru3	x	x	x	o	▽	△	△	△	▽	△	▽	△
Unpr	x	x	x	x	o	△	△	△	△	△	▽	△
Unpr13	x	x	x	x	x	o	▽	▽	▽	▽	▽	▽
Unpr1	x	x	x	x	x	x	o	▽	▽	▽	▽	▽
Unpr3	x	x	x	x	x	x	x	o	▽	△	▽	▽
Unpr-Pr	x	x	x	x	x	x	x	x	o	△	▽	△
Unpr-Pr13	x	x	x	x	x	x	x	x	x	o	▽	▽
Unpr-Pr1	x	x	x	x	x	x	x	x	x	x	o	△
Unpr-Pr3	x	x	x	x	x	x	x	x	x	x	x	o
Rank Médio	**5.184**	6.368	7.263	6.263	**5.658**	7.526	7.526	6.868	6.184	7.158	5.684	6.316

Tabela A.8: Revocação e Teste *post-hoc* obtidos nos experimentos

Datasets	Pru	Pru(13)	Pru(1)	Pru(3)	Unpr	Unpr(13)	Unpr(1)	Unpr(3)	Unpr-Pr	Unpr-Pr(13)	Unpr-Pr(1)	Unpr-Pr(3)
Recall												
allhyper	**0.972**	0.970	**0.972**	0.966	0.959	0.970	0.970	0.971	**0.972**	0.971	**0.972**	0.969
medical	0.271	0.227	0.197	**0.296**	0.172	0.228	0.220	0.182	0.186	0.219	0.232	0.193
allhypo	**0.968**	0.920	0.922	0.949	0.936	0.927	0.922	0.936	0.928	0.931	0.922	0.932
genbase	0.385	0.313	0.303	**0.399**	0.270	0.229	0.237	0.281	0.251	0.246	0.247	0.249
ann-thyroid	0.993	**0.994**	0.993	0.993	0.993	**0.994**	**0.994**	0.993	0.993	0.990	0.993	0.993
lympoma	0.411	0.532	0.433	0.465	0.478	0.476	0.500	0.510	0.430	0.553	0.472	**0.578**
ecoli	0.681	**0.711**	0.610	0.445	0.684	0.647	0.636	0.588	0.648	0.597	0.660	0.696
glass	0.677	0.677	0.677	**0.678**	0.677	0.673	0.676	0.677	0.676	0.677	**0.678**	0.673
lymph	0.757	0.757	0.751	0.769	0.765	0.715	0.743	0.751	0.751	0.748	0.730	**0.788**
dermatology	0.261	0.612	0.372	0.668	0.362	0.606	0.384	**0.720**	0.406	0.613	0.323	0.706
postoperative	**0.711**	0.555	**0.711**	0.133	0.644	0.666	0.700	0.533	**0.711**	0.544	**0.711**	0.633
breast	0.41	0.5383	0.54	0.591	0.491	0.448	0.534	0.441	0.573	0.553	**0.610**	0.598
scene	0.332	0.308	0.294	0.247	0.317	0.316	0.36	0.301	**0.353**	0.293	0.342	0.274
splice	0.898	0.897	0.897	0.885	**0.904**	0.892	0.900	0.886	0.897	0.884	0.898	0.898
emotions	0.601	**0.656**	0.648	0.568	0.533	0.612	0.528	0.532	0.509	0.633	0.573	0.522
lung cancer	0.308	0.308	0.316	0.400	0.308	0.291	0.225	**0.425**	0.275	0.374	0.299	0.258
contraceptive	0.482	0.450	0.479	0.466	0.468	0.463	0.447	0.458	0.478	0.450	**0.488**	0.484
wine	0.746	0.734	0.727	**0.756**	0.729	0.734	0.746	0.728	0.729	0.745	0.729	0.655
iris	0.586	**0.700**	0.599	0.513	0.626	0.633	0.593	0.646	0.58	0.606	0.56	0.626
Resultado do Teste *post-hoc*												
Pru	o	o	△	△	△	△	△	△	△	△	△	△
Pru13	x	o	△	△	△	△	△	△	△	△	△	△
Pru1	x	x	o	▽	△	△	△	△	△	△	▽	△
Pru3	x	x	x	o	△	△	△	△	△	△	▽	△
Unpr	x	x	x	x	o	△	△	▽	△	▽	▽	▽
Unpr13	x	x	x	x	x	o	▽	▽	▽	▽	▽	▽
Unpr1	x	x	x	x	x	x	o	▽	▽	▽	▽	▽
Unpr3	x	x	x	x	x	x	x	o	△	▽	▽	▽
Unpr-Pr	x	x	x	x	x	x	x	x	o	▽	▽	▽
Unpr-Pr13	x	x	x	x	x	x	x	x	x	o	▽	▽
Unpr-Pr1	x	x	x	x	x	x	x	x	x	x	o	△
Unpr-Pr3	x	x	x	x	x	x	x	x	x	x	x	o
Rank Médio	**5.316**	**5.421**	6.579	6.158	6.474	7.263	7.263	6.947	7.105	7.079	5.947	6.447

Tabela A.9: Medida -F e Teste *post-hoc* obtidos nos experimentos

Datasets	Pru	Pru(13)	Pru(1)	Pru(3)	Unpr	Unpr(13)	Unpr(1)	Unpr(3)	Unpr-Pr	Unpr-Pr(13)	Unpr-Pr(1)	Unpr-Pr(3)
Medida -F												
allhyper	**0,972**	0,970	**0,972**	0,965	0,959	0,970	0,970	0,970	**0,972**	0,971	**0,972**	0,968
medical	**0,299**	0,099	0,108	0,316	0,247	0,079	0,084	0,212	0,280	0,090	0,106	0,226
allhypo	**0,968**	0,920	0,922	0,949	0,936	0,927	0,922	0,918	0,928	0,931	0,922	0,929
genbase	0,234	0,089	0,081	0,237	**0,297**	0,108	0,108	0,310	0,279	0,131	0,137	0,275
ann-thyroid	0,993	**0,994**	0,993	0,993	0,993	**0,994**	**0,994**	0,993	0,993	0,990	0,993	0,993
lympoma	0,411	0,530	0,423	0,498	0,478	0,484	0,500	0,506	0,430	0,553	0,472	**0,565**
ecoli	0,681	**0,711**	0,610	0,443	0,684	0,647	0,636	0,586	0,648	0,597	0,660	0,691
glass	0,677	0,677	0,677	**0,678**	0,677	0,673	0,676	0,677	0,676	0,677	**0,678**	0,673
lymph	0,768	0,729	0,738	0,771	0,759	0,688	0,709	0,749	0,751	0,689	0,688	**0,775**
dermatology	0,328	0,612	0,372	0,668	0,362	0,606	0,384	**0,720**	0,406	0,613	0,362	0,706
postoperative	**0,711**	0,555	**0,711**	0,130	0,644	0,666	0,700	0,533	**0,711**	0,544	**0,711**	0,633
breast	0,410	0,538	0,540	0,591	0,491	0,448	0,534	0,441	0,573	0,553	**0,610**	0,598
scene	0,342	0,318	0,304	0,256	0,326	0,325	**0,369**	0,315	0,365	0,303	0,352	0,285
emotions	0,554	0,581	0,568	0,528	0,529	0,549	0,509	0,515	0,500	**0,569**	0,549	0,500
lung cancer	0,308	0,308	0,316	0,400	0,308	0,291	0,225	**0,425**	0,275	0,374	0,299	0,258
contraceptive	0,482	0,450	0,479	0,466	0,468	0,463	0,447	0,458	0,478	0,450	**0,488**	0,484
wine	0,746	0,734	0,727	**0,756**	0,729	0,734	0,746	0,728	0,729	0,745	0,729	0,655
iris	0,586	**0,700**	0,599	0,513	0,626	0,633	0,593	0,646	0,580	0,606	0,560	0,626
Resultado do Teste *post-hoc*												
Pru	o	△	△	△	△	△	△	△	△	△	△	△
Pru.13.	x	o	△	▽	▽	△	△	△	△	△	▽	△
Pru.1.	x	x	o	▽	▽	△	△	△	▽	△	▽	▽
Pru.3.	x	x	x	o	▽	△	△	△	△	△	▽	△
Unpr	x	x	x	x	o	△	△	△	△	△	▽	△
Unpr.13.	x	x	x	x	x	o	▽	▽	▽	▽	▽	▽
Unpr.1.	x	x	x	x	x	x	o	▽	▽	▽	▽	▽
Unpr.3.	x	x	x	x	x	x	x	o	▽	△	▽	▽
Unpr.Pr	x	x	x	x	x	x	x	x	o	△	▽	△
Unpr.Pr.13.	x	x	x	x	x	x	x	x	x	o	▽	▽
Unpr.Pr.1.	x	x	x	x	x	x	x	x	x	x	o	△
Unpr.Pr.3.	x	x	x	x	x	x	x	x	x	x	x	o
Rank Médio	**5,500**	6,286	6,571	6,190	6,143	7,810	7,238	6,833	6,310	6,905	**5,833**	6,381

