UNIVERSIDADE TECNOLÓGICA FEDERAL DO PARANÁ

RAIMUNDO OSVALDO VIEIRA

UM MÉTODO PARA SELEÇÃO DE ATRIBUTOS EM BASES DE DADOS DE CLASSIFICAÇÃO HIERÁRQUICA MULTIRRÓTULO

PONTA GROSSA
2022

RAIMUNDO OSVALDO VIEIRA

UM MÉTODO PARA SELEÇÃO DE ATRIBUTOS EM BASES DE DADOS DE CLASSIFICAÇÃO HIERÁRQUICA MULTIRRÓTULO

A method for feature selection on hierarchical multilabel classification datasets

Dissertação apresentada como requisito para obtenção do título de Mestre em Ciência da Computação da Universidade Tecnológica Federal do Paraná (UTFPR).

Orientador(a): Profa. Dra. Helyane Bronoski Borges.

PONTA GROSSA

2022

Ministério da Educação
Universidade Tecnológica Federal do Paraná
Campus Ponta Grossa

RAIMUNDO OSVALDO VIEIRA

UM MÉTODO PARA SELEÇÃO DE ATRIBUTOS EM BASES DE DADOS DE CLASSIFICAÇÃO HIERÁRQUICA MULTIRRÓTULO

Trabalho de pesquisa de mestrado apresentado como requisito para obtenção do título de Mestre Em Ciência Da Computação da Universidade Tecnológica Federal do Paraná (UTFPR). Área de concentração: Sistemas E Métodos De Computação.

Data de aprovação: 07 de Julho de 2022

Dra. Helyane Bronoski Borges, Doutorado - Universidade Tecnológica Federal do Paraná

Dr. Bruno Feres De Souza, Doutorado - Universidade Federal do Maranhão (Ufma)

Dra. Simone Nasser Matos, Doutorado - Universidade Tecnológica Federal do Paraná

Dedico este trabalho a todos aqueles que me incentivaram a chegar até aqui.

AGRADECIMENTOS

A gratidão, mais que uma atitude, é uma linguagem que pode ser utilizada para expressar amor. Sabendo disso, quero aqui registrar minha gratidão àqueles que contribuíram para que este trabalho fosse realizado.

Agradeço, antes de tudo, a Deus, por ter me criado, por Sua decisão irrevogável por mim e por Seu Amor e Providência que sustentam a minha vida e me conduziram nessa jornada.

À minha orientadora Profa. Dra. Helyane Bronoski Borges, por ter me acolhido entre os seus alunos e por toda paciência, delicadeza, sabedoria e segurança com que me guiou nesse caminho tão árduo.

Aos meus queridos irmãos e amigos da Comunidade Católica Shalom – Missão de Ponta Grossa (PR), por todo o amor com que me acolheram durante o tempo em que estive na cidade, por terem me apoiado nas dificuldades e serem suporte espiritual e fraterno em todos os momentos. De fato, nunca estive sozinho. De modo particular, quero deixar registrado meu reconhecimento às autoridades da missão, em especial Alexandre Aguiar, por todo zelo com que cuidaram de mim durante a realização deste trabalho.

Às minhas queridas irmãs, formadoras e amigas da Comunidade Católica Shalom – Missão de São Luís (MA) Lucilene e Jaciara Cardoso, por me encorajarem quando eu pensava que não seria possível e por me fazerem entender que Deus tem planos que ultrapassam a razão humana.

À minha colega do PPGCC Tathiana Mikamura Barchi, por todo apoio, incentivo e socorro nas horas de aflição.

À profa. Dra. Simone Nasser, por sua paciência, generosidade, firmeza e grande incentivo nos momentos difíceis.

A Clau, Lydia, Januária, Fabíola, Angélica, Ana Maria e Alex Lourenço por me ajudarem nos momentos mais angustiantes, por me aconselharem, por terem acreditado e, simplesmente, por terem parado muitas vezes para me ouvir.

Aos meus colegas do Departamento de Computação do IFMA Campus Monte Castelo, pelo incentivo e apoio. Em especial Evaldinolia, Karla Fook, Jeane, Lourdinha, Eveline, João Carlos e Josenildo, por nunca terem deixado de acreditar.

Gostaria de deixar registrado também, o meu reconhecimento à minha família, pois sem ela eu não teria chegado tão longe.

Enfim, a todos os que por algum motivo contribuíram para a realização desta pesquisa.

“Porque vistos de um jeito certo, os erros, eles nos preparam para nossas vitórias e conquistas futuras porque não há aprendizado na vida que não passe pelas experiências dos erros” (O CADERNO, 2008).

RESUMO

Problemas de classificação hierárquica multirrótulo normalmente precisam lidar com conjuntos de dados que possuem grande número de atributos e rótulos, o que pode interferir de forma negativa no desempenho do classificador. A aplicação de métodos de redução de dimensionalidade pode prover uma melhora significativa no desempenho dos classificadores. A seleção de atributos é um dos métodos de redução de dimensionalidade em bases de dados e compreende a escolha dos atributos mais relevantes a partir dos originais. Três abordagens principais para a seleção de atributos podem ser utilizadas: filtro, *wrapper* e embutida. De modo particular, a abordagem filtro faz a seleção baseado apenas nas características dos próprios dados e de maneira independente do algoritmo de treinamento. No contexto da classificação hierárquica multirrótulo, alguns métodos de seleção de atributos têm sido propostos. Estes métodos fazem uso de técnicas consolidadas em contextos de classificação plana e classificação monorrótulo, apresentando bons resultados. Neste sentido, este trabalho verificou a aplicabilidade da medida *Fisher Score* para a seleção de atributos em cenários de classificação hierárquica multirrótulo e propôs um método para esta tarefa utilizando a abordagem filtro. O método FSF-HMC consiste em avaliar os atributos a partir do cálculo individual do *Fisher Score*. Este cálculo foi adaptado para considerar a hierarquia de classes. Os atributos avaliados com pontuação acima do valor médio de *Fisher Score* apurado para todos os atributos são selecionados para compor o conjunto de dados reduzido que será utilizado para avaliação do classificador. Para validação do método proposto foram realizados experimentos com 10 bases de dados da *Gene Ontology*. Tais experimentos consistiram em avaliar o desempenho de dois classificadores hierárquicos multirrótulo, Clus-HMC e MHC-CNN, em termos da medida AUPRC, sendo realizada uma comparação dos resultados produzidos a partir dos conjuntos de dados originais e dos conjuntos de dados reduzidos. Os resultados dos experimentos demonstram que houve um ganho em termos do percentual de redução do número de atributos sobre os dados originais e que o desempenho dos classificadores foi estatisticamente equivalente para os conjuntos de dados originais e reduzidos.

Palavras-chave: seleção de atributos; Fisher Score; redução de dimensionalidade; classificação hierárquica multirrótulo.

ABSTRACT

Hierarchical multi-label classification problems usually need to deal with datasets that have a large number of attributes and labels, which can negatively interfere with the performance of the classifier. The application of dimensionality reduction methods can provide a significant improvement in the performance of classifiers. Feature selection is one of the dimensionality reduction methods in databases and comprises choosing the most relevant attributes from the originals. Three main approaches to feature selection can be used: filter, wrapper and embedded. In particular, the filter approach makes the selection based only on the characteristics of the data itself and independently of the training algorithm. In the context of hierarchical multi-label classification, some feature selection methods have been proposed. These methods make use of consolidated techniques in contexts of flat classification and single-label classification, showing good results. In this sense, this work investigated the applicability of the Fisher Score measure for the feature selection in hierarchical multi-label classification scenarios and proposed a method for this task using the filter approach. The FSF-HMC method consists of evaluating the attributes from the individual calculation of the Fisher Score. This calculation has been adapted to consider the class hierarchy. The attributes evaluated with a score above the average Fisher Score calculated for all attributes are selected to compose the reduced dataset that will be used to evaluate the classifier. To validate the proposed method, experiments were performed with 10 Gene Ontology databases. These experiments consisted of evaluating the performance of two multi-label hierarchical classifiers, Clus-HMC and MHC-CNN, in terms of the AUPRC measure, with a comparison of the results produced from the original datasets and the reduced datasets. The results of the experiments demonstrate that there was a gain in terms of the percentage of reduction in the number of attributes over the original data and that the performance of the classifiers was statistically equivalent for the original and reduced datasets.

Keywords: feature selection; Fisher Score; dimensionality reduction; hierarchical multilabel classification.

LISTA DE FIGURAS

LISTA DE GRÁFICOS

LISTA DE QUADROS

LISTA DE TABELAS

LISTA DE ABREVIATURAS E SIGLAS

AD	Árvore de Decisão
AG	Algoritmo Genético
AM	Aprendizagem de Máquina
AUPRC	*Area Under the Precision Recall Curve*
BR	*Binary Relevance*
BR-χ^2	*Binary Relevance with Chi-Square*
BR-GR	*Binary Relevance with Gain Ratio*
BR-IG	*Binary Relevance with Information Gain*
BR-RF	*Binary Relevance with Relief*
DAG	*Directed Acyclic Graph*
FS	*Fisher Score*
FSF-HMC	*Feature Selection based on Fisher score for Hierarchical Multi-Label Classification*
FSSS	*Feature Selection based on Semantic and Structural Information of Labels*
FSW-HMC	*Feature Selection based on Wrapper approach for Hierarchical Multi-label Classification*
GO	*Gene Ontology*
GR	*Gain Ratio*
HMC	*Hierarchical Multi-label Classification*
HMC-GA	*Hierarchical Multi-Label Classification with Genetic Algorithm*
HMC-LMLP	*Hierarchical Multi-Label Classification with Local Multi-Layer Perceptron*
IG	*Information Gain*
IG-BR	*Information Gain based on the Binary Relevance transformation*
IG-LP	*Information Gain based on the Label Powerset transformation*
LP	*Label Powerset*
LP-χ^2	*Label Powerset with Chi-Square*
LP-GR	*Label Powerset with Gain Ratio*
LP-IG	*Label Powerset with Information Gain*
LP-RF	*Label Powerset with Relief*
MDS	*Multidimensional Scaling*
MHCAIS	*Multi-label Hierarchical Classification with an Artificial Immune System*
MHC-CNN	*Multi-label Hierarchical Classification - Competitive Neural Network*

PCA	*Principal Component Analysis*
PCT	*Predictives Cluster Trees*
RF-BR	*ReliefF based on the Binary Relevance transformation*
RF-LP	*ReliefF based on the Label Powerset transformation*
RNA	Redes Neurais Artificais
SIA	Sistemas Imunológicos Artificiais
SOM	*Self Organizing Map*
SVM	*Support Vector Machines*

SUMÁRIO

1 INTRODUÇÃO

O problema de classificação é um dos mais importantes no campo da Aprendizagem de Máquina (AM) e consiste em atribuir um rótulo a um elemento do conjunto de dados alvo da tarefa de classificação (HAN; KAMBER; PEI, 2012) (RUSSELL; NORVIG, 2013). De modo particular, há contextos nos quais um objeto de um conjunto pode ser associado a mais de uma classe e, além disso, é possível, ainda, que as classes mantenham entre si relações taxonômicas. Tem-se, nestes casos, cenários de classificação multirrótulo e classificação hierárquica, respectivamente (FENG; ZHAO; FU, 2020).

Alguns domínios de aplicação exigem a aplicação simultânea da classificação multirrótulo e da classificação hierárquica. Neste caso, tem-se a classificação hierárquica multirrótulo, do inglês *Hierarchical Multi-label Classification* (HMC), em que exemplos podem ser associados a múltiplas classes e, além disso, ao associar um exemplo a uma classe da hierarquia, também associará esse exemplo a todas as suas classes ancestrais. A HMC pode ser aplicada na categorização de texto, processamento de imagem, predição de proteínas, entre outros (BORGES, 2012) (DIMITROVSKI *et al.*, 2012) (FENG; ZHAO; FU, 2020). As classes nesses domínios podem ter sua hierarquia estruturada como árvore ou como um Grafo Acíclico Dirigido, do inglês *Directed Acyclic Graph* (DAG), sendo que a classificação para estruturas do tipo DAG é mais complexa, pois, nesse caso, toda a estrutura hierárquica é respeitada.

Em geral, aplicações deste tipo possuem um grande volume de dados e um número elevado de atributos e rótulos, o que interfere de forma negativa no desempenho do classificador (HUANG *et al*, 2020). Para lidar com este problema, é necessária a utilização de técnicas para reduzir a dimensionalidade dos dados sem com isso alterar seu significado intrínseco. As abordagens mais conhecidas são a seleção e a extração de atributos e visam melhorar o desempenho da tarefa de classificação (FACELI *et al.,* 2011) (BORGES, 2012).

De modo particular, a seleção de atributos compreende a escolha dos atributos mais relevantes a partir do conjunto de dados original (TANG; ALELYANI; LIU, 2014), sendo, portanto, excluídos os atributos irrelevantes e/ou redundantes. O estado da arte para o problema da seleção de atributos em bases de dados de classificação hierárquica multirrótulo, conforme apontado pelo mapeamento sistemático de literatura realizado neste trabalho, evidencia que, na última década,

têm sido produzidos alguns estudos voltados para o desenvolvimento de métodos de seleção de atributos específicos para o domínio da HMC.

O mapeamento indicou que foram utilizadas nesses estudos, técnicas tradicionais para seleção de atributos em contextos de classificação plana e/ou monorrótulo adaptadas para a HMC, obtendo-se resultados relevantes quanto à redução do número de atributos nas bases de dados e quanto à eficiência da tarefa de classificação nas bases de dados reduzidas. Entretanto, a quantidade de estudos é pequena e não engloba todas as abordagens e técnicas para seleção de atributos existentes.

Este trabalho propõe um método para seleção de atributos em problemas de classificação hierárquica multirrótulo: *Feature Selection based on Fisher score for Hierarchical Multi-Label Classification* (FSF-HMC). O objetivo do método é selecionar o menor subconjunto de atributos que seja capaz de melhorar o desempenho da tarefa de classificação e que possa ser utilizado com qualquer algoritmo para a classificação hierárquica multirrótulo.

O método FSF-HMC foi avaliado através da medida AUPRC (*Area Under the Precision Recall Curve*) (VENS *et* al., 2008) produzida pelos classificadores hierárquicos multirrótulo Clus-HMC e MHC-CNN (*Multi-label Hierarchical Classification - Competitive Neural Network*) quando aplicados sobre os conjuntos de dados originais e os conjuntos de dados reduzidos gerados pela aplicação do método. Para os experimentos foram utilizados 10 conjuntos de dados da *Gene Ontolgy* (GO) estruturados na forma de DAG. Os resultados dos experimentos demonstram que o desempenho dos classificadores em termos da medida AUPRC foi estatisticamente equivalente para os conjuntos de dados originais e reduzidos. Entretanto, houve um ganho em termos do percentual de redução do número de atributos sobre os dados originais, o que confere ao classificador um menor tempo de processamento quando aplicado sobre o conjunto de dados reduzido.

1.1 Definição do problema, contexto e motivação

Na AM a seleção de atributos é utilizada com o propósito de reduzir o tempo de execução dos classificadores, além de aumentar sua capacidade preditiva e de geração de um modelo de aprendizagem que represente de forma mais compacta o

conceito aprendido. Três diferentes abordagens para a seleção de atributos podem ser adotadas: a abordagem filtro, que se baseia nas características dos próprios dados de maneira independente do algoritmo de treinamento; a abordagem *wrapper*, que busca pelo subconjunto de atributos ótimo adaptado a um algoritmo de aprendizagem específico; e a abordagem embutida, que realiza a seleção dos atributos no próprio processo de treinamento (FACELI *et al.*, 2011) (KUMAR; MINZ, 2014).

Problemas de HMC possuem desafios próprios relacionados às características dos dados, sobretudo quanto à organização das classes em estruturas hierárquicas. Além disso, é necessário lidar com a possibilidade de classificação de um exemplo em mais de uma classe simultaneamente. Este cenário afeta o processo de seleção de atributos, evidenciando a necessidade do desenvolvimento de métodos específicos para lidar com esta categoria de problema.

