UNIVERSIDADE FEDERAL DE OURO PRETO

# Desenvolvimento de Técnicas de Seleção de Atributos no Contexto da Classificação Hierárquica Monorrótulo

Thieres Nardy Dias
Universidade Federal de Ouro Preto

Orientador: Prof. Dr. Luiz Henrique de Campos Merschmann

Dissertação submetida ao Instituto de Ciências Exatas e Biológicas da Universidade Federal de Ouro Preto para obtenção do título de Mestre em Ciência da Computação.

Ouro Preto, Dezembro de 2015

# Desenvolvimento de Técnicas de Seleção de Atributos no Contexto da Classificação Hierárquica Monorrótulo

Thieres Nardy Dias
Universidade Federal de Ouro Preto

Orientador: Prof. Dr. Luiz Henrique de Campos Merschmann

UFOP

Universidade Federal
de Ouro Preto

D541d Dias, Thieres.
Desenvolvimento de técnicas de seleção de atributos no contexto da classificação hierárquica monorrótulo [manuscrito] / Thieres Dias. - 2015.
66f.: il.: color; tabs.

Orientador: Prof. Dr. Luiz Henrique Merschmann.

Dissertação (Mestrado) - Universidade Federal de Ouro Preto. Instituto de Ciências Exatas e Biológicas. Departamento de Computação. Programa de Pós-Graduação em Ciência da Computação.
Área de Concentração: Ciência da Computação.

1. Sistemas auto-organizadores. 2. Programação heurística. I. Merschmann, Luiz Henrique. II. Universidade Federal de Ouro Preto. III. Titulo.

CDU: 004.023

Catalogação: www.sisbin.ufop.br

**Ata da Defesa Pública de Dissertação de Mestrado**

Aos 18 dias do mês de dezembro de 2015, às 15 horas na Sala de Seminários do DECOM no Instituto de Ciências Exatas e Biológicas (ICEB), reuniram-se os membros da banca examinadora composta pelos professores: **Prof. Dr. Luiz Henrique de Campos Merschmann (presidente e orientador), Prof. Dr. Haroldo Gambini Santos e Prof. Dr. Alexandre Plastino de Carvalho**, aprovada pelo Colegiado do Programa de Pós-Graduação em Ciência da Computação, a fim de arguirem o mestrando **Thieres Nardy Dias**, com o título **"Desenvolvimento de Técnicas de Seleção de Atributos no Contexto da Classificação Hierárquica Monorrótulo"**. Aberta a sessão pelo presidente, coube ao candidato, na forma regimental, expor o tema de sua dissertação, dentro do tempo regulamentar, sendo em seguida questionado pelos membros da banca examinadora, tendo dado as explicações que foram necessárias.

Recomendações da Banca:

(X) Aprovada sem recomendações

( ) Reprovada

( ) Aprovada com recomendações: _____

Banca Examinadora:

_____
Prof. Dr. Luiz Henrique de Campos Merschmann

_____
Prof. Dr. Haroldo Gambini Santos

_____
Prof. Dr. Alexandre Plastino de Carvalho

_____
Prof. Dr. Luiz Henrique de Campos Merschmann
Coordenador do Programa de Pós-Graduação em Ciência da Computação
DECOM/ICEB/UFOP

**Ouro Preto, 18 de dezembro de 2015.**

*Dedico este trabalho a minha querida mãe Maria, que sempre me ensina o caminho certo para se seguir; e a minha namorada Camylla, que sinto muito carinho e orgulho por fazer parte da minha vida.*

# Desenvolvimento de Técnicas de Seleção de Atributos no Contexto da Classificação Hierárquica Monorrótulo

## Resumo

A seleção de atributos, tradicionalmente adotada como uma etapa de pré-processamento dos dados, tem como objetivo principal identificar os atributos relevantes para a tarefa de classificação. No entanto, para o cenário de classificação hierárquica, onde as classes a serem preditas estão estruturadas de acordo com uma hierarquia, poucos trabalhos na literatura apresentam propostas de técnicas de seleção de atributos. Mais especificamente, para problemas de classificação hierárquica monorrótulo, não foram encontradas na literatura técnicas de seleção de atributos que possam ser utilizadas em conjunto com classificadores hierárquicos globais, ou seja, classificadores que são treinados levando-se em consideração toda a hierarquia de classes de uma só vez.

Desse modo, neste trabalho propomos uma adaptação da medida Incerteza Simétrica (*Symmetrical Uncertainty* – $SU$) para permitir que ela possa ser utilizada em técnicas de seleção de atributos para problemas de classificação hierárquica monorrótulo que usam classificadores hierárquicos globais. Posteriormente, utilizamos essa adaptação proposta, denominada Incerteza Simétrica Hierárquica (*Hierarchical Symmetrical Uncertainty* – $SU_H$), em duas técnicas distintas de seleção de atributos: uma que faz uso da abordagem Filtro e outra que segue uma abordagem Híbrida (Filtro e *Wrapper*). A técnica que implementa a abordagem Híbrida corresponde a uma heurística que utiliza o classificador hierárquico *Global-Model Naive Bayes* ($GMNB$) para avaliar os subconjuntos de atributos.

A partir das duas técnicas de seleção de atributos propostas neste trabalho, pudemos verificar a adequação da adaptação da medida $SU$ para o cenário hierárquico. Além disso,

o método heurístico proposto, nomeado como *Hybrid Feature Selection for Hierarchical Classification* (*HFS4HC*), apresentou resultados bastante promissores para o contexto da classificação hierárquica monorrótulo.

**Palavras-chave:** Aprendizado de Máquina, Classificação Hierárquica Monorrótulo, Seleção de Atributos, Heurística.

# Development of Techniques for Feature Selection in the Context of Hierarchical Single Label Classification

## Abstract

Feature selection, usually adopted as a preprocessing step, aims at identifying as much relevant features as possible with the goal of improving classification accuracy. However, for hierarchical classification scenario, where the classes to be predicted are arranged in a hierarchy, there are few studies in literature that address feature selection techniques. More specifically, for hierarchical single-label classification problems, to the best of our knowledge, there is no work in the literature that addresses feature selection in conjunction with global hierarchical classifiers.

Thus, in this work we propose an adaptation of the measure Symmetrical Uncertainty (SU) to allow it to be used in feature selection techniques for hierarchical single label classification problems using global hierarchical classifiers. Thereafter, we used this adaptation proposal called Hierarchical Symmetrical Uncertainty ($SU_H$) in two distinct techniques for feature selection: one makes use of the filter approach and another follows a hybrid approach (filter and wrapper). The technique that implements a hybrid approach corresponds to a heuristic that uses the hierarchical classifier Global-Model Naive Bayes (GMNB) for assessing the feature subsets.

From the two feature selection techniques proposed in this work, we could verify the appropriateness of the measure $SU$ tailored to hierarchical context. Besides, the proposed heuristic method, called Hybrid Feature Selection for Hierarchical Classification (HFS4HC), presented promising results for the context of hierarchical single label classification.

**Keywords:** Machine Learning, Hierarchical Single Label Classification, Feature Selection, Heuristic.

# Declaração

Esta dissertação é resultado de meu próprio trabalho, exceto onde referência explícita é feita ao trabalho de outros, e não foi submetida para outra qualificação nesta nem em outra universidade.

Thieres Nardy Dias

# Agradecimentos

Primeiramente, agradeço o meu orientador Luiz Henrique de Campos Merschmann por sua paciência em me ensinar, por sua disponibilidade em realizar várias reuniões durante o mestrado, por seu pragmatismo na condução das pesquisas, e por todo o esforço que dedicou à conclusão deste trabalho.

Agradeço aos professores Alan Robert Resende de Freitas, Anderson Almeida Ferreira e Haroldo Gambini Santos por tudo que me ensinaram durante minha jornada no curso de mestrado em Ciência da Computação da UFOP. Agradeço, também, a todos os funcionários do DECOM, principalmente a Mariana Ferreira Lanna. Agradeço à Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES) pelo suporte financeiro.

Gostaria de agradecer, também, aos amigos Luciano Perdigão Cota e André Luís Gomes Carvalho Pires que diretamente contribuíram para a realização deste trabalho.

# Sumário

# Lista de Figuras

# Lista de Tabelas

# Lista de Algoritmos

*"Eu tive muitas coisas que guardei em minhas mãos, e as perdi. Mas tudo o que eu guardei nas mãos de Deus, eu ainda possuo."*

— Martin Luther King

# Nomenclatura

| | |
|---|---|
| CFS | *Correlation-Based Feature Subset Selection* |
| GDA | Grafo Direcionado Acíclico |
| GMNB | *Global-Model Naive Bayes* |
| hF | *hierarchical F-measure* |
| HFS4HC | *Hybrid Feature Selection for Hierarchical Classification* |
| IG | *Information Gain* |
| IWSS | *Incremental Wrapper Subset Selection* |
| k-NN | *k-Nearest Neighbors* |
| LRC | Lista Restrita de Candidatos |
| MLNP | *Mandatory Leaf Node Prediction* |
| NMLNP | *Non-Mandatory Leaf Node Prediction* |
| SBFS | *Sequential Backward Floating Selection* |
| SBS | *Sequential Backward Selection* |
| SFFS | *Sequential Forward Floating Selection* |
| SFS | *Sequential Forward Selection* |
| SU | *Symmetrical Uncertainty* |
| SVM | *Support Vector Machine* |
| WEKA | *Waikato Environment for Knowledge Analysis* |

# Capítulo 1

# Introdução

A crescente quantidade de dados disponíveis em ambientes computacionais propiciou o surgimento da área de pesquisa e aplicação em Ciência da Computação conhecida como Mineração de Dados (Fayyad et al., 1996). De forma simples, tarefas em Mineração de Dados podem ser definidas como processos automatizados de descoberta de novas informações a partir de grandes massas de dados.

Dentre todas as tarefas da área de Mineração de Dados, uma de grande destaque e importância é a de classificação de dados (Han et al., 2011; Witten et al., 2011). Ela tem por objetivo predizer a classe de uma nova instância, utilizando, para essa finalidade, as características (os valores dos atributos preditores) daquela instância em particular.

Grande parte dos esforços em pesquisas nas áreas de Mineração de Dados e Aprendizado de Máquina são direcionados aos problemas de classificação plana, onde cada instância é associada a uma ou mais classes pertencentes a um conjunto predefinido de classes, sendo que não há qualquer tipo de relacionamento entre elas. No entanto, existem problemas de classificação mais complexos, onde as classes a serem preditas estão estruturadas de acordo com uma hierarquia, tal como uma árvore ou um grafo direcionado acíclico (GDA) (Silla Jr. e Freitas, 2011). Para resolver esses problemas, conceitos e técnicas de classificação hierárquica vêm sendo apresentados em trabalhos nessa área de pesquisa que é relativamente nova e, portanto, ainda muito pouco explorada.

Existem na literatura diferentes abordagens para tratar problemas de classificação hierárquica: abordagens locais e abordagem global (Silla Jr. e Freitas, 2011). Enquanto as abordagens locais empregam um conjunto de classificadores locais (planos), na abordagem global somente um classificador é construído e utilizado com o objetivo de explorar,

de uma única vez, os relacionamentos entre as classes. Além disso, os problemas de classificação hierárquica podem ser monorrótulo ou multirrótulo. Nos problemas de classificação hierárquica monorrótulo apenas uma classe pode ser atribuída a uma nova instância. Já nos problemas de classificação hierárquica multirrótulo pode-se atribuir a uma determinada instância uma ou mais classes.

Os problemas de classificação hierárquica são encontrados em diversas áreas, tais como bioinformática (Secker et al., 2010), classificação de documentos (Koller e Sahami, 1997; Mladenic e Grobelnik, 2003), processamento de imagens e áudio (Barutcuoglu e DeCoro, 2006; Silla Jr. e Kaestner, 2013), dentre outras.

A seleção de atributos é uma área de pesquisa ativamente estudada pelas comunidades de Mineração de Dados e de Aprendizado de Máquina (Guyon e Elisseeff, 2006; Liu et al., 2010). Tradicionalmente adotada como uma etapa de pré-processamento dos dados, o seu principal objetivo é identificar os atributos relevantes para a tarefa de classificação visando os seguintes benefícios (Blum e Langley, 1997; Hall e Smith, 1999): (i) melhora da capacidade preditiva dos classificadores; (ii) redução do tempo gasto no processo de classificação; e (iii) obtenção de modelos de classificação mais compactos e, portanto, mais fáceis de serem interpretados.

Apesar da comprovada importância da seleção de atributos para a tarefa de classificação, poucos trabalhos na literatura apresentam propostas de técnicas de seleção de atributos para o contexto da classificação hierárquica. Mais especificamente, para problemas de classificação hierárquica monorrótulo, não foram encontradas na literatura propostas de técnicas de seleção de atributos que possam ser utilizadas em conjunto com classificadores hierárquicos globais, ou seja, classificadores que são treinados levando-se em consideração toda a hierarquia de classes de uma só vez.

Assim, neste trabalho, propomos uma adaptação da medida Incerteza Simétrica (*Symmetrical Uncertainty – SU*) para permitir que ela possa ser utilizada em técnicas de seleção de atributos para problemas de classificação hierárquica monorrótulo que utilizam classificadores hierárquicos globais. Em seguida, utilizamos essa adaptação proposta, denominada Incerteza Simétrica Hierárquica (*Hierarchical Symmetrical Uncertainty – $SU_H$*), em duas técnicas de seleção de atributos: uma que utiliza a abordagem Filtro e outra que implementa uma abordagem Híbrida (Filtro e *Wrapper*). A técnica que implementa a abordagem Híbrida corresponde a uma heurística, também proposta neste trabalho, que utiliza o classificador hierárquico *Global-Model Naive Bayes* (*GMNB*), proposto pelos autores Silla Jr. e Freitas (2009), para avaliar os subconjuntos de atributos.

A partir das duas técnicas de seleção de atributos utilizadas neste trabalho pudemos verificar a adequação da adaptação da medida *SU* para o contexto hierárquico. Além disso, o método heurístico proposto, chamado de *Hybrid Feature Selection for Hierarchical Classification* (*HFS4HC*), apresentou resultados bastante promissores para o cenário de classificação hierárquica monorrótulo.

O restante deste trabalho encontra-se organizado da forma descrita a seguir. O Capítulo 2 apresenta os trabalhos relacionados à seleção de atributos no contexto de classificação hierárquica. Já o Capítulo 3 tem como objetivo fornecer o referencial teórico que embasou o desenvolvimento deste trabalho. A adaptação da medida Incerteza Simétrica e a sua utilização em dois métodos de seleção de atributos propostos para o contexto da classificação hierárquica monorrótulo são apresentados no Capítulo 4. Em seguida, no Capítulo 5, são descritos os experimentos computacionais realizados para validar a medida e os métodos de seleção de atributos aqui propostos e discutidos os resultados obtidos. Por fim, o Capítulo 6 apresenta a conclusão deste trabalho e indica alguns direcionamentos para trabalhos futuros.