Tabela A.10: Número de Nós e Teste *post-hoc* obtidos nos experimentos

Datasets	Pru	Pru(13)	Pru(1)	Pru(3)	Unpr	Unpr(13)	Unpr(1)	Unpr(3)	Unpr-Pr	Unpr-Pr(13)	Unpr-Pr(1)	Unpr-Pr(3)
Número de Nós												
emotions	69.8	53.6	65.8	56.2	73.8	57.2	76.4	62.4	68.2	**49**	66	62.4
genbase	26.2	6.6	5.9	169.1	10	4.5	**4.1**	5.1	9.7	7.4	4.8	17.4
scene	152.4	154	156.4	141.2	157.6	143.2	161.4	122.2	149.6	147.2	177	**130**
medical	56.6	6	6.2	113.2	46.1	22.2	15.8	94.9	32.9	8.0	**4.6**	76.4
lymph	27.3	**20.6**	22.3	30.0	38.7	24.4	33.2	38.6	26.9	24.9	26.8	26.1
ecoli	11.2	14.0	**6.7**	12.4	11.4	17.6	10.8	17.0	13.1	10.8	10.4	13.0
allhypo	15.0	10.5	**1.0**	20.6	4.4	12.2	**1.0**	19.1	3.6	11.8	**1.0**	15.7
allhyper	2.0	10.5	1.5	26.0	18.4	10.9	3.5	34.2	2.0	10.3	**1.1**	25.8
postoperative	1.2	15.0	**1.1**	18.7	4.0	13.2	6.7	20.7	**1.1**	15.6	1.2	15.2
lympoma	5.6	**4.1**	6.0	4.3	5.4	3.7	5.1	5.0	6.4	4.5	4.4	4.6
lung cancer	13.3	10.3	12.6	**8.6**	18.3	11.8	18.8	12.8	14.6	11.1	13.0	8.9
ann-thyroid	19.8	**17.0**	19.2	17.2	20.8	19.8	21.4	21.8	19.8	17.2	20.0	18.2
contraceptive	72.2	253.2	**56.4**	257.8	320.6	406.0	315.2	430.0	80.2	246.8	63.8	251.8
dermatology	**5.0**	17.2	**5.0**	19.6	**5.0**	24.0	**5.0**	34.1	**5.0**	18.2	**5.0**	22.1
breast	5.2	6.8	**5.0**	5.4	6.4	6.0	6.4	7.0	5.0	5.4	5.4	6.6
wine	11.6	**6.6**	11.8	9.4	11.6	7.4	13.0	11.0	11.0	10.0	10.8	10.4
glass	**4.0**	**4.0**	**4.0**	**4.0**	**4.0**	**4.0**	**4.0**	**4.0**	**4.0**	**4.0**	**4.0**	**4.0**
iris	3.4	5.2	**3.0**	4.2	3.8	4.0	3.4	4.6	3.6	5.0	3.0	5.2
splice	**144.7**	182.4	147.2	176.3	426.2	346.6	430.4	353.7	147.5	180.5	149.0	178.4
Resultado do Teste *post-hoc*												
Pru	o	▽	▽	△	△	▽	△	△	▽	▽	▽	△
Pru13	x	o	▽	△	△	△	△	▲	△	▽	▽	△
Pru1	x	x	o	△	▲	△	△	▲	△	△	△	△
Pru3	x	x	x	o	△	▽	△	△	▽	▽	▽	△
Unpr	x	x	x	x	o	▽	▽	△	▽	▽	▼	▽
Unpr13	x	x	x	x	x	o	△	△	▽	▽	▽	△
Unpr1	x	x	x	x	x	x	o	△	▽	▽	▽	▽
Unpr3	x	x	x	x	x	x	x	o	▽	▼	▼	▽
Unpr.Pr	x	x	x	x	x	x	x	x	o	▽	▽	△
Unpr-Pr13	x	x	x	x	x	x	x	x	x	o	▽	△
Unpr-Pr1	x	x	x	x	x	x	x	x	x	x	o	△
Unpr-Pr3	x	x	x	x	x	x	x	x	x	x	x	o
Rank Médio	6.421	5.684	**4.289**	6.895	8.421	6.342	7.447	9.132	6.053	5.500	**4.632**	7.184

Apêndice

B

Taxa de acerto das árvores de decisão de rótulos - Etapa 1

A primeira etapa, como já mencionado anteriormente, tem como objetivo descobrir as relações entre os rótulos baseado nas árvores induzidas, por isso é importante saber se essas árvores tem um bom desempenho para não prejudicar a elaboração das relações dos rótulos.

A seguir podemos observar que nas tabelas são mostrados as taxas de acerto das árvores de cada rótulos geradas na primeira etapa das abordagens BR-RTb pru e BR-RTb unpr-pru1, respectivamente, no qual as linhas correspondem aos rótulos e as colunas corresponde as bases de dados.

Tabela B.1: Taxa de acerto das AD de rótulos - BR-RTb pru - Primeiro Nível da Hierarquia

Rótulos	Pheno	Seq	Church	CellCycle	Derisi	Eisen	Gasch2	SPO	Gach1	Expr	Média
1.0.0.0	75,6	80,1	82,3	82,2	77,8	83,2	82,3	82,3	82,3	82,3	81,0
2.0.0.0	92,3	93,3	93,3	93,3	93,2	93,3	93,3	93,3	93,3	93,3	93,2
3.0.0.0	81,7	83,0	85,1	85,1	82,2	86,0	85,1	85,0	85,1	85,2	84,4
4.0.0.0	77,9	81,4	83,5	83,5	76,3	82,9	83,6	83,4	83,5	83,6	81,9
5.0.0.0	100,0	95,1	95,6	95,6	100,0	96,5	95,6	95,7	95,6	95,6	96,5
6.0.0.0	88,4	86,1	88,4	88,3	86,7	87,2	88,4	88,4	88,3	88,4	87,9
8.0.0.0	90,3	90,1	91,2	91,1	90,0	91,5	91,2	91,3	91,2	91,2	90,9
10.0.0.0	98,1	98,5	98,4	98,4	98,4	99,5	98,4	98,4	98,4	98,4	98,5
11.0.0.0	100,0	100,0	100,0	100,0	100,0	100,0	100,0	100,0	100,0	100,0	100,0
13.0.0.0	100,0	100,0	100,0	100,0	100,0	100,0	100,0	100,0	100,0	100,0	100,0
14.0.0.0	100,0	100,0	100,0	100,0	100,0	100,0	100,0	100,0	100,0	100,0	100,0
30.0.0.0	91,6	94,7	95,5	95,4	94,6	95,0	95,4	95,4	95,4	95,4	94,8
40.0.0.0	67,6	67,8	68,9	68,8	68,4	76,5	68,8	68,6	68,8	68,8	69,3
62.0.0.0	100,0	100,0	100,0	100,0	100,0	100,0	100,0	100,0	100,0	100,0	100,0
63.0.0.0	100,0	100,0	100,0	100,0	100,0	100,0	100,0	100,0	100,0	100,0	100,0
67.0.0.0	94,1	94,3	95,1	95,2	94,0	95,5	95,1	95,2	95,1	95,2	94,9

Tabela B.2: Taxa de acerto das AD de rótulos - BR-RTb unpr-pru1 - Primeiro Nível da Hierarquia

Rótulos	Pheno	Seq	Church	CellCycle	Derisi	Eisen	Gasch2	SPO	Gach1	Expr	Média
1.0.0.0	73,48	78,48	79,30	80,66	73,74	79,79	81,10	80,46	79,96	79,90	78,687

Continua na próxima página

Tabela B.2 – *Continuação da página anterior*

Rótulos	Pheno	Seq	Church	CellCycle	Derisi	Eisen	Gasch2	SPO	Gach1	Expr	Média
2.0.0.0	80,45	82,78	83,63	83,48	80,14	84,82	83,74	83,83	83,93	82,60	82,941
3.0.0.0	70,71	74,23	75,47	75,46	70,26	74,64	75,34	74,53	75,29	74,62	74,055
4.0.0.0	75,11	77,94	79,33	78,59	74,30	78,76	80,17	78,41	78,98	78,95	78,055
5.0.0.0	100,00	89,77	90,62	90,01	100,00	91,42	90,29	89,65	90,91	90,55	92,322
6.0.0.0	76,56	78,00	80,79	79,55	75,75	80,21	79,44	79,49	79,03	79,98	78,878
8.0.0.0	79,70	82,04	84,99	83,53	78,46	85,20	84,56	83,80	84,20	84,21	83,067
10.0.0.0	84,92	77,49	80,36	80,13	76,47	85,98	79,25	79,57	78,90	80,20	80,326
11.0.0.0	100,00	100,00	100,00	100,00	100,00	100,00	100,00	100,00	100,00	100,00	100,000
13.0.0.0	100,00	100,00	100,00	100,00	100,00	100,00	100,00	100,00	100,00	100,00	100,000
14.0.0.0	100,00	100,00	100,00	100,00	100,00	100,00	100,00	100,00	100,00	100,00	100,000
30.0.0.0	72,97	71,20	72,23	71,69	71,28	75,84	72,36	73,37	71,50	72,54	72,497
40.0.0.0	68,70	67,08	67,55	67,84	66,40	69,53	68,95	68,17	67,10	68,44	67,976
62.0.0.0	100,00	100,00	100,00	100,00	100,00	100,00	100,00	100,00	100,00	100,00	100,000
63.0.0.0	100,00	100,00	100,00	100,00	100,00	100,00	100,00	100,00	100,00	100,00	100,000
67.0.0.0	82,02	81,56	82,62	82,58	82,61	87,34	83,90	84,20	83,72	83,31	83,386

Tabela B.3: Taxa de acerto das AD de rótulos - BR-RTb pru - Segundo Nível da Hierarquia