A escolha por uma abordagem específica não é uma tarefa fácil, pois depende das características dos conjuntos de dados, incluindo a alta dimensionalidade em termos do número de amostras e atributos. Como a avaliação de eficácia desses métodos não é fácil e depende do desempenho do mecanismo de aprendizado e do conhecimento sobre as características dos dados, métodos para seleção de atributos têm sido propostos, utilizando diferentes estratégias, como a criação de um novo método, a combinação e/ou adaptação de técnicas existentes, combinação de métodos existentes (KUMAR; MINZ, 2014).

Dentre as abordagens utilizadas para a seleção de atributos em contextos de HMC há uma prevalência da abordagem filtro sobre as abordagens *wrapper* e embutida. Isto pode ser explicado pelo fato de a abordagem filtro ser independente de algoritmo e não possuir tantos algoritmos para HMC disponíveis. Além disso, para conjuntos de dados com alta dimensionalidade, a abordagem filtro é mais adequada por apresentar menor complexidade computacional.

Dentre as técnicas do tipo filtro, têm sido utilizadas, neste contexto, algumas mais tradicionais como *Relief*, *Information Gain* (IG), *Gain Ratio* (GR) e *Chi-Square* (χ^2). Há outras técnicas que são utilizadas com eficácia em domínios de classificação plana e/ou monorrótulo, como, por exemplo, a medida *Fisher Score* (FS) (GU; LI; HAN, 2012). Esta técnica, que funciona como métrica de avaliação da qualidade de atributos, é capaz de avaliar independentemente os atributos numa base de dados, podendo ser utilizada para ranqueamento dos melhores atributos (SUN *et al.*, 2021).

Uma questão que pode ser levantada é se a medida FS pode ser utilizada para pontuar a importância de atributos em contextos de classificação hierárquica multirrótulo, pois, em sua concepção original, esta medida pontua um atributo com base nos valores que esse atributo assume para cada classe a que está associado e, na classificação hierárquica multirrótulo, o mapeamento de um atributo a uma classe exige que esse mapeamento seja estendido às suas classes ancestrais.

A principal motivação deste trabalho é buscar uma resposta para esta questão, uma vez que não há estudos que indiquem se a adoção da medida FS é adequada para domínios de HMC. Diante disso, é relevante a proposição de um método que permita avaliar a aplicabilidade da medida FS para a seleção de atributos em bases de dados de classificação hierárquica multirrótulo.

A abordagem a adotada para o método deve ser a abordagem filtro, pois permite a utilização da métrica FS para avaliar os atributos de maneira independente e, em seguida selecionar aqueles que tenham recebido as melhores pontuações. Além disso, a adoção de uma abordagem independente de algoritmo justifica-se porque tal abordagem demanda baixo custo computacional, uma vez que todo o processo é realizado antes da aplicação do algoritmo de aprendizagem, o que concede maior flexibilidade e adaptabilidade a diferentes tipos de classificadores.

Dessa maneira, motivado pela possibilidade de aplicar a medida FS na seleção de atributos em contextos de classificação hierárquica multirrótulo, este trabalho apresenta um novo método de filtro para seleção de atributos, baseado na medida FS.

1.2 Objetivos

Esta Seção apresenta o objetivo geral e os objetivos específicos deste trabalho.

1.2.1 Objetivo geral

Investigar a aplicabilidade da medida *Fisher Score* para o problema de seleção de atributos em contextos de classificação hierárquica multirrótulo.

1.2.2 Objetivos específicos

- Analisar o contexto da redução de dimensionalidade em bases de dados de classificação hierárquica multirrótulo;
- Propor um método para seleção de atributos em contextos de classificação hierárquica multirrótulo;
- Avaliar o método proposto em bases de classificação hierárquica multirrótulo da *Gene Ontology*.

1.3 Organização do trabalho

Este documento está organizado em 7 capítulos. O Capítulo 2 aborda os principais conceitos e abordagens de classificação, destacando a classificação hierárquica multirrótulo. No capítulo 3 são apresentados os conceitos e abordagens relativos à redução de dimensionalidade, enfatizando-se a seleção de atributos e a abordagem filtro. O Capítulo 4 detalha o mapeamento sistemático de literatura, cujo objetivo é descrever o estado da arte da área de aplicação deste trabalho. O Capítulo 0 descreve o método proposto para a seleção de atributos em bases de dados de classificação hierárquica multirrótulo. No Capítulo 6 os experimentos realizados são descritos e os resultados obtidos são apresentados. Por fim, o Capítulo 7 é destinado às conclusões do trabalho e à proposição de trabalhos futuros.

2 ABORDAGENS DE CLASSIFICAÇÃO DE DADOS

Este Capítulo apresenta os conceitos básicos sobre classificação de dados e algumas abordagens para esta tarefa. A Seção 2.1 apresenta uma visão geral sobre o processo de classificação. A Seção 2.2 detalha os conceitos relativos à classificação multirrótulo, a Seção 2.3 descreve a classificação hierárquica, a Seção 2.4 aborda a classificação hierárquica multirrótulo, apresentando os principais conceitos, abordagens, algoritmos e medidas de avaliação para este tipo de classificação. Por fim, a Seção 2.5 expõe as considerações finais do capítulo.

2.1 Visão geral sobre a tarefa de classificação

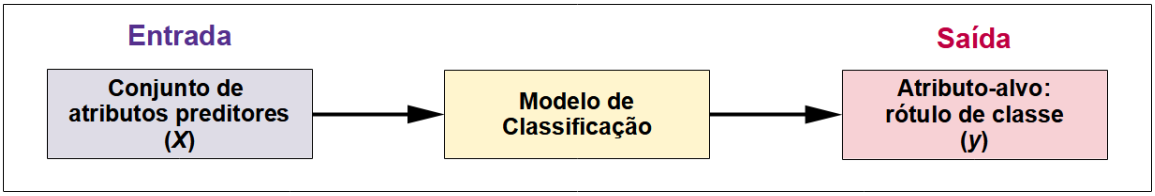
A tarefa de classificação consiste na organização de um conjunto de objetos (instâncias) em categorias pré-definidas (TAN; STEINBACH; KUMAR, 2009). Em outras palavras, pode-se dizer que tal tarefa equivale a associar instâncias a uma categoria, ou classe, o que é feito a partir do conjunto de atributos que caracterizam cada objeto individualmente.

De modo particular na Aprendizagem de Máquina (AM), a classificação faz parte de uma categoria de aprendizagem indutiva conhecida como aprendizagem supervisionada (MITCHELL, 1997). Segundo Goldschmidt, Passos e Bezerra (2015) a aprendizagem indutiva corresponde à capacidade que alguns algoritmos têm de aprender a partir de exemplos. Ou seja, esses algoritmos aprendem os relacionamentos que possam existir entre dados e geram um modelo para representar o conhecimento aprendido. Neste caso, diz-se que o modelo de aprendizagem é induzido pelo algoritmo a partir do conjunto de exemplos.

Em especial, a aprendizagem supervisionada compreende a indução de um modelo de aprendizagem a partir de exemplos ou instâncias de treinamento representados como pares ordenados (X_i, y_i), onde X_i corresponde a um conjunto, ou vetor, de atributos preditores, isto é, os valores dos atributos de entrada do algoritmo; y_i é o atributo-alvo, equivalente à saída desejada, que, na classificação, é um valor discreto e indica o rótulo de classe; e $i = 1, 2, \ldots, n$, em que n é o número total de instâncias de treinamento (CARVALHO; FREITAS, 2009) (GOLDSCHMIDT; PASSOS; BEZERRA, 2015).

Na tarefa de classificação é comum chamar o modelo de aprendizagem produzido de modelo de classificação. Uma vez que o modelo de classificação tenha sido obtido, este pode ser utilizado para prever rótulos de classe para objetos não rotulados. Assim, conforme ilustrado na Figura 1, o modelo recebe um conjunto de valores dos atributos X de um objeto não rotulado e atribui um rótulo de classe y para este objeto.

Figura 1 - Classificação como tarefa de mapear um conjunto de atributos no seu rótulo de classe



Fonte: Adaptado de Tan, Steinbach e Kumar (2009, p. 172)

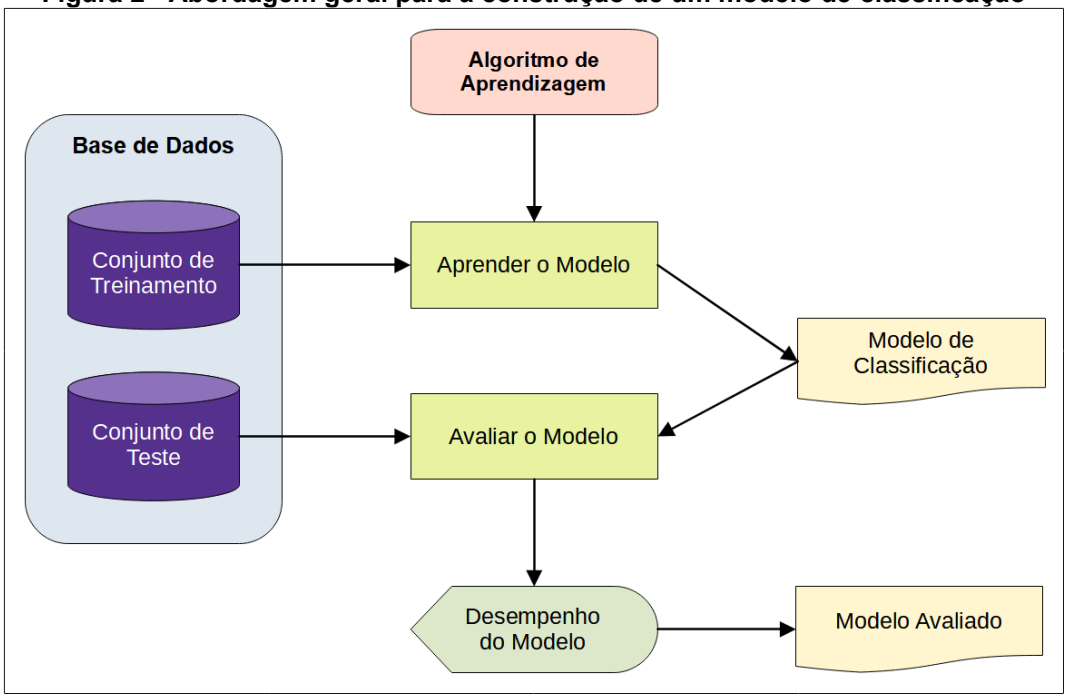
Formalmente, um modelo de classificação pode ser entendido como uma função f que mapeia cada conjunto de atributos preditores x_i para sua respectiva classe associada y_i. Neste sentido, a tarefa de classificação pode ser interpretada como a obtenção de uma função h que seja uma aproximação de f, onde h é a hipótese ou modelo de f (CARVALHO; FREITAS, 2009) (TAN; STEINBACH; KUMAR, 2009) (FACELI *et al*., 2011) (GOLDSCHMIDT; PASSOS; BEZERRA, 2015).

Diferentes técnicas podem ser utilizadas para a indução de classificadores a partir de um conjunto de exemplos, destacando-se Árvores de Decisão (AD), Redes Neurais Artificiais (RNA), classificadores *bayesianos* e *Support Vector Machines* (SVM) (TAN; STEINBACH; KUMAR, 2009) (FACELI *et al*., 2011) (RUSSEL; NORVIG, 2013). Entretanto, qualquer que seja a técnica adotada, a resolução de problemas de classificação segue uma abordagem geral (TAN; STEINBACH; KUMAR, 2009) (ALVES, 2010) como ilustrado na Figura 2.

Conforme se pode observar na Figura 2, os dados disponíveis para a geração de um modelo de classificação são separados em dois conjuntos mutuamente exclusivos: um para treinamento e outro para teste. O conjunto de treinamento consiste em exemplos cujos rótulos sejam conhecidos e é utilizado pelo algoritmo de aprendizagem no processo de indução do modelo de aprendizagem (aprender o modelo). O modelo aprendido é, então, avaliado mediante sua aplicação no conjunto de teste. Neste caso, o atributo-alvo fica indisponível e o classificador faz a previsão dos rótulos de classe para cada instância do conjunto de testes. Medidas de

desempenho comumente utilizadas são a precisão ou taxa de acerto do classificador e a taxa de erro ou taxa de classificação incorreta (MONARD; BARANAUSKAS, 2005). Ambas as medidas são baseadas na contagem de previsões corretas e incorretas e fornecem uma avaliação justa do erro e nível de generalização do modelo aprendido. O nível de generalização do modelo aprendido corresponde ao seu desempenho ao classificar instâncias de teste que não tenham sido utilizadas durante a etapa de treinamento (GEVERT *et al.*, 2010).

Figura 2 - Abordagem geral para a construção de um modelo de classificação



Fonte: Adaptado de Tan, Steinbach e Kumar (2009, p. 175)

A abordagem descrita, que consiste na associação de um exemplo a uma única classe de um conjunto de classes independentes, pode ser entendida como um caso de classificação plana e monorrótulo. De acordo com Faceli *et al.* (2011) e Borges (2012), é possível tipificar problemas de classificação de dados de duas maneiras: 1) com base na dependência entre as classes; e 2) de acordo com a possibilidade de associar um exemplo a apenas uma ou a mais de uma classe.

No primeiro caso, a classificação pode ser plana (tradicional) ou hierárquica. Na classificação plana as classes são independentes, enquanto na classificação hierárquica as classes são organizadas obedecendo a um conjunto de relações taxonômicas. No segundo caso, diz-se que um problema de classificação pode ser monorrótulo ou multirrótulo. Na classificação monorrótulo uma instância é mapeada

para apenas uma das possíveis classes ao passo que na classificação multirrótulo um exemplo pode ser categorizado em mais de uma classe (BORGES, 2012).

Na Seção 2.2 e na Seção 2.3 são apresentados detalhes sobre a classificação multirrótulo e a classificação hierárquica. Tais detalhes são importantes para a apresentação e discussão da classificação hierárquica multirrótulo, abordada na Seção 2.4.

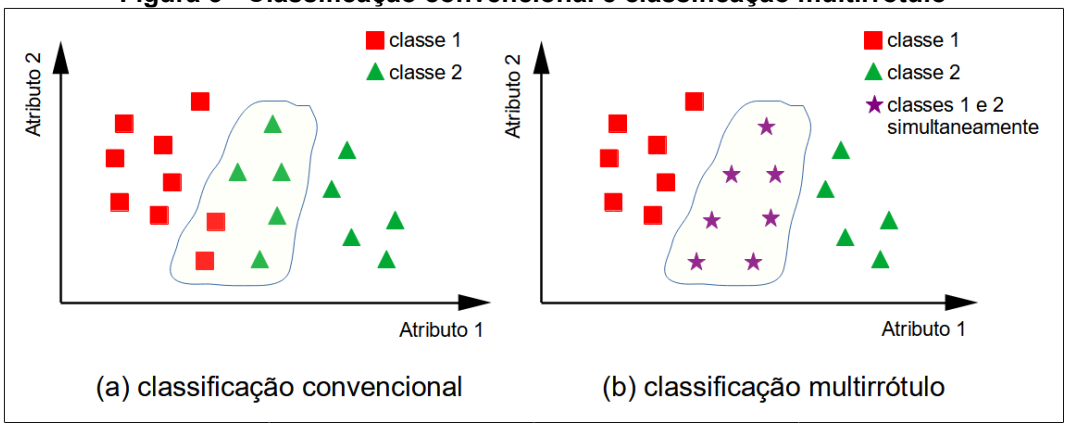
2.2 Classificação multirrótulo

A classificação multirrótulo pode ser entendida como uma modalidade de classificação na qual um exemplo pode pertencer simultaneamente a mais de uma classe, ou seja, cada instância é associada a vários rótulos (SIBLINI; KUNTZ; MEYER, 2019).

Problemas deste tipo ocorrem em diferentes domínios, como bioinformática, diagnóstico médico e classificação de imagens. A principal motivação para o desenvolvimento de métodos de classificação multirrótulo advém das tarefas de classificação de texto, onde um mesmo documento pode pertencer concomitantemente a duas categorias distintas (TSOUMAKAS; KATAKIS, 2009) (CARVALHO; FREITAS, 2009) (FACELI *et al.*, 2011). Por exemplo, uma matéria jornalística pode ser categorizada como sendo de ciência e de esporte.