# Capítulo 2

# Trabalhos Relacionados

No cenário da classificação hierárquica, onde as classes a serem preditas estão estruturadas de acordo com uma hierarquia preestabelecida, poucos trabalhos na literatura apresentam propostas de técnicas de seleção de atributos.

No trabalho de Koller e Sahami (1997), o problema de classificação de documentos (cujas classes representam uma hierarquia de tópicos) foi tratado por meio da abordagem de classificação local por nó em conjunção com a seleção de atributos, chamado pelos autores de *Framework* Probabilístico, pois o foco é nos métodos probabilísticos para seleção de atributos e para classificação (Koller e Sahami, 1997). Nesse, um classificador binário é construído para cada nó da hierarquia de classes e um método de seleção de atributos é aplicado visando identificar os atributos mais relevantes para a construção de cada um dos classificadores locais. O método de seleção de atributos utiliza uma medida da teoria da informação que foi proposta pelos autores Koller e Sahami (1995). Como resultado dessa aplicação, além da melhora da acurácia preditiva, a redução do número de atributos possibilitou a utilização de classificadores mais robustos e complexos.

Em Secker et al. (2010), os autores resolveram um importante problema de predição de funções de proteínas realizando a seleção de atributos de forma conjunta com a abordagem de classificação hierárquica local por nó. Nesse trabalho os autores utilizaram a estratégia de classificação hierárquica *top-down* com seleção de classificador e seleção de atributos para cada base de dados e para cada nó da hierárquica. Assim, em cada nó onde um classificador é construído, uma etapa de seleção de atributos é executada objetivando reduzir a dimensionalidade dos dados da base daquele nó específico. O método de seleção

de atributos proposto utiliza a medida *Correlation-based Feature Selection* (*CFS*), bem como o algoritmo de busca *Best First* – ambos disponíveis no *toolkit* de mineração de dados *WEKA* (Garner, 1995; Hall et al., 2009). A investigação foi conduzida com o intuito de responder se a seleção de atributos pode melhorar a eficiência computacional sem comprometer a acurácia no contexto da predição de funções de proteína. Os experimentos computacionais demonstraram que esse sistema *top-down* proposto reduz significativamente o tempo necessário para treinar e testar o modelo de classificação, ainda que mantendo a acurácia preditiva.

Em Paes et al. (2014), os autores exploraram o uso de técnicas de seleção de atributos visando melhorar o desempenho preditivo da classificação. Foram avaliadas duas abordagens de classificação hierárquica distintas: local por nó pai e local por nível. Esse trabalho propôs um método de seleção de atributos que produz um *ranking* dos atributos através da medida *Information Gain (IG)* (Cover e Thomas, 1991). Após formado o *ranking*, os $n$ melhores atributos são selecionados, com $n$ sendo um parâmetro de entrada do método. Além disso, esse trabalho também avaliou para o contexto da classificação hierárquica uma estratégia *lazy* de seleção de atributos (Pereira et al., 2008, 2011a,b), a qual posterga a seleção dos atributos ao momento da classificação de uma nova instância. Os experimentos foram conduzidos a partir de bases de dados provenientes da área de bioinformática. Os resultados desse trabalho apontam que os melhores resultados obtidos pelos classificadores ocorreram quando alguma estratégia de seleção de atributos foi adotada.

Vale observar que, em todos os trabalhos supramencionados, a aplicação de técnicas de seleção de atributos e a construção de classificadores foram realizados decompondo-se o problema de classificação hierárquica em vários problemas de classificação plana, o que os permitiu utilizar técnicas de seleção de atributos e algoritmos de classificação tradicionalmente adotados em trabalhos de classificação plana. Até onde conhecemos, apenas o trabalho Slavkov et al. (2014), relacionado com classificação hierárquica multirrótulo, propôs uma técnica de seleção de atributos capaz de lidar com a hierarquia de classes como um todo, ou seja, sem a decomposição do problema hierárquico em diversos problemas de classificação plana. Esse trabalho realizou uma adaptação do algoritmo *ReliefF* (Robnik-Sikonja e Kononenko, 2003) para o contexto hierárquico multirrótulo.

No entanto, no cenário da classificação hierárquica monorrótulo, não foram encontradas na literatura técnicas de seleção de atributos que possam ser utilizadas em conjunto com classificadores hierárquicos globais. Portanto, assim como realizado em Slavkov et al.

(2014) para o contexto hierárquico multirrótulo, neste trabalho propomos a adaptação de uma medida de avaliação da capacidade preditiva de um atributo (Incerteza Simétrica – $SU$) para o contexto hierárquico monorrótulo e, em seguida, utilizamos essa medida em duas abordagens de seleção de atributos.

# Capítulo 3

# Referencial Teórico

Este capítulo tem por finalidade apresentar o referencial teórico que foi utilizado no desenvolvimento deste trabalho. A Seção 3.1 trata da classificação hierárquica e suas definições gerais. A Seção 3.2 apresenta o classificador hierárquico utilizado neste trabalho, o *Global-Model Naive Bayes* (*GMNB*). Em seguida, a Seção 3.3 mostra uma métrica comumente adotada na avaliação do desempenho preditivo de classificadores hierárquicos. Ainda tratando de avaliação de classificadores, o método $k$-validação cruzada é descrito na Seção 3.4. Por fim, a Seção 3.5 finaliza o capítulo com uma breve discussão sobre a importância da seleção de atributos para a tarefa de classificação e as abordagens existentes na literatura para esse propósito.

## 3.1 Classificação Hierárquica

A maior parte das pesquisas nas áreas de Mineração de Dados e Aprendizado de Máquina está relacionada com problemas de classificação plana, onde para cada nova instância a ser classificada são atribuídas uma ou mais classes que não possuem relacionamentos entre si. Todavia, muitas aplicações do mundo real são naturalmente expressas como problemas de classificação hierárquica, onde as classes envolvidas estão relacionadas de acordo com uma hierarquia (Silla Jr., 2011). Para resolver problemas dessa natureza, métodos de classificação hierárquica vêm sendo apresentados na literatura,

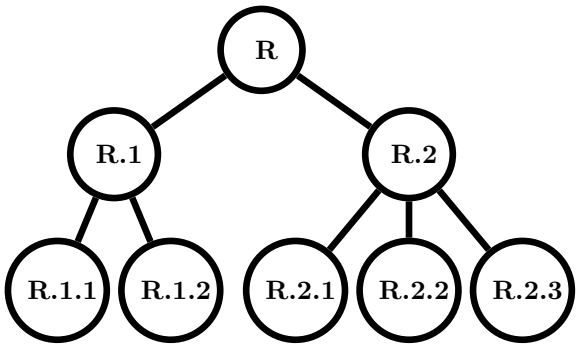
tais como (Silla Jr. e Freitas, 2009), (Chen et al., 2009) e (Silla Jr. e Freitas, 2011).

Os métodos de classificação hierárquica podem ser caracterizados de acordo com diferentes aspectos (Silla Jr. e Freitas, 2011). O primeiro deles está relacionado com o tipo de estrutura hierárquica (árvore ou grafo direcionado acíclico – GDA) com que o método é capaz de lidar. Essa estrutura representa os relacionamentos entre as classes do problema a ser resolvido. A Figura 3.1 apresenta essas estruturas. Basicamente, em uma árvore (Figura 3.1(a)) cada nó (classe) encontra-se associado a no máximo um nó (classe) pai, enquanto em um GDA (Figura 3.1(b)) um nó (classe) filho pode estar associado a vários nós (classes) pais. No caso deste trabalho, a estrutura hierárquica envolvida corresponde a uma árvore.

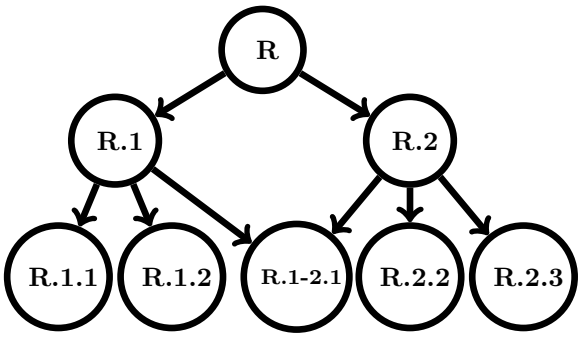
O segundo aspecto diz respeito à profundidade na estrutura hierárquica na qual a classificação é realizada. Um método pode realizar predições utilizando somente as classes dos nós folha da hierarquia (*Mandatory Leaf Node Prediction – MLNP*) ou utilizando classes referentes a qualquer nó (interno ou folha) da estrutura hierárquica (*Non-Mandatory Leaf Node Prediction – NMLNP*). Neste trabalho, a classificação é feita utilizando-se qualquer classe da estrutura hierárquica (*NMLNP*).

O terceiro aspecto diz respeito ao número de diferentes ramos de classes da hierarquia que um método pode dar como resposta para a classificação de uma instância. Um método pode ser capaz de predizer múltiplas classes para uma determinada instância (multirrótulo), envolvendo, portanto, múltiplos ramos da hierarquia (*multiple paths of labels*) ou somente uma classe (monorrótulo), a qual estará vinculada a somente um ramo da hierarquia (*single path of labels*). As soluções propostas neste trabalho foram projetadas para o cenário de classificação monorrótulo.

Por fim, o quarto aspecto está relacionado com a maneira adotada pelos métodos para manipular a estrutura hierárquica. Três abordagens diferentes são apresentadas na literatura: (i) abordagem por classificação plana, na qual a hierarquia é ignorada e as predições são realizadas considerando-se somente as classes dos nós folha da hierarquia; (ii) abordagens locais, que utilizam um conjunto de classificadores planos tradicionais; e (iii) abordagem global, onde um único modelo é construído considerando toda a hierarquia de classes durante uma única execução do algoritmo de classificação. As técnicas de seleção de atributos propostas neste trabalham visam o cenário da classificação hierárquica que faz uso de classificadores globais.




(a) Árvore




(b) Grafo direcionado acíclico (GDA)

**Figura 3.1:** Tipos de estruturas hierárquicas.

### 3.1.1 Abordagem por Classificação Plana

A abordagem por classificação plana é a mais simples para lidar com problemas de classificação hierárquica. Ela consiste na transformação do problema de classificação hierárquica original em um problema de classificação plana e, portanto, na utilização de algoritmos de classificação plana para resolução do problema transformado.

O classificador é construído ignorando-se completamente a hierarquia de classes, realizando predições considerando somente as classes dos nós folha da hierarquia, como mostrado na Figura 3.2 (o retângulo pontilhado representa o classificador plano que realiza predições somente nos nós folha da estrutura hierárquica). Desse modo, ela fornece uma solução indireta para o problema de classificação hierárquica, dado que, se uma classe de um nó folha é atribuída a uma instância, todas as suas classes ancestrais também são implicitamente atribuídas a essa instância.

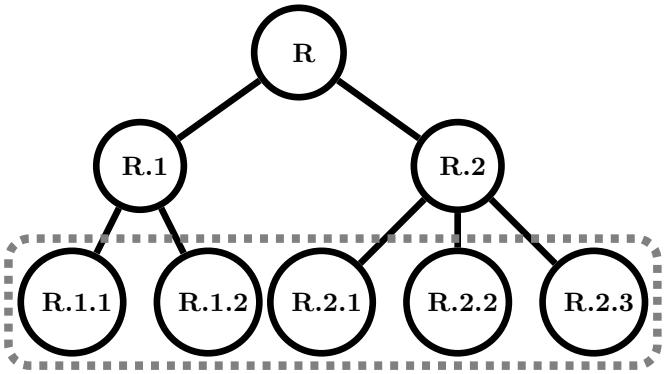



**Figura 3.2:** Abordagem por classificação plana.

Apesar da simplicidade dessa abordagem, ela apresenta algumas desvantagens. Uma delas está relacionada com o fato de o classificador perder a oportunidade de explorar os relacionamentos entre as classes presentes na hierarquia. Além disso, essa abordagem também é incapaz de lidar com problemas onde se deseja realizar predições em qualquer nível da hierarquia, uma vez que somente classes dos nós folha podem ser atribuídas às instâncias.

### 3.1.2 Abordagens Locais

Nas abordagens locais vários classificadores planos são construídos, cada um com uma visão local do problema, ou seja, a hierarquia de classes é explorada segundo uma perspectiva local (Silla Jr. e Freitas, 2011). De acordo com as diferentes maneiras de se utilizar essa informação local, os classificadores podem ser agrupados nas seguintes categorias: abordagem local por nó, abordagem local por nó pai e abordagem local por nível.

Na abordagem local por nó um classificador binário é criado para cada nó da hierarquia de classes (exceto para o nó raiz). Cada classificador prediz se uma instância pertence ou não à classe a qual ele está associado. Os retângulos pontilhados da Figura 3.3 representam os classificadores. Observe que essa abordagem permite que uma instância seja associada a classes que pertencem a diferentes ramos da hierarquia, o que pode resultar numa inconsistência se o problema de classificação em questão for monorrótulo. Por exemplo, na Figura 3.3, na classificação de uma instância, os classificadores binários podem dar resultado positivo para as classes $R.2$, $R.1.1$ e $R.2.1.2$. Nesse caso, temos uma inconsistência entre a classe predita no nível 2 (classe $R.1.1$) e aquelas preditas nos níveis 1 e 3 (classes $R.2$ e $R.2.1.2$, respectivamente). Para resolver esse problema, vários métodos de tratamento ou correção de inconsistências foram propostos na literatura (Barutcuoglu e DeCoro, 2006; Valentini, 2011; Wu et al., 2005).