Rótulos	Pheno	Seq	Church	CellCycle	Derisi	Eisen	Gasch2	SPO	Gach1	Expr	Média
1.1.0.0	100,00	100,00	100,00	100,00	100,00	100,00	100,00	100,00	100,00	100,00	100,00
1.2.0.0	100,00	100,00	100,00	100,00	100,00	100,00	100,00	100,00	100,00	100,00	100,00
1.20.0.0	100,00	100,00	100,00	100,00	100,00	100,00	100,00	100,00	100,00	100,00	100,00
1.3.0.0	99,94	99,95	99,95	99,95	99,95	100,00	99,95	99,95	99,95	99,95	99,95
1.4.0.0	99,94	99,97	99,97	99,97	99,97	100,00	99,97	99,97	99,97	99,97	99,97
1.5.0.0	99,94	99,97	99,97	99,97	99,97	99,96	99,97	99,97	99,97	99,97	99,97
1.6.0.0	99,94	99,97	99,97	99,97	99,97	100,00	99,97	99,97	99,97	99,97	99,97
1.7.0.0	100,00	99,95	99,95	99,95	99,95	100,00	99,95	99,95	99,95	99,95	99,96
2.1.0.0	98,81	99,11	99,07	99,07	99,04	98,80	99,08	99,06	99,07	99,08	99,02
2.10.0.0	99,25	99,36	99,34	99,34	99,33	99,30	99,34	99,33	99,34	99,34	99,33
2.11.0.0	100,00	100,00	100,00	100,00	100,00	100,00	100,00	100,00	100,00	100,00	100,00
2.13.0.0	97,61	97,76	97,95	97,95	97,99	97,65	97,97	97,98	97,96	97,97	97,88
2.16.0.0	98,99	99,16	99,12	99,12	99,17	99,50	99,13	99,16	99,13	99,13	99,16
2.19.0.0	98,99	99,06	99,02	99,02	98,98	99,13	99,02	99,00	99,02	99,02	99,03
2.22.0.0	99,81	99,85	99,84	99,84	99,84	99,79	99,84	99,84	99,84	99,84	99,83
2.25.0.0	99,87	99,82	99,81	99,81	99,81	99,83	99,82	99,81	99,81	99,82	99,82
2.7.0.0	99,69	99,77	99,76	99,76	99,76	99,67	99,76	99,76	99,76	99,76	99,75
2.99.0.0	99,43	99,59	99,57	99,58	99,57	99,63	99,58	99,57	99,58	99,58	99,57
3.1.0.0	100,00	99,97	99,97	99,97	99,97	100,00	99,97	99,97	99,97	99,97	99,98
3.3.0.0	100,00	100,00	100,00	100,00	100,00	100,00	100,00	100,00	100,00	100,00	100,00
3.99.0.0	99,75	99,85	99,87	99,84	99,87	99,96	99,84	99,87	99,87	99,84	99,85
4.1.0.0	100,00	99,97	99,97	99,97	99,97	100,00	99,97	99,97	99,97	99,97	99,98
4.3.0.0	100,00	99,97	99,97	99,97	99,97	100,00	99,97	99,97	99,97	99,97	99,98
4.5.0.0	100,00	99,97	99,97	99,97	99,97	99,96	99,97	99,97	99,97	99,97	99,97
4.7.0.0	99,50	99,39	99,36	99,36	99,36	99,17	99,37	99,35	99,36	99,37	99,36
4.99.0.0	98,56	98,55	98,51	98,51	98,53	99,01	98,52	98,52	98,52	98,52	98,57
5.1.0.0	95,23	94,53	94,53	94,63	94,88	92,82	94,45	94,85	94,46	94,45	94,48
5.10.0.0	99,31	99,06	99,02	99,02	98,98	98,68	99,02	98,98	99,02	99,02	99,01
5.4.0.0	98,18	98,47	98,43	98,41	98,42	97,81	98,42	98,44	98,41	98,42	98,34
5.7.0.0	98,93	99,21	99,18	99,18	99,17	98,93	99,18	99,16	99,18	99,18	99,13
5.99.0.0	99,56	99,59	99,57	99,58	99,57	99,42	99,58	99,57	99,58	99,58	99,56
6.1.0.0	98,56	98,58	98,54	98,54	98,47	98,31	98,55	98,46	98,54	98,55	98,51
6.10.0.0	97,42	97,58	97,50	97,50	97,45	96,66	97,52	97,47	97,51	97,52	97,41
6.13.0.0	99,12	99,31	99,28	99,28	99,30	99,59	99,29	99,30	99,28	99,29	99,30
6.4.0.0	97,55	97,28	97,26	97,13	97,24	96,74	97,15	97,25	97,24	97,12	97,20
6.7.0.0	96,61	95,75	95,59	95,56	95,58	94,35	95,59	95,55	95,57	95,59	95,57
6.99.0.0	99,81	99,80	99,81	99,81	99,79	99,83	99,82	99,81	99,81	99,82	99,81
8.1.0.0	98,68	98,98	98,94	98,94	98,93	98,72	98,94	98,92	98,94	98,94	98,89
8.10.0.0	99,94	99,80	99,79	99,79	99,79	99,71	99,79	99,78	99,79	99,79	99,80
8.13.0.0	98,68	98,93	98,96	98,94	98,95	98,60	98,94	98,95	98,97	98,94	98,89
8.16.0.0	99,18	99,03	99,02	98,99	99,01	98,89	99,00	99,00	99,02	99,00	99,01
8.19.0.0	98,56	98,58	98,54	98,54	98,53	97,69	98,55	98,52	98,54	98,55	98,46
8.22.0.0	99,18	99,31	99,28	99,28	99,28	99,26	99,29	99,27	99,28	99,29	99,27
8.4.0.0	98,30	98,68	98,72	98,78	98,74	98,72	98,79	98,73	98,78	98,79	98,70
8.7.0.0	97,74	97,84	97,77	97,77	97,78	97,85	97,78	97,76	97,75	97,76	97,78
8.99.0.0	99,37	99,39	99,36	99,36	99,41	99,46	99,37	99,41	99,36	99,37	99,39
10.1.0.0	100,00	100,00	100,00	100,00	100,00	100,00	100,00	100,00	100,00	100,00	100,00
10.5.0.0	99,81	99,90	99,89	99,89	99,89	99,96	99,89	99,89	99,89	99,89	99,89
11.1.0.0	100,00	100,00	100,00	100,00	100,00	100,00	100,00	100,00	100,00	100,00	100,00
11.10.0.0	100,00	100,00	100,00	100,00	100,00	100,00	100,00	100,00	100,00	100,00	100,00
11.7.0.0	100,00	99,95	99,95	99,95	99,95	100,00	99,95	99,95	99,95	99,95	99,96
11.99.0.0	99,94	99,77	99,76	99,76	99,76	99,71	99,76	99,76	99,76	99,76	99,77
13.1.0.0	100,00	99,97	99,97	99,97	99,97	100,00	99,97	99,97	99,97	99,97	99,98
13.11.0.0	100,00	100,00	100,00	100,00	100,00	100,00	100,00	100,00	100,00	100,00	100,00

Continua na próxima página

Tabela B.3 – *Continuação da página anterior*

Rótulos	Pheno	Seq	Church	CellCycle	Derisi	Eisen	Gasch2	SPO	Gach1	Expr	Média
14.1.0.0	100,00	99,97	99,97	99,97	99,97	100,00	99,97	99,97	99,97	99,97	99,98
14.10.0.0	99,69	99,77	99,76	99,76	99,76	99,79	99,76	99,76	99,76	99,76	99,76
14.20.0.0	99,81	99,90	99,89	99,89	99,89	99,92	99,89	99,89	99,89	99,89	99,89
14.4.0.0	100,00	100,00	100,00	100,00	100,00	100,00	100,00	100,00	100,00	100,00	100,00
29.99.0.0	99,87	99,92	99,92	99,92	99,92	100,00	99,92	99,92	99,92	99,92	99,92
30.1.0.0	94,22	97,28	97,26	97,18	97,13	96,53	97,20	97,20	97,22	97,20	96,84
30.10.0.0	99,56	99,67	99,65	99,65	99,65	99,75	99,66	99,65	99,66	99,66	99,66
30.16.0.0	99,62	99,57	99,63	99,60	99,65	99,67	99,60	99,65	99,60	99,60	99,62
30.19.0.0	99,94	99,95	99,95	99,95	99,95	99,92	99,95	99,95	99,95	99,95	99,94
30.2.0.0	100,00	99,97	99,97	99,97	99,97	99,96	99,97	99,97	99,97	99,97	99,97
30.22.0.0	99,94	99,97	99,97	99,97	99,97	99,96	99,97	99,97	99,97	99,97	99,97
30.25.0.0	99,62	99,54	99,52	99,52	99,52	99,42	99,52	99,51	99,52	99,52	99,52
30.3.0.0	100,00	99,97	99,97	99,97	99,97	99,96	99,97	99,97	99,97	99,97	99,97
30.4.0.0	99,25	99,41	99,39	99,39	99,41	99,38	99,39	99,41	99,39	99,39	99,38
30.7.0.0	99,94	99,97	99,97	99,97	99,97	99,96	99,97	99,97	99,97	99,97	99,97
30.8.0.0	99,87	99,92	99,92	99,92	99,92	99,88	99,92	99,92	99,92	99,92	99,91
30.9.0.0	100,00	99,97	99,97	99,97	99,97	99,96	99,97	99,97	99,97	99,97	99,97
30.99.0.0	99,75	99,82	99,81	99,81	99,81	99,96	99,82	99,81	99,81	99,82	99,82
40.1.0.0	99,25	99,03	99,02	98,99	99,01	99,38	99,00	99,00	99,02	99,00	99,07
40.10.0.0	84,11	82,27	81,58	81,62	81,48	79,62	81,65	81,48	81,63	81,67	81,71
40.16.0.0	94,47	94,00	94,05	94,05	94,00	93,98	94,03	94,04	94,06	94,03	94,07
40.19.0.0	99,12	99,41	99,39	99,39	99,36	99,13	99,39	99,35	99,39	99,39	99,33
40.2.0.0	96,98	97,56	97,48	97,45	97,45	96,91	97,47	97,44	97,45	97,47	97,37
40.22.0.0	99,87	99,72	99,71	99,71	99,71	99,63	99,71	99,70	99,71	99,71	99,72
40.25.0.0	98,37	98,88	98,91	98,88	98,90	98,64	98,92	98,89	98,91	98,89	98,82
40.27.0.0	99,69	99,49	99,50	99,47	99,46	99,38	99,47	99,46	99,47	99,47	99,49
40.3.0.0	90,14	90,28	89,88	89,88	89,76	88,99	89,91	89,81	89,90	89,91	89,85
40.30.0.0	100,00	99,97	99,97	99,97	99,97	100,00	99,97	99,97	99,97	99,97	99,98
40.4.0.0	97,17	97,71	97,63	97,64	97,70	97,61	97,65	97,71	97,64	97,65	97,61
40.5.0.0	99,37	99,21	99,18	99,20	99,22	99,22	99,18	99,22	99,18	99,18	99,22
40.7.0.0	96,55	96,16	96,01	95,99	95,98	94,76	96,01	95,98	96,00	96,01	95,95
40.8.0.0	97,99	97,99	97,90	97,90	97,91	97,69	97,91	97,90	97,91	97,91	97,90
40.9.0.0	99,25	98,93	98,88	98,88	98,90	98,93	98,89	98,89	98,89	98,89	98,93
40.99.0.0	99,69	99,80	99,79	99,79	99,79	99,79	99,79	99,78	99,79	99,79	99,78
62.2.0.0	100,00	100,00	100,00	100,00	100,00	100,00	100,00	100,00	100,00	100,00	100,00
63.1.0.0	100,00	99,97	99,97	99,97	99,97	100,00	99,97	99,97	99,97	99,97	99,98
63.9.0.0	100,00	99,97	99,97	99,97	99,97	99,96	99,97	99,97	99,97	99,97	99,97
67.1.0.0	100,00	100,00	100,00	100,00	100,00	100,00	100,00	100,00	100,00	100,00	100,00
67.10.0.0	99,50	99,36	99,34	99,34	99,33	99,09	99,34	99,33	99,34	99,34	99,33
67.11.0.0	100,00	99,97	99,97	99,97	99,97	100,00	99,97	99,97	99,97	99,97	99,98
67.13.0.0	99,87	99,82	99,81	99,81	99,81	99,79	99,82	99,81	99,81	99,82	99,82
67.16.0.0	99,50	99,62	99,60	99,60	99,62	99,75	99,60	99,62	99,60	99,60	99,61
67.19.0.0	99,87	99,77	99,76	99,76	99,76	99,83	99,76	99,76	99,76	99,76	99,78
67.28.0.0	99,06	99,11	99,07	99,10	99,09	99,67	99,08	99,08	99,07	99,08	99,14
67.4.0.0	100,00	99,97	99,97	99,97	99,97	100,00	99,97	99,97	99,97	99,97	99,98
67.50.0.0	100,00	100,00	100,00	100,00	100,00	100,00	100,00	100,00	100,00	100,00	100,00
67.7.0.0	99,37	99,14	99,10	99,10	99,06	99,09	99,10	99,08	99,10	99,10	99,12
67.99.0.0	98,68	98,93	98,94	98,94	98,87	99,42	98,94	98,92	98,94	98,94	98,95