A Figura 3 ilustra os problemas de classificação convencional e multirrótulo.

Figura 3 - Classificação convencional e classificação multirrótulo



Fonte: Adaptado de Faceli *et al.* (2011, p. 289)

Considerando, por exemplo, a tarefa de categorização de matérias jornalísticas (reportagens), a Figura 3(a) mostra um contexto em que reportagens

podem ser classificadas em apenas duas classes disponíveis (por exemplo: ciência ou esportes), enquanto a Figura 3(b) ilustra um cenário em que reportagens podem ser classificadas em ambas as classes (ciência e esporte). Vê-se em destaque na Figura 3 exemplos de matérias jornalísticas que podem ser rotuladas simultaneamente como sendo de ciência e de esportes.

De acordo com Cherman, Monard e Metz (2011) a classificação multirrótulo pode ser definida formalmente da seguinte maneira: seja D um conjunto de treinamento com n exemplos $E_i = (X_i, Y_i)$, onde $i = 1, 2, \ldots, n$. Cada instância E_i está associada com um vetor de m características $X_i = (x_{i1}, x_{i2}, \ldots, x_{im})$ e um subconjunto de rótulos de classe $Y_i \subseteq L$, onde $L = \{y_j : j = 1, \ldots, q\}$ é o conjunto de q rótulos possíveis. A Figura 4 ilustra este cenário, onde a instância E_1, por exemplo, está associada com o conjunto de atributos $(x_{11}, x_{12}, \ldots, x_{1m})$ e com o subconjunto Y_1 de rótulos de classe.

Figura 4 - Dados e classificação multirrótulo

D		X				Y
		x_1	x_2	$\cdots$	x_m	
E_1	X_1	x_{11}	x_{12}	$\cdots$	x_{1m}	Y_1
E_2	X_2	x_{21}	x_{22}	$\cdots$	x_{2m}	Y_2
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$	Y_3
E_n	X_n	x_{n1}	x_{n2}	$\cdots$	x_{nm}	Y_1

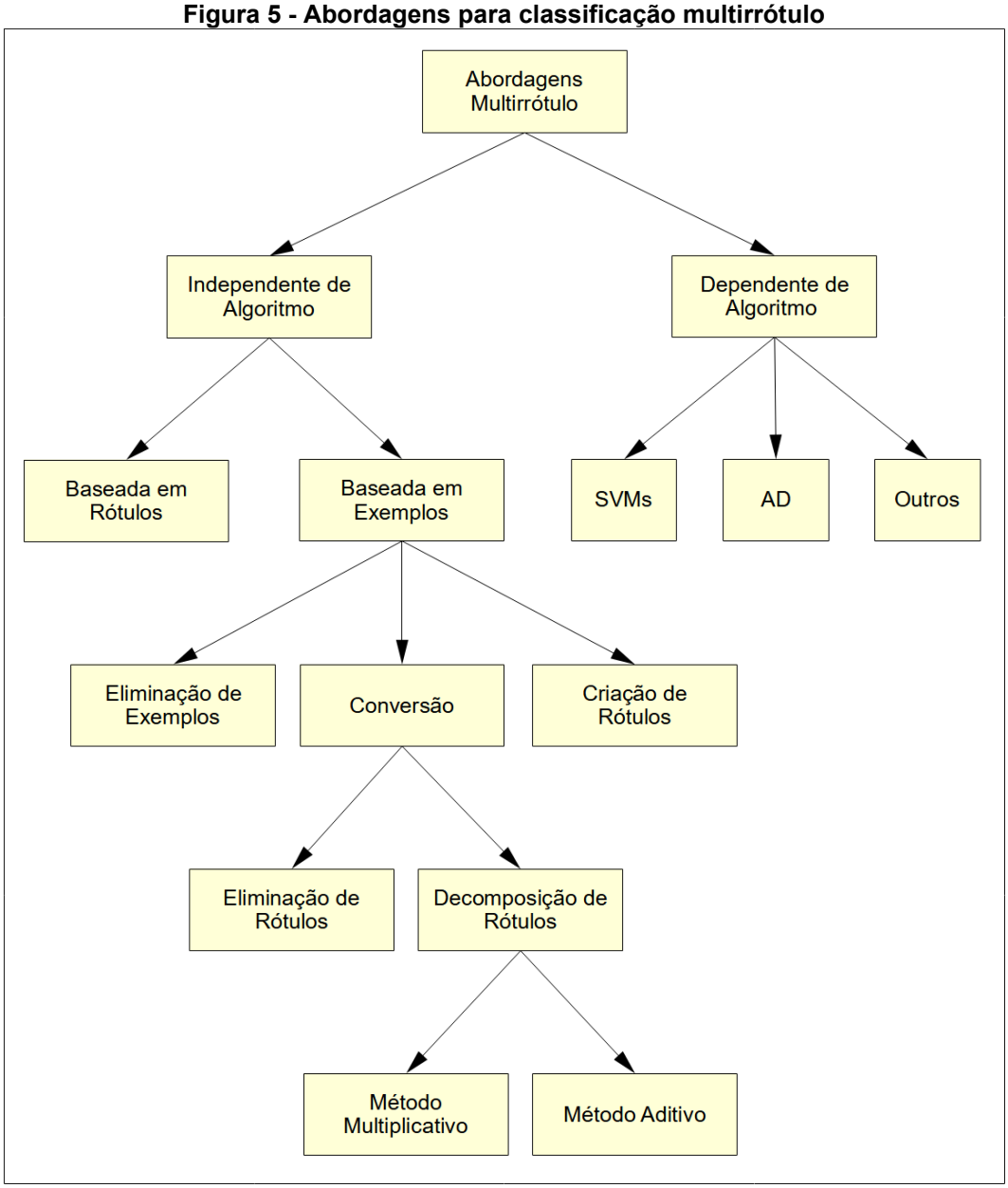
Fonte: Cherman, Monard e Metz (2011, p. 2)

Diante do exposto, pode-se afirmar que o propósito de um algoritmo de aprendizagem multirrótulo é induzir um classificador que, dado um exemplo não rotulado, $E = (x, ?)$, seja capaz de prever seu subconjunto de rótulos de classe Y. Isto é, $f(E) \rightarrow Y$, onde Y é o conjunto dos rótulos associados à instância E. É comum representar o subconjunto Y como uma lista dos rótulos de classe separados pelo símbolo "@" (BORGES, 2012). Neste caso, considerando $Y = \{AA, BB, CC\}$, sua representação é $AA@BB@CC$.

Diferentes abordagens têm sido propostas na literatura para lidar com problemas multirrótulo (HERRERA *et al.*, 2016). Estas, por sua vez, podem ser organizados em duas categorias principais, conforme mostrado na Figura 5: 1)

métodos independentes de algoritmo ou de transformação de dados; e 2) métodos dependentes de algoritmos ou de adaptação de algoritmos (TSOUMAKAS; KATAKIS, 2009) (CERRI, 2010) (ANSARI; SHAHANE, 2019).

Figura 5 - Abordagens para classificação multirrótulo



Fonte: Carvalho e Freitas (2009, p. 3)

Métodos independentes de algoritmo abordam a transformação do problema original em um ou mais problemas de rótulo único, deste modo podem ser utilizados algoritmos tradicionais para a tarefa de classificação (CARVALHO; FREITAS, 2009) (CERRI, 2010). Por outro lado, métodos dependentes de algoritmo concentram-se no desenvolvimento de algoritmos específicos para lidar diretamente com o problema da classificação multirrótulo (CARVALHO; FREITAS, 2009).

Os métodos de transformações podem ser baseados nos rótulos das classes ou nos exemplos de treinamento (CARVALHO; FREITAS, 2009). Duas técnicas bastante conhecidas são a técnica *Binary Relevance* (BR), também chamada de um-contra-todos ou binária, e a técnica *Label Powerset* (LP) (TSOUMAKAS; KATAKIS, 2009) (CHERMAN; MONARD; METZ, 2011) (HERRERA *et al.*, 2016) (ANSARI; SHAHANE, 2019).

A técnica BR é baseada em rótulos e consiste na decomposição do problema multirrótulo em q problemas binários, onde q é o número de classes do problema. Neste caso, para cada classe $y_i \in Y$, cria-se um problema binário de modo que todas as instâncias que pertencem à classe y_i são consideradas exemplos positivos, enquanto que as demais instâncias são consideradas exemplos negativos (TAN; STEINBACH; KUMAR, 2009) (ANSARI; SHAHANE, 2019). Um classificador binário é construído para separar instâncias da classe y_i das demais classes (CHERMAN; MONARD; METZ, 2011). Como para cada classe é construído um classificador, "cada classificador torna-se especializado na classificação de uma classe particular" (CERRI, 2010, p. 32).

A maior vantagem da técnica BR é a baixa complexidade computacional quando comparada a outras técnicas multirrótulo, sendo apropriada quando o valor de q não é muito grande. Entretanto, em diversos domínios é possível que haja um número elevado de rótulos de classes. Além disso, uma desvantagem da técnica é supor que os rótulos sejam independentes, deixando, portanto, de considerar informações importantes de relacionamentos entre rótulos (CHERMAN; MONARD; METZ, 2011).

LP é uma técnica de combinação de rótulos que consiste na transformação de um problema multirrótulo em um problema de classificação multiclasse de rótulo único, mapeando cada subconjunto Y de vários rótulos em um valor de classe única (CHERMAN; MONARD; METZ, 2011) (SPOLAÔR *et al.*, 2013a) (COSTA JUNIOR *et al,* 2017). Isto é, cada combinação de rótulos no conjunto de dados original é transformada em um único rótulo. Em outras palavras, cada $E_i = (X_i, Y_i), i = 1, 2, \ldots, n$ é transformado em $E_i = (X_i, r_i)$, onde r_i é o rótulo único que representa um subconjunto distinto de rótulos. A ideia expressa nesta definição é apresentada na Figura 6, na qual os atributos foram omitidos, já que o processo de transformação ocorre apenas no espaço de rótulos. Na Figura 6(a) tem-se um conjunto de exemplos multirrótulo no formato original. Na Figura 6(b) é mostrado o mesmo conjunto de

exemplos após transformação LP. A notação $y_{i,j,\ldots,k}$ indica que a instância é rotulada com a conjunção $y_i \wedge y_j \wedge \ldots \wedge j_k$ (CHERMAN; MONARD; METZ, 2011). O conjunto de rótulos originais, mostrado na Figura 6(a), é $L = \{y_1, y_2, y_3, y_4\}$. Após a transformação LP, o conjunto de rótulos multiclasse gerado é $L = \{y_{2,3}, y_{1,3,4}, y_4, y_{2,3}\}$, conforme mostrado na Figura 6(b).

Figura 6 - Exemplo de transformação LP num conjunto de dados multirrótulo

D	Y
E_1	$Y_1 = \{y_2, y_3\}$
E_2	$Y_2 = \{y_1, y_3, y_4\}$
E_3	$Y_3 = \{y_4\}$
E_4	$Y_1 = \{y_2, y_3\}$

(a)

D	Y
E_1	$y_{2,3}$
E_2	$y_{1,3,4}$
E_3	y_4
E_4	$y_{2,3}$

(b)

Fonte: Cherman, Monard e Metz (2011, p. 3)

A técnica LP considera as correlações de rótulos e, por ser simples, é possível, após a transformação, utilizar qualquer algoritmo multiclasse para a tarefa de classificação. A principal desvantagem da técnica evidencia-se quando alguns rótulos de classe no conjunto de dados multiclasse estão associados a um número pequeno de exemplos, causando o desbalanceamento desse conjunto de dados (CHERMAN; MONARD; METZ, 2011) (SPOLAÔR *et al.*, 2013a).

Na abordagem dependente de algoritmo, novos algoritmos são desenvolvidos para tratar problemas de classificação multirrótulo (CERRI, 2010, p. 37). Em geral, esses algoritmos são baseados em modificações ou generalizações de métodos de classificação monorrótulo existentes, como AD, SVM, kNN (*k-Nearest Neighbor*) e RNA (CERRI, 2010) (HERRERA *et al.*, 2009) (ANSARI; SHAHANE, 2019). A principal dificuldade na utilização desta abordagem é construir um modelo que seja capaz de prever várias saídas simultaneamente, visto que algumas abordagens, como kNN, são facilmente adaptáveis, enquanto outras, como SVM, exigem maior esforço (HERRERA *et al.*, 2016).

Herrera *et. al.* (2016) e Siblini, Kuntz e Meyer (2019) acrescentam uma terceira categoria para abordar o problema da classificação multirrótulo: a combinação de classificadores (*ensembles*). De modo particular, combinações de classificadores binários têm sido utilizados para a classificação multiclasse (GALAR *et al.*, 2011).

Ademais, o uso de *ensembles* tem sido aplicado para enfrentar o problema do desbalanceamento (GALAR *et al.,* 2012). Melo e Paulheim (2019) tratam essa abordagem como a utilização de diversos métodos de transformação e de adaptação de algoritmos como base para o processo de classificação, combinando as saídas dos diferentes modelos para realizar a previsão.

Um aspecto peculiar na classificação multirrótulo, e que requer atenção, é a quantidade de exemplos e de classes em uma base de dados. Pode acontecer de o número de instâncias de uma classe ser pequeno em relação ao total de exemplos, ou, em outros casos, essa quantidade pode ser grande. E isto pode interferir no desempenho dos métodos de classificação. Deste modo, devem ser utilizados, além de medidas de desempenho, os conceitos de cardinalidade de rótulo e densidade de rótulo (Faceli *et al.*, 2011).

Outro ponto importante diz respeito às medidas de avaliação de classificadores multirrótulo. Tais medidas devem ser diferentes daquelas aplicadas a classificadores monorrótulo, pois neste caso um exemplo é classificado apenas de forma correta ou errada. Na classificação multirrótulo, dado que um exemplo está associado a diversas classes, este pode ser classificado de maneira parcialmente correta ou parcialmente errada. Isto é, um classificador pode atribuir a uma instância uma ou mais classes a que ela pertence deixando de atribuir uma ou mais classes a que essa instância também esteja associada. Do mesmo modo um classificador pode designar a um exemplo uma ou mais classes a que ele não esteja vinculado (CERRI, 2010).

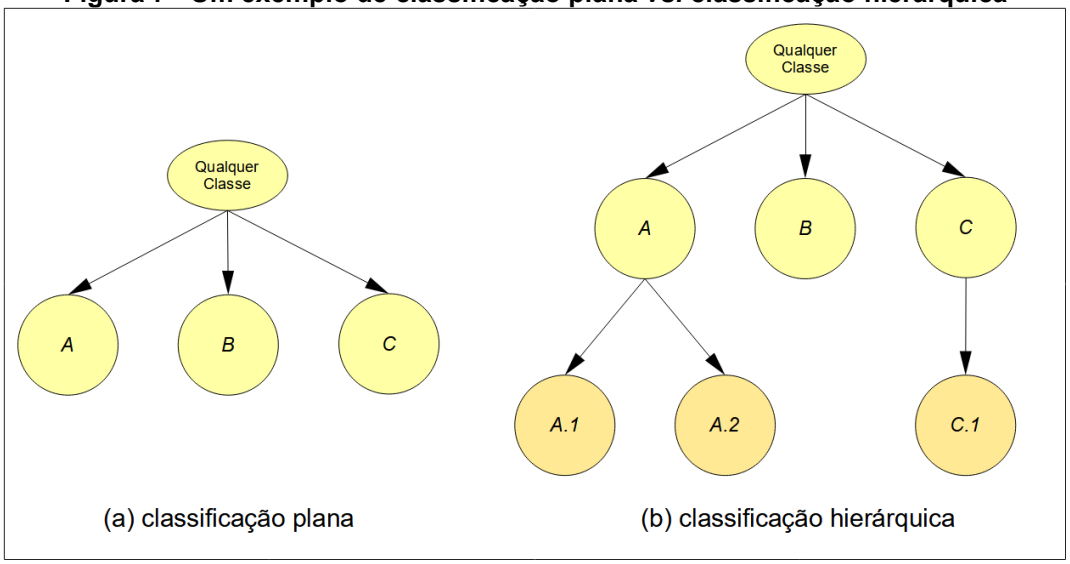
De modo geral, a avaliação de um classificador pode ser realizada com base na atribuição dos rótulos realizada pelo modelo de classificação, ou a partir de uma função de *ranking*, onde o classificador gera um *ranking* de rótulos para cada instância (CERRI, 2010).