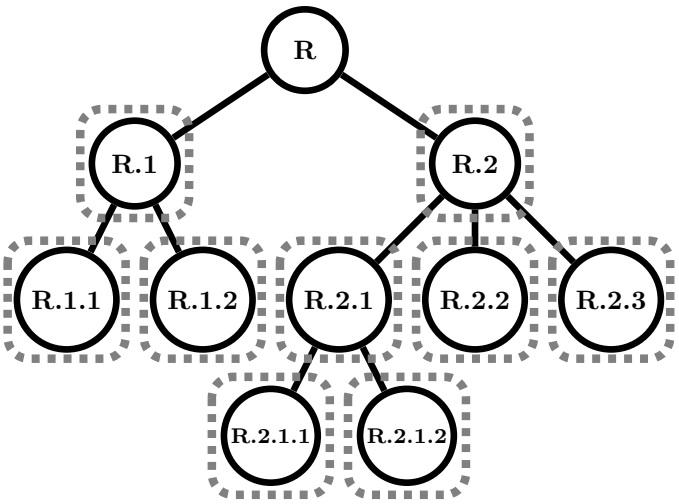



**Figura 3.3:** Abordagem local por nó.

Já na abordagem local por nó pai um classificador plano é treinado para cada nó pai da hierarquia de classes. Desse modo, para cada classificador construído, somente as classes associadas aos seus nós filhos são consideradas durante o processo de classificação. Essa abordagem é frequentemente usada com a estratégia *"top-down"* para a classificação de novas instâncias. Para ilustrar essa estratégia, considere a hierarquia apresentada na na Figura 3.4, onde cada retângulo pontilhado representa um classificador que é usado para predizer uma das classes associadas a seus nós filhos. Suponha que a nova instância seja classificada como classe $R.2$ pelo classificador associado ao nó raiz da hierarquia. Então, no primeiro nível da hierarquia, o classificador relacionado com a classe $R.2$ irá classificar essa mesma instância em uma das classes associadas com seus nós filhos ($R.2.1$, $R.2.2$ ou $R.2.3$) e assim por diante, até que a instância seja classificada num determinado nível da hierarquia.

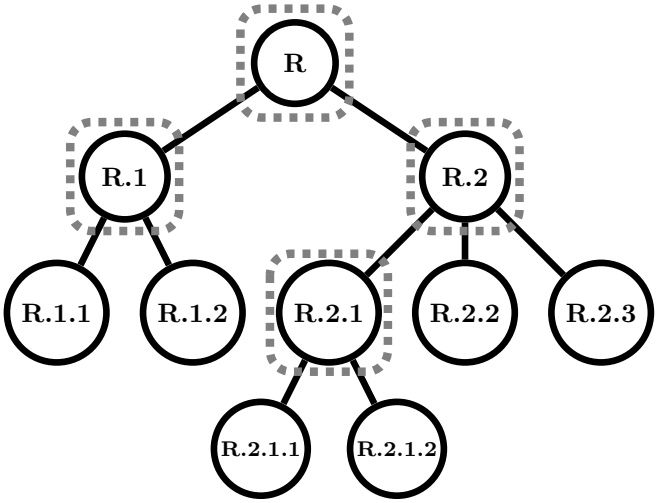



**Figura 3.4:** Abordagem local por nó pai.

Por fim, na abordagem local por nível um classificador plano tradicional é construído para cada nível da hierarquia. Essa é a abordagem hierárquica menos difundida na literatura (Silla Jr., 2011). Sua maior desvantagem é a possibilidade de geração de inconsistências na resolução de problemas de classificação monorrótulo. Por exemplo, considerando a hierarquia de classes mostrada na Figura 3.5, três classificadores seriam treinados, um para cada nível da hierarquia de classes (representados pelos retângulos pontilhados). Então, dada uma instância para ser classificada, a seguinte predição poderia ser realizada: classe $R.2$ no nível 1, classe $R.1.2$ no nível 2 e classe $R.2.1.1$ no nível 3. Claramente, a classe predita $R.1.2$ não é consistente com as classes $R.2$ e $R.2.1.1$. Desse modo, esse tipo de abordagem exige um pós-processamento visando a correção das

inconsistências.

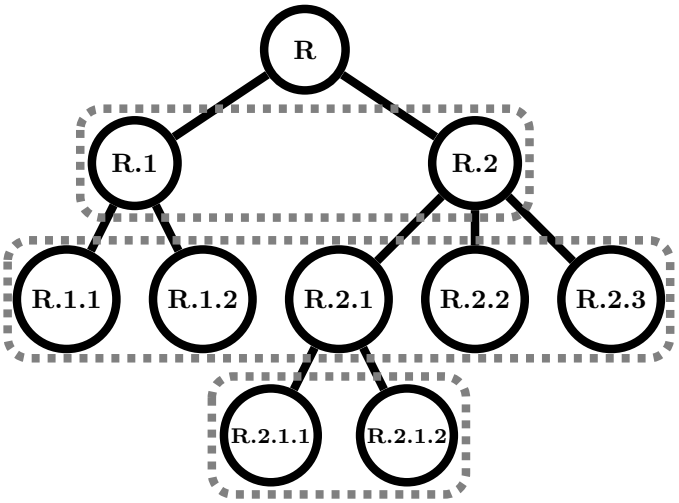



**Figura 3.5:** Abordagem local por nível.

### 3.1.3 Abordagem Global

Na abordagem global, ao invés de termos um conjunto de classificadores, um único classificador é construído considerando-se toda a hierarquia de classes numa única execução do algoritmo de classificação. Portanto, dada uma nova instância para ser classificada, o classificador global é capaz de apresentar como resultado uma classe de qualquer nível da hierarquia. Essa abordagem é ilustrada na Figura 3.6, onde o único retângulo pontilhado representa o classificador.

Enquanto a abordagem local com a estratégia de predição *top-down* tem a desvantagem de propagar o erro de classificação cometido num determinado nível da hierarquia para os demais níveis mais profundos, a abordagem global evita esse problema realizando a classificação em uma única etapa utilizando um único classificador.

Vale observar que a abordagem global perde a natureza modular das abordagens locais, ou seja, a característica de dividir a fase de treinamento em diversos processos, cada um considerando parte da hierarquia de classes. Portanto, o único classificador construído na abordagem global tende a ser mais complexo do que cada classificador individual construído nas abordagens locais. No entanto, essa natureza modular das abordagens

locais não necessariamente implica numa superioridade de desempenho em termos de acurácia preditiva em relação à abordagem global. O classificador hierárquico *Global Model Naive Bayes (GMNB)* (Silla Jr. e Freitas, 2009) adotado para realização dos experimentos computacionais deste trabalho segue a abordagem global.

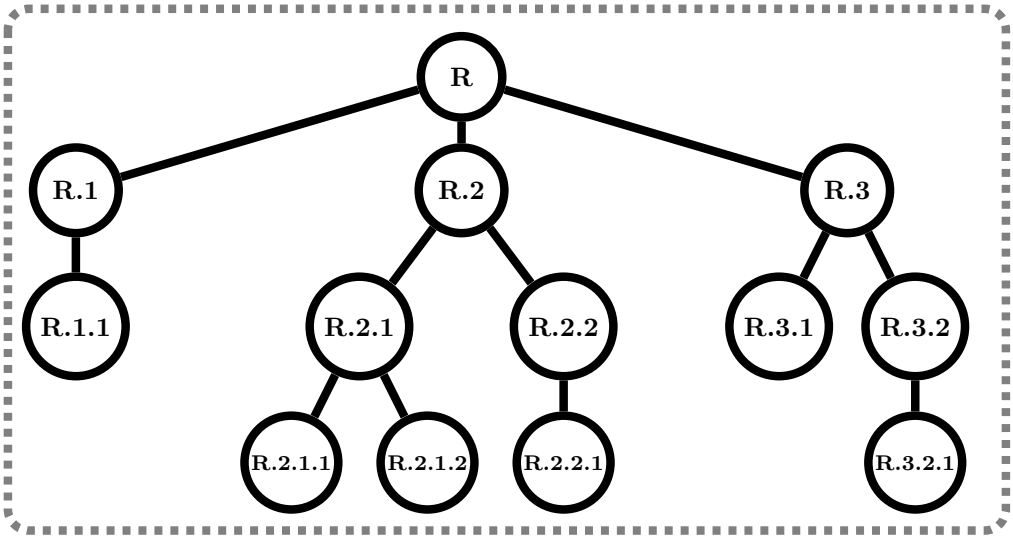



**Figura 3.6:** Abordagem global (ou *Big Bang*).

## 3.2 Global-Model Naive Bayes

A maioria das pesquisas realizadas na área de Mineração de Dados foca apenas no desenvolvimento e no aperfeiçoamento dos algoritmos de aprendizado voltados aos problemas de classificação plana. Além disso, dos poucos trabalhos voltados para problemas de classificação hierárquica disponíveis na literatura, a maior parte deles emprega as abordagens de classificação hierárquica local (Silla Jr. e Freitas, 2011). Dessa forma, a abordagem de classificação hierárquica global é ainda muito pouco explorada na literatura.

Tendo em vista esse cenário, no trabalho (Silla Jr. e Freitas, 2009) foi proposto um algoritmo de classificação que segue a abordagem de classificação hierárquica global. Essa proposta corresponde a uma extensão do classificador plano *Naive Bayes* (Duda e Hart, 1973), denominada *Global Model Naive Bayes (GMNB)*, que permite levar em consideração os relacionamentos existentes entre as classes do problema.

Para explicar o seu funcionamento, considere um problema com a hierarquia de classes apresentada na Figura 3.7, que corresponde a uma árvore com os seguintes nós (classes): $R.1$, $R.2$, $R.1.1$, $R.1.2$, $R.2.1$ e $R.2.2$. Assim como no *Naive Bayes*, dada uma nova instância $X = \{x_1, x_2, \ldots, x_m\}$ para ser classificada, onde $x_1, x_2, \ldots, x_m$ correspondem aos valores dos atributos preditores $A_1, A_2, \ldots, A_m$, respectivamente, o classificador atribui a essa nova instância a classe que tem a maior probabilidade *a posteriori* $P(C_i|\, X) \propto P(X|\, C_i)P(C_i)$, onde $P(X|\, C_i) = \prod_{j=1}^{m} P(x_j|\, C_i)$. Desse modo, a principal diferença para o classificador *Naive Bayes* está na maneira de calcular as probabilidades $P(C_i)$ e $P(x_j|\, C_i)$, uma vez que no *GMNB* elas são calculadas levando-se em consideração a hierarquia de classes.

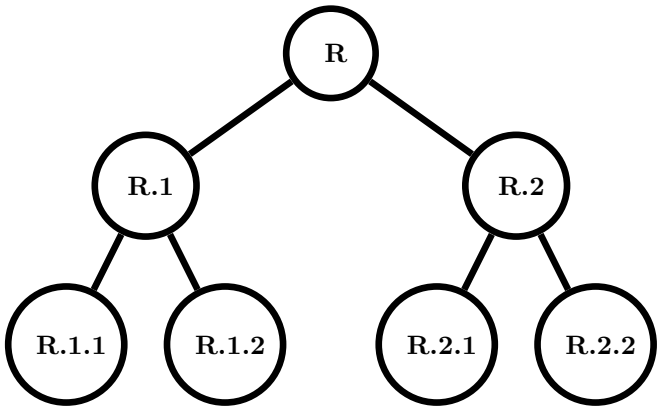



**Figura 3.7:** Hierarquia de classes.

Mais especificamente, no cálculo dessas probabilidades, o *GMNB* considera que qualquer instância da classe $C_i$ também pertence a todas as suas classes ancestrais. Por exemplo, se uma instância de treinamento pertence à classe $R.2.2$, então ela será utilizada no cálculo de todas as probabilidades envolvendo a classe $R.2.2$ ($P(R.2.2)$ e $P(x_j|\, R.2.2)$) e também de todas as probabilidades relacionadas com sua classe ancestral $R.2$ ($P(R.2)$ e $P(x_j|\, R.2)$).

Assim, para o *GMNB*, a probabilidade a priori $P(C_i)$ para cada uma das classes do conjunto de dados de treinamento é dada pela soma do número de instâncias do conjunto de treinamento $Tr$ pertencente à classe $C_i$ (de forma idêntica ao *Naive Bayes* tradicional) e do número de instâncias no conjunto de treinamento $Tr$ cuja classe é descendente de $C_i$, dividido pelo número total de instâncias do conjunto de treinamento $Tr$, conforme mostra a Equação 3.1.

$$P(C_i) = \frac{|C_i| + | \Downarrow C_i|}{|Tr|}, \tag{3.1}$$

onde $|C_i|$ é o número total de instâncias do conjunto de treinamento $Tr$ cuja mais específica classe é $C_i$; $| \Downarrow C_i|$ é o número total de instâncias do conjunto de treinamento $Tr$ cuja mais específica classe é descendente de $C_i$ e $|Tr|$ é o número total de instâncias do conjunto de treinamento $Tr$.

De forma análoga ao cálculo da probabilidade a priori, o cálculo das probabilidades condicionais $P(x_j|\, C_i)$ também é adaptado considerando que qualquer instância que pertence à classe $C_i$ também pertence a todas as suas classes ancestrais na hierarquia. Dessa maneira, quando uma instância de treinamento da classe $C_i$ é processada, a contagem de frequência para seus pares atributo-valor é adicionada à contagem de frequência da classe $C_i$ e de todas as suas classes ancestrais. Por exemplo, se a instância de treinamento pertence à classe $R.2.1$, as contagens de frequência de cada par atributo-valor são adicionadas às contagens de frequência de ambas as classes $R.2.1$ e $R.2$. Tais modificações durante a fase de treinamento vão permitir ao algoritmo predizer classes em qualquer nível da hierarquia durante a fase de classificação de novas instâncias.

## 3.3 F-Measure Hierárquica

Para expressar o desempenho preditivo de um classificador hierárquico, Kiritchenko et al. (2006) propuseram uma adaptação da métrica *F-measure* para o contexto de classificação hierárquica, denominada *F-measure* hierárquica ($hF$).

A *F-measure* hierárquica é computada como

$$hF = \frac{2 \times hP \times hR}{hP + hR}, \tag{3.2}$$

onde $hP$ e $hR$ correspondem à precisão hierárquica e revocação hierárquica, respectivamente. Considerando que $P_i$ é o conjunto formado pelas classes mais específicas preditas para uma instância de teste $i$ e todas as suas classes ancestrais e que $V_i$ é o conjunto composto pelas classes verdadeiras mais específicas da instância de teste $i$ e todas as suas classes

ancestrais, $hP$ e $hR$ foram definidas como:

$$hP = \frac{\sum_i |P_i \cap V_i|}{\sum_i |P_i|} \tag{3.3}$$

e

$$hR = \frac{\sum_i |P_i \cap V_i|}{\sum_i |V_i|} \tag{3.4}$$

O emprego da métrica *F-measure* hierárquica para avaliar o desempenho preditivo de classificadores hierárquicos é recomendado pelos autores Silla Jr. e Freitas (2011), visto que ela pode ser efetivamente aplicada em qualquer cenário de classificação hierárquica. Vale ressaltar que, no cenário de classificação hierárquica NMLNP, as métricas $hF$, $hP$ e $hR$ consideram os erros de generalização e de especialização.

O erro de generalização se refere ao caso onde a classe mais específica predita para uma instância de teste é mais genérica do que a classe verdadeira mais específica associada com aquela instância. Por exemplo, predizer a classe $R.1$ para um instância cuja classe verdadeira mais específica é $R.1.1$.