Tabela B.4: Taxa de acerto das AD de rótulos - BR-RTb unpr-pru1 - Segundo Nível da Hierarquia

Rótulos	Pheno	Seq	Church	CellCycle	Derisi	Eisen	Gasch2	SPO	Gach1	Expr	Média
1.1.0.0	100,00	100,00	100,00	100,00	100,00	100,00	100,00	100,00	100,00	100,00	100,00
1.2.0.0	100,00	100,00	100,00	100,00	100,00	100,00	100,00	100,00	100,00	100,00	100,00
1.20.0.0	100,00	100,00	100,00	100,00	100,00	100,00	100,00	100,00	100,00	100,00	100,00
1.3.0.0	88,19	87,53	87,38	88,87	89,23	100,00	87,75	88,60	88,36	88,46	89,44
1.4.0.0	88,25	87,94	88,33	87,70	88,34	100,00	88,09	88,65	88,84	88,78	89,49
1.5.0.0	100,00	100,00	99,97	100,00	99,97	100,00	99,97	100,00	100,00	99,97	99,99
1.6.0.0	88,76	87,59	88,41	88,02	89,09	100,00	87,99	89,22	88,47	89,17	89,67
1.7.0.0	100,00	99,08	99,18	99,04	99,06	100,00	99,18	98,98	98,97	98,94	99,24
2.1.0.0	93,59	84,00	83,02	84,52	83,55	89,23	85,08	84,15	83,62	83,13	85,39
2.10.0.0	85,11	84,58	85,30	85,34	84,32	91,42	85,53	85,28	84,52	85,03	85,64
2.11.0.0	100,00	100,00	100,00	100,00	100,00	100,00	100,00	100,00	100,00	100,00	100,00
2.13.0.0	85,87	88,76	86,95	86,85	86,28	92,57	86,03	86,50	86,66	86,35	87,28
2.16.0.0	85,43	83,74	86,42	86,35	85,21	92,95	85,16	85,98	86,45	85,69	86,34
2.19.0.0	83,61	81,18	81,85	82,44	83,20	84,57	82,89	82,70	81,60	82,36	82,64
2.22.0.0	97,24	89,26	89,40	89,59	89,07	97,15	88,80	89,08	89,45	89,97	90,90
2.25.0.0	99,56	99,19	99,07	99,28	99,22	99,46	99,16	99,33	99,15	99,23	99,27
2.7.0.0	87,75	86,54	87,22	86,43	86,66	91,58	86,88	86,42	86,53	86,40	87,24
2.99.0.0	84,80	84,89	85,68	85,13	86,23	91,25	85,69	85,63	85,58	84,95	85,98
3.1.0.0	100,00	100,00	100,00	100,00	100,00	100,00	100,00	100,00	100,00	100,00	100,00
3.3.0.0	100,00	100,00	100,00	100,00	100,00	100,00	100,00	100,00	100,00	100,00	100,00
3.99.0.0	87,63	86,49	87,27	87,04	88,21	92,74	87,22	86,66	86,53	87,01	87,68
4.1.0.0	100,00	99,49	99,47	99,26	99,04	100,00	99,34	99,30	99,28	99,42	99,46

Continua na próxima página

Tabela B.4 – *Continuação da página anterior*

Rótulos	Pheno	Seq	Church	CellCycle	Derisi	Eisen	Gasch2	SPO	Gach1	Expr	Média
4.3.0.0	100,00	99,44	99,73	99,71	99,44	100,00	99,23	99,60	99,60	99,66	99,64
4.5.0.0	100,00	93,59	93,09	92,72	92,10	90,47	92,74	93,05	92,55	93,08	93,34
4.7.0.0	93,09	91,55	90,54	90,25	90,73	90,14	91,10	90,78	91,22	90,73	91,01
4.99.0.0	81,16	78,58	79,83	78,86	79,47	81,19	78,85	79,06	79,16	79,03	79,52
5.1.0.0	90,33	91,40	91,95	91,02	91,93	94,97	91,52	91,78	91,33	90,39	91,66
5.10.0.0	88,44	86,67	85,54	85,21	85,37	90,10	85,29	86,06	86,03	85,95	86,47
5.4.0.0	87,88	86,01	86,95	86,53	87,97	89,48	86,77	87,17	86,64	86,51	87,19
5.7.0.0	87,06	87,64	86,47	86,75	85,88	87,79	86,11	85,82	86,37	86,72	86,66
5.99.0.0	90,52	80,95	82,25	81,67	81,78	85,15	82,44	81,86	82,79	82,23	83,16
6.1.0.0	89,07	83,74	83,15	81,73	81,19	86,22	81,49	81,29	81,57	81,62	83,11
6.10.0.0	85,43	84,94	84,27	83,35	81,73	83,21	83,39	82,64	84,12	83,44	83,65
6.13.0.0	87,25	80,74	81,21	81,25	81,99	82,05	81,12	81,02	82,18	82,15	82,10
6.4.0.0	89,20	86,85	86,69	85,76	86,31	89,81	85,66	86,15	86,03	86,19	86,86
6.7.0.0	75,13	74,21	74,78	73,84	74,52	80,98	74,25	73,99	73,97	74,47	75,01
6.99.0.0	99,18	87,64	87,62	88,55	88,13	99,55	87,17	87,84	87,25	88,80	90,17
8.1.0.0	88,63	85,63	85,76	86,11	86,33	88,04	85,93	86,04	86,29	86,32	86,51
8.10.0.0	99,87	96,34	96,28	96,65	96,78	99,67	96,54	96,55	96,53	96,33	97,15
8.13.0.0	92,78	94,05	92,59	92,93	93,27	93,65	93,53	93,67	92,60	92,55	93,16
8.16.0.0	85,87	83,97	84,45	83,40	84,70	89,65	82,84	85,42	83,93	83,05	84,73
8.19.0.0	94,72	93,49	93,14	93,23	92,82	93,36	92,76	93,07	92,87	93,03	93,25
8.22.0.0	97,05	95,60	95,11	95,43	94,91	96,74	95,04	95,28	95,57	95,38	95,61
8.4.0.0	96,36	94,73	93,89	94,42	95,31	97,36	93,61	95,44	94,01	94,88	95,00
8.7.0.0	87,63	89,47	89,05	88,34	88,50	94,47	88,25	87,63	88,44	88,38	89,02
8.99.0.0	87,12	86,98	86,93	87,97	86,92	92,08	86,98	86,36	87,86	86,93	87,61
10.1.0.0	100,00	100,00	100,00	100,00	100,00	100,00	100,00	100,00	100,00	100,00	100,00
10.5.0.0	99,12	99,52	99,34	99,10	99,41	99,30	99,34	99,22	99,34	99,34	99,30
11.1.0.0	100,00	100,00	100,00	100,00	100,00	100,00	100,00	100,00	100,00	100,00	100,00
11.10.0.0	100,00	100,00	100,00	100,00	100,00	100,00	100,00	100,00	100,00	100,00	100,00
11.7.0.0	100,00	100,00	99,92	100,00	100,00	100,00	99,97	99,95	99,97	99,95	99,98
11.99.0.0	94,60	82,63	82,04	81,49	82,40	85,81	81,25	82,02	83,67	81,36	83,72
13.1.0.0	100,00	100,00	100,00	100,00	100,00	100,00	100,00	100,00	99,97	100,00	100,00
13.11.0.0	100,00	100,00	100,00	100,00	100,00	100,00	100,00	100,00	100,00	100,00	100,00
14.1.0.0	100,00	100,00	100,00	99,97	100,00	100,00	100,00	100,00	100,00	99,97	99,99
14.10.0.0	88,51	85,83	86,63	86,93	86,39	90,76	85,77	87,36	86,24	86,14	87,06
14.20.0.0	86,93	87,05	87,38	87,33	86,98	92,74	87,25	87,47	87,86	87,43	87,84
14.4.0.0	100,00	100,00	100,00	100,00	100,00	100,00	100,00	100,00	100,00	100,00	100,00
29.99.0.0	88,13	86,98	88,97	88,39	88,24	100,00	88,54	88,73	88,10	89,25	89,53
30.1.0.0	78,39	77,13	79,22	78,43	78,16	81,56	77,19	77,87	78,66	78,27	78,49
30.10.0.0	91,96	87,26	87,59	87,46	87,08	92,78	87,11	87,17	87,09	87,40	88,29
30.16.0.0	86,37	86,08	87,16	87,49	86,92	90,59	86,53	86,60	87,38	87,06	87,22
30.19.0.0	99,94	99,92	99,97	99,89	99,92	99,88	99,92	99,95	99,87	99,89	99,92
30.2.0.0	100,00	99,90	99,81	99,84	99,87	99,83	99,74	99,68	99,87	99,74	99,83
30.22.0.0	100,00	99,95	99,97	99,97	99,95	100,00	100,00	99,97	100,00	99,97	99,98
30.25.0.0	91,27	96,03	96,76	96,68	96,60	94,76	96,49	96,55	96,71	96,17	95,80
30.3.0.0	100,00	99,97	100,00	100,00	100,00	99,96	99,97	100,00	99,97	99,95	99,98
30.4.0.0	84,11	84,05	84,80	85,42	84,99	91,17	84,34	84,10	83,64	85,03	85,17
30.7.0.0	100,00	99,97	100,00	100,00	100,00	100,00	100,00	99,97	99,97	100,00	99,99
30.8.0.0	99,94	99,92	99,92	99,89	99,87	99,83	99,82	99,92	99,95	99,84	99,89
30.9.0.0	100,00	99,82	99,63	99,76	99,79	99,75	99,71	99,54	99,50	99,55	99,70
30.99.0.0	86,81	85,73	87,03	87,65	87,67	100,00	86,74	88,03	86,51	86,93	88,31
40.1.0.0	86,24	86,87	86,05	85,74	87,35	91,50	86,61	87,12	86,61	87,33	87,14
40.10.0.0	78,58	77,33	78,42	79,15	79,50	83,91	77,58	78,06	79,22	79,56	79,13
40.16.0.0	84,80	83,52	83,39	83,35	84,14	88,04	84,21	83,96	82,64	84,18	84,22
40.19.0.0	83,35	86,95	86,77	86,35	86,66	89,56	85,48	85,82	86,45	86,11	86,35
40.2.0.0	83,98	82,63	84,03	83,27	82,88	87,95	83,10	83,42	82,58	82,60	83,64
40.22.0.0	100,00	95,47	95,51	95,01	95,39	99,88	95,80	95,44	95,36	95,35	96,32
40.25.0.0	84,80	87,76	86,90	87,17	87,70	92,53	86,37	86,44	87,62	87,35	87,47
40.27.0.0	87,37	83,85	84,56	85,23	86,07	90,92	83,84	84,99	85,45	83,52	85,58
40.3.0.0	83,61	79,95	79,99	79,87	78,56	83,75	79,48	79,03	78,87	79,77	80,29
40.30.0.0	100,00	87,79	87,80	88,74	88,26	100,00	88,51	88,60	88,89	88,30	90,69
40.4.0.0	82,73	83,90	83,10	83,00	84,99	87,21	83,68	84,15	83,88	84,84	84,15
40.5.0.0	87,00	86,57	87,46	86,37	85,74	91,38	87,99	87,17	87,38	87,22	87,43
40.7.0.0	77,83	80,18	79,94	78,35	78,64	78,14	79,32	78,81	78,47	79,38	78,91
40.8.0.0	85,80	89,09	89,74	88,23	88,72	91,96	89,15	88,22	88,68	88,91	88,85
40.9.0.0	98,30	99,36	99,26	99,44	99,33	99,50	99,05	99,16	99,28	99,26	99,20
40.99.0.0	88,51	86,98	87,78	86,96	87,43	91,79	87,09	87,60	87,20	87,48	87,88
62.2.0.0	100,00	100,00	100,00	100,00	100,00	100,00	100,00	100,00	100,00	100,00	100,00
63.1.0.0	100,00	100,00	99,97	100,00	100,00	100,00	100,00	100,00	99,97	100,00	99,99
63.9.0.0	100,00	87,84	88,07	88,05	88,18	92,33	88,01	88,57	88,71	88,49	89,82
67.1.0.0	100,00	100,00	100,00	100,00	100,00	100,00	100,00	100,00	100,00	100,00	100,00
67.10.0.0	87,31	87,94	87,94	89,51	88,69	91,54	88,75	88,54	87,22	88,65	88,61
67.11.0.0	100,00	87,59	88,39	88,63	88,24	100,00	88,49	88,73	89,21	88,72	90,80
67.13.0.0	98,24	97,58	98,03	98,59	97,94	98,43	97,94	97,71	97,88	97,89	98,02
67.16.0.0	92,78	95,70	95,40	96,15	95,10	99,17	95,46	95,36	96,29	95,48	95,69
67.19.0.0	88,51	87,38	87,51	88,26	87,81	91,87	87,88	88,65	88,63	88,20	88,47
67.28.0.0	98,68	94,45	94,45	94,74	94,75	92,82	94,48	93,99	94,25	94,77	94,74
67.4.0.0	100,00	99,95	99,87	99,95	99,97	100,00	99,92	99,92	99,95	99,89	99,94
67.50.0.0	100,00	100,00	100,00	100,00	100,00	100,00	100,00	100,00	100,00	100,00	100,00