2.3 Classificação hierárquica

Apesar da classificação plana ser o tipo mais comum de classificação, em um número significativo de problemas uma ou mais classes podem ser divididas em subclasses ou agrupadas em superclasses (CERRI, 2010) (ALVES, 2010). Neste caso, exemplos estão associados a classes dispostas taxonomicamente. A Figura 7

ilustra a classificação plana e a classificação hierárquica. Cada nó, exceto a raiz, encontra-se rotulado com o número de uma classe.

Figura 7 - Um exemplo de classificação plana *vs.* classificação hierárquica



Fonte: Adaptado de Freitas e Carvalho (2007, p. 180)

Observa-se que na Figura 7(a) as classes estão organizadas em um nível único, não mantendo nenhuma relação entre si, enquanto na Figura 7(b) as classes $A.1$ e $A.2$ estão vinculadas de forma direta à classe A, ocupando um nível abaixo. O mesmo pode ser dito das classes $C.1$ e C. Neste caso, tem-se uma relação de dependência e uma ordem taxonômica entre essas classes, ou seja, as classes possuem entre si relacionamentos de subclasse e superclasse (FREITAS; CARVALHO, 2007).

De acordo com Alves (2010), em estruturas hierárquicas as classes podem apresentar a seguinte relação: $y_i \preccurlyeq_h y_j$, onde $\preccurlyeq_h$ significa “superclasse de”. Neste sentido, de acordo com a Figura 7(b), pode-se dizer que a classe A é “superclasse da” classe $A.1$. Outra relação possível entre as classes é “subclasse de”, como é o caso, na Figura 7(b), em que a classe $C.1$ é “subclasse da” classe C. Nós que não possuem relação do tipo “superclasse de” com nenhum outro nó é chamado de nó-folha e os demais podem ser chamados de nós internos.

Em geral, essas relações também podem ser expressas como “pai de” e “filho de” para indicar “superclasse de” e “subclasse de”, respectivamente. Pode-se usar, ainda, as relações “ancestral de” e “descendente de” para indicar as relações “superclasse de” e “subclasse de”, respectivamente. Porém deve-se destacar que estas relações, “ancestral de” e “descendente de” têm uma característica mais

genérica, visto que são válidas para situações em que as classes relacionadas não são diretamente “pai de” e “filha de”, como na Figura 10, onde a classe AA é classe “ancestral da” classe DD, mas também é “ancestral da” classe HH. Neste trabalho, utilizar-se apenas a nomenclatura ancestral e descendente para designar as classes da hierarquia, considerando-se quaisquer graus de ancestralidade entre as classes.

A partir do que foi exposto, pode-se entender que na classificação plana ou tradicional um exemplo é atribuído a um nó folha, ou seja, a uma classe mais específica. Em problemas de classificação hierárquica, classificar apenas utilizando nós-folha pode gerar perda de informações. Portanto, um algoritmo de classificação hierárquica deve levar em consideração a estrutura hierárquica das classes, o que proporciona a indução de classificadores com melhor desempenho (FACELI *et al*., 2011).

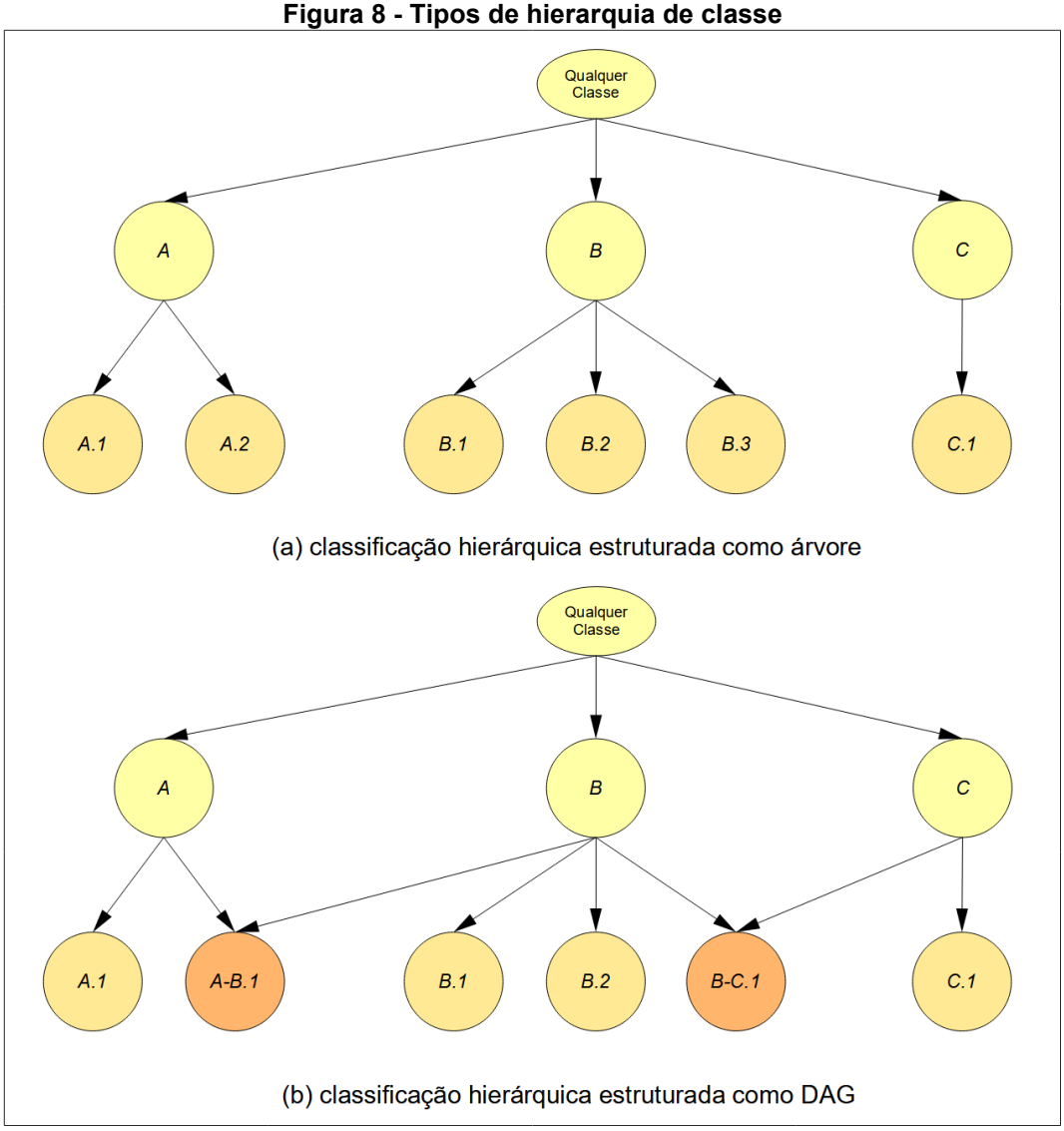
Problemas de classificação hierárquica podem ser caracterizados a partir de três aspectos básicos: o tipo de hierarquia utilizado para implementar o relacionamento entre as classes; se os dados podem seguir apenas um ou mais de um caminho na hierarquia; e o nível hierárquico onde as predições são realizadas (SILLA JUNIOR; FREITAS, 2010) (COSTA, 2008). Referente ao primeiro aspecto, o relacionamento entre as classes pode ser representado como uma árvore ou como um DAG. A Figura 8 mostra esses dois tipos de estruturas. A diferença essencial entre a estrutura árvore e a estrutura DAG reside no fato de que em uma árvore (Figura 8(a)) cada nó, exceto a raiz, possui um, e somente um, nó ancestral direto (pai), enquanto na estrutura DAG (Figura 8(b)), cada nó, exceto a raiz, pode ter mais de um nó pai.

O segundo aspecto, caminho das rotulações, consiste nos casos em que as previsões podem seguir mais de um caminho na hierarquia. Por exemplo, uma instância pode pertencer simultaneamente às classes $A.2$ e $B.2$ da estrutura da Figura 8(a). Neste caso, tem-se um problema de classificação hierárquica múltipla ou multirrótulo.

O terceiro aspecto, nível hierárquico onde as predições são realizadas, corresponde à profundidade das rotulações dos dados. Em geral, problemas de classificação hierárquica visam a classificação de novas instâncias em um nó folha. Quando todos os exemplos possuem rótulos correspondentes a nós folha, diz-se que se tem rotulação completa em profundidade (FACELI *et al*., 2011). Porém, quanto mais profundo for o nível, mais difícil se torna a predição, visto que a quantidade de

exemplos em níveis mais profundos em geral é menor e, devido a isso, a confiabilidade do classificador induzido também é menor (COSTA, 2008). Portanto, é conveniente e mais seguro realizar a classificação em níveis mais elevados. Neste caso, tem-se uma rotulação parcial em profundidade (FACELI *et al.*, 2011).

Figura 8 - Tipos de hierarquia de classe



Fonte: Adaptado de Freitas e Carvalho (2007, p. 191-192)

Um problema de classificação hierárquica é caracterizado por uma combinação dos aspectos destacados anteriormente (FACELI *et al.*, 2011). Por exemplo, um problema pode possuir estrutura hierárquica do tipo DAG, onde os dados possuem rótulos que podem seguir caminhos diversos na hierarquia (multirrótulo), podendo ser rotulados com nós internos. Essa caracterização é importante, pois a partir dela é que se torna possível seleciona a abordagem mais apropriada à solução do problema.

Em relação às abordagens que podem ser adotadas para a classificação hierárquica, Freitas e Carvalho (2007) elenca três principais: classificação hierárquica plana, classificação hierárquica local ou *top-down* e classificação hierárquica global ou *one-shot*.

A classificação hierárquica plana consiste na transformação de um problema de classificação hierárquica em um problema de classificação plana, desprezando-se a hierarquia de classes e realizando a predição apenas de nós folha. De acordo com Borges (2012), esta abordagem é semelhante à classificação plana convencional e pode ser aplicada tanto em estruturas do tipo árvore quanto do tipo DAG.

A classificação hierárquica local, segundo Costa (2008), consiste no treinamento de um ou mais classificadores para cada um dos nós da hierarquia, o que produz uma árvore de classificadores como resultado, na qual os elementos do conjunto de testes são rotulados de forma iterativa e *top-down* ao longo da hierarquia. Esse processo funciona da seguinte forma: sempre que um exemplo é rotulado, ele é submetido a um classificador, a fim de que este prediga a qual subclasse da classe predita no nível anterior este exemplo pertence. Esta ação é repetida até que seja atingido um nó folha (predição obrigatória em nós folha) ou até que o exemplo seja associado a um nó interno da hierarquia (predição opcional em nós folha).

A abordagem de classificação global considera toda a hierarquia durante o processo de treinamento, o que resulta na indução de um único modelo de classificação, que tende a ser mais complexo do que os classificadores gerados na abordagem local (FACELI *et al.*, 2011). Neste caso, o modelo concebido pode ser utilizado na predição de qualquer classe da hierarquia. Ou seja, quando o classificador recebe um exemplo como entrada, a predição é realizada num passo único, a partir de informações de toda a hierarquia (COSTA, 2008).

Não existe consenso quanto às medidas de avaliação de desempenho para modelos hierárquicos, uma vez que as medidas tradicionais não se mostram adequadas por não considerarem a estrutura hierárquica de tais problemas (CERRI, 2010). O que torna necessária a adaptação dessas medidas para o tipo de avaliação que se almeja fazer. Um aspecto fundamental a ser considerado na avaliação de qualquer modelo é a forma de obtenção do resultado. Neste caso, uma taxa de desempenho é obtida para todo o modelo, para cada nível ou para cada classe (BORGES, 2012). Entre as medidas de desempenho específicas para classificadores hierárquicos propostas na literatura, destacam-se as seguintes: baseadas em

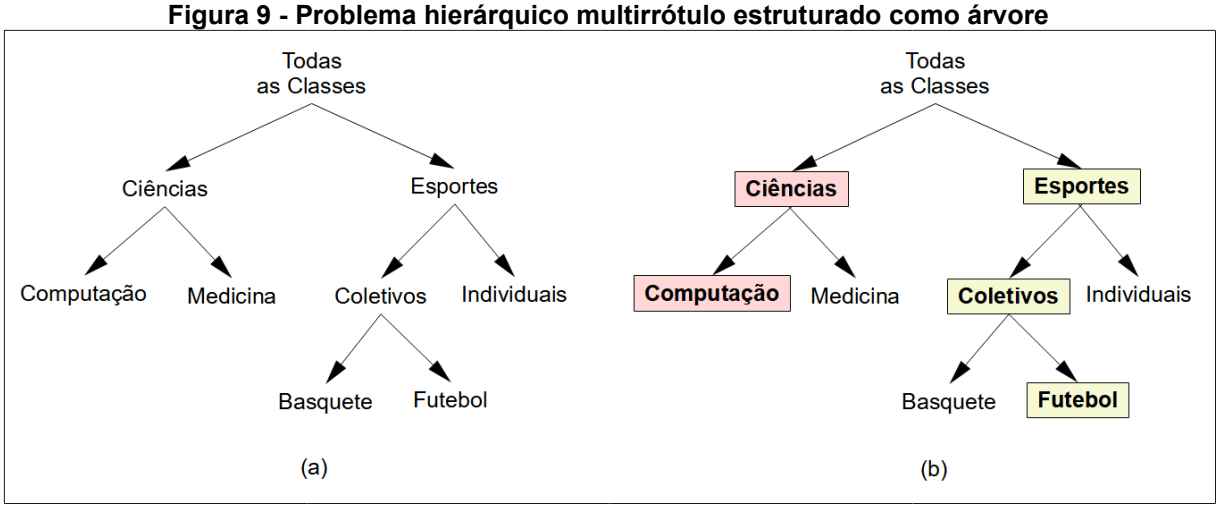
distância, baseadas nas relações de descendência e/ou ancestralidade, baseadas em similaridade, baseadas em profundidade, baseadas em matriz de custos, baseadas na curva de precisão e revocação (CERRI, 2010) (BORGES, 2012).

De modo particular, este trabalho tem por interesse problemas de classificação hierárquica multirrótulo. Tais problemas possuem características tanto de classificação multirrótulo quanto de classificação hierárquica e são detalhados na Seção 2.4.

2.4 Classificação Hierárquica Multirrótulo

De acordo Melo e Paulheim (2019), problemas de HMC são caracterizados tanto como um problema de classificação hierárquica quanto um problema de classificação multirrótulo. Isto quer dizer que em um problema de classificação hierárquica multirrótulo, um exemplo pode pertencer simultaneamente a mais de uma classe e essas classes são organizadas obedecendo relações hierárquicas entre si (CERRI, 2010). A Figura 9 ilustra um problema típico de HMC cuja hierarquia é organizada como uma árvore.

Figura 9 - Problema hierárquico multirrótulo estruturado como árvore



Fonte: Cerri (2010, p. 44)

Uma matéria jornalística, por exemplo, pode abordar um assunto relacionado à ciência da computação e ao futebol. Neste caso, a Figura 9(a) ilustra a hierarquia de classes para este domínio e a

Figura 9(b) mostra que tal matéria pode ser classificada tanto como “Ciências/ Computação” quanto como “Esportes/ Coletivos/ Futebol”. O resultado do processo

de predição é expresso por meio de uma subárvore da árvore original, conforme destacado na Figura 9(b).