Por outro lado, o erro de especialização se remete ao caso onde a classe mais específica predita para uma instância de teste é mais específica do que a classe verdadeira mais específica daquela instância. Como exemplo, considere a predição da classe $R.1.1$ para uma instância de teste cuja classe verdadeira mais específica é $R.1$.

Neste trabalho adotou-se-se a métrica *F-measure* hierárquica nas avaliações do classificador hierárquico utilizado, a saber, o *Global-Model Naive Bayes (GMNB).*

## 3.4 *k*-Validação Cruzada

Tradicionalmente, a avaliação de classificadores é realizada utilizando-se o método denominado de $k$-validação cruzada (Kohavi, 1995). Esse procedimento divide o conjunto de dados em $k$ partições mutuamente exclusivas e, a cada iteração do método, uma partição é utilizada para teste e as remanescentes são usadas para o treinamento do

classificador. No final, o desempenho do classificador é dado pelo valor médio dos desempenhos obtidos nas $k$ iterações.

Por exemplo, num procedimento de 3-validação cruzada, as instâncias de treinamento são divididas em três partições contendo cada uma o mesmo (ou aproximadamente o mesmo) número de instâncias. A cada iteração da validação cruzada 2 partições são utilizadas para treinamento e 1 para teste. Esse procedimento é repetido por três vezes e, no final, cada partição foi empregada exatamente uma vez para teste. A Figura 3.8 ilustra o procedimento para o exemplo em questão.

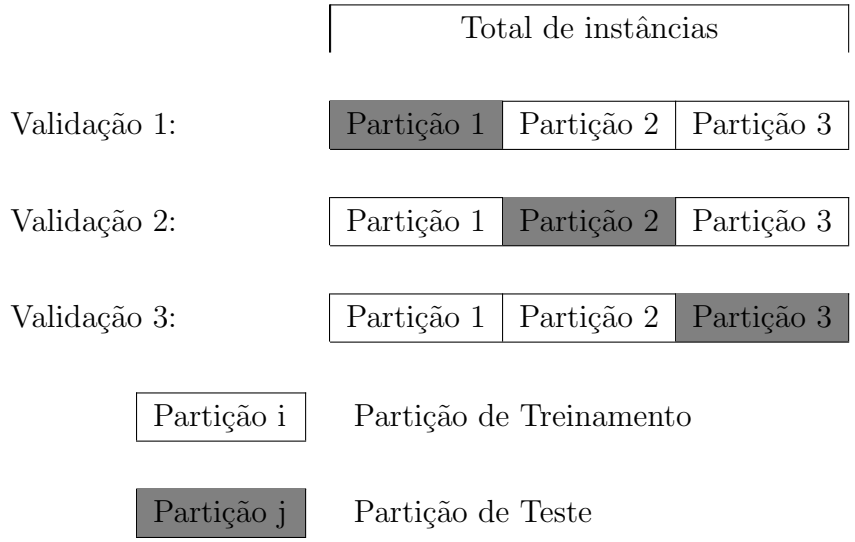



**Figura 3.8:** Exemplificação do método 3-validação cruzada para avaliar a acurácia preditiva de classificadores.

## 3.5 Seleção de Atributos

A seleção de atributos é um processo adotado em aplicações das áreas de Mineração de Dados e Aprendizado de Máquina, sendo frequentemente empregado como uma etapa de pré-processamento de dados para a tarefa de classificação (Guyon e Elisseeff, 2006). Para se alcançar o sucesso em aplicações envolvendo tal tarefa, o emprego da seleção de atributos é, em muitos casos, uma etapa fundamental (Liu et al., 2010).

O objetivo primário dos métodos de seleção de atributos é identificar os atributos que

são relevantes para a tarefa de classificação. Portanto, o emprego dos algoritmos para seleção de atributos tem como meta identificar e remover os atributos irrelevantes e/ou redundantes das bases de dados (Hall e Holmes, 2003).

Em decorrência da enorme quantidade de dados disponível para análise nos dias atuais, a seleção de atributos se torna uma etapa de grande importância, visto que, a permanência de atributos irrelevantes e/ou redundantes no processo de construção do modelo de classificação pode resultar num classificador com baixa capacidade preditiva, assim como aumentar o esforço computacional gasto no processo de aprendizado (Hall, 1998; Liu e Yu, 2003). Vários estudos e pesquisas já demonstraram que, em determinados conjuntos de dados, alguns dos atributos podem ser removidos do conjunto de atributos sem, com isso, incorrer na deterioração da acurácia preditiva (Blum e Langley, 1997).

Os métodos de seleção de atributos são capazes de processar conjuntos de dados que possuem instâncias pré-classificadas, parcialmente pré-classificadas e as que não possuem classes, o que leva ao desenvolvimento de algoritmos supervisionados, semi-supervisionados e não-supervisionados, respectivamente (Liu et al., 2010). Um algoritmo para seleção de atributos supervisionado determina a relevância dos atributos avaliando a correlação existente entre eles e o atributo classe. Neste trabalho, o objetivo é trabalhar com métodos de seleção de atributos supervisionados e, portanto, esse será o cenário considerado daqui em diante.

Na prática, o uso da seleção de atributos nas tarefas de classificação pode resultar nos seguintes benefícios (Liu e Motoda, 2007): (i) melhora da capacidade preditiva dos classificadores; (ii) redução do tempo gasto no processo de classificação; e (iii) obtenção de modelos de classificação mais compactos e, portanto, mais fáceis de serem interpretados.

Dependendo de como e quando a qualidade dos atributos é avaliada, diferentes abordagens podem ser consideradas, que em termos gerais recaem nas seguintes categorias: Embutida, Filtro, *Wrapper* e Híbrida (envolvendo combinações de abordagens) (Liu et al., 2010).

Basicamente, na abordagem Embutida, as técnicas de seleção de atributos estão incorporadas ao algoritmo de indução do modelo de classificação. Exemplos típicos são os algoritmos de indução de árvores de decisão, que realizam a seleção dos atributos no próprio processo de definição dos nós que irão formar as árvores de decisão (Chen et al., 2009; Costa et al., 2007; Quinlan, 1986, 1993; Vens et al., 2008).

Já as técnicas do tipo Filtro são independentes do algoritmo de classificação aplicado

na etapa de mineração de dados. Elas utilizam as informações da própria base de dados e, a partir de alguma medida, avaliam a qualidade de atributos ou de subconjuntos de atributos da base. Nesse tipo de abordagem, as técnicas são divididas em dois grupos: as que avaliam cada atributo individualmente, como por exemplo *Information Gain Attribute Ranking* (Yang e Pedersen, 1997) e *ReliefF* (Kira e Rendell, 1992); e as que avaliam subconjuntos de atributos, como por exemplo *Correlation-based Feature Selection* (*CFS*) (Hall, 2000) e *Consistensy-based Feature Selection* (Liu e Setiono, 1996).

Na abordagem *Wrapper*, ilustrada na Figura 3.9, as técnicas avaliam a qualidade dos subconjuntos de atributos com o próprio classificador que será utilizado na etapa de classificação. Portanto, o mérito de um determinado subconjunto de atributos é mensurado através da avaliação do classificador treinado utilizando apenas os atributos incluídos naquele subconjunto (Kohavi e John, 1997). As técnicas de seleção de atributos baseadas nessa abordagem precisam percorrer o espaço de soluções formado pelos possíveis subconjuntos de atributos para escolher aqueles que serão avaliados pelo classificador. Cada subconjunto de atributo escolhido é então utilizado para treinar um classificador, cuja avaliação indicará a qualidade daquele subconjunto de atributos. O processo de percorrer o espaço de soluções (realizado por um algoritmo de busca) prossegue até que um critério de parada seja alcançado, quando o algoritmo de seleção retorna o subconjunto de atributos que alcançou a melhor avaliação.

Uma maneira de avaliar os subconjuntos de atributos é percorrer, de modo exaustivo, todo o espaço de busca do problema. No entanto, devido ao custo computacional, essa estratégia é claramente proibitiva, mesmo para conjuntos de dados que não contenham uma grande quantidade de atributos, pois o número de avaliações de subconjuntos cresce exponencialmente em relação ao número de atributos da base de dados (Hall, 1998). Desse modo, heurísticas e metaheurísticas, que exploram um espaço de soluções reduzido, são comumente utilizadas na tarefa de selecionar um subconjunto de atributos. Estratégias metaheurísticas comumente citadas na literatura como método de busca em técnicas de seleção de atributos são Busca Tabu (Yusta, 2009), as estratégias Evolucionárias (Yusta, 2009), os algoritmos Meméticos (Kannan e Ramaraj, 2010), GRASP (Greedy Randomized Adaptive Search Procedure) (Bermejo et al., 2011; Esseghir, 2010), dentre outras.

As técnicas que seguem a abordagem *Wrapper* geralmente produzem resultados melhores do que aqueles obtidos pelas técnicas que se baseiam na abordagem Filtro, uma vez que a avaliação dos atributos é conduzida pelo próprio algoritmo de classificação que será posteriormente utilizado na etapa de mineração de dados (Hall e Smith, 1999; Inza et al., 2004). Contudo, como nas técnicas baseadas na abordagem *Wrapper* o classificador

deve ser executado várias vezes durante o processo de seleção de atributos, elas podem ter um custo computacional muito elevado, tornando-se até mesmo impraticáveis em conjuntos de dados que possuem muitos atributos (Bermejo et al., 2011).

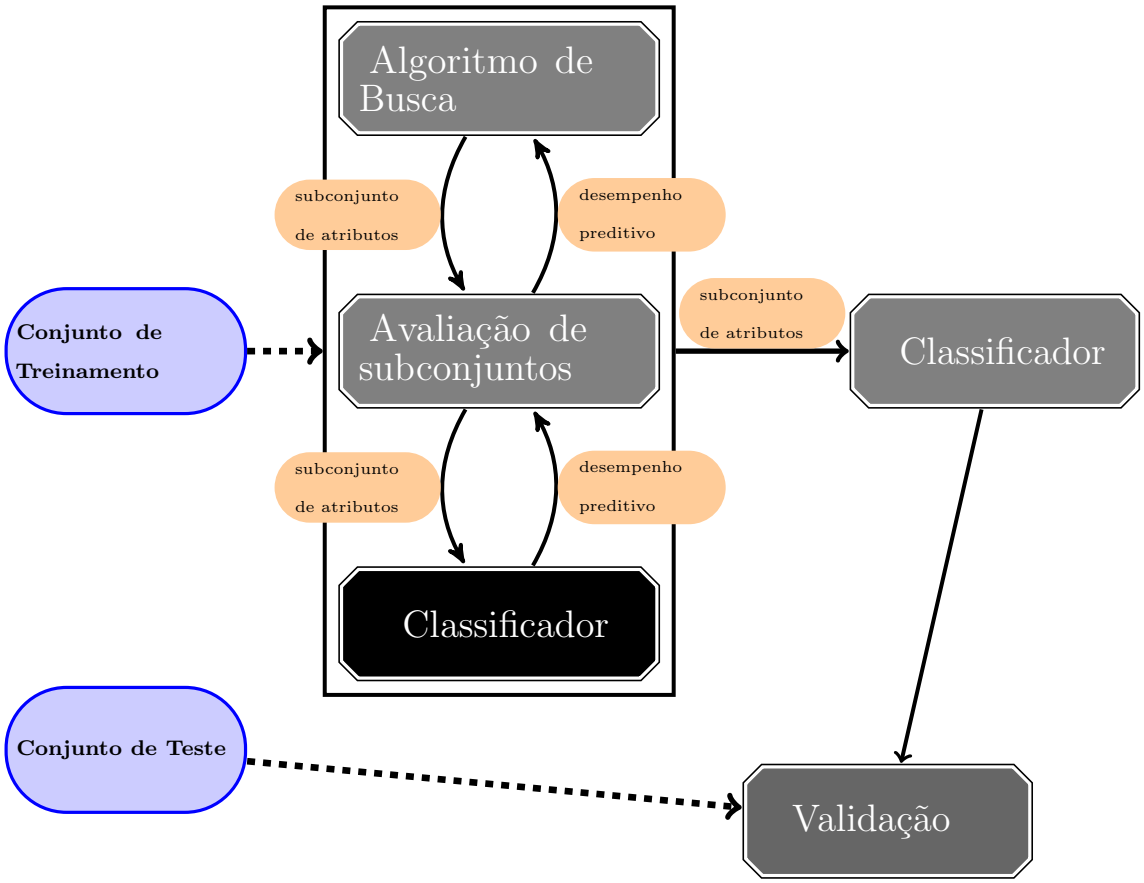



**Figura 3.9:** O funcionamento básico da abordagem *Wrapper* para seleção de subconjuntos de atributos.

Tendo isso em vista, recentemente surgiram propostas de técnicas de seleção de atributos que seguem uma abordagem híbrida. Essas técnicas visam se beneficiar dos pontos positivos das abordagens Filtro e *Wrapper*, ou seja, a eficiência computacional da abordagem Filtro e a precisão preditiva da abordagem *Wrapper* (Liu e Yu, 2005).

Esseghir (2010) propôs um novo e efetivo esquema de hibridização das abordagens Filtro e *Wrapper* incorporado aos componentes básicos da metaheurística *GRASP*. No primeiro momento, uma técnica que segue a abordagem Filtro é empregada para avaliar cada atributo individualmente e, assim, formar um *ranking* dos atributos. Após formado o *ranking* dos atributos, a etapa construtiva do *GRASP* prossegue com a seleção de $d$ atributos, sendo $d$ um parâmetro ao método.

No trabalho Bermejo et al. (2011), os autores também usaram a metaheurística *GRASP* em uma abordagem híbrida (Filtro e *Wrapper*) objetivando a seleção de atributos em conjuntos de dados que possuem grandes quantidades de atributos. O objetivo principal dessa proposta era acelerar o processo de seleção de atributos reduzindo o número de avaliações *Wrapper*. Os resultados mostraram que essa estratégia é comparável, em termos de desempenho preditivo e cardinalidade de subconjunto selecionado, com estratégias anteriores da literatura, mas requer uma quantidade significativamente menor de avaliações de subconjuntos.

Assim, considerando o fato de que vários estudos e pesquisas recentes vêm demonstrando ganhos com a utilização de técnicas que seguem a abordagem Híbrida, neste trabalho é proposto um método heurístico que adota essa abordagem visando a seleção de atributos no contexto da classificação hierárquica monorrótulo.

# Capítulo 4

# Seleção de Atributos para Classificação Hierárquica Monorrótulo

Este capítulo apresenta na Seção 4.1 a proposta de adaptação da medida Incerteza Simétrica (*Symmetrical Uncertainty – SU*), que visa permitir a utilização da mesma em técnicas de seleção de atributos para problemas de classificação hierárquica monorrótulo que utilizam classificadores hierárquicos globais. Em seguida, na Seção 4.2, apresenta-se um algoritmo de ranqueamento para seleção de atributos que utiliza a abordagem Filtro com a medida Incerteza Simétrica adaptada para o contexto hierárquico. Por fim, na Seção 4.3, é descrita a proposta de uma heurística híbrida (Filtro e *Wrapper*) para seleção de atributos que também utiliza a medida apresentada na Seção 4.1.