Continua na próxima página

Tabela B.4 – *Continuação da página anterior*

Rótulos	Pheno	Seq	Church	CellCycle	Derisi	Eisen	Gasch2	SPO	Gach1	Expr	Média
67.7.0.0	86,43	85,30	85,78	86,83	85,96	92,57	86,40	86,33	86,77	85,50	86,79
67.99.0.0	84,74	86,31	86,26	87,41	87,22	94,27	85,87	86,23	86,03	86,08	87,04

Tabela B.5: Taxa de acerto das AD de rótulos - BR-RTb pru - Terceiro Nível da Hierarquia

Rótulos	Pheno	Seq	Church	CellCycle	Derisi	Eisen	Gasch2	SPO	Gach1	Expr	Média
1.1.1.0	96,42	97,23	97,10	97,10	97,11	96,70	97,12	97,09	97,11	97,12	97,01
1.1.10.0	98,99	99,16	99,12	99,12	99,06	99,09	99,13	99,08	99,13	99,13	99,10
1.1.4.0	98,99	99,31	99,31	99,28	99,28	98,89	99,29	99,30	99,31	99,29	99,22
1.1.7.0	98,99	99,41	99,39	99,39	99,38	99,17	99,39	99,38	99,39	99,39	99,33
1.1.99.0	99,87	99,87	99,87	99,87	99,87	99,88	99,87	99,87	99,87	99,87	99,87
1.2.1.0	98,99	99,03	98,99	98,99	98,95	99,17	99,00	99,00	98,99	99,00	99,01
1.2.4.0	99,31	99,34	99,34	99,31	99,33	99,17	99,31	99,33	99,34	99,34	99,31
1.20.1.0	100,00	100,00	100,00	100,00	100,00	100,00	100,00	100,00	100,00	100,00	100,00
1.20.17.0	100,00	100,00	100,00	100,00	100,00	100,00	100,00	100,00	100,00	100,00	100,00
1.3.1.0	98,81	98,96	98,91	98,94	98,95	98,89	98,92	98,98	98,91	98,92	98,92
1.3.10.0	99,75	99,80	99,79	99,79	99,79	99,75	99,79	99,78	99,79	99,79	99,78
1.3.13.0	99,69	99,67	99,65	99,65	99,65	99,67	99,66	99,65	99,66	99,66	99,66
1.3.16.0	99,25	99,41	99,44	99,44	99,46	99,26	99,45	99,43	99,44	99,45	99,40
1.3.19.0	99,56	99,64	99,63	99,63	99,65	99,79	99,63	99,65	99,63	99,63	99,64
1.3.4.0	99,31	99,36	99,34	99,34	99,33	98,97	99,34	99,33	99,34	99,34	99,30
1.3.7.0	99,50	99,72	99,71	99,71	99,71	99,67	99,71	99,70	99,71	99,71	99,68
1.3.99.0	99,87	99,85	99,81	99,81	99,81	99,83	99,82	99,81	99,81	99,82	99,83
1.4.1.0	99,69	99,64	99,63	99,63	99,62	99,55	99,63	99,62	99,63	99,63	99,63
1.4.4.0	99,69	99,80	99,79	99,79	99,79	99,67	99,79	99,78	99,79	99,79	99,77
1.4.7.0	99,69	99,75	99,73	99,73	99,73	99,75	99,74	99,73	99,73	99,74	99,73
1.4.99.0	100,00	99,97	99,97	99,97	99,97	100,00	99,97	99,97	99,97	99,97	99,98
1.5.1.0	92,09	93,39	93,12	93,09	93,06	93,65	93,13	93,10	93,11	93,13	93,09
1.5.4.0	96,61	96,95	96,81	96,81	96,70	96,74	96,83	96,77	96,82	96,83	96,79
1.5.7.0	99,43	98,93	98,88	98,88	98,85	98,72	98,89	98,87	98,89	98,89	98,92
1.5.99.0	100,00	99,95	99,95	99,95	99,95	99,96	99,95	99,95	99,95	99,95	99,95
1.6.1.0	96,11	97,05	96,94	96,95	96,86	96,82	96,96	96,90	96,95	96,96	96,85
1.6.10.0	99,50	99,49	99,50	99,50	99,49	99,30	99,47	99,49	99,50	99,47	99,47
1.6.13.0	99,37	99,47	99,44	99,44	99,44	99,42	99,45	99,46	99,44	99,45	99,44
1.6.4.0	99,12	99,36	99,34	99,34	99,36	99,71	99,34	99,35	99,34	99,34	99,36
1.6.7.0	99,69	99,34	99,31	99,31	99,30	99,01	99,31	99,30	99,31	99,31	99,32
1.6.99.0	99,56	99,67	99,68	99,65	99,65	99,79	99,66	99,65	99,66	99,66	99,66
1.7.1.0	98,37	98,40	98,33	98,33	98,29	98,23	98,34	98,30	98,33	98,34	98,32
1.7.10.0	99,94	99,92	99,92	99,92	99,92	99,92	99,92	99,92	99,92	99,92	99,92
1.7.4.0	99,87	99,82	99,81	99,81	99,81	99,79	99,82	99,81	99,81	99,82	99,82
1.7.7.0	99,87	99,92	99,92	99,92	99,92	99,88	99,92	99,92	99,92	99,92	99,91
1.7.99.0	99,87	99,80	99,79	99,79	99,79	99,92	99,79	99,78	99,79	99,79	99,81
2.11.5.0	100,00	99,97	99,97	99,97	99,97	100,00	99,97	99,97	99,97	99,97	99,98
2.11.99.0	100,00	99,97	99,97	99,97	99,97	100,00	99,97	99,97	99,97	99,97	99,98
3.1.3.0	97,99	97,61	97,56	97,56	97,53	96,82	97,52	97,52	97,53	97,52	97,52
3.1.5.0	98,24	98,35	98,27	98,27	98,29	98,64	98,28	98,27	98,28	98,28	98,32
3.1.9.0	100,00	100,00	100,00	100,00	100,00	100,00	100,00	100,00	100,00	100,00	100,00
3.3.1.0	91,02	91,91	91,58	91,58	91,61	91,79	91,60	91,56	91,60	91,60	91,59
3.3.2.0	97,42	97,35	97,26	97,24	97,27	98,18	97,25	97,22	97,27	97,25	97,37
3.3.3.0	98,68	99,11	99,07	99,07	99,04	99,01	99,08	99,06	99,07	99,08	99,03
3.3.4.0	100,00	100,00	100,00	100,00	100,00	100,00	100,00	100,00	100,00	100,00	100,00
4.1.1.0	99,69	99,57	99,55	99,55	99,54	99,30	99,55	99,54	99,55	99,55	99,54
4.1.4.0	98,56	98,40	98,35	98,35	98,31	97,77	98,36	98,33	98,36	98,36	98,32
4.1.99.0	99,69	99,85	99,84	99,84	99,84	99,83	99,84	99,84	99,84	99,84	99,83
4.3.1.0	99,69	99,57	99,55	99,55	99,54	99,30	99,55	99,54	99,55	99,55	99,54
4.3.3.0	99,31	99,08	99,07	99,07	99,06	98,64	99,05	99,06	99,07	99,05	99,05
4.3.6.0	99,50	99,54	99,52	99,52	99,52	99,59	99,52	99,51	99,52	99,52	99,53
4.3.99.0	99,94	99,90	99,89	99,89	99,89	99,83	99,89	99,89	99,89	99,89	99,89
4.5.1.0	99,81	99,77	99,76	99,76	99,76	99,75	99,76	99,76	99,76	99,76	99,77
4.5.5.0	98,87	99,01	98,96	98,96	98,95	98,68	98,97	98,95	98,97	98,97	98,93
4.5.99.0	99,87	99,77	99,79	99,79	99,79	99,92	99,76	99,78	99,79	99,76	99,80
5.4.1.0	99,94	99,95	99,95	99,95	99,95	100,00	99,95	99,95	99,95	99,95	99,95
5.4.2.0	100,00	99,95	99,95	99,95	99,95	99,92	99,95	99,95	99,95	99,95	99,95
6.13.1.0	98,18	97,51	97,40	97,40	97,37	96,58	97,41	97,36	97,40	97,41	97,40
6.13.4.0	99,25	99,54	99,52	99,52	99,52	99,46	99,52	99,51	99,52	99,52	99,49
6.13.99.0	99,56	99,62	99,60	99,60	99,60	99,59	99,60	99,60	99,60	99,60	99,60
6.7.1.0	99,94	99,97	99,97	99,97	99,97	100,00	99,97	99,97	99,97	99,97	99,97
6.7.11.0	99,94	99,97	99,97	99,97	99,97	100,00	99,97	99,97	99,97	99,97	99,97
6.7.2.0	99,94	99,95	99,95	99,95	99,95	99,96	99,95	99,95	99,95	99,95	99,95
6.7.3.0	99,75	99,85	99,84	99,84	99,84	100,00	99,84	99,84	99,84	99,84	99,85
6.7.4.0	99,87	99,90	99,89	99,89	99,89	100,00	99,89	99,89	99,89	99,89	99,90
6.7.99.0	99,87	99,85	99,84	99,84	99,84	99,96	99,84	99,84	99,84	99,84	99,86
10.1.1.0	99,87	99,90	99,89	99,89	99,89	100,00	99,89	99,89	99,89	99,89	99,90
10.1.5.0	100,00	100,00	100,00	100,00	100,00	100,00	100,00	100,00	100,00	100,00	100,00
10.1.9.0	99,94	99,97	99,97	99,97	99,97	99,96	99,97	99,97	99,97	99,97	99,97
10.1.99.0	98,81	99,06	99,02	99,02	98,95	99,59	99,02	99,00	99,02	99,02	99,05