Segundo Vens *et al.* (2008) e Stojanova *et al.* (2013), a tarefa de classificação hierárquica multirrótulo pode ser formalmente definida como encontrar a função $f: X \longrightarrow \wp(L)$ que associa cada instância $X_i \in X$ a um conjunto de classes $Y_i \in \wp(L)$, onde X é o espaço de instâncias, $\wp(L)$ é o conjunto potência de L e $L = \{y_1, y_2, \ldots, y_q\}$ é o conjunto de todos os rótulos de classes possíveis, que, por sua vez, são organizadas hierarquicamente segundo uma ordem parcial $\preccurlyeq_h$, que representa o relacionamento de superclasse, isto é, $\forall y_1, y_2 \in L: y_1 \preccurlyeq_h y_2$ se, e somente se, y_1 é superclasse de y_2. Assim, dado um conjunto de exemplos D, em que cada exemplo tem a forma $E_i = (X_i, Y_i)$, a função f deve ser tal que $y \in f(x) \Rightarrow \forall y' \preccurlyeq_h y: y' \in f(x)$.

Para ilustrar essa definição, é utilizada uma base de dados fictícia com as seguintes características: 11 classes, cuja hierarquia é ilustrada na Figura 10, 9 atributos e 6 instâncias, mostrados na Tabela 1.

Tabela 1 - Base de dados fictícia para classificação hierárquica multirrótulo

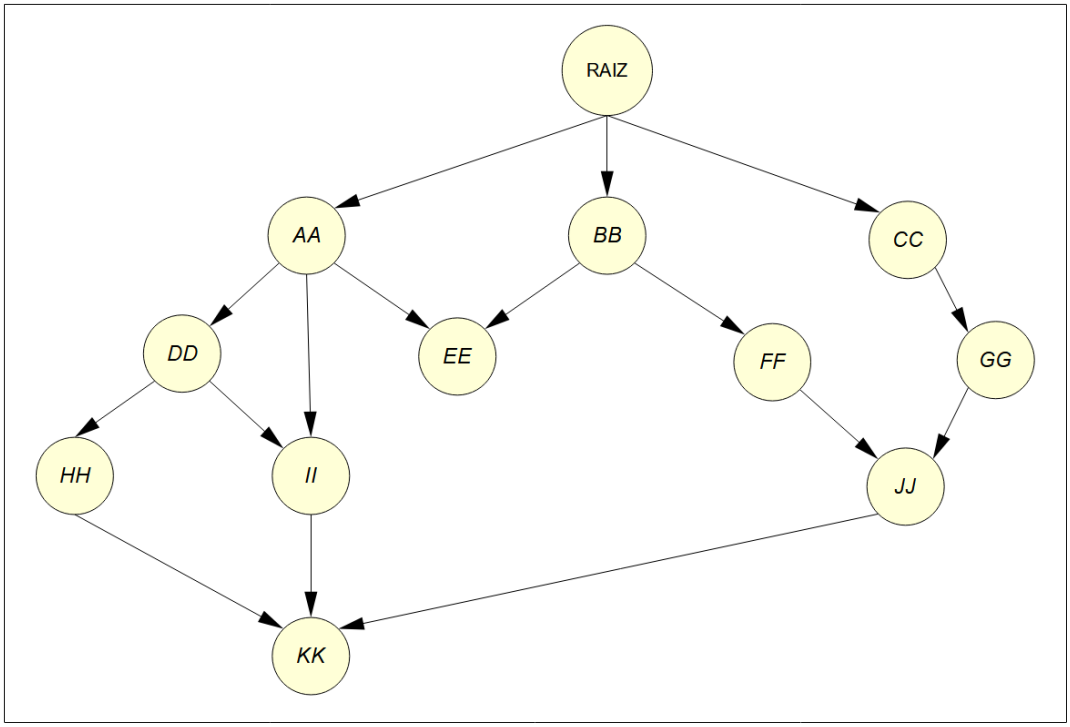
$\boldsymbol{D}$	$\boldsymbol{x_1}$	$\boldsymbol{x_2}$	$\boldsymbol{x_3}$	$\boldsymbol{x_4}$	$\boldsymbol{x_5}$	$\boldsymbol{x_6}$	$\boldsymbol{x_7}$	$\boldsymbol{x_8}$	$\boldsymbol{x_9}$	$\boldsymbol{Y}$
E_1	0,45	0,42	0,56	0,92	0,95	0,34	0,21	0,88	0,16	$CC@EE@II$
E_2	0,84	0,37	0,10	0,35	0,47	0,78	0,92	0,81	0,01	$BB@HH@II$
E_3	0,27	0,49	0,04	0,41	0,83	0,90	0,36	0,67	0,91	JJ
E_4	0,85	0,11	0,95	0,89	0,69	0,47	0,16	0,35	0,67	$BB@II$
E_5	0,26	0,82	0,53	0,95	0,74	0,89	0,33	0,41	0,49	CC
E_6	0,36	0,76	0,88	0,04	0,11	0,61	0,77	0,58	0,24	$HH@JJ$

Fonte: Autoria Própria (2022)

O número de instâncias de treinamento é $n = 6$. O conjunto de exemplos de treinamento é dado por $D = \{E_1, E_2, E_3, E_4, E_5, E_6\}$, em que cada exemplo corresponde a um vetor de valores de atributos e um conjunto de rótulos de classe. Neste caso, para o exemplo E_4, tem-se que $E_4 = (X_4, Y_4)$, onde $X_4 = (0{,}85; 0{,}11; 0{,}95; 0{,}89; 0{,}69; 0{,}47; 0{,}16; 0{,}35; 0{,}67)$ é o vetor dos valores dos atributos e $Y_4 = \{BB, II\}$ é conjunto de rótulos de classe associados ao exemplo, que também pode ser escrito como $Y_4 = BB@II$. O conjunto $X = \{X_1, \ldots, X_6\}$ contém os vetores que representam o conjunto de atributos para cada um dos exemplos e o conjunto $Y = \{Y_1, \ldots, Y_6\}$ compõe-se de todos os conjuntos de rótulos de classe associados a cada uma das instâncias da base.

O conjunto de todos os possíveis rótulo de classe L pode ser obtido a partir da Figura 10 e é representado por $L = \{AA, BB, CC, DD, EE, FF, GG, HH, II, JJ, KK\}$. A definição apresentada para a classificação hierárquica multirrótulo, indica que a função f induzida deve associar um conjunto de atributos X_i a um conjunto de rótulos de classe Y_i, onde $Y_i \subseteq \wp(L)$, onde $\wp(L)$ é o conjunto potência de L, cuja quantidade de elementos é dada por $\wp(L) = 2^{11} = 2048$. Deste modo, $\wp(L)$ contém todos os possíveis 2048 subconjuntos de L, e, por motivo de simplicidade, não é representado de forma exaustiva, mas apenas exemplificativa. Assim, $\wp(L) = \{\emptyset, \{AA\}, \{BB\}, \dots, \{JJ\}, \dots, \{AA, BB\}, \dots, \{BB, II\}, \dots \{CC, EE, II\}, \dots, L\}$.

Figura 10 - Hierarquia das classes da base de dados fictícia



Fonte: Autoria Própria (2022)

A hierarquia de classes pode ser representada de acordo com a seguinte notação *ascendente/descendente*. Deste modo, tem-se, os seguintes relacionamentos: $RAIZ/AA$, $RAIZ/BB$, $RAIZ/CC$, AA/DD, AA/EE, AA/II, BB/EE, BB/FF, CC/GG, DD/HH, DD/II, FF/JJ, GG/JJ, HH/KK, II/KK e JJ/KK. Outrossim, como a organização hierárquica corresponde a uma ordem parcial $\preccurlyeq_h$, conforme outrora apresentado, para cada classe que esteja associada a um exemplo, suas respectivas classes ancestrais também estão associadas ao mesmo exemplo. Neste sentido, considerando a instância E_4, tem-se que suas classes associadas são $Y_4 =$

$\{BB, II\}$ e, devido a organização das classes como uma hierarquia, tem-se que as classes AA e DD também estão associadas a E_4, pois são ancestrais da classe II. Isto é, como $II \in f(x)$, então $DD \in f(x)$ e $AA \in f(x)$.

De acordo com Cerri (2010), há duas estratégias para tratar problemas de HMC: 1) através do aprendizado de um classificador binário para cada classe da hierarquia; e 2) através do desenvolvimento de técnicas específicas para este tipo de problema. A primeira estratégia segue a técnica um-contra-todos apresentada na Seção 2.1, enquanto que a segunda, consiste em desenvolver algoritmos específicos para lidar com o problema da HMC, podendo-se adaptar estratégias utilizadas para lidar com problemas multirrótulo e com problemas hierárquicos.

Por causa das limitações da primeira estratégia, elencadas na Seção 2.1, grande parte dos trabalhos existentes na literatura seguem a segunda estratégia (CERRI, 2010), elaborando técnicas específicas para tratar o problema da classificação hierárquica multirrótulo. O desenvolvimento de tais técnicas pode seguir duas abordagens: global ou local. A abordagem global considera todas as classes de uma única vez para o treinamento do classificador, o que pode gerar um aumento na complexidade da tarefa de classificação (BORGES, 2012). A abordagem local considera as classes por nível da hierarquia, descendo na hierarquia nível a nível, considerando em cada nó apenas algumas classes (CERRI, 2010). Entretanto, nessa abordagem um erro de classificação tende a ser propagado para os níveis inferiores da hierarquia (FREITAS; CARVALHO, 2007).

Na Subseção 2.4.1 são apresentados alguns algoritmos de HMC disponíveis na literatura e a Subseção 2.4.2 aborda as medidas de avaliação de desempenho definidas para este domínio de problema.

2.4.1 Algoritmos de Classificação Hierárquica Multirrótulo

Na literatura, existem alguns trabalhos que propõem métodos para o tratamento de tarefas de classificação hierárquica multirrótulo em diferentes domínios, como classificação de imagens (DIMITROVSKI *et al*., 2011) (YAN; WONG, 2017), bioinformática (BLOCKEEL *et al*., 2006) (VENS *et al*., 2008) (BORGES, 2012) (MELO; PAULHEIM, 2019) (CERRI *et al*., 2018) e classificação de texto (GARGIULO *et al*., 2019) (PRABOWO; IBROHIM; BUDI, 2019) (ALJEDANI; ALOTAIBI; TAILEB, 2021).

Entretanto, não existe consenso sobre que abordagem utilizar para lidar com problemas de classificação hierárquica multirrótulo.

Nesse contexto, alguns algoritmos de classificação hierárquica multirrótulo foram propostos, destacando-se o Clus-HMC (BLOCKEEL *et al*., 2002) (VENS *et al.,* 2008), o *Multi-label Hierarchical Classification with na Artificial Immune System* (MHCAIS) (ALVES, 2010), o *Hierarchical Multi-Label Classification with Genetic Algorithm* (HMC-GA) (CERRI; BARROS; CARVALHO, 2012), e o *Multi-label Hierarchical Classification using a Competitive Neural Network* (MHC-CNN) (BORGES, 2012).

O algoritmo MHCAIS, proposto por Alves (2010), é baseado em Sistemas Imunológicos Artificiais (SIA). Os classificadores hierárquicos multirrótulo gerados são representados por meio de regras SE-ENTÃO. O método possui dois procedimentos fundamentais, um para Extração Sequencial de Regras e outro para Evolução das Regras (ALVES, 2010).

O HMC-GA é um método global para induzir regras de classificação hierárquica multirrótulo por meio do uso de um Algoritmo Genético (AG). Neste caso, o AG evolui os antecedentes das regras de classificação. Os antecedentes otimizados, por sua vez, são selecionados para produzir o conjunto de classes a serem previstas, que correspondem ao consequente das regras (CERRI; BARROS; CARVALHO, 2012)

Neste trabalho, são utilizados o Clus-HMC e MHC-CNN para a realização de testes experimentais com o método proposto. Estes algoritmos foram escolhidos para esta tarefa pela facilidade de uso e por terem resultados publicados com o mesmo conjunto de dados utilizados neste trabalho. Uma breve descrição destes algoritmos é apresentada na Subseção 2.4.1.1 e na Subseção 2.4.1.2.

2.4.1.1 Clus-MHC

O Clus-HMC é um algoritmo para indução de classificadores baseados em AD proposto por Blockeel *et al.* (2002). Trata-se de um método fundamentado em *Predictives Cluster Trees* (PCT) que gera uma Árvore de Decisão considerando toda hierarquia de classes de um problema. A versão mais utilizada gera classificadores

globais para a classificação hierárquica multirrótulo e é capaz de lidar com estruturas hierárquicas organizadas tanto em árvores quanto em DAGs (VENS *et al.*, 2008).

De modo geral, o algoritmo Clus-HMC possui o seguinte mecanismo: as árvores de decisão são vistas como hierarquia de grupos capazes de prever um conjunto de classes. A ideia geral é separar o conjunto de classes em grupos, de modo que a distância entre os grupos seja minimizada (BLOCKEEL *et al.*, 2006). Neste sentido, o nó raiz corresponde a um grupo com todas as instâncias de treinamento. Na medida em que a hierarquia é percorrida em direção aos nós-folha, as instâncias são divididas recursivamente em grupos menores. Ao final, uma AD única é gerada através da combinação dos grupos formados ao longo do processo (BLOCKEEL *et al.*, 2006) (VENS *et al.*, 2008).

2.4.1.2 MHC-CNN

O algoritmo MHC-CNN utiliza a abordagem de classificação global baseada em uma RNA competitiva composta por uma camada de entrada e uma de saída, cujo treinamento é realizado por meio do cálculo das distâncias entre os nós da hierarquia e os exemplos de treinamento (BORGES, 2012).

O método é baseado na aprendizagem competitiva, na qual os neurônios que compõem a camada de saída da rede competem entre si para serem ativados. Neste caso, os neurônios que apresentam as menores distâncias são considerados vencedores, tendo seus pesos atualizados, juntamente com seus vizinhos (BORGES, 2012).

2.4.2 Medidas de avaliação para a classificação hierárquica multirrótulo

Quaisquer das medidas de desempenho citadas na Seção 2.2 e na Seção 2.3 podem ser utilizadas para a avaliação de classificadores hierárquicos multirrótulo, entretanto não há consenso de qual medida deve ser adotada como regra para essa avaliação (CERRI, 2010).

De modo particular, o algoritmo Clus-HMC adota a medida AUPRC. O algoritmo MHC-CNN utiliza, além da AUPRC, as medidas de distância e medida-*Fh* para avaliação de seu desempenho na tarefa de classificação.

Nas Subseção 2.4.2.1 e na Subseção 2.4.2.2 são descritas as medidas AUPRC e de distância.

2.4.2.1 Medida AUPRC

A medida baseada na curva de precisão e revocação (curvas PR) foi proposta Vens *et al.* (2008) e consiste na escolha de um conjunto de limiares entre 0 e 1, onde cada limiar corresponde a um ponto no espaço da curva PR. Uma curva PR descreve a precisão de um modelo em função de sua revocação. Os limiares escolhidos podem ser interpretados como a probabilidade de atribuição de uma certa classe a um exemplo (VENS et al., 2008). Quanto menor for o valor do limiar mais exemplos são atribuídos a uma classe aumentando a medida de revocação (BORGES, 2012) e, normalmente, diminuído a precisão (VENS *et al.*, 2008).

Em cenários multirrótulo curvas PR podem ser construídas para cada classe individual, sendo necessário combinar os desempenhos de cada classe para quantificar o desempenho geral, o que pode ser feito por meio de duas abordagens: a área abaixo da curva PR (AUPRC) e a área média abaixo da curva PR (Vens *et al.*, 2008).

A primeira abordagem consiste em construir uma curva PR geral, a partir da transformação do problema multirrótulo em um problema binário. Isto é, um classificador binário recebe um par de entrada $(instância, classe)$ e prevê se a $instância$ pertence ou não à $classe$ (VENS *et al.*, 2008). Um classificador que prevê a probabilidade de uma instância pertencer a uma classe pode ser transformado num classificador binário, escolhendo-se um valor de limiar. Para um determinado limiar, calcula-se um ponto de precisão e revocação $(\overline{Prec}, \overline{Rev})$ no espaço PR através da Equação (1) e da Equação (2), nas quais VP corresponde à quantidade de verdadeiros positivos, FP é a quantidade de falsos positivos, FN indica a quantidade de falsos negativos e i representa as classes.

$$\overline{Prec} = \frac{\sum_i VP_i}{\sum_i VP_i + \sum_i FP_i} \tag{1}$$

$$\overline{Rev} = \frac{\sum_i VP_i}{\sum_i VP_i + \sum_i FN_i} \tag{2}$$

Variando-se esse limiar, obtém-se a curva PR média. A área abaixo dessa curva é denotada por AU($\overline{\text{PRC}}$).