## 4.1 Adaptação da Medida Incerteza Simétrica

Diversas medidas já foram propostas na literatura para avaliar a qualidade de um atributo para a tarefa de classificação. A Incerteza Simétrica (*Symmetrical Uncertainty – SU*), uma medida de correlação não linear bastante utilizada na avaliação de atributos é calculada utilizando-se outras duas medidas provenientes da área de teoria da informação, denominadas ganho de informação e entropia.

Apesar de a medida de ganho de informação ser utilizada em técnicas de seleção

de atributos, ela é tendenciosa em favor de atributos que possuem muitos valores possíveis (Liu e Motoda, 2007). Já a medida *SU*, além de não apresentar essa tendência, realiza uma normalização de modo a apresentar sempre resultados no intervalo $[0, 1]$ (Han et al., 2011). Sendo assim, ela foi a medida escolhida para ser adaptada para o contexto de classificação hierárquica monorrótulo.

A medida *SU* proposta para a seleção de atributos no cenário de classificação plana é calculada conforme a Equação 4.1 onde, para um dado atributo preditor $A$, o valor de $SU(A, C)$ quantifica o grau de correlação entre o atributo preditor $A$ e o atributo classe $C$.

$$SU(A, C) = 2 \times \left( \frac{GI(C, A)}{E(C) + E(A)} \right), \tag{4.1}$$

onde $GI(C, A)$ é o ganho de informação, ou seja, a redução causada na entropia do atributo classe $C$ devido à informação adicional fornecida pelo atributo $A$; $E(C)$ é a entropia do atributo classe $C$ e $E(A)$ é a entropia do atributo preditor $A$. O cálculo do ganho de informação é expresso conforme a Equação 4.2.

$$GI(C, A) = E(C) - E(C \mid A), \tag{4.2}$$

onde $E(C)$ é a entropia do atributo classe $C$ e $E(C \mid A)$ é a entropia do atributo classe $C$ depois de observado o atributo $A$. A entropia do atributo classe $C$, dado por $E(C)$, representa o grau de aleatoriedade (impureza) na distribuição das classes num conjunto de dados $D$ (Cover e Thomas, 1991; Quinlan, 1986). Seu cálculo é mostrado na Equação 4.3.

$$E(C) = - \sum_{c \in C} P(c) \times \log_2 P(c), \tag{4.3}$$

onde $P(c)$ é a probabilidade de ocorrência da classe $c$ no conjunto de dados $D$, expressa como sendo a razão entre o número de instâncias em $D$ que pertencem à classe $c$ e o número total de instâncias em $D$.

Já a entropia do atributo classe $C$ após observados os valores do atributo preditor $A$ num conjunto de dados $D$, expressa como $E(C \mid A)$, é definida conforme a Equação 4.4.

$$E(C \mid A) = - \sum_{a \in A} P(a) \times \sum_{c \in C} P(c \mid a) \times \log_2 P(c \mid a), \tag{4.4}$$

com $P(a)$ sendo a probabilidade do valor $a$ do atributo $A$ ocorrer no conjunto de dados $D$ e $P(c \mid a)$ a probabilidade de ocorrência da classe $c$ no conjunto de dados $D$ dada a ocorrência do valor $a$ do atributo $A$. Assim, se os valores observados do atributo classe $C$ são particionados conforme os valores do atributo preditor $A$ e a entropia de $C$ nas partições induzidas por $A$ é menor que a entropia de $C$ a priori, então existe algum nível de correlação entre $C$ e $A$. A quantidade pela qual a entropia de $C$ diminui reflete informação adicional fornecida pelo atributo preditor $A$ (Hall, 2000).

A motivação para a adaptação da medida $SU$ (ver Equação 4.1) para o contexto da classificação hierárquica monorrótulo surgiu do fato de essa medida não ser capaz de lidar com conjuntos de dados que possuem suas classes estruturadas de acordo com uma hierarquia. Para exemplificar, considere o conjunto de dados apresentado na Figura 4.1.

| $A_1$ | ... | Classe |
|---|---|---|
| 1 | | R.1.1 |
| 1 | | R.1.1 |
| 2 | | R.2.1 |
| 2 | | R.2.1 |
| 2 | | R.2.2 |
| 2 | | R.2.2 |

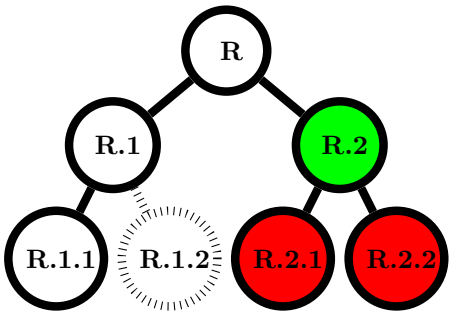



**Figura 4.1:** Correlações existentes entre o atributo preditor $A_1$ e a estrutura hierárquica de classes do atributo classe $C$.

No caso desse exemplo da Figura 4.1, o valor de $SU(A_1, C)$ tradicional é igual a $0,73$. Isso ocorre porque, apesar de termos uma partição dos dados ($A_1 = 1$) totalmente pura (somente classe $R.1.1$), a segunda partição ($A_1 = 2$) contém duas classes completamente distintas $R.2.1$ e $R.2.2$ quando consideramos o contexto da classificação plana. No entanto, no contexto hierárquico, as classes $R.2.1$ e $R.2.2$ não são completamente distintas, uma vez que elas compartilham a mesma classe pai ($R.2$), conforme pode ser visto pela

Figura 4.1. Portanto, a adaptação da medida *SU* para o contexto hierárquico visa levar em consideração essa informação inerente à hierarquia de classes.

Em Chen et al. (2009), os autores propuseram um algoritmo para indução de árvores de decisão capaz de lidar com bases de dados cujas classes estão estruturadas segundo uma hierarquia. A fim de escolher o atributo que compõe cada nó da árvore de decisão, o classificador proposto utilizou uma adaptação da medida de ganho de informação para o contexto de classificação hierárquica, descrita pela Equação 4.5. Nessa equação, o $GI_H(C, A)$ corresponde à redução obtida na entropia hierárquica do atributo classe $C$ após a observação dos valores do atributo $A$.

$$GI_H(C, A) = E_H(C) - E_H(C \mid A), \tag{4.5}$$

onde $E_H(C)$ é a entropia hierárquica do atributo classe $C$ e $E_H(C \mid A)$ é a entropia hierárquica do atributo classe $C$ depois de observado o atributo $A$.

O cálculo da entropia hierárquica, também proposto em (Chen et al., 2009), corresponde a uma média ponderada da entropia calculada para cada um dos níveis da hierarquia de classes. Mais especificamente, considere a estrutura hierárquica de classes $HC = \{N_1, N_2, ..., N_h\}$, onde $N_i$ denota o $i$-ésimo nível hierárquico e $h$ o número total de níveis da hierarquia. Além disso, considere que o nível $N_1$ é aquele que contém os nós filhos do nó raiz da estrutura hierárquica. Desse modo, a entropia é calculada para cada um dos níveis da hierarquia de classes a partir do primeiro nível ($i = 1$). Quanto maior (mais profundo) o nível hierárquico, mais específicas são as classes pertencentes àquele nível. Por fim, considere também $N_i = \{N_{(i,j)} \mid j = 1, ..., m_i\}$, onde $N(i, j)$ corresponde ao $j$-ésimo nó (classe) do nível $i$ e $m_i$ é o número total de nós (classes) do nível $i$. Portanto, na equação da entropia hierárquica do atributo classe $C$ apresentada a seguir (Equação 4.6), $p(i, j)$ é a probabilidade de ocorrência da $j$-ésima classe do nível $i$.

$$E_H(C) = -\sum_{i=1}^{h}\sum_{j=1}^{m_i}(P_{(i,j)} \times \log_2 P_{(i,j)}) \times w_i, \tag{4.6}$$

onde $w_i$, o peso atribuído ao nível $i$ da hierarquia, é calculado conforme a Equação 4.7.

$$w_i = (h - i + 1) \times \left( \frac{2}{h \times (h+1)} \right) , onde \quad i \geq 1 \tag{4.7}$$

Vale ressaltar que $\sum_{i=1}^{h} wi = 1$. Além disso, pode-se verificar a partir da Equação 4.7 que $w_1, w_2, ..., w_h$ corresponde a uma série aritmética onde os maiores pesos são atribuídos aos níveis menos profundos da hierarquia de classes.

Neste trabalho, propõe-se a utilização das medidas de ganho de informação hierárquica (Equação 4.5) e entropia hierárquica (Equação 4.6) para realizar a adaptação da medida SU para o contexto hierárquico. Sendo assim, a Incerteza Simétrica Hierárquica (*Hierarchical Symmetrical Uncertainty* – $SU_H$) é expressa de acordo com a Equação 4.8.

$$SU_H(C, A) = 2 \times \left( \frac{GI_H(C, A)}{E_H(C) + E(A)} \right), \tag{4.8}$$

onde $GI_H(C, A)$ é o ganho de informação hierárquico, ou seja, a redução causada na entropia hierárquica do atributo classe $C$ devido à informação adicional fornecida pelo atributo $A$; $E_H(C)$ é a entropia hierárquica do atributo classe $C$ e $E(A)$ é a entropia do atributo preditor $A$.

Voltando ao exemplo das Figura 4.1, onde o valor de $SU(A_1, C)$ tradicional era igual a $0,73$, ao calcularmos o valor da medida Incerteza Simétrica Hierárquica $SU_H(C, A_1)$, chegamos ao valor de $0,89$. O aumento observado no valor da medida $SU_H$, quando comparado ao resultado obtido com da medida $SU$ tradicional, se dá pelo fato de a $SU_H$ levar em consideração que, no contexto hierárquico, as classes $R.2.1$ e $R.2.2$ não são completamente distintas, uma vez que compartilham a mesma classe pai $R.2$. Ou seja, nesse exemplo, se considerarmos apenas o primeiro nível da hierarquia, temos duas partições de dados ($A_1 = 1$ e $A_1 = 2$) totalmente puras (somente classe $R.1$ na primeira partição e somente a classe $R.2$ na segunda partição). Alguma impureza aparece apenas quando analisamos os dados considerando-se o segundo nível da hierarquia, quando para a segunda partição ($A_1 = 2$), temos instâncias associadas a duas classes, a $R.2.1$ e a $R.2.2$. Portanto, pode-se observar que a medida proposta ($SU_H$) é capaz de considerar os relacionamentos entre as classes.

Nas seções a seguir, com objetivo de avaliar a utilização da medida $SU_H$, serão

apresentadas duas técnicas de seleção de atributos que farão uso da mesma.

## 4.2 Método de Ranqueamento

Esta seção apresenta um algoritmo para a criação de um *ranking* de atributos que pode ser empregado para a seleção de atributos no contexto da classificação hierárquica monorrótulo.

De maneira formal, a partir de uma medida de avaliação individual de atributos, que no caso deste trabalho é a medida $SU_H$, os atributos pertencentes a um conjunto de dados $D$ são avaliados e pontuados conforme a relevância que cada um deles possui para a tarefa de classificação. A partir dessa avaliação, um *ranking* de atributos é criado de modo que os atributos mais relevantes fiquem nas primeiras posições desse *ranking*.

**Algoritmo 1:** Procedimento incremental para a criação e avaliação de *rankings* gerados utilizando a medida $SU_H$.

**Entrada:** Conjunto de Dados $D$, $k$
**Saída:** $S$: subconjunto contendo $k$ atributos

```
1 para cada A_i ∈ D faça
2 |  pontuacao[i] = SU_H(A_i, C) //  C = Atributo Classe
3 fim
4 Rank[ ] = criar o ranking dos atributos usando a pontuação
5 S = ∅
6 para i = 1 to k faça
7 |  S = S ∪ Rank[i]
8 fim
9 retorna S
```

Assim, o Algoritmo 1 pode ser adotado como um método de seleção de atributos que implementa a abordagem Filtro no cenário da classificação hierárquica monorrótulo. A partir de uma base de dados $D$ e do tamanho do subconjunto de atributos desejado ($k$), o método retorna os $k$ melhores atributos avaliados de acordo com a medida $SU_H$. Apesar da simplicidade, a grande vantagem desse método é a sua eficiência, dado que sua complexidade computacional é $O(n)$, sendo $n$ o número de atributos do conjunto de dados $D$. Neste trabalho, o Algoritmo 1 foi utilizado apenas com o propósito de realizarmos uma primeira avaliação sobre a adequação da medida proposta ($SU_H$) para o contexto da classificação hierárquica monorrótulo.

## 4.3 Método Heurístico

Mais recentemente, algoritmos de seleção de atributos híbridos, que combinam as abordagens Filtro e *Wrapper*, vêm ganhando destaque em pesquisas no contexto da classificação plana (Bermejo et al., 2014, 2011; Hu et al., 2015; Taheri e Nezamabadi-pour, 2014). Isso se deve ao fato de eles conseguirem juntar as vantagens das abordagens Filtro (baixo custo computacional) e *Wrapper* (otimizar a precisão da classificação). Basicamente, a ideia é utilizar alguma técnica que implementa a abordagem Filtro e, em seguida, usar a sua saída para guiar um método que segue a abordagem *Wrapper*.

Como não foram encontradas na literatura propostas de métodos de seleção de atributos para problemas de classificação hierárquica monorrótulo que usam classificadores globais, neste trabalho, apresentamos um método heurístico, denominado *Hybrid Feature Selection for Hierarchical Classification – HFS4HC*, que implementa uma abordagem híbrida de seleção de atributos para o cenário de classificação hierárquica supramencionado.

A hibridização de abordagens existente no método proposto ocorre da seguinte maneira: num primeiro momento, um *ranking* dos atributos é construído a partir de uma abordagem Filtro. Em seguida, esse *ranking* é utilizado para guiar a busca pelos subconjuntos que são avaliados a partir da abordagem *Wrapper*. Essa mesma estratégia já foi adotada em trabalhos de seleção de atributos para o contexto de classificação plana (Bermejo et al., 2014, 2011).

O Algoritmo 2 apresenta o método heurístico proposto. Seja $D$ um conjunto de dados, $Max\_Iter$ o número de iterações a serem realizadas pelo método e $SU_H$ a medida Incerteza Simétrica Hierárquica (ver Seção 4.1). O método aqui proposto possui duas etapas principais: uma etapa denominada Filtro e outra chamada de Incremental *Wrapper*. Na etapa Filtro (linhas 6 a 9), cada atributo é avaliado individualmente utilizando-se a medida $SU_H$ (linha 7) e, a partir da pontuação recebida por cada um deles, calcula-se a sua probabilidade de ser selecionado durante o processo de construção do *ranking* de atributos, que acontece na segunda etapa do método (Incremental *Wrapper*).