Continua na próxima página

Tabela B.5 – *Continuação da página anterior*

Rótulos	Pheno	Seq	Church	CellCycle	Derisi	Eisen	Gasch2	SPO	Gach1	Expr	Média
11.10.3.0	100,00	99,97	99,97	99,97	99,97	100,00	99,97	99,97	99,97	99,97	99,98
11.7.1.0	99,75	99,90	99,89	99,89	99,95	99,92	99,89	99,95	99,89	99,89	99,89
11.7.99.0	100,00	99,95	99,95	99,95	99,95	100,00	99,95	99,95	99,95	99,95	99,96
13.1.1.0	99,94	99,97	99,97	99,97	99,97	99,96	99,97	99,97	99,97	99,97	99,97
13.1.3.0	100,00	100,00	100,00	100,00	100,00	100,00	100,00	100,00	100,00	100,00	100,00
13.11.3.0	100,00	100,00	100,00	100,00	100,00	100,00	100,00	100,00	100,00	100,00	100,00
14.1.3.0	99,37	99,59	99,57	99,58	99,57	99,88	99,58	99,57	99,58	99,58	99,59
14.10.2.0	100,00	100,00	100,00	100,00	100,00	100,00	100,00	100,00	100,00	100,00	100,00
14.4.3.0	100,00	100,00	100,00	100,00	100,00	100,00	100,00	100,00	100,00	100,00	100,00
30.10.3.0	99,69	99,54	99,55	99,55	99,54	99,46	99,52	99,54	99,52	99,52	99,54
40.10.3.0	98,74	98,88	98,86	98,83	98,85	98,89	98,84	98,84	98,86	98,84	98,84
62.2.5.0	99,62	99,67	99,65	99,65	99,65	99,88	99,66	99,65	99,66	99,66	99,67
67.1.1.0	99,75	99,87	99,84	99,87	99,84	99,88	99,84	99,87	99,87	99,84	99,85
67.4.1.0	100,00	100,00	100,00	100,00	100,00	100,00	100,00	100,00	100,00	100,00	100,00
67.4.7.0	99,62	99,67	99,65	99,65	99,65	99,59	99,66	99,65	99,66	99,66	99,65
67.50.22.0	98,49	98,86	98,94	98,91	98,93	98,60	98,92	98,92	98,94	98,92	98,84
67.50.25.0	99,25	99,29	99,26	99,26	99,25	99,34	99,26	99,25	99,26	99,26	99,27
67.50.99.0	99,94	99,97	99,97	99,97	99,97	100,00	99,97	99,97	99,97	99,97	99,97
67.7.99.0	100,00	99,95	99,95	99,95	99,95	100,00	99,95	99,95	99,95	99,95	99,96

Tabela B.6: Taxa de acerto das AD de rótulos - BR-RTb unpr-pru1 - Terceiro Nível da Hierarquia