A segunda abordagem corresponde a calcular a média ponderada das áreas sob as curvas PR para cada classe individual (VENS *et al.* 2008). Esse cálculo é feito por meio da Equação (3), onde w_i são os pesos e L é o conjunto de todas as classes.

$$\overline{AUPRC}_{w_1,\ldots,w_{|L|}} = \sum_i w_i \cdot AUPRC_i \tag{3}$$

Essa abordagem pode ser utilizada de duas formas. De acordo com Borges (2012), na primeira abordagem, indicada por $\overline{\text{AUPRC}}$, os pesos são inicializados com o valor $1/|L|$. A segunda, denominada $\overline{\text{AUPRC}}_w$, consiste em atribuir um peso para uma classe de acordo com sua frequência. Este peso é calculado por $w_i = v_i/\sum_j v_j$, onde v_i é frequência da classe c_i no conjunto de dados (VENS *et al.* 2018).

2.4.2.2Medida baseada em distância

De acordo com Borges (2012, p. 36), “esta medida consiste em atribuir um custo que é proporcional à distância entre o exemplo da classe predita e a classe verdadeira”. A definição dessa distância considera a quantidade de ligações entre os nós no menor caminho que liga a classe verdadeira à classe predita. Esse tipo de medida pode ser independente da profundidade ou dependente da profundidade.

No primeiro caso, a medida de distância é calculada sem considerar o nível das duas classes na estrutura hierárquica, isto significa que a distância entre dois nós na hierarquia é definida como “o número de extremidades no caminho mais curto que os conecta” (BORGES, 2012, p. 36). Esse cálculo é mais complexo quando a hierarquia é organizada como um DAG, pois pode haver mais de um caminho entre dois nós, tornando necessário definir um critério para selecionar o caminho entre o nó da classe verdadeira e o nó da classe predita. De acordo com Borges (2012), um possível critério é a escolha do menor caminho entre os dois nós.

No segundo caso, o nível da hierarquia é importante durante o cálculo da medida de distância. Para isso, consideram-se o número de conexões entre as duas classes e a profundidade das classes na hierarquia (BORGES, 2012). Deve-se

destacar que um erro de predição num nível mais geral é considerado mais grave do que um erro de predição num nível mais específico. Devido a isto, faz-se necessário definir custos para cada nível da hierarquia, que pode ser feito por meio da atribuição de pesos para as conexões, onde as conexões recebem peso menor que as conexões localizadas em níveis menos profundos (BLOCKEEL *et al*, 2002). Neste sentido, no cálculo da distância de um caminho da hierarquia são utilizados os pesos das conexões para determinar a distância ponderada entre as classes.

De acordo com Sun e Lim (2001), uma forma de se obter a precisão ($Prec$) e a revocação (Rev) para a classificação hierárquica é considerar os erros de predição baseada na distância. Para isso, deve-se calcular a contribuição de falso positivo ($FpCon_i$) para cada classe c_i da hierarquia. Este cálculo é feito por meio da Equação (4), onde d é um exemplo de entrada, C_p é a classe predita e $RCon$ é a contribuição refinada, que é utilizada para normalizar a contribuição de cada exemplo no intervalo $[-1; 1]$.

$$FpCon_i = \sum_{d \in FP_i} RCon(d, C_p) \tag{4}$$

O valor da contribuição refinada é calculado pela Equação (5), em que C_v é a classe verdadeira.

$$RCon(d, C_p) = \min\left(1, \max\left(-1, Con(d, C_v)\right)\right) \tag{5}$$

Para o cálculo de $RCon$ é exigido o valor de contribuição de falsos positivos Con para cada classe. A Equação (6) mostra a fórmula utilizada para calcular este valor.

$$Con(d, C_p) = 1.0 - \frac{Dis(C_p, C_v)}{Dis_\theta} \tag{6}$$

onde $Dis(C_p, C_v)$ corresponde à distância entre a classe predita e a classe verdadeira e Dis_θ é a distância aceitável. A distância aceitável é especificada pelo usuário e seu valor deve ser maior do que zero.

O cálculo da contribuição de falsos negativos ($FnCon_i$) é realizado de maneira semelhante, conforme Equação (7), Equação (8) e Equação (9).

$$FnCon_i = \sum_{d \in FN_i} RCon(d, C_v) \tag{7}$$

$$RCon(d, C_v) = \min\left(1, \max\left(-1, Con\left(d, C_p\right)\right)\right) \tag{8}$$

$$Con(d, C_v) = 1.0 - \frac{Dis\left(C_p, C_v\right)}{Dis_\theta} \tag{9}$$

Uma vez que se tenha calculado as contribuições $FpCon_i$ e $FnCon_i$, os valores de precisão e revocação podem ser obtidos pela Equação (10) e pela Equação (11), respectivamente.

$$Prec = \frac{\max(0, |VP_i| + FpCon_i + FnCon_i)}{|VP_i| + |FP_i| + FnCon_i} \tag{10}$$

$$Rev = \frac{max(0, |VP_i| + FpCon_i + FnCon_i)}{|VP_i| + |FN_i| + FpCon_i} \tag{11}$$

A principal desvantagem das medidas baseadas em distância é o fato destas não levarem em consideração que classificações em níveis mais profundos da hierarquia são mais difíceis e conduzem a informações mais específicas do que classificações realizadas em níveis mais altos (ALMEIDA, 2018). Tal desvantagem pode ser contornada, aplicando-se custos mais altos para classificações erradas que ocorram em níveis mais elevados da hierarquia.

Neste trabalho, adota-se a medida AUPRC para avaliação da tarefa de classificação com os conjuntos de dados reduzidos resultantes da aplicação do método proposto. Escolheu-se esta medida porque são utilizados os classificadores Clus-HMC e a MHC-CNN na etapa de classificação e estes algoritmos a utilizam para medir seus desempenhos.

2.5 Considerações finais do capítulo

Neste capítulo foram apresentados os conceitos básicos sobre classificação e algumas abordagens para esta tarefa, destacando-se a classificação hierárquica

multirrótulo. Nesta abordagem, diferente dos métodos tradicionais de classificação, um exemplo pode ser categorizado em mais de uma classe simultaneamente e, além disso, as classes são organizadas segundo uma ordem taxonômica.

Devido a esta dupla característica, os métodos desenvolvidos para classificação hierárquica multirrótulo são, em geral, baseados em métodos existentes para a classificação multirrótulo e para a classificação hierárquica. Existem trabalhos disponíveis na literatura aplicados em diferentes domínios, destacando-se a classificação de texto e a bioinformática. Tais trabalhos utilizam uma abordagem global para considerar a organização hierárquica das classes.

É importante destacar que o desempenho dos métodos de classificação para problemas de classificação hierárquica global depende do número de exemplos, atributos e classes. Muitas vezes, tais problemas apresentam alta dimensionalidade de atributos, escassez de exemplos de treinamento e variação no número de classes às quais cada exemplo pertence. Neste caso, o uso de métodos de redução de dimensionalidade se faz necessário para transpor tais barreiras. Esses métodos são apresentados no Capítulo 3, sendo enfatizada a seleção de atributos.

3 REDUÇÃO DE DIMENSIONALIDADE E SELEÇÃO DE ATRIBUTOS

Este Capítulo apresenta os conceitos básicos sobre redução de dimensionalidade, destacando-se a seleção de atributos. A Seção 3.1 apresenta uma visão geral sobre a tarefa de redução de dimensionalidade e suas principais abordagens. A Seção 3.2 detalha os principais conceitos e técnicas de seleção de atributos. A Seção 3.3 descreve a abordagem filtro e apresenta as principais técnicas utilizadas. Por fim, a Seção 3.4 aborda as considerações finais do capítulo.

3.1 Redução de dimensionalidade

A redução de dimensionalidade consiste na redução do número de atributos, rótulos ou ambos, com o objetivo de melhorar a performance dos classificadores (GHODSI, 2006) (MANIKANDAN; ABIRAMI, 2018). Em outras palavras, reduzir a dimensionalidade é encontrar uma representação significativa em dimensionalidade reduzida para dados de alta dimensão, mantendo um número de mínimo de parâmetros que preservam as propriedades observadas nos dados. Esta tarefa facilita a classificação, a visualização e a compreensão dos dados (VAN DER MAATEN; POSTMA; VAN DEN HERIK, 2009).

Aplicar técnicas de redução de dimensionalidade proporciona, ainda, outros benefícios, como a redução de ruído e de redundância entre os atributos, promovendo um incremento no potencial para aprendizagem (MAATEN; POSTMA; HERIK, 2009). Além do mais, a redução de dimensionalidade pode levar a um aumento na capacidade de generalização dos métodos de aprendizagem de máquina quando os dados utilizados possuem grande quantidade de atributos (DUDA; HART; STORK, 2001). Este é o caso de problemas de classificação hierárquica multirrótulo, cuja quantidade excessiva de atributos em muitos domínios prejudica a extração de conhecimento do conjunto de dados (HUANG *et al*, 2020).

Segundo Faceli *et al.* (2011) *e* Borges (2012) a redução de dimensionalidade pode ser realizada por meio da extração de atributos ou pela seleção de atributos. A extração de atributos é um processo que, a partir do conjunto de dados original, cria novas características por meio da transformação ou combinação de características do conjunto original (GHODSI, 2006). Essas novas características tendem a ser mais

expressivas e representam com mais qualidade a variabilidade dos dados. Para (JAIN; DUIN; MAO, 2000), dado um espaço de características de dimensão m, os métodos de extração de características determinam um subespaço apropriado de dimensionalidade d tal que $d < m$. Diversas técnicas podem ser aplicadas para a redução de dimensionalidade por meio da extração de atributos, destacando-se: Análise de Componentes Principais (*Principal Component Analysis* – PCA) (GHODSI, 2006), *Multidimensional Scaling* – MDS (KRUSKAL, 1964) e *Self Organizing Map* – SOM (KOHONEN, 1990).

A seleção de atributos consiste na identificação dos relacionamentos entre os atributos de um conjunto de dados e na escolha daqueles que sejam mais significativos para compor um conjunto de dados simplificado capaz de produzir resultados iguais ou muito próximos aos obtidos a partir da análise do conjunto completo de dados para uma dada tarefa (GU; LI; HAN, 2012) (HUANG; LIU, 2020).

A escolha entre uma das duas abordagens está condicionada ao domínio de aplicação, bem como depende dos dados de treinamento disponíveis (SIQUEIRA, 2019). De modo particular, este trabalho versa sobre a utilização da seleção de atributos para redução de dimensionalidade em bases de dados de classificação hierárquica multirrótulo. Por este motivo, a Seção 3.2 é dedicada à exposição dos principais conceitos, abordagens e técnicas de seleção de atributos.

3.2 Seleção de atributos

De acordo com Lin e Gunopulos (2003), a seleção de atributos é o processo em que se escolhe um subconjunto ideal de d atributos em um conjunto original de alta dimensão de m atributos $(d < m)$, reduzindo a quantidade de atributos com base em um determinado critério. O processo de seleção de atributos objetiva encontrar os atributos que possuem relacionamentos relevantes, para formar um novo subconjunto. Nesse processo é desejável que atributos irrelevantes sejam desconsiderados (KUMAR; MINZ, 2014) (VENKATESH; ANURADHA, 2019).

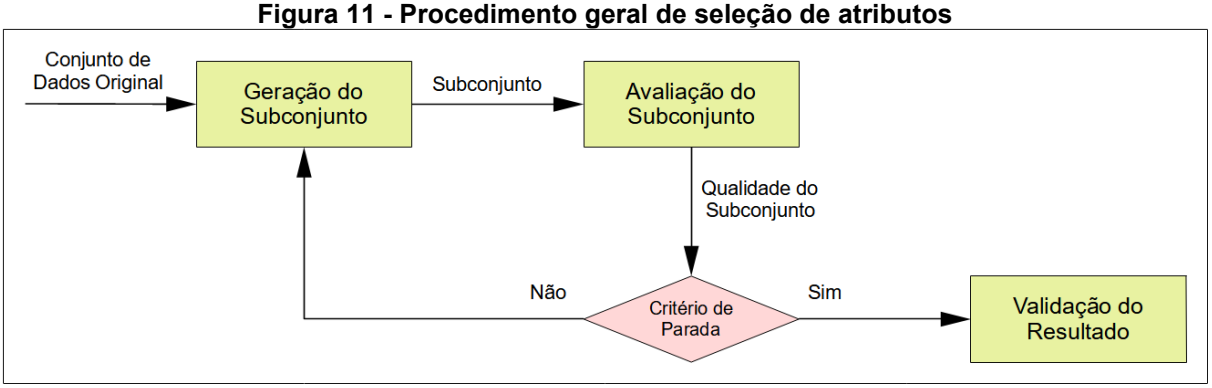
A seleção de atributos foi definida formalmente nos trabalhos de Kudo e Sklansky (1998) e Kumar e Minz (2014). Tomando por base tais definições, apresenta-se a seguir uma definição formal para esta tarefa.

Seja um conjunto original de atributos X, em que $|X| = m$, e seja a função $J(\cdot)$ um critério de avaliação a ser maximizado e definido como $J: X' \subseteq X \rightarrow \mathcal{R}$. A tarefa de seleção de atributos pode ser definida com base nos seguintes objetivos: (1) encontrar o melhor subconjunto X' com um determinado tamanho $d < m$. Isto quer dizer que, $J(X')$ é maximizado quando $d < m$ e $X' \subseteq X$; (2) encontrar o menor subconjunto X' com $J(X') > \theta$. Ou seja, definir um limiar θ para encontrar o menor subconjunto de atributos de modo que $d < m$ e $J(X') > \theta$; (3) encontrar a função de otimização $J(X')$ com o subconjunto de atributos ótimo.

É possível afirmar, portanto, que o subconjunto ótimo X' é, na verdade, apenas um dos subconjuntos ótimos, visto que a definição apresentada não garante que o subconjunto de atributos ótimo seja exclusivo (KUMAR; MINZ, 2014). Neste sentido, o subconjunto de atributos ótimo deve ser definido em termos do desempenho do classificador induzido (KOHAVI; JOHN, 1997): dado um indutor $\mathcal{I}$ e um conjunto de dados D, definido pelos atributos $(X_1, X_2, \ldots, X_d)$ de uma distribuição $\mathfrak{D}$ sobre o espaço de instâncias rotulado, um subconjunto ótimo de atributos, X_{opt}, para o qual a precisão do classificador induzido $\mathcal{C} = \mathcal{I}(\mathfrak{D})$ é máxima.

Em geral, na literatura, a tarefa de seleção de atributos é tratada a partir de um modelo geral, que é mostrado na Figura 11. Entretanto, Faceli *et al*. (2011) estabelece duas categorias de técnicas para seleção de atributos: técnicas de seleção de subconjuntos e técnicas de ordenação.