Após concluída a etapa Filtro, inicia-se a etapa Incremental *Wrapper*, que é executada $Max\_Iter$ vezes (linhas 13 a 30). Dessa forma, a cada iteração, uma nova solução (subconjunto de atributos) é construída e avaliada por um classificador. Essa etapa, que foi baseada no algoritmo *Incremental Wrapper Subset Selection* proposto em Ruiz et al.

(2006), é detalha a seguir.

Inicialmente, um *ranking* dos atributos da base é construído utilizando-se o método roleta (Goldberg, 1989), onde a chance de um atributo ser selecionado é proporcional à sua avaliação ($SU_H$) em comparação com a avaliação de todos os demais atributos preditivos da base. Ou seja, os atributos melhor avaliados (maior $SU_H$) terão mais chance de serem selecionados nas primeiras rodadas do método roleta, ocupando as posições iniciais do *ranking*. A diversidade das soluções avaliadas (subconjuntos de atributos) se dá pela construção de um novo *ranking* de atributos a cada nova iteração do método.

Numa determinada iteração, após a obtenção do *ranking*, inicia-se a etapa de construção de um subconjunto de atributos que será avaliado por um classificador. A construção do subconjunto ($S'$) começa com a seleção do primeiro atributo do *ranking* (linha 16) e a sua avaliação é realizada por um classificador (linha 17). Em seguida, iterativamente e seguindo a ordem estabelecida no *ranking*, novos atributos são adicionados à solução corrente ($S'$) se a adição dos mesmos resultar em um subconjunto melhor avaliado pelo classificador (linhas 20 a 26). Vale ressaltar que a avaliação feita pelo classificador utiliza o método 5-valiadação cruzada e a medida $hF$ (*hierarchical F-measure*). A comparação entre dois subconjuntos (linhas 23 e 27) é realizada por uma função denominada *relevância*, que será explica nos próximos parágrafos. A cada iteração do método, terminada a construção do subconjunto, se ele for melhor do que todos aqueles construídos nas iterações anteriores, então ele é armazenado como o melhor subconjunto já construído ($S^*$) até o momento (linhas 27 a 29). Finalizadas as iterações, o método retorna o melhor subconjunto de atributos construído.

A função *relevância*, utilizada na comparação entre dois subconjuntos de atributos, é apresentada no Algoritmo 3. Essa função recebe como parâmetro, além dos subconjuntos a serem comparados ($S'$ e $S''$), um valor $k$ que indica a quantidade mínima de partições (geradas pelo método 5-validação cruzada) nas quais um subconjunto deve ter obtido desempenho superior ao do outro para ser considerado melhor. Além da quantidade de partições, para ser considerado melhor, o subconjunto também deve apresentar uma avaliação média ($hF$) superior ao do outro subconjunto. O valor do parâmetro $k$ ($2 \leq k \leq 4$) é aleatoriamente escolhido a cada iteração do método de seleção de atributos (linha 15 do Algoritmo 2).

**Algoritmo 2:** Heurística de Seleção de Atributos para Classificação Hierárquica Monorrótulo.

**Entrada:** Conjunto de Dados $D$, $Max\_Iter$
**Saída:** Subconjunto de atributos: $S^*.subcj$
$S^*.subcj = \emptyset$
$S^*.h\vec{F} = \vec{0}$
$soma = 0$
**para cada** $A_i \in D$ **faça**
    $pontuacao[i] = SU_H(A_i, C)$ // C = Atributo Classe
    $soma+ = pontuacao[i]$
**fim**
**para cada** $A_i \in D$ **faça**
    $prob[i] = pontuacao[i]/soma$
**fim**
**para** $i = 1$ **to** $Max\_Iter$ **faça**
    $Rank[\,] =$ roleta() //atributos ranqueados a partir de uma seleção probabilística
    $k = random(2, 4)$ // número aleatório $\in$ [2..4]
    $S'.subcj = \{Rank[1]\}$
    $S'.h\vec{F} = Classificador(S'.subcj, D)$
    $S''.subcj = \emptyset$
    $S''.h\vec{F} = \vec{0}$
    **para** $w = 2$ **to** $Rank.size$ **faça**
        $S''.subcj = S'.subcj \cup \{Rank[w]\}$
        $S''.h\vec{F} = Classificador(S''.subcj, D)$
        **se** $relevancia(S'', S', k)$ **então**
            $S' = S''$
        **fim**
    **fim**
    **se** $relevancia(S', S^*, k)$ **então**
        $S^* = S'$
    **fim**
**fim**
**retorna** $S^*.subcj$

Portanto, a função *relevância* começa com o cálculo do valor médio de $hF$ de cada um dos dois subconjuntos (linhas 1 e 2) considerando-se as partições da base obtidas pelo método 5-validação cruzada. Em seguida, verifica-se se o desempenho médio de um subconjunto é superior ao do outro (linha 4) e, além disso, se a quantidade de partições nas quais ele é superior é pelo menos igual a $k$ (linhas 5 a 12). Se as duas condições forem verdadeiras, a função retorna o valor *verdadeiro*, indicando a superioridade de um subconjunto em relação ao outro.

**Algoritmo 3:** Função *relevancia*, utilizada na comparação entre dois subconjuntos de atributos.

**Entrada:** $S'$, $S''$, $k$

```
1  S'.hF = ( Σ_{i=1}^{5} S'.hF[i] )/5
2  S''.hF = ( Σ_{i=1}^{5} S''.hF[i])/5
3  cont = 0
4  se S'.hF > S''.hF então
5  |  para i = 1 to 5 faça
6  |  |  se S'.hF[i] > S''.hF[i] então
7  |  |  |  cont++
8  |  |  fim
9  |  fim
10 |  se cont ≥ k então
11 |  |  retorna Verdadeiro
12 |  fim
13 fim
14 retorna Falso
```

# Capítulo 5

# Experimentos Computacionais

Este capítulo apresenta os experimentos computacionais realizados para avaliar a medida $SU_H$ a partir de dois métodos de seleção de atributos propostos neste trabalho. Todas as avaliações foram conduzidas para o contexto da classificação hierárquica monorrótulo empregando-se um classificador hierárquico global. Na Seção 5.1 são descritas as características dos conjuntos de dados utilizados nos experimentos e o pré-processamento aplicado nos mesmos. Em seguida, as configurações dos experimentos realizados com o método de ranqueamento e os resultados obtidos são mostrados na Seção 5.2. Por fim, a Seção 5.3 apresenta os experimentos conduzidos com o método heurístico *HFS4HC* e discute os resultados obtidos.

## 5.1 Bases de Dados

Todos os experimentos foram conduzidos a partir de dez bases de dados relacionadas com classificação de funções de genes. Nessas bases, os atributos preditores incluem diversos tipos de dados da área de bioinformática, tais como: estrutura secundária da sequência, fenótipo, homologia, estatísticas da sequência e outros. Essas bases de dados[1] foram inicialmente apresentadas em Clare e King (2003) e depois utilizadas em outros trabalhos, como por exemplo em Clare (2004), Vens et al. (2008) e Merschmann e Freitas

[1] `https://dtai.cs.kuleuven.be/clus/hmcdatasets/`

(2013). Vale ressaltar que as bases originalmente propostas em Clare e King (2003) são multirrótulo, ou seja, cada instância está associada a uma ou mais classes da estrutura hierárquica.

Como neste trabalho estamos lidando com problemas de classificação hierárquica monorrótulo, nessas bases mencionadas anteriormente, as instâncias multirrotuladas foram convertidas para instâncias associadas a uma única classe através do seguinte pré-processamento: inicialmente contabilizou-se a frequência de todas as classes do conjunto de dados original e, em seguida, para cada instância associada a mais de uma classe, manteve-se apenas a classe que era a mais frequente na base original. Dessa forma, todas as bases de dados se tornaram monorrótulo.

Uma segunda etapa de pré-processamento foi realizada para substituição dos valores ausentes de atributos dessas bases. O procedimento descrito a seguir foi adotado para a substituição dos valores ausentes. Quando identificado um valor ausente para um determinado atributo $A_i$ de uma instância associada à classe $C_j$, calcula-se a média dos valores conhecidos do atributo $A_i$ de todas as demais instâncias da base associadas à classe $C_j$ e, em seguida, utiliza-se essa média para substituição do valor ausente. Se para a classe $C_j$ nenhuma instância possuir valor conhecido para o atributo $A_i$, calcula-se a média dos valores conhecidos do atributo $A_i$ de todas as instâncias da base associadas às classes descendentes de $C_j$ na hierarquia e, em seguida, utiliza-se essa média para substituição do valor ausente. Em último caso, se a classe $C_j$ não possuir classes descendentes ou se para as classes descendentes de $C_j$ nenhuma instância possuir valor conhecido para o atributo $A_i$, então substitui-se o valor ausente pela média global do atributo $A_i$.

Numa terceira etapa de pré-processamento, para garantir a representatividade de todas as classes da base de dados, realizou-se um pré-processamento de modo que toda classe representada por menos de dez instâncias foi agregada com a sua classe pai, sendo esse processo repetido até que todas as classes da hierarquia fossem representadas por pelo menos dez instâncias. Quando nesse processo de agregação a classe mais específica da instância tornou-se a raiz da hierarquia, então essa instância foi eliminada da base de dados.

Por fim, numa quarta etapa de pré-processamento, todos os atributos contínuos foram discretizados utilizando-se o método de discretização não supervisionado *Equal Frequency Binning* (com 20 intervalos). A Tabela 5.1 mostra as principais características das bases de dados após as etapas de pré-processamento citadas nos parágrafos anteriores. Essa tabela apresenta, para cada base de dados, o seu número de atributos, número de

instâncias e o número de classes em cada nível da hierarquia ($1^\circ/2^\circ/3^\circ/4^\circ$ níveis).

**Tabela 5.1:** Características das bases de dados.

| Conjunto de dados | # Atributos | # Instâncias | # Classes/Nível |
|---|---|---|---|
| CellCycle | 78 | 3651 | 4/24/30/16 |
| Church | 28 | 3659 | 4/24/31/16 |
| Derisi | 64 | 3596 | 4/23/29/15 |
| Eisen | 80 | 2308 | 4/15/19/14 |
| Expr | 552 | 3681 | 5/23/31/16 |
| Gasch1 | 174 | 3676 | 4/24/31/16 |
| Gash2 | 53 | 3691 | 5/23/31/16 |
| Phenotype | 69 | 1520 | 5/17/16/9 |
| Sequence | 479 | 3813 | 5/23/31/16 |
| SPO | 81 | 3592 | 4/24/29/15 |

## 5.2 Avaliação do Método de Ranqueamento

Esse conjunto de experimentos foi realizado com objetivo de se obter uma primeira avaliação sobre a adequação da medida Incerteza Simétrica Hierárquica (*Hierarchical Symmetrical Uncertainty* – $SU_H$) para o contexto de classificação hierárquica monorrótulo. Para isso, avaliou-se a capacidade da medida $SU_H$ na geração de um bom *ranking* de atributos. A avaliação aqui conduzida parte do pressuposto que um bom *ranking* de atributos é aquele em que os atributos mais relevantes aparecem nas primeiras posições do mesmo. Por outro lado, um *ranking* pode ser considerado ruim quando se tem uma distribuição uniforme dos atributos mais relevantes ao longo das posições do mesmo, ou seja, um *ranking* aleatório (Slavkov et al., 2014).

Desse modo, a avaliação da medida $SU_H$ foi realizada a partir de execuções do método de ranqueamento apresentado no Algoritmo 1. Basicamente, dada uma base de dados, a ideia é utilizar o método de ranqueamento para construir um *ranking* dos atributos preditores dessa base e, em seguida, avaliar o desempenho de um classificador hierárquico global treinado a partir dessa mesma base contendo somente os $k$ melhores atributos desse *ranking*. Considerando-se que a base de dados contém $n$ atributos preditores, por

meio de um procedimento incremental, avalia-se $n$ vezes o desempenho do classificador, ou seja, uma vez para cada valor de $k$, com $k$ variando de 1 a $n$. Como base de comparação, esse mesmo procedimento incremental foi executado a partir de *rankings* de atributos gerados aleatoriamente. Portanto, se a medida $SU_H$ for realmente capaz de construir *rankings* onde os atributos mais relevantes aparecem nas primeiras posições dos mesmos, a expectativa é de que o classificador avaliado obtenha desempenhos preditivos melhores do que aqueles alcançados a partir dos *rankings* gerados de modo aleatório.

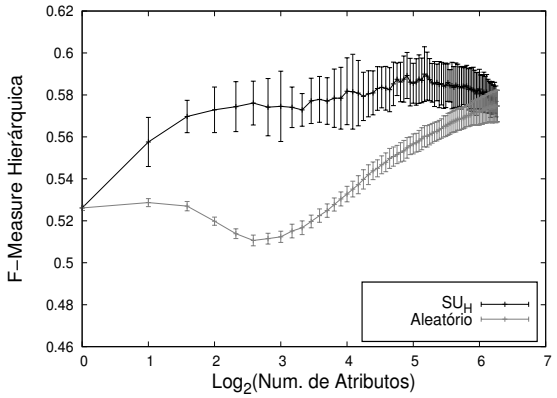
Para avaliação comparativa entre o *ranking* de atributos gerado pela medida $SU_H$ e o *ranking* gerado aleatoriamente, o procedimento incremental descrito anteriormente usou o método de classificação hierárquica proposto em Silla Jr. e Freitas (2009), denominado *Global-Model Naive Bayes* (*GMNB*). Esse classificador, que segue a abordagem global, foi descrito na Seção 3.2 do Capítulo 3.

O método $k$-validação cruzada ($k$=10) foi utilizado na avaliação do desempenho preditivo do *GMNB*. Além disso, para cada base de dados, os mesmos dez pares de bases (treinamento e teste) foram utilizados na avaliação dos dois *rankings* de atributos que estão sendo comparados. Para expressar o desempenho preditivo do classificador hierárquico *GMNB* adotou-se a métrica *F-measure* hierárquica (*hF*).