Rótulos	Pheno	Seq	Church	CellCycle	Derisi	Eisen	Gasch2	SPO	Gach1	Expr	Média
1.1.1.0	71,23	71,18	72,58	71,61	73,26	72,69	71,61	71,94	71,50	73,36	72,09
1.1.10.0	81,03	72,83	75,05	73,71	74,52	73,51	74,62	74,53	73,44	72,56	74,58
1.1.4.0	73,56	80,82	79,96	80,48	80,41	79,41	80,59	79,76	79,77	81,01	79,58
1.1.7.0	77,32	74,21	75,21	75,35	74,57	73,93	74,02	75,50	75,21	75,42	75,07
1.1.99.0	78,96	74,76	73,96	76,02	75,29	75,17	74,99	75,90	75,64	75,20	75,59
1.2.1.0	82,91	79,01	78,63	77,61	80,06	92,82	78,16	77,76	77,36	77,19	80,15
1.2.4.0	84,67	78,68	78,37	80,21	78,59	78,96	78,90	78,30	78,39	79,64	79,47
1.20.1.0	100,00	100,00	100,00	100,00	100,00	100,00	100,00	100,00	100,00	100,00	100,00
1.20.17.0	100,00	100,00	100,00	100,00	100,00	100,00	100,00	100,00	100,00	100,00	100,00
1.3.1.0	75,25	71,99	74,28	73,33	73,58	73,35	72,72	73,99	73,49	72,09	73,41
1.3.10.0	99,81	80,97	81,34	80,45	80,36	82,71	81,36	80,81	79,85	80,43	82,81
1.3.13.0	89,01	85,12	86,61	86,48	85,88	79,33	86,03	85,85	85,34	87,25	85,69
1.3.16.0	77,83	73,75	74,52	75,35	75,13	75,83	74,91	75,12	74,95	75,28	75,27
1.3.19.0	76,13	74,97	76,69	76,33	77,12	76,98	76,16	75,80	75,85	76,31	76,23
1.3.4.0	75,63	74,94	75,55	75,88	75,96	76,82	75,65	77,65	75,66	75,34	75,91
1.3.7.0	75,06	80,18	81,16	80,58	80,60	86,14	79,72	80,35	80,54	80,41	80,47
1.3.99.0	77,58	75,04	76,24	75,38	77,17	77,76	76,18	76,17	75,90	77,13	76,46
1.4.1.0	77,14	71,97	74,01	72,24	72,56	74,30	72,85	73,53	73,78	72,88	73,53
1.4.4.0	77,01	74,92	75,68	75,86	77,04	76,40	75,81	75,71	76,19	77,13	76,18
1.4.7.0	99,81	99,80	99,71	99,73	99,73	99,92	99,84	99,70	99,89	99,74	99,79
1.4.99.0	100,00	75,12	77,09	75,70	75,91	100,00	76,68	76,60	77,01	76,52	81,06
1.5.1.0	72,42	70,44	69,81	70,78	71,14	70,54	70,11	71,62	70,39	70,74	70,80
1.5.4.0	69,35	67,85	69,07	69,16	69,29	70,71	69,69	70,24	67,63	68,00	69,10
1.5.7.0	79,02	71,76	74,14	73,65	72,91	73,14	73,17	73,72	73,91	72,56	73,80
1.5.99.0	100,00	75,58	76,93	76,25	76,71	78,18	76,42	77,12	76,48	74,99	78,87
1.6.1.0	74,94	71,79	72,79	72,59	72,83	73,18	71,72	72,53	72,64	72,46	72,75
1.6.10.0	80,59	73,70	74,78	74,69	74,06	76,11	75,36	74,58	74,81	73,59	75,23
1.6.13.0	76,13	77,33	78,32	77,40	78,19	87,67	78,29	80,00	77,94	77,42	78,87
1.6.4.0	75,31	72,25	72,39	72,16	72,35	73,18	72,30	72,29	71,50	71,80	72,55
1.6.7.0	72,61	67,13	67,58	67,92	67,93	70,38	68,34	68,49	68,85	69,10	68,83
1.6.99.0	76,70	74,61	75,29	75,43	75,59	76,24	76,26	76,79	75,21	75,63	75,77
1.7.1.0	75,00	73,06	75,31	73,52	75,19	74,67	74,76	74,72	74,15	74,41	74,48
1.7.10.0	77,14	75,60	76,08	75,56	76,74	77,97	76,79	76,28	75,27	75,60	76,30
1.7.4.0	77,64	74,21	76,51	75,67	77,04	79,37	75,97	75,96	76,33	76,13	76,48
1.7.7.0	77,76	81,23	81,48	81,30	81,38	82,01	81,25	80,84	81,44	81,65	81,03
1.7.99.0	76,19	75,53	75,98	76,28	77,14	77,19	75,81	76,74	76,01	76,60	76,35
2.11.5.0	100,00	75,04	76,75	76,12	76,66	100,00	75,44	75,80	76,70	77,13	80,96
2.11.99.0	100,00	74,56	76,99	76,47	76,37	100,00	76,42	76,52	76,54	76,89	81,08
3.1.3.0	73,81	71,20	71,41	70,94	71,84	72,07	70,90	72,32	70,94	71,11	71,65
3.1.5.0	75,44	71,38	73,08	73,89	73,61	73,80	73,83	73,02	72,61	72,78	73,34
3.1.9.0	100,00	100,00	100,00	100,00	100,00	100,00	100,00	100,00	100,00	100,00	100,00
3.3.1.0	68,78	67,85	68,32	67,81	68,22	69,84	68,63	67,74	67,97	67,10	68,23
3.3.2.0	72,05	69,22	68,67	69,56	69,45	70,09	69,47	68,65	69,22	68,39	69,48
3.3.3.0	71,92	72,73	72,15	72,22	70,15	72,77	70,90	70,84	72,35	71,38	71,74
3.3.4.0	100,00	100,00	100,00	100,00	100,00	100,00	100,00	100,00	100,00	100,00	100,00
4.1.1.0	78,02	78,91	79,64	79,20	78,67	78,55	78,80	78,19	79,00	79,27	78,83
4.1.4.0	75,19	72,42	75,02	74,34	73,93	74,50	72,54	73,75	73,78	73,54	73,90
4.1.99.0	76,51	75,91	77,81	76,23	76,34	75,74	76,23	76,77	75,53	76,66	76,37
4.3.1.0	99,81	97,18	97,98	97,72	97,83	97,24	97,68	97,39	97,75	97,83	97,84
4.3.3.0	78,77	73,54	73,98	72,96	73,87	72,90	73,04	73,83	73,01	74,10	74,00
4.3.6.0	75,57	75,12	76,11	76,23	76,66	77,19	77,53	76,71	75,90	76,74	76,37

Continua na próxima página

Tabela B.6 – *Continuação da página anterior*

Rótulos	Pheno	Seq	Church	CellCycle	Derisi	Eisen	Gasch2	SPO	Gach1	Expr	Média
4.3.99.0	77,20	73,26	73,43	74,26	74,09	74,46	73,41	73,80	72,69	73,62	74,02
4.5.1.0	74,94	74,71	76,43	76,63	78,51	76,65	76,89	76,39	75,82	78,03	76,50
4.5.5.0	75,88	71,84	71,83	71,69	72,91	72,90	73,54	73,02	71,66	72,30	72,76
4.5.99.0	76,26	74,51	75,66	76,04	76,13	75,95	75,81	75,88	76,72	76,31	75,93
5.4.1.0	76,51	75,15	76,59	76,25	76,47	100,00	76,29	76,17	76,88	76,05	78,64
5.4.2.0	100,00	74,15	77,73	76,33	76,26	76,53	75,36	76,87	75,61	76,00	78,48
6.13.1.0	72,99	70,06	71,80	71,50	72,05	71,78	71,82	72,88	71,74	71,72	71,84
6.13.4.0	76,13	75,10	76,38	75,75	77,22	77,02	77,00	77,33	75,61	75,84	76,34
6.13.99.0	76,82	75,27	75,50	76,33	76,47	76,86	76,76	76,47	75,66	76,05	76,22
6.7.1.0	77,39	74,76	77,52	76,92	76,05	100,00	76,26	76,04	77,25	76,58	78,88
6.7.11.0	76,95	74,59	77,54	76,73	76,29	100,00	75,81	76,01	77,47	76,50	78,79
6.7.2.0	77,26	72,14	73,85	72,86	72,80	96,70	73,46	73,75	72,77	72,70	75,83
6.7.3.0	76,26	74,94	76,24	76,04	77,14	100,00	76,37	76,31	76,46	76,39	78,62
6.7.4.0	78,33	74,82	76,59	76,02	76,63	100,00	77,11	76,31	75,87	76,31	78,80
6.7.99.0	100,00	75,22	75,74	76,28	76,82	99,96	76,79	76,52	76,14	76,13	80,96
10.1.1.0	77,45	75,48	76,35	76,18	76,50	100,00	76,52	77,57	75,85	76,29	78,82
10.1.5.0	100,00	100,00	100,00	100,00	100,00	100,00	100,00	100,00	100,00	100,00	100,00
10.1.9.0	99,94	100,00	100,00	99,97	99,97	100,00	100,00	100,00	99,97	99,97	99,98
10.1.99.0	75,31	70,31	70,24	70,81	70,87	80,12	70,74	71,89	71,24	69,92	72,15
11.10.3.0	100,00	99,62	99,71	99,68	99,89	100,00	99,74	99,89	99,92	99,79	99,82
11.7.1.0	99,37	99,36	99,28	99,28	99,01	100,00	99,08	98,95	99,15	99,29	99,28
11.7.99.0	100,00	99,87	99,76	99,95	99,92	100,00	99,89	99,95	99,92	99,92	99,92
13.1.1.0	76,38	75,02	77,31	75,56	76,02	77,64	75,94	76,58	77,01	76,66	76,41
13.1.3.0	100,00	100,00	100,00	100,00	100,00	100,00	100,00	100,00	100,00	100,00	100,00
13.11.3.0	100,00	100,00	100,00	100,00	100,00	100,00	100,00	100,00	100,00	100,00	100,00
14.1.3.0	90,83	92,93	93,49	93,63	93,30	97,19	92,63	93,26	93,45	93,50	93,42
14.10.2.0	100,00	100,00	100,00	100,00	100,00	100,00	100,00	100,00	100,00	100,00	100,00
14.4.3.0	100,00	100,00	100,00	100,00	100,00	100,00	100,00	100,00	100,00	100,00	100,00
30.10.3.0	78,45	78,25	76,96	77,93	76,66	80,24	78,98	77,82	79,16	78,29	78,27
40.10.3.0	72,99	70,67	72,02	72,01	71,95	73,60	71,72	72,40	72,00	71,96	72,13
62.2.5.0	90,08	85,37	84,61	85,71	85,37	97,65	85,40	85,15	86,06	85,77	87,12
67.1.1.0	78,14	74,82	76,27	76,84	76,82	76,57	75,97	76,47	76,72	75,07	76,37
67.4.1.0	100,00	100,00	100,00	100,00	100,00	100,00	100,00	100,00	100,00	100,00	100,00
67.4.7.0	76,95	77,54	76,69	77,72	75,64	75,62	76,82	76,71	77,33	76,89	76,79
67.50.22.0	77,07	75,93	76,03	75,03	76,90	77,72	76,13	76,98	76,56	74,81	76,32
67.50.25.0	78,08	75,45	75,63	75,75	76,34	76,90	76,55	75,93	76,25	75,20	76,21
67.50.99.0	77,20	75,43	76,99	76,10	76,55	100,00	75,42	76,39	76,86	77,19	78,81
67.7.99.0	100,00	74,87	77,89	76,04	76,69	100,00	76,26	76,25	76,80	75,84	81,06

Tabela B.7: Taxa de acerto das AD de rótulos - BR-RTb pru - Quarto Nível da Hierarquia