Figura 11 - Procedimento geral de seleção de atributos



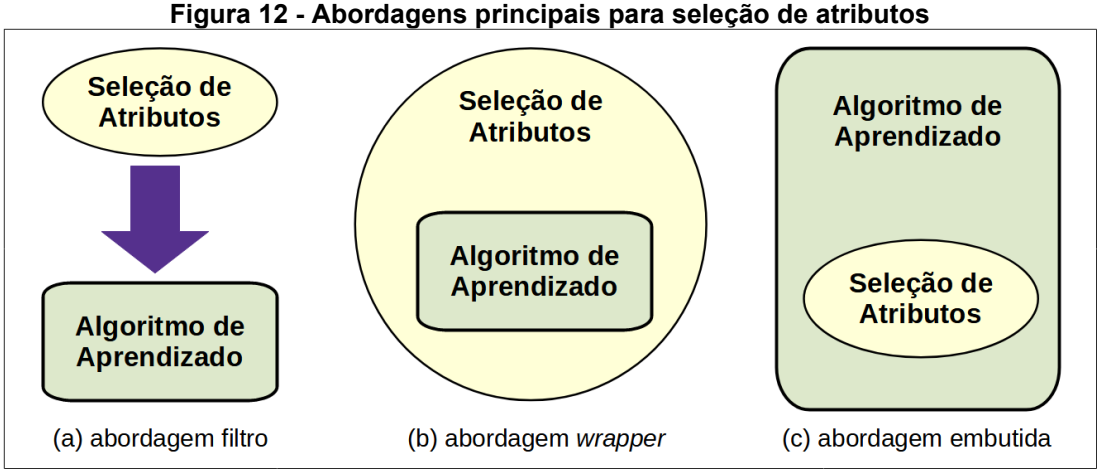
Fonte: Adaptado de Dash e Liu (1997, p. 133)

A primeira categoria corresponde ao modelo geral da Figura 11, onde um subconjunto é selecionado a partir dos atributos originais. Neste caso, a tarefa de seleção de atributos é tratada como um problema de busca. A segunda categoria é

caracterizada pela avaliação individual de cada atributos com base em sua relevância para separar o conjunto de dados nas diferentes classes.

Seja qual for a técnica adotada para a seleção de atributos, uma (ou uma combinação) das seguintes abordagens deve ser utilizada: filtro, *wrapper* e embutida (CHANDRASHEKAR; SAHIN, 2014) (JOVIĆ; BRKIĆ; BOGUNOVIĆ, 2014) (VENKATESH; ANURADHA, 2019). A Figura 12 ilustra as principais abordagens para a tarefa de seleção de atributos, destacando a participação do algoritmo de classificação. A seguir, cada uma dessas abordagens é descrita.

Figura 12 - Abordagens principais para seleção de atributos



Fonte: Covões (2010, p. 12)

A abordagem filtro, conforme ilustrado na Figura 12(a), consiste na escolha de um subconjunto de atributos por meio da estimação da qualidade dos atributos com base apenas nos dados, o que é feito antes da aplicação do algoritmo de indução (JOVIĆ; BRKIĆ; BOGUNOVIĆ, 2014) (VENKATESH; ANURADHA, 2019). Por isso, essa abordagem é dita independente do algoritmo de aprendizagem e as características dos dados do conjunto de treinamento são utilizadas para selecionar alguns recursos e descartar outros. Este método, apesar de ser computacionalmente eficiente, pode levar à perda de recursos que não sejam úteis por si mesmos, mas que podem ser úteis se combinados com outros (KUMAR; MINZ, 2014).

A abordagem *wrapper* é dita dependente do algoritmo de aprendizagem e seu objetivo é descobrir o conjunto de atributos que melhor se adeque a ele. A ideia geral é a avaliação do conjunto de atributos usando um algoritmo que funciona como uma caixa preta para encontrar os melhores subconjuntos de atributos (KUMAR; MINZ, 2014) (VENKATESH; ANURADHA, 2019). Em outras palavras, conforme representado na Figura 12(b), o algoritmo de classificação é executado para cada

subconjunto e sua avaliação é feita a partir da acurácia preditiva do algoritmo (ALMEIDA, 2018). Neste caso, o melhor subconjunto de atributos é aquele que produzir o melhor desempenho de aprendizado.

A abordagem embutida (do inglês *embedding*), ilustrada na Figura 12(c), considera que alguns algoritmos indutores conseguem realizar sua própria seleção de atributos de forma dinâmica enquanto buscam por uma solução (CHANDRASHEKAR; SAHIN, 2014). Em outras palavras, técnicas de seleção de atributos que usam a abordagem embutida selecionam o melhor subconjunto de atributos no próprio processo de indução do classificador, durante a fase de treinamento. Por exemplo, árvores de decisão utilizam uma função de avaliação para selecionar atributos que tenham melhor poder de discriminar entre as classes do problema.

Segundo Faceli *et al.* (2011), as abordagens filtro e *wrapper* podem ser empregadas na seleção de atributos tanto por ordenação quanto por seleção de subconjuntos. Já a abordagem embutida pode ser empregada apenas com a técnica de seleção de subconjuntos. Mais detalhes sobre essas categorias de técnicas de seleção de atributos são apresentados na Subseção 3.2.1 e na Subseção 3.2.2.

3.2.1 Técnicas de seleção de subconjuntos

Em geral, a tarefa de seleção de atributos é descrita conforme um modelo geral, que corresponde à seleção de subconjuntos. Esse modelo é apresentado na Figura 11 e sua descrição, que é apresentada em seguida, está baseada nos trabalhos de Dash e Liu (1997), Liu e Motoda (1998), Boz (2002) e Kumar e Minz (2014).

Há quatro etapas básicas para a seleção de atributos: geração de subconjunto, avaliação do subconjunto, critério de parada e validação do resultado. A etapa de geração do subconjunto pode ser traduzida como uma tarefa de busca para se produzir subconjuntos de atributos que serão avaliados. O processo é iniciado com a escolha de um subconjunto inicial, através da organização e seleção de atributos por meio de uma estratégia específica. Esse subconjunto é, então, avaliado para se garantir a qualidade dos atributos selecionados. No momento em que o critério de parada for satisfeito, a seleção é finalizada no procedimento de validação, caso contrário as etapas de geração e avaliação do subconjunto são realizadas novamente.

A geração dos subconjuntos se configura como uma busca heurística (KUMAR; MINZ, 2014). Diante disso, dois aspectos devem ser considerados para se especificar um subconjunto de atributos: a organização ou estratégia da busca e o ponto de partida e direção da busca ou geração de sucessores (FACELI *et al.*, 2011) (VENKATESH; ANURADHA, 2019) (JOVIĆ; BRKIĆ; BOGUNOVIĆ, 2014).

Referente ao primeiro aspecto, organização da busca, há três estratégias citadas na literatura: busca exponencial, busca sequencial e busca randômica (KUMAR; MINZ, 2014) (VENKATESH; ANURADHA, 2019).

A busca exponencial, também chamada de busca completa, avalia todos os possíveis subconjuntos (Faceli *et al.*, 2011). Trata-se de uma pesquisa otimizada que garante a melhor solução. Entretanto, esta é uma estratégia de busca exaustiva que requer 2^n combinações para seleção de atributos para um conjunto original com n atributos (VENKATESH; ANURADHA, 2019). Isto significa que se trata de um problema NP-Completo, o que se configura como uma desvantagem dessa estratégia.

A estratégia randômica foi introduzida para lidar com esse problema. Esta estratégia configura-se como uma abordagem não-determinística, em que os subconjuntos são gerados de maneira estocástica. Nesta abordagem não é garantido que se encontre uma solução ótima, entretanto é possível que seja encontrada uma boa solução sem que seja necessário percorrer todo o espaço de busca (Faceli *et al.*, 2011).

A busca sequencial seleciona apenas um entre todos os sucessores, adicionando a um conjunto inicialmente vazio ou removendo de um conjunto inicialmente completo (KUMAR; MINZ, 2014). Isto é feito de forma iterativa e o número de passos possíveis é $O(n)$ (VENKATESH; ANURADHA, 2019). O fato de os atributos serem adicionados ou removidos sequencialmente pode levar a mínimos locais (JOVIĆ; BRKIĆ; BOGUNOVIĆ, 2014). Podem ser utilizadas regras para guiar o processo de busca, porém não se assegura que se alcance uma solução ótima.

No que concerne ao segundo aspecto, direção da busca ou geração de sucessores, quatro estratégias principais são elencadas na literatura: geração para frente (*forward selection*), geração para trás (*backward elimination*), geração bidirecional e geração estocástica ou randômica (FACELI *et al.*, 2011) (JOVIĆ; BRKIĆ; BOGUNOVIĆ, 2014) (KUMAR; MINZ, 2014) (VENKATESH; ANURADHA, 2019). A estratégia *forward selection* começa com um conjunto de atributos vazio e, em seguida, um ou mais atributos são adicionados recursivamente ao conjunto a cada

iteração. Na estratégia *backward elimination*, inicia-se com o conjunto de atributos completo e, a cada iteração, atributos são removidos até que se tenha o subconjunto desejado. A geração bidirecional consiste em iniciar a busca por qualquer ponto e atributos podem ser adicionados ou removidos do subconjunto. Na geração estocástica, o ponto de partida e os atributos a serem adicionados ou removidos são definidos de maneira estocástica em cada iteração.

Os melhores atributos são escolhidos com base nos critérios definidos para avaliação dos subconjuntos. Há dois critérios de avaliação amplamente utilizados: independentes e dependentes de algoritmo. A abordagem filtro é utilizado como critério independente de algoritmo, pois considera apenas as características dos dados de treinamento a fim de avaliar a qualidade do subconjunto, sem que seja necessário utilizar nenhum algoritmo de aprendizagem (KUMAR; MINZ, 2014). As abordagens *wrapper* e embutida são utilizadas como critério dependente de algoritmo e requerem que seja utilizado um algoritmo de aprendizagem específico. Neste caso, a avaliação da qualidade do subconjunto de atributos é baseada no desempenho do algoritmo de aprendizagem (JOVIĆ; BRKIĆ; BOGUNOVIĆ, 2014). Isto determina quais atributos serão selecionados.

É necessário definir critérios de parada para finalizar o processo de busca pelo melhor subconjunto de atributos. Existem alguns critérios gerais de parada que foram elencados por Kumar e Minz (2014) e Venkatesh e Anuradha (2019), dentre os quais destacam-se: número máximo de iterações, quantidade de atributos pré-definida, taxa mínima de erro de classificação, busca concluída e percentual de avanço em duas iterações sucessivas. Quando o critério de parada é satisfeito, o melhor subconjunto de atributos de atributos foi selecionado. Este deve, então, ser validado. A validação pode ser realizada usando dados sintéticos ou conjuntos de dados do mundo real

3.2.2 Técnicas de ordenação

Na seleção de atributos por técnicas de ordenação “os atributos são ordenados de acordo com sua relevância para um dado critério” (FACELI *et al.*, 2011, p. 49). Esse critério pode ser, por exemplo, a classificação dos objetos nas diferentes classes. Neste caso, são selecionados os atributos situados no topo da ordenação,

ou seja, aqueles que são mais bem avaliados. Isto quer dizer que para cada atributo é calculado um valor de relevância e este valor é utilizado no processo de seleção (BOZ, 2002). Para medir a relevância ou qualidade de um atributo, pode-se, por exemplo, avaliar seu grau de associação com a classe (ALMEIDA, 2018). Para isto, diversas métricas foram propostas, as quais são abordadas na Seção 3.3.

Em geral, a seleção de atributos por ordenação é realizada através da abordagem filtro, embora também seja possível aplicar a abordagem *wrapper*, conforme já destacado. O uso da abordagem filtro é conveniente, pois, não requisita a participação de um algoritmo de aprendizagem no processo de seleção, focando apenas nas características dos dados de treinamento para avaliar a qualidade dos atributos. Na Seção 3.3 são apresentados detalhes sobre a abordagem filtro, incluindo detalhes sobre as medidas de avaliação da relevância dos atributos.

É importante destacar que em ambas as categorias de técnicas para seleção de atributos, o resultado do processo de seleção é um subconjunto reduzido. O que as diferencia é que no caso da seleção de subconjuntos, os atributos escolhidos não são necessariamente aqueles com maiores pontuações, que estejam no topo da ordenação. Neste caso, avalia-se quão bem dois ou mais atributos são significativos quando “atuam de forma coletiva, em conjunto” (FACELI *at al.*, 2011, p. 50).

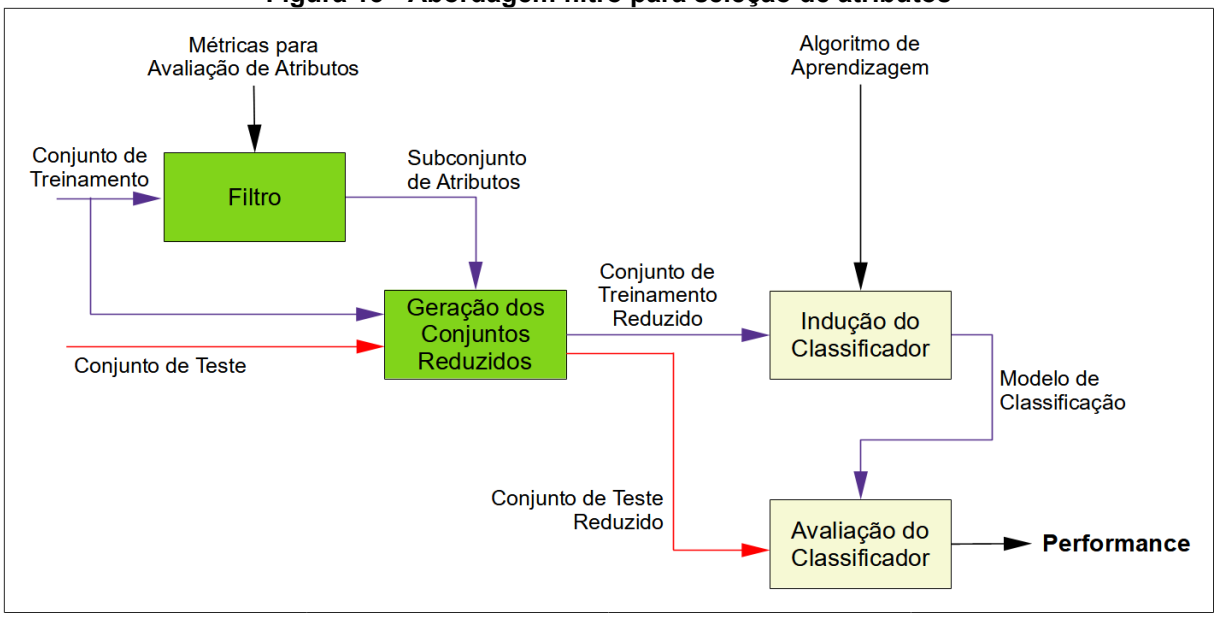
3.3 Técnicas da abordagem filtro

Conforme já exposto, a abordagem filtro caracteriza-se por ser independente do algoritmo de aprendizagem, ou seja, a seleção acontece previamente à etapa de indução do classificador, isto é, na etapa de pré-processamento. Neste caso, a seleção dos melhores atributos considera apenas as características dos próprios dados, sendo utilizada a técnica de ordenação. A Figura 13 ilustra, de forma geral, o processo de seleção de atributos na abordagem filtro.

Como se observa na Figura 13, somente após a geração dos conjuntos de dados de treinamento e teste reduzidos é que ocorre a indução e a avaliação do classificador. A terminologia filtro provém, portanto, da ideia de que os atributos irrelevantes são filtrados da base de dados antes da aplicação do algoritmo de classificação (BLUM; LANGLEY, 1997). Nesta abordagem, pode-se distinguir duas atividades básicas, as quais estão destacadas (LAZAR *et al.*, 2012). A primeira,

identificada como Filtro, corresponde à seleção de atributos propriamente dita e consiste em pontuar os atributos, baseado em algum critério de avaliação, e escolher aqueles que forem mais bem avaliados para compor o subconjunto de atributos. A segunda atividade é a geração dos conjuntos de dados reduzidos a partir dos conjuntos de dados originais e do subconjunto de atributos selecionados.

Figura 13 - Abordagem filtro para seleção de atributos



Fonte: Adaptado de Lee (2005, p. 25)

Um dos aspectos mais importantes da abordagem filtro é discriminar os melhores atributos. Para isso, é preciso definir critérios de avaliação precisos, isto é, medidas para determinar a importância dos atributos e, assim, descartar os menos significativos e selecionar os melhores. Essas medidas podem ser organizadas nas seguintes categorias: distância, informação, dependência e consistência (LEE, 2005).