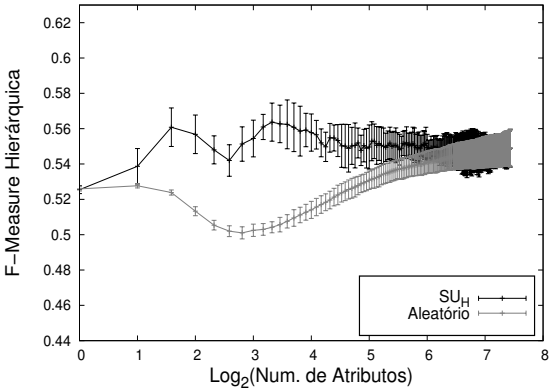
Vale ressaltar que, para o *ranking* aleatório, os resultados de desempenho do classificador *GMNB* correspondem a uma média obtida a partir de 50 rankings aleatórios construídos para cada base de dados avaliada de acordo com o procedimento incremental descrito anteriormente.

Ao final da execução do procedimento incremental, para cada base de dados avaliada, um vetor de tamanho $n$, onde $n$ é o número de atributos preditores da base, é produzido contendo o desempenho preditivo de cada um dos modelos de classificação induzidos. A partir desses vetores, um gráfico é construído para permitir a visualização e análise do desempenho obtido a partir dos *rankings* de atributos gerados utilizando-se a medida $SU_H$ e de modo aleatório.

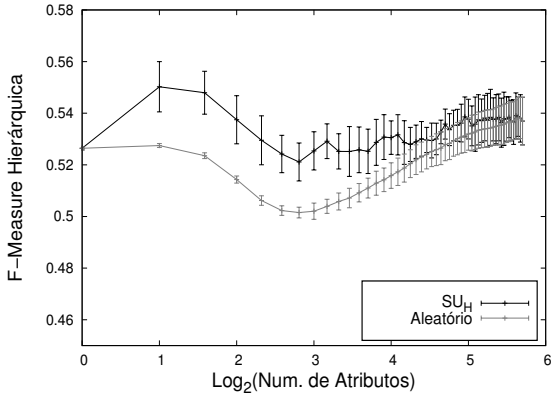
As Figuras 5.1 e 5.2 apresentam os gráficos obtidos para cada uma das dez bases de dados de funções de genes. Em cada gráfico, o eixo $X$ representa o número de atributos preditores da base de dados utilizada no treinamento do classificador *GMNB* e o eixo $Y$ apresenta o desempenho preditivo médio do mesmo em termos de *hF*. As barras verticais correspondem ao desvio padrão da média.



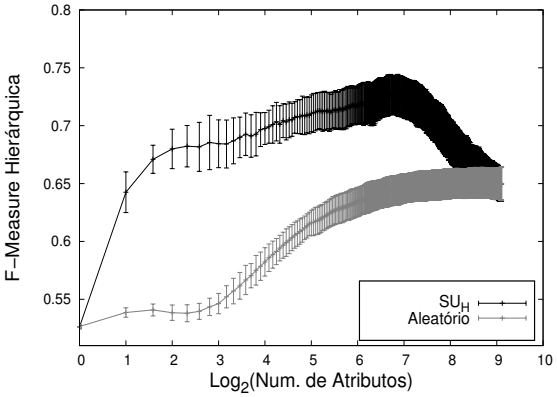

(a) Desempenho do classificador para a base CellCycle.




(b) Desempenho do classificador para a base Gasch1.

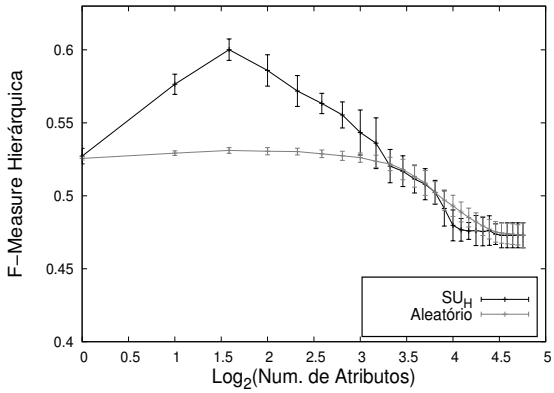



(c) Desempenho do classificador para a base Gasch2.

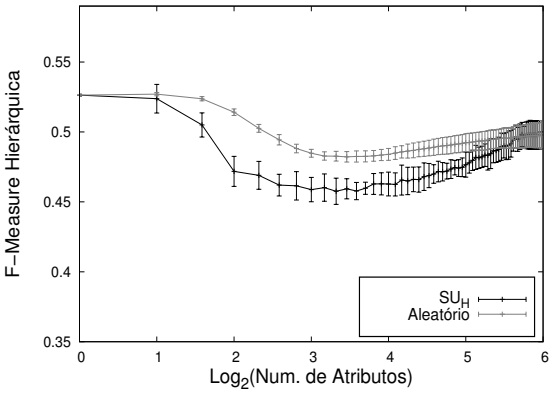



(d) Desempenho do classificador para a base Expr.

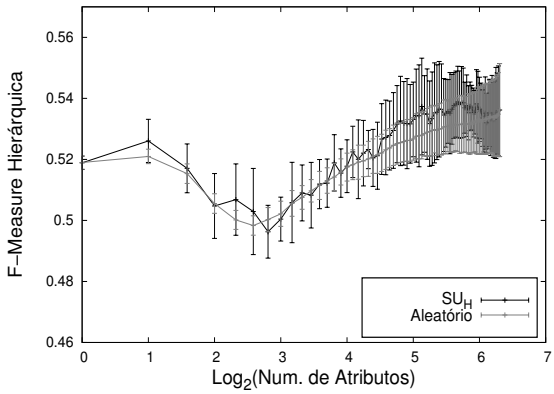
**Figura 5.1:** Gráficos de desempenho do classificador *GMNB* para as bases de dados CellCycle, Gasch1, Gasch2 e Expr.



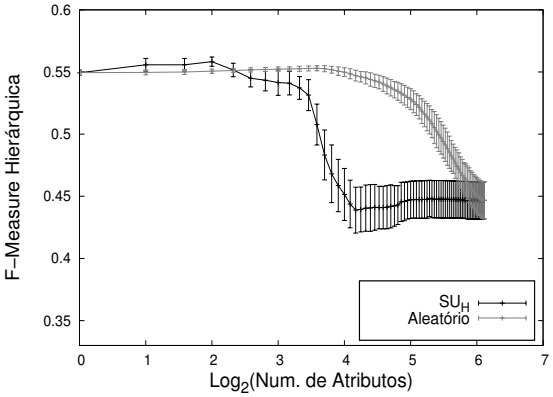

(a) Desempenho do classificador para a base Church.



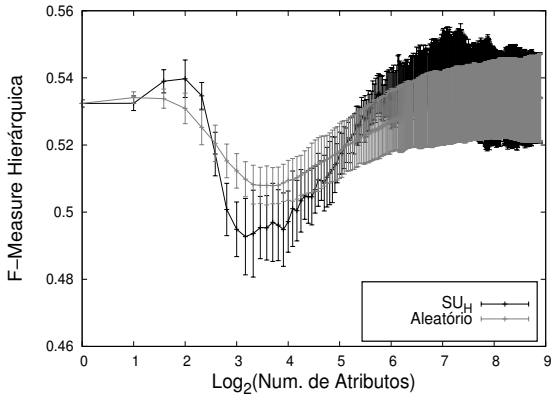

(b) Desempenho do classificador para a base Derisi.



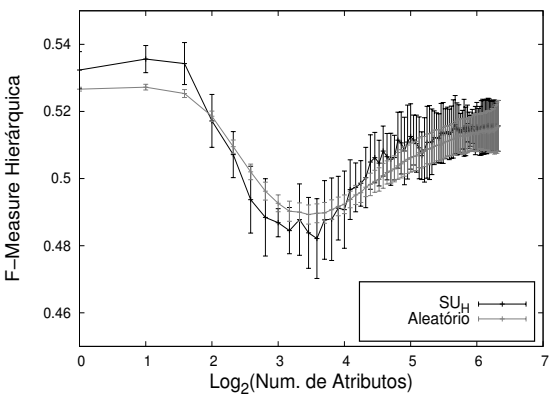

(c) Desempenho do classificador para a base Eisen.




(d) Desempenho do classificador para a base Phenotype.




(e) Desempenho do classificador para a base Sequence.




(f) Desempenho do classificador para a base SPO.

**Figura 5.2:** Gráficos de desempenho do classificador *GMNB* para as bases de dados Church, Derisi, Eisen, Phenotype, Sequence e SPO.

Inicialmente, podemos observar que, nos gráficos da Figura 5.1, a curva de desempenho do classificador para os subconjuntos de atributos obtidos a partir do *ranking* construído utilizando-se a medida $SU_H$ está sempre acima daquela que representa o desempenho do classificador para os subconjuntos gerados a partir do *ranking* aleatório. Isso significa que, de fato, a medida $SU_H$ consegue medir a qualidade preditiva dos atributos e, portanto, permite a geração de *rankings* onde os atributos mais relevantes aparecem nas primeiras posições do mesmos.

Já nos gráficos de desempenho da Figura 5.2, a curva de desempenho do classificador para os subconjuntos gerados a partir do *ranking* construído por meio da medida $SU_H$ está acima ou coincidente com aquela que representa o desempenho do classificador para os subconjuntos gerados a partir do *ranking* aleatório apenas para os menores subconjuntos (primeiros atributos do *ranking*). A explicação para esse fenômeno, comprovada a partir dos experimentos apresentados na próxima seção, é que essas bases possuem muito poucos atributos relevantes, fazendo com que para subconjuntos de atributos maiores, haja uma alternância do melhor desempenho do classificador ora em favor dos subconjuntos de atributos obtidos a partir do *ranking* construído utilizando-se a medida $SU_H$ e ora em favor dos subconjuntos gerados a partir do *ranking* aleatório.

## 5.3 Avaliação do Método Heurístico

Um segundo conjunto de experimentos foi realizado para avaliar o desempenho do método heurístico proposto neste trabalho, o *HFS4HC*, que também utiliza a medida $SU_H$ em uma de suas etapas.

Dada a inexistência na literatura de propostas de métodos de seleção de atributos para problemas classificação hierárquica monorrótulo que utilizam classificadores globais, o método aqui proposto foi avaliado comparando-se o desempenho do classificador hierárquico *GMNB* obtido a partir das bases de dados completas (sem seleção de atributos) e reduzidas (contendo somente os atributos selecionados pelo método proposto).

Além de ser adotado na avaliação do método de seleção de atributos, o classificador *GMNB* também é utilizado na segunda etapa (Incremental Wrapper) do método *HFS4HC* (ver Algoritmo 2). Em todos os testes executados com a heurística *HFS4HC*, o parâmetro $Max\_Iter$ (ver Algoritmo 2) foi definido com o valor de 50 iterações. O procedimento de

$k$-validação cruzada ($k$=10) foi utilizado na avaliação do desempenho preditivo do *GMNB* (exceto na segunda etapa do *HFS4HC*, quando adotou-se $k$=5). Para cada base de dados, as mesmas instâncias dos dez pares de bases (treinamento e teste) foram utilizadas na avaliação do classificador para os dois tipos de bases avaliadas (completa e reduzida). Para expressar o desempenho preditivo do classificador hierárquico *GMNB*, adotou-se a métrica *F-measure* hierárquica ($hF$). Além disso, para cada base de dados, de modo a determinar se existe diferença com significância estatística entre os resultados apresentados pelo classificador para as bases completa e reduzida (após a seleção), o teste estatístico *Wilcoxon's Signed-Rank Test* (*two-sided test*) foi utilizado (Japkowicz e Shah, 2011).

A Tabela 5.2 mostra os resultados obtidos pelo classificador *GMNB* a partir das bases de dados completas (sem seleção de atributos) e reduzidas (após a seleção de atributos com o *HFS4HC*). Nessa tabela, as duas primeiras colunas apresentam os nomes das bases de dados e a quantidade de atributos das mesmas (sem a seleção de atributos). Na terceira coluna, têm-se os resultados de desempenho ($hF$ médio e desvio-padrão) do *GMNB* para as bases de dados completas. Em seguida, as próximas duas colunas mostram os resultados de desempenho ($hF$ médio e desvio-padrão) do *GMNB* para as bases de dados reduzidas (após a seleção de atributos) e o número de atributos das mesmas. Em cada linha, o melhor desempenho preditivo está destacado em negrito. Por fim, a última coluna apresenta qual estratégia (sem seleção ou com seleção pelo *HFS4HC*) obteve o melhor desempenho em termos de $hF$ com significância estatística ou o símbolo (–) indicando que a diferença entre elas não é estatisticamente significativa.

**Tabela 5.2:** Resultados comparativos dos experimentos.

| Bases de dados | # Atributos | Sem Seleção hF (desvio padrão) | *HFS4HC* hF (desvio padrão) | # Atributos Selecionados | Resultados do Teste Estatístico ($\alpha = 0.05$) |
|---|---|---|---|---|---|
| CellCycle | 77 | **0,57** (0,007) | **0,57** (0,012) | 18 | – |
| Church | 27 | 0,47 (0,008) | **0,60** (0,007) | 3 | ***HFS4HC*** |
| Derisi | 63 | 0,49 (0,010) | **0,53** (0,001) | 2 | ***HFS4HC*** |
| Eisen | 79 | **0,53** (0,015) | 0,52 (0,005) | 2 | – |
| Expr | 551 | 0,64 (0,014) | **0,73** (0,015) | 37 | ***HFS4HC*** |
| Gasch1 | 173 | 0,54 (0,010) | **0,57** (0,011) | 21 | ***HFS4HC*** |
| Gash2 | 52 | 0,53 (0,009) | **0,55** (0,008) | 2 | ***HFS4HC*** |
| Phenotype | 69 | 0,44 (0,014) | **0,56** (0,004) | 6 | ***HFS4HC*** |
| Sequence | 478 | 0,53 (0,013) | **0,54** (0,006) | 2 | ***HFS4HC*** |
| SPO | 80 | 0,51 (0,007) | **0,53** (0,007) | 3 | ***HFS4HC*** |

A partir dos resultados apresentados na Tabela 5.2 observa-se que, para oito bases de dados, o classificador alcançou desempenho preditivo significativamente superior

(com 95% de confiança) após a seleção de atributos realizada pelo método heurístico proposto neste trabalho. Para outras duas bases, não houve diferença estatisticamente significativa entre os resultados do classificador para as versões completa e reduzida da base. Além disso, para todas as bases, o método de seleção foi capaz de obter uma redução significativa da quantidade de atributos.

Portanto, com base nos resultados apresentados, verifica-se que o método de seleção de atributos apresentado neste trabalho foi capaz de selecionar subconjuntos de atributos que mantiveram ou melhoraram o desempenho preditivo do classificador hierárquico global *GMNB*, além de alcançar uma redução significativa do número de atributos para todas as bases de dados avaliadas.

# Capítulo 6

# Conclusão

Apesar da importância da seleção de atributos para a tarefa de classificação, até onde se tem conhecimento, para problemas de classificação hierárquica monorrótulo, não existem na literatura propostas de técnicas de seleção de atributos que possam ser utilizadas em conjunto com classificadores hierárquicos globais.