Rótulos	Pheno	Seq	Church	CellCycle	Derisi	Eisen	Gasch2	SPO	Gach1	Expr	Média
1.1.1.11	100,00	99,95	99,95	99,95	99,97	99,96	99,95	99,97	99,95	99,95	99,96
1.1.1.15	100,00	99,97	99,97	99,97	99,97	99,96	99,97	99,97	99,97	99,97	99,97
1.1.1.7	99,94	99,97	99,97	99,97	99,97	99,96	99,97	99,97	99,97	99,97	99,97
1.1.10.1	99,94	99,97	99,97	99,97	99,97	99,96	99,97	99,97	99,97	99,97	99,97
1.1.10.5	100,00	99,97	99,97	99,97	99,97	100,00	99,97	99,97	99,97	99,97	99,98
1.20.1.7	99,94	99,97	99,97	99,97	99,97	100,00	99,97	99,97	99,97	99,97	99,97
1.20.17.3	99,75	99,90	99,89	99,89	99,89	99,83	99,89	99,89	99,89	99,89	99,87
1.3.16.1	99,87	99,90	99,89	99,89	99,89	100,00	99,89	99,89	99,89	99,89	99,90
1.5.1.3	100,00	99,97	99,97	99,97	99,97	100,00	99,97	99,97	99,97	99,97	99,98
1.6.1.1	99,94	99,95	99,95	99,95	99,95	100,00	99,95	99,95	99,95	99,95	99,95
1.6.1.3	100,00	99,97	99,97	99,97	99,97	100,00	99,97	99,97	99,97	99,97	99,98
1.6.1.7	100,00	99,97	99,97	99,97	99,97	100,00	99,97	99,97	99,97	99,97	99,98
3.1.5.1	97,74	97,76	97,69	97,72	97,70	96,95	97,70	97,68	97,69	97,70	97,63
3.1.9.5	99,12	99,19	99,15	99,15	99,14	98,84	99,16	99,14	99,15	99,16	99,12
3.3.1.3	99,25	99,31	99,28	99,28	99,28	99,17	99,29	99,27	99,28	99,29	99,27
3.3.4.3	100,00	99,97	99,97	99,97	99,97	100,00	99,97	99,97	99,97	99,97	99,98
3.3.4.5	99,94	99,95	99,95	99,95	99,95	100,00	99,95	99,95	99,95	99,95	99,95
4.5.1.1	99,12	98,27	98,19	98,19	98,18	97,61	98,20	98,19	98,20	98,20	98,24
4.5.1.4	90,33	91,50	91,15	91,16	91,05	89,56	91,21	91,02	91,17	91,23	90,94
4.5.5.1	98,12	97,43	97,61	97,61	97,67	97,24	97,62	97,65	97,61	97,62	97,62
4.5.5.5	100,00	99,95	99,95	99,95	99,95	99,96	99,95	99,95	99,95	99,95	99,95
6.13.1.1	100,00	99,97	99,97	99,97	99,97	100,00	99,97	99,97	99,97	99,97	99,98
6.13.4.2	99,94	99,97	99,97	99,97	99,97	100,00	99,97	99,97	99,97	99,97	99,97
10.1.5.5	99,69	99,69	99,68	99,68	99,68	99,96	99,68	99,68	99,68	99,68	99,71
10.1.9.11	100,00	99,97	99,97	99,97	99,97	100,00	99,97	99,97	99,97	99,97	99,98
13.1.1.1	98,74	98,80	98,75	98,75	98,74	98,93	98,76	98,73	98,75	98,76	98,77
13.1.1.3	99,18	99,31	99,36	99,36	99,36	99,17	99,37	99,35	99,36	99,37	99,32
13.1.1.99	99,31	99,41	99,39	99,39	99,38	99,34	99,39	99,38	99,39	99,39	99,38
13.1.3.1	99,87	99,92	99,92	99,92	99,92	99,92	99,92	99,92	99,92	99,92	99,92
13.1.3.3	99,75	99,82	99,81	99,81	99,81	99,83	99,82	99,81	99,81	99,82	99,81
13.1.3.5	100,00	99,95	99,95	99,95	99,95	99,96	99,95	99,95	99,95	99,95	99,95
13.1.3.99	100,00	99,97	99,97	99,97	99,97	99,96	99,97	99,97	99,97	99,97	99,97
13.11.3.1	99,12	99,41	99,39	99,39	99,38	99,63	99,39	99,38	99,39	99,39	99,39
13.11.3.13	99,43	99,57	99,55	99,55	99,52	99,88	99,55	99,54	99,55	99,55	99,57

Continua na próxima página

Tabela B.7 – *Continuação da página anterior*

Rótulos	Pheno	Seq	Church	CellCycle	Derisi	Eisen	Gasch2	SPO	Gach1	Expr	Média
13.11.3.7	99,50	99,39	99,36	99,36	99,33	99,46	99,37	99,33	99,36	99,37	99,38
14.1.3.99	99,69	99,69	99,68	99,68	99,68	99,59	99,68	99,68	99,68	99,68	99,67
14.10.2.1	99,94	99,97	99,97	99,97	99,97	100,00	99,97	99,97	99,97	99,97	99,97
14.4.3.1	95,04	95,96	95,85	95,83	95,87	95,38	95,85	95,88	95,86	95,85	95,74
14.4.3.3	96,98	96,72	96,63	96,60	96,54	95,92	96,59	96,55	96,61	96,59	96,57
14.4.3.5	96,48	97,10	96,97	96,97	96,95	97,61	96,99	96,93	96,98	96,99	97,00
67.4.1.1	99,50	99,41	99,39	99,39	99,36	99,22	99,39	99,35	99,20	99,39	99,36
67.4.1.2	98,99	99,16	99,20	99,20	99,20	98,93	99,21	99,19	99,39	99,21	99,17

Tabela B.8: Taxa de acerto das AD de rótulos - BR-RTb unpr-pru1 - Quarto Nível da Hierarquia

Rótulos	Pheno	Seq	Church	CellCycle	Derisi	Eisen	Gasch2	SPO	Gach1	Expr	Média
1.1.1.11	100,00	64,46	65,40	65,21	66,32	66,79	65,12	65,90	64,74	64,80	68,87
1.1.1.15	100,00	64,03	64,68	65,10	65,73	66,58	65,57	65,58	65,32	65,59	68,82
1.1.1.7	65,08	64,49	64,68	65,15	65,68	66,63	65,62	65,61	65,48	65,36	65,38
1.1.10.1	64,57	64,11	64,82	65,15	65,86	67,86	65,33	65,71	65,32	65,41	65,41
1.1.10.5	100,00	63,65	65,19	65,13	65,84	100,00	65,78	65,34	64,37	65,43	72,07
1.20.1.7	65,20	64,46	64,55	64,94	66,69	100,00	64,88	65,61	65,16	65,70	68,72
1.20.17.3	65,14	63,16	64,31	64,83	62,89	66,05	64,43	64,45	64,10	63,69	64,31
1.3.16.1	65,89	64,05	64,02	65,68	64,28	100,00	65,33	65,44	65,11	65,33	68,51
1.5.1.3	100,00	64,21	64,50	65,37	66,59	100,00	65,75	66,01	64,66	65,51	72,26
1.6.1.1	66,58	74,31	74,70	74,24	74,73	100,00	74,23	74,69	74,73	74,23	76,24
1.6.1.3	100,00	100,00	100,00	100,00	99,97	100,00	100,00	100,00	99,97	99,97	99,99
1.6.1.7	100,00	64,51	64,36	65,18	65,59	100,00	64,91	65,69	65,38	64,62	72,02
3.1.5.1	61,75	60,24	60,27	61,59	60,74	64,48	61,08	61,46	61,27	60,92	61,38
3.1.9.5	71,73	67,39	67,29	67,52	67,36	69,35	67,68	68,22	66,89	67,15	68,06
3.3.1.3	64,38	61,56	63,03	62,18	61,09	63,28	61,53	63,05	61,77	62,56	62,44
3.3.4.3	100,00	64,41	64,31	65,31	65,57	100,00	65,01	65,77	65,35	65,04	72,08
3.3.4.5	100,00	74,94	74,70	74,24	74,73	100,00	74,23	74,18	74,20	74,23	79,55
4.5.1.1	61,75	60,67	59,74	60,50	61,04	61,72	61,39	60,24	60,95	60,95	60,89
4.5.1.4	58,92	56,02	56,05	56,79	56,59	57,67	57,27	56,74	56,57	57,38	57,00
4.5.5.1	64,70	64,51	65,00	63,90	63,91	64,77	64,54	64,07	63,65	63,61	64,27
4.5.5.5	100,00	64,84	65,59	65,84	64,55	67,29	64,80	64,53	65,40	64,22	68,71
6.13.1.1	100,00	64,56	64,26	65,34	66,45	100,00	64,96	65,93	65,27	65,33	72,21
6.13.4.2	66,14	64,26	64,76	65,13	65,68	100,00	66,09	65,34	64,74	65,41	68,75
10.1.5.5	88,88	93,77	93,75	93,81	93,73	99,75	94,03	93,42	94,19	94,14	93,95
10.1.9.11	100,00	64,08	65,08	64,94	65,76	100,00	66,33	65,20	64,77	65,14	72,13
13.1.1.1	67,78	70,11	69,31	69,99	69,51	76,20	68,58	70,43	68,77	69,40	70,01
13.1.1.3	83,10	80,64	82,30	83,08	82,64	80,24	83,05	82,99	83,78	83,71	82,55
13.1.1.99	75,13	83,21	83,10	82,79	84,62	79,33	82,97	82,61	83,06	83,13	81,99
13.1.3.1	66,14	64,51	65,32	65,23	64,58	65,06	65,22	64,37	65,01	64,09	64,95
13.1.3.3	65,45	64,67	64,92	66,27	64,34	66,71	65,43	65,55	65,24	64,75	65,33
13.1.3.5	100,00	74,28	74,59	74,56	74,68	67,57	74,15	74,58	74,68	74,73	76,38
13.1.3.99	100,00	99,97	99,87	99,92	99,92	99,96	99,89	99,97	99,95	99,95	99,94
13.11.3.1	80,03	82,70	81,29	81,86	82,05	77,06	81,01	81,56	80,51	81,17	80,93
13.11.3.13	99,56	99,80	99,81	99,73	99,68	99,96	99,84	99,70	99,81	99,71	99,76
13.11.3.7	92,09	93,92	95,14	94,74	94,35	93,52	94,56	94,61	94,43	95,01	94,24
14.1.3.99	96,48	90,71	90,09	90,01	90,09	94,64	90,39	89,54	90,77	90,47	91,32
14.10.2.1	100,00	100,00	99,97	100,00	100,00	100,00	99,97	100,00	100,00	100,00	99,99
14.4.3.1	60,87	63,01	62,93	63,43	62,81	62,62	63,45	62,99	62,67	63,22	62,80
14.4.3.3	81,28	75,96	77,15	76,68	78,32	76,16	76,89	76,25	76,96	76,29	77,19
14.4.3.5	63,07	59,88	59,85	60,48	59,43	69,27	58,73	59,95	59,73	59,49	60,99
67.4.1.1	86,93	90,56	90,88	91,08	90,43	95,96	91,18	90,62	90,40	90,68	90,87
67.4.1.2	90,52	88,91	88,31	88,95	88,00	88,53	89,31	86,98	87,80	87,88	88,52