As medidas de distância, também chamadas de medidas de separabilidade, divergência ou discriminação, indicam que na diferença entre dois valores de distância, se deve dar preferência ao atributo cujo valor absoluto resultante seja maior (KUMAR; MINZ, 2014). Quanto às medidas de informação, a escolha de um atributo é definida por seu valor de ganho de informação em relação ao valor do ganho de informação de outros atributos (LORENA; CARVALHO; LORENA, 2015). As medidas de dependência indicam o quanto dois atributos estão correlacionados, permitindo determinar quão redundantes esses atributos são (LEE, 2005). As medidas de consistência “encontram o subconjunto mínimo de atributos que satisfaz a proporção de inconsistência aceita, geralmente definida pelo usuário” (LEE, 2005).

Outro aspecto importante na abordagem filtro diz respeito à forma como a avaliação da qualidade dos atributos é feita. As técnicas do tipo filtro podem avaliar atributos individualmente (filtro univariado) ou avaliar subconjuntos de atributos (filtro multivariado) (JOVIĆ; BRKIĆ; BOGUNOVIĆ, 2014) (ALMEIDA, 2018). Para a tarefa de classificação, técnicas filtro univariado geralmente abordam o ranqueamento dos atributos, enquanto que as técnicas filtro multivariado, em geral, envolvem dependência e consistência.

Embora existam muitas técnicas para a abordagem filtro descritas na literatura, destacam-se as seguintes: como técnicas filtro univariado – *Information Gain* (IG) (AZHAGUSUNDARI; THANAMANI, 2013) (HOQUE; BHATTACHARYYA; KALITA, 2014) (PEREIRA *et al.*, 2015) (ALJEDANI; ALOTAIB; TAILEB, 2021) (SILVA; CERRI, 2021) (OMUYA; OKEYO; KIMWELE, 2021), *Gain Ratio* (GR) (WITTEN; FRANK, 2011) (PRIYADARSINI; VALARMATHI; SIVAKUMARI, 2011) (NAGPAL; GAUR, 2015) (ALJEDANI; ALOTAIB; TAILEB, 2021) (CHEN; HU; ZHANG, 2021), *Chi-Square* (χ^2) (WITTEN; FRANK, 2011) (JIN, *et al.*, 2015) (ZHAI *et al.*, 2018) (ALJEDANI; ALOTAIB; TAILEB, 2021) (PUTRI; RUSTAM; SARWINDA, 2019), *Relief* e suas variantes (LIU; MOTODA, 2008) (SPOLAÔR, N. *et al.* 2013b) (SLAVKOV *et al.*, 2014) (SLAVKOV *et al.*, 2018) (URBANOWICZ *et al*., 2018) (SILVA; CERRI, 2021) e *Fisher Score* (FS) (GU; LI; HAN, 2012) (SINGH *et al.*, 2014) (PÉREZ-ORTIZ *et al.*, 2016) (SUN *et al.*, 2021); e como técnicas filtro multivariado – *Correlation-based Feature Selection* (MICHALAK; KWASNICKA, 2010) (WITTEN; FRANK, 2011) (GOPIKA; KOWSHALAYA, 2018), *Fast Correlation-based Filter* (YU; LIU, 2003) e *Consistency-based Feature Selection* (DASH; LIU; MOTODA, 2000) (SHIN; XU, 2009).

O Quadro 1 resume as principais técnicas do tipo filtro para a seleção de atributos aplicadas à tarefa de classificação, indicando a forma como realizam a avaliação dos atributos e o tipo de medida de avaliação em que são categorizadas.

De modo particular, este trabalho aborda uma medida baseada no conceito de distância para avaliar atributos individualmente. Neste sentido, algumas das técnicas listadas anteriormente são brevemente descritas a seguir, destacando-se a medida *Fisher Score*, objeto de estudo do trabalho. Mais detalhes podem ser encontrados nos trabalhos de referência indicados.

Quadro 1 - Técnicas da abordagem filtro

Forma de Avaliação	Tipo de Medida de Avaliação	Técnica
Filtro Univariado	Informação	*Information Gain*
		Gain Ratio
	Distância	*Chi-Square*
		Relief
		Fisher Score
Filtro Multivariado	Distância	*Correlation-based feature selection*
	Informação	*Fast Correlation-based Filter*
	Consistência	*Consistency-based Feature Selection*

Fonte: Autoria Própria (2022)

3.3.1 *Information Gain* (IG)

O IG é uma técnica do tipo filtro univariado, baseada no conceito de entropia (PEREIRA *et al.*, 2015), que calcula informações mútuas para cada atributo e classe, de modo a produzir um *ranking* de todos os atributos (REMESEIRO; BOLON-CANEDO, 2019). Mais formalmente, seja um atributo X e uma classe Y, o IG para o atributo X dada a classe Y pode ser calculado como a diferença entre a entropia da classe e a entropia condicional da classe dado o atributo X, de acordo com a Equação (12) (OMUYA; OKEYO; KIMWELE, 2021).

$$IG(X/Y) = H(Y) - H(Y/X) \tag{12}$$

onde $H(Y)$ é a entropia da classe Y antes da observação do atributo X e $H(Y/X)$ é a entropia condicional da classe Y após a observação do atributo X.

Considerando Y uma variável aleatória que pode assumir qualquer valor no espaço de classes $L = \{y_1, y_2, \dots, y_q\}$, com $q > 1$, e supondo que a probabilidade de se observar cada um desses valores seja $p_1, p_2, \dots, p_q$, a entropia da classe Y antes da observação do atributo X é calculada pela Equação (13) (AZHAGUSUNDARI; THANAMANI, 2013).

$$H(Y) = -\sum_{j=1}^{q} p(y_j) \cdot \log_2 p(y_j) \tag{13}$$

onde $p(y_j)$ é dado pela razão entre o número de instâncias em que o valor y_j da classe ocorre e o número total de instâncias na base de dados.

Seja X uma variável aleatória, que pode assumir qualquer valor no espaço de atributos $F = \{x_1, x_2, \dots, x_m\}$ e seja $p(y_j|x_i)$ a probabilidade condicional de se observar o valor de classe y_j dado o valor de atributo x_i. A entropia condicional da classe Y, dado o atributo X, é calculada conforme a Equação (14) (ALMEIDA, 2018).

$$H(Y|X) = -\sum_{j=1}^{q}\sum_{i=1}^{m}\left[p(y_j|x_i)\cdot \log_2\left(\frac{p(y_j|x_i)}{p(x_i)}\right)\right] \quad (14)$$

Quanto mais informativo um atributo X é em relação à classe Y, menor é a sua entropia condicional $H(Y|X)$. Além disso, a análise da Equação (13) e da Equação (14) permite deduzir que $H(Y|X) > H(Y)$, uma vez que o fato de se conhecer o valor de X conduz à determinação do valor de Y. Isto justifica a fórmula da Equação (12).

3.3.2 *Gain Ratio* (GR)

Um problema enfrentado pela medida de IG é que a qualidade dos atributos pode ser superestimada (QUINLAN, 1986). Uma forma de contornar este problema e consiste em ponderar a fórmula original, calculando a razão do ganho de informação do atributo X em relação à classe Y e a entropia do atributo, conforme Equação (15).

$$GR(Y|X) = \frac{IG(Y|X)}{H(X)} \quad (15)$$

A medida RG expressa a proporção de informação gerada pela partição do conjunto de dados que aparenta ser útil para a classificação, para isso, leva em consideração as informações intrínsecas da partição (PRIYADARSINI; VALARMATHI; SIVAKUMARI, 2011).

3.3.3 *Chi-Square* (χ^2)

A métrica *Chi-Square* mede a relevância entre um atributo X e uma classe Y (ZHAI *et al.*, 2018). Assim sendo, a qualidade do atributo é avaliada de acordo com sua correlação com a classe, que é feito por meio de um teste estatístico χ^2 (PUTRI; RUSTAM; SARWINDA, 2019).

Segundo Almeida (2018), a métrica χ^2 é calculada pela Equação :

$$\chi^2 = \sum_{i=1}^{m}\sum_{j=1}^{q}\left(\frac{o_{ij} - e_{ij}}{e_{ij}}\right) \tag{16}$$

Em que o_{ij} é a frequência observada da combinação $\left(X = x_i; Y = y_j\right)$ e e_{ij} é a frequência esperada para a mesma combinação, onde x_i é o valor do atributo X, com $1 \leq i \leq m$, e y_j é o valor da classe Y, com $1 \leq j \leq q$.

A frequência observada o_{ij} para cada combinação de $X = x_i$ e $Y = y_j$ é calculada pela Equação (17).

$$o_{ij} = \frac{count\left(X = x_i;\ Y = y_j\right)}{n} \tag{17}$$

onde $count\left(X = x_i; Y = y_j\right)$ é o número de instâncias em que ocorre o valor i do atributo X simultaneamente com o valor j da classe Y e n é o número total de instâncias da base de dados.

A frequência esperada para cada combinação de $X = x_i$ e $Y = y_j$ é calculada pela Equação (18).

$$e_{ij} = \frac{count(X = x_i) \cdot count\left(Y = y_j\right)}{n} H(X) + H(Y) \tag{18}$$

onde $count(X = x_i)$ é o número de instâncias em que ocorre o valor x_i do atributo X, $count\left(Y = y_j\right)$ é o número de instâncias associadas à classe y_j e n é o número total de instâncias da base de dados.

Aplica-se, então, o teste estatístico para determinar se o atributo X e a classe Y são independentes. Se essa hipótese puder ser rejeitada, conforme um nível de significância estatística pré-determinado e uma distribuição, então o atributo é fortemente relacionado à classe (ALMEIDA, 2018).

3.3.4 *Relief* (RF)

Relief é um algoritmo que implementa uma métrica de avaliação univariada de atributos, que considera não somente a diferença de valores dos atributos e a diferença de classes, mas também as distâncias entre as instâncias (URBANOWICZ *et al*., 2018). Essas distâncias são calculadas no espaço de atributos, o que indica que instâncias similares estão próximas entre si e instâncias dissimilares estão mais afastadas (SPOLAÔR *et al.* 2013b).

O mecanismo do algoritmo é o seguinte (LIU; MOTODA, 2008): dado um conjunto de dados com n instâncias, para cada instância E_k de um subconjunto aleatório de t instâncias ($t \leq n$), calculam-se a instância mais próxima (E_H) e a instância mais distante (E_M) da mesma classe. Em seguida, atualiza-se a medida de qualidade W para cada atributo, por meio da Equação (19). Essa medida de qualidade considera a capacidade do atributo de diferenciar entre as classes.

$$W_i = \frac{W_i - diff(i, E_k, E_H)}{t} + \frac{diff(i, E_K, E_M)}{t} \tag{19}$$

onde $diff(i, E_j, E_k)$ é uma função que calcula a diferença entre os valores de duas instâncias para cada atributo x_i, a qual é definida pela Equação (20), quando o atributo é numérico e pela Equação (21), se o atributo é nominal.

$$diff(i, E_j, E_k) = \frac{|E_{j,i} - E_{k,i}|}{\max(x_i) - \min(x_i)} \tag{20}$$

$$diff(i, E_j, E_k) = \begin{cases} 0 & \text{se } E_{j,i} = E_{k,i} \\ 1 & \text{se } E_{j,i} \neq E_{k,i} \end{cases} \tag{21}$$

A medida *Relief*, embora seja mais elaborada, não consegue lidar com o problema de dados desbalanceados.

3.3.5 *Fisher Score* (FS)

Fisher score é uma métrica eficiente para redução de dimensionalidade, proposta como uma estratégia heurística para seleção de atributos (PÉREZ-ORTIZ *et al.,* 2016), que calcula uma pontuação independente para cada atributo usando a razão de Fisher (DUDA; HART; STORK, 2001). Essa medida indica que os atributos de boa qualidade devem possuir valores semelhantes para instâncias de uma mesma classe e valores diferentes para instâncias de classes distintas.

De acordo com Gu, Li e Han (2012) e Sun *et al*. (2021), a ideia central é encontrar um subconjunto de atributos tal que no espaço que corresponde aos atributos selecionados as distâncias entre pontos de dados de mesma classe sejam o menor possível e a distância entre pontos de dados de classes diferentes seja o maior possível. Formalmente, dado um conjunto de dados treinamento $D \in \mathbb{R}^{m \times q}$, onde m é o número de atributos e q é número de rótulos de classes, deseja-se selecionar d atributos de modo que $D \in \mathbb{R}^{m \times q}$ é reduzido para $T' \in \mathbb{R}^{d \times q}$.

Em geral, para reduzir o custo computacional, calcula-se a medida para cada atributo individualmente. Deste modo, o FS do k-ésimo atributo, considerando a j-ésima classe, é calculado pela Equação (22) (SUN *et al.*, 2021):

$$FS(x_k) = \frac{\sum_{j=1}^{q} n_j \left(\mu_k^j - \mu_k\right)^2}{\sum_{j=1}^{q} S_t^j(x_k)} \tag{22}$$

onde, n_j é o número de exemplos da j-ésima classe; μ_k^j e μ_k correspondem, respectivamente à média do k-ésimo atributo, considerando a j-ésima classe e à média do k-ésimo atributo em relação a todo o conjunto de dados; e $S_t^j(x_k)$ é a matriz de dispersão da classe do k-ésimo atributo em relação à j-ésima classe. Seu valor é calculado conforme a Equação (23):

$$S_t^j(x_k) = \sum_{i=1}^{n_j} \left(x_{ik}^j - \mu_k^j\right)^2 \tag{23}$$

onde x_{ik}^{j} é o valor do k-ésimo atributo para a i-ésima instância na j-ésima classe.

Uma vez que se tenha calculado o valor do FS para cada um dos atributos no conjunto de treinamento, são selecionados os d atributos com os maiores valores. Pode-se escolher o valor de d de acordo com a quantidade de atributos que se deseje selecionar ou utilizar algum outro critério como, por exemplo, considerar o valor médio das medidas calculadas como limiar (GÜNES *et al*., 2010). Neste caso, os atributos que possuem o valor de FS acima da média aritmética dos valores de FS entre todos os atributos são incluídos no subconjunto de dados reduzido.

A medida calculada representa o potencial do k-ésimo atributo em discriminar entre duas classes, ou seja, quanto maior o valor de FS, mais discriminativo é o atributo.

3.4 Considerações finais do capítulo

Neste capítulo foi apresentado o problema da redução de dimensionalidade em bases de dados, destacando-se como este problema pode ser abordado através de técnicas de seleção de atributos. O principal objetivo da aplicação de tais técnicas é melhorar o desempenho do classificador, aumentando sua capacidade preditiva.

A seleção de atributos consiste na identificação e seleção do subconjunto de atributos mais relevantes, o que é feito a partir da avaliação de qualidade dos atributos presentes no conjunto original de dados, podendo ser utilizadas as abordagens filtro, *wrapper* ou embutida. De modo particular, a abordagem filtro pode ser aplicada para medir a relevância dos atributos de modo individual ou em subconjuntos. Nesta abordagem, aplicam-se métricas de avaliação que consideram apenas a natureza dos dados para atribuir uma pontuação para cada atributo, a fim de se escolher aqueles que sejam mais bem avaliados.

As técnicas apresentadas neste capítulo têm sido aplicadas para seleção de atributos em contextos de classificação hierárquica multirrótulo. Este cenário é apresentado no capítulo 4, onde é apresentado o estado da arte para o problema.

Neste trabalho, utiliza-se a abordagem filtro e a medida FS como métrica para avaliação da qualidade dos atributos. A partir do valor FS de cada atributo, busca-se

selecionar aqueles mais bem avaliados para compor o subconjunto de dados reduzidos, que, por sua vez, são utilizados na tarefa de HMC.

4 MAPEAMENTO SISTEMÁTICO DE LITERATURA

Este Capítulo apresenta o levantamento bibliográfico que foi realizado para estabelecer o estado da arte atual a respeito das técnicas de redução de dimensionalidade em bases de dados de HMC. A Seção 4.1 apresenta o método utilizado para a realização do mapeamento sistemático de literatura. A Seção 4.2 aborda a etapa de planejamento inicial do mapeamento. A Seção 4.3 descreve a efetuação da etapa de buscas. A Seção 4.4 apresenta os resultados obtidos com o mapeamento sistemático. A Seção 4.5 aborda as considerações finais do capítulo.

4.1 Descrição do método de mapeamento sistemático