Portanto, este trabalho apresenta as seguintes contribuições: (1) proposta e avaliação de uma adaptação da medida Incerteza Simétrica (*Symmetrical Uncertainty – SU*) para permitir que ela possa ser utilizada em técnicas de seleção de atributos para problemas de classificação hierárquica monorrótulo que utilizam classificadores hierárquicos globais; (2) avaliação dessa adaptação proposta, denominada Incerteza Simétrica Hierárquica (*Hierarchical Symmetrical Uncertainty* – $SU_H$), em uma técnica de seleção de atributos que utiliza a abordagem Filtro e (3) proposta e avaliação de um método heurístico de seleção de atributos, denominado *Hybrid Feature Selection for Hierarchical Classification – HFS4HC*, que implementa uma abordagem híbrida (Filtro e *Wrapper*) e utiliza a medida $SU_H$ proposta.

A avaliação da medida $SU_H$ para o contexto hierárquico utilizando um método de ranqueamento que segue a abordagem Filtro, mostrou a adequação da adaptação proposta para o cenário de classificação hierárquica. A avaliação foi realizada comparando-se o *ranking* de atributos gerado a partir da medida $SU_H$ e outro gerado de modo aleatório. Nessa avaliação comparativa, o *ranking* construído utilizando-se a medida $SU_H$ resultou em desempenhos preditivos do classificador hierárquico *GMNB* sempre superiores àqueles alcançados pelo classificador a partir do *ranking* aleatório.

Um segundo conjunto de experimentos foi realizado para avaliar o desempenho do método heurístico proposto neste trabalho, o *Hybrid Feature Selection for Hierarchical Classification – HFS4HC*, que também utilizou a medida $SU_H$ em uma de suas etapas. Das dez bases de dados utilizadas nos experimentos, para oito delas o classificador hierárquico *GMNB* alcançou desempenho preditivo significativamente superior após a seleção de atributos realizada pelo método *HFS4HC*. Para as outras duas, não houve diferença de desempenho com significância estatística. Além disso, a heurística proposta alcançou uma redução significativa do número de atributos para todas as bases de dados avaliadas. Portanto, a partir desses resultados podemos concluir que o método proposto, que fez uso da medida $SU_H$, apresentou resultados bastante promissores para o cenário de classificação hierárquica monorrótulo.

Os resultados apresentados neste trabalho são importantes por evidenciarem o potencial de utilização da medida $SU_H$ e da heurística *HFS4HC* no contexto de classificação hierárquica monorrótulo para os casos onde se deseja utilizar um classificador hierárquico global. No entanto, uma vez que não há garantia de que classificadores hierárquicos globais alcancem desempenhos preditivos superiores aos de conjuntos de classificadores locais (abordagens locais), um trabalho futuro relevante corresponde a uma avaliação comparativa entre os resultados apresentados neste trabalho e aqueles que seriam obtidos a partir da medida $SU$ tradicional avaliada com um conjunto de classificadores locais planos. Além disso, trabalhos futuros envolvendo a utilização da medida $SU_H$ em outras abordagens de seleção de atributos correspondem a uma continuidade natural desta pesquisa.

# Referências Bibliográficas


Barutcuoglu, Z. e DeCoro, C.: 2006, Hierarchical shape classification using bayesian aggregation, *Proceedings of the IEEE International Conference on Shape Modeling and Applications 2006*, IEEE Computer Society.

Bermejo, P., Gámez, J. A. e Puerta, J. M.: 2014, Speeding up incremental wrapper feature subset selection with naive bayes classifier, *Knowledge-Based Systems* pp. 140–147.

Bermejo, P., Gámez, J. A. e Puerta, J. M.: 2011, A grasp algorithm for fast hybrid (filter-wrapper) feature subset selection in high-dimensional datasets, *Pattern Recognition Letters* pp. 701–711.

Blum, A. L. e Langley, P.: 1997, Selection of relevant features and examples in machine learning, *Artificial Intelligence* pp. 245–271.

Chen, Y.-L., Hu, H.-W. e Tang, K.: 2009, Constructing a decision tree from data with hierarchical class labels, *Expert Systems with Applications* pp. 4838–4847.

Clare, A.: 2004, *Machine learning and data mining for yeast functional genomics*, PhD thesis, University of Wales Aberystwyth, United Kingdom.

Clare, A. e King, R. D.: 2003, Predicting gene function in saccharomyces cerevisiae, *Bioinformatics* pp. 42–49.

Costa, E. P., Lorena, A. C., Carvalho, A. C. P. L. F., A.Freitas, A. e Holden, N.: 2007, Comparing several approaches for hierarchical classification of proteins with decision trees, *Advances in Bioinformatics and Computational Biology*, Springer Berlin Heidelberg, pp. 126–137.

Cover, T. M. e Thomas, J. A.: 1991, *Elements of Information Theory*, Wiley-Interscience.

Duda, R. O. e Hart, P. E.: 1973, *Pattern recognition and scene analysis*, Vol. 32, Wiley, New York.

Esseghir, M. A.: 2010, Effective wrapper-filter hybridization through grasp schemata, *Journal of Machine Learning Research - Proceedings Track* pp. 45–54.

Fayyad, U. M., Piatetsky-Shapiro, G. e Smyth, P.: 1996, Advances in knowledge discovery and data mining, American Association for Artificial Intelligence, Menlo Park, CA, USA, chapter From Data Mining to Knowledge Discovery: An Overview, pp. 1–34.

Garner, S.: 1995, Weka: The waikato environment for knowledge analysis, *Proceedings of the New Zealand Computer Science Research Students Conference*, pp. 57–64.

Goldberg, D. E.: 1989, *Genetic Algorithms in Search, Optimization and Machine Learning*, Addison-Wesley Longman Publishing Co., Inc., Boston, MA, USA.

Guyon, I. e Elisseeff, A.: 2006, An introduction to feature extraction, *Feature Extraction*, Springer Berlin Heidelberg, pp. 1–25.

Hall, M. A.: 1998, *Correlation-based Feature Subset Selection for Machine Learning*, PhD thesis, University of Waikato, Hamilton, New Zealand.

Hall, M. A.: 2000, Correlation-based feature selection for discrete and numeric class machine learning, *Proceedings of the Seventeenth International Conference on Machine Learning*, Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, pp. 359–366.

Hall, M. A. e Holmes, G.: 2003, Benchmarking attribute selection techniques for discrete class data mining, *IEEE Transactions on Knowledge and Data Engineering* pp. 1437–1447.

Hall, M. A. e Smith, L. A.: 1999, Feature selection for machine learning: Comparing a correlation-based filter approach to the wrapper, *Proceedings of the Twelfth International Florida Artificial Intelligence Research Society Conference, May 1-5, 1999, Orlando, Florida, USA*, pp. 235–239.

Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P. e Witten, I. H.: 2009, The weka data mining software: An update, *ACM SIGKDD Explorations Newsletter* pp. 10–18.

Han, J., Kamber, M. e Pei, J.: 2011, *Data Mining: Concepts and Techniques*, 3rd edn, Morgan Kaufmann Publishers Inc., San Francisco, CA, USA.

Hu, Z., Bao, Y., Xiong, T. e Chiong, R.: 2015, Hybrid filter–wrapper feature selection for short-term load forecasting, *Engineering Applications of Artificial Intelligence* pp. 17–27.

Inza, I., Larrañaga, P., Blanco, R. e Cerrolaza, A. J.: 2004, Filter versus wrapper gene selection approaches in dna microarray domains, *Artificial Intelligence in Medicine* pp. 91–103.

Japkowicz, N. e Shah, M.: 2011, *Evaluating Learning Algorithms: A Classification Perspective*, Cambridge University Press, New York, USA.

Kannan, S. S. e Ramaraj, N.: 2010, A novel hybrid feature selection via symmetrical uncertainty ranking based local memetic search algorithm, *Knowledge-Based Systems* pp. 580–585.

Kira, K. e Rendell, L. A.: 1992, A practical approach to feature selection, *Proceedings of the Ninth International Workshop on Machine Learning*, Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, pp. 249–256.

Kiritchenko, S., Matwin, S., Nock, R. e Famili, A. F.: 2006, Learning and evaluation in the presence of class hierarchies: Application to text categorization, *Proceedings of the 19th International Conference on Advances in Artificial Intelligence: Canadian Society for Computational Studies of Intelligence*, Springer-Verlag, Berlin, Heidelberg, pp. 395–406.

Kohavi, R.: 1995, A study of cross-validation and bootstrap for accuracy estimation and model selection, *International Joint Conference on Artificial Intelligence*, pp. 1137–1143.

Kohavi, R. e John, G. H.: 1997, Wrappers for feature subset selection, *Artificial Intelligence* pp. 273–324.

Koller, D. e Sahami, M.: 1995, Toward optimal feature selection, *In 13th International Conference on Machine Learning*, pp. 284–292.

Koller, D. e Sahami, M.: 1997, Hierarchically classifying documents using very few words, *Proceedings of the Fourteenth International Conference on Machine Learning*, Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, pp. 170–178.

Liu, H. e Motoda, H.: 2007, *Computational Methods of Feature Selection (Chapman & Hall/Crc Data Mining and Knowledge Discovery Series)*, Chapman & Hall/CRC.

Liu, H. e Setiono, R.: 1996, A probabilistic approach to feature selection - a filter solution, Morgan Kaufmann, pp. 319–327.

Liu, H. e Yu, L.: 2003, Feature selection for high-dimensional data: A fast correlation-based filter solution, *Correlation-Based Filter Solution". In Proceedings of The Twentieth International Conference on Machine Leaning (ICML-03)*, pp. 856–863.

Liu, H. e Yu, L.: 2005, Toward integrating feature selection algorithms for classification and clustering, *IEEE Transactions on Knowledge and Data Engineering* pp. 491–502.

Liu, H., Motoda, H., Setiono, R. e Zhao, Z.: 2010, Feature selection: An ever evolving frontier in data mining, *FSDM*, pp. 4–13.

Merschmann, L. H. C. e Freitas, A. A.: 2013, An extended local hierarchical classifier for prediction of protein and gene functions, *Data Warehousing and Knowledge Discovery*, Springer Berlin Heidelberg, pp. 159–171.

Mladenic, D. e Grobelnik, M.: 2003, Feature selection on hierarchy of web documents, *Decision Support Systems* pp. 45–87.

Paes, B. C., Plastino, A. e Freitas, A. A.: 2014, Exploring attribute selection in hierarchical classification, *Journal of Information and Data Management* pp. 124–133.

Pereira, R. B., Plastino, A., Zadrozny, B., Merschmann, L. H. C. e Freitas, A. A.: 2008, Seleção lazy de atributos - uma nova perspectiva, Anais do IV Workshop em Algoritmos e Aplicações de Mineracao de Dados, Campinas, SP, Brasil.

Pereira, R. B., Plastino, A., Zadrozny, B., Merschmann, L. H. C. e Freitas, A. A.: 2011a, Improving lazy attribute selection, *Journal of Information and Data Management* pp. 447–462.

Pereira, R. B., Plastino, A., Zadrozny, B., Merschmann, L. H. C. e Freitas, A. A.: 2011b, Lazy attribute selection: Choosing attributes at classification time, *Intelligent Data Analysis* pp. 715–732.

Quinlan, J. R.: 1986, Induction of decision trees, *Machine Learning* pp. 81–106.

Quinlan, J. R.: 1993, *C4.5: Programs for Machine Learning*, Morgan Kaufmann Publishers Inc., San Francisco, CA, USA.

Robnik-Sikonja, M. e Kononenko, I.: 2003, Theoretical and empirical analysis of relieff and rrelieff, *Machine Learning* pp. 23–69.

Ruiz, R., Riquelme, J. C. e Aguilar-Ruiz, J. S.: 2006, Incremental wrapper-based gene selection from microarray data for cancer classification, *Pattern Recognition* pp. 2383–2392.

Secker, A., Davies, M. N., Freitas, A. A., Clark, E. B., Timmis, J. e Flower, D. R.: 2010, Hierarchical classification of g-protein-coupled receptors with data-driven selection of attributes and classifiers, *International Journal Data Mining and Bioinformatics* pp. 191–210.

Silla Jr., C. N.: 2011, *Novel Approaches for Hierarchical Classification With Case Studies in Protein Function Prediction*, PhD thesis, Computer Science at the School of Computing, The University of Kent at Canterbury, United Kingdom.

Silla Jr., C. N. e Freitas, A. A.: 2009, A global-model naive bayes approach to the hierarchical prediction of protein functions, *Proceedings of the 2009 Ninth IEEE International Conference on Data Mining*, IEEE Computer Society, pp. 992–997.

Silla Jr., C. N. e Freitas, A. A.: 2011, A survey of hierarchical classification across different application domains, *Data Mining and Knowledge Discovery* pp. 31–72.

Silla Jr., C. N. e Kaestner, C. A. A.: 2013, Hierarchical classification of bird species using their audio recorded songs, *Systems, Man, and Cybernetics (SMC), 2013 IEEE International Conference on*, pp. 1895–1900.

Slavkov, I., Karcheska, J., Kocev, D., Kalajdziski, S. e Džeroski, S.: 2014, Relieff for hierarchical multi-label classification, *New Frontiers in Mining Complex Patterns*, Springer International Publishing, pp. 148–161.

Taheri, N. e Nezamabadi-pour, H.: 2014, A hybrid feature selection method for high-dimensional data, *Computer and Knowledge Engineering (ICCKE), 2014 4th International eConference on*, pp. 141–145.

Valentini, G.: 2011, True path rule hierarchical ensembles for genome-wide gene function prediction, *IEEE/ACM Transactions on Computational Biology and Bioinformatics* pp. 832–847.

Vens, C., Struyf, J., Schietgat, L., Džeroski, S. e Blockeel, H.: 2008, Decision trees for hierarchical multi-label classification, *Machine Learning* pp. 185–214.

Witten, I. H., Frank, E. e Hall, M. A.: 2011, *Data Mining: Practical Machine Learning Tools and Techniques*, 3rd edn, Morgan Kaufmann Publishers Inc., San Francisco, CA, USA.

Wu, F., Zhang, J. e Honavar, V.: 2005, Learning classifiers using hierarchically structured class taxonomies, *Proceedings of the 6th International Conference on Abstraction, Reformulation and Approximation*, Springer-Verlag, pp. 313–320.

Yang, Y. e Pedersen, J. O.: 1997, A comparative study on feature selection in text categorization, *Proceedings of the Fourteenth International Conference on Machine Learning*, Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, pp. 412–420.

Yusta, S. C.: 2009, Different metaheuristic strategies to solve the feature selection problem, *Pattern Recognition Letters* pp. 525–534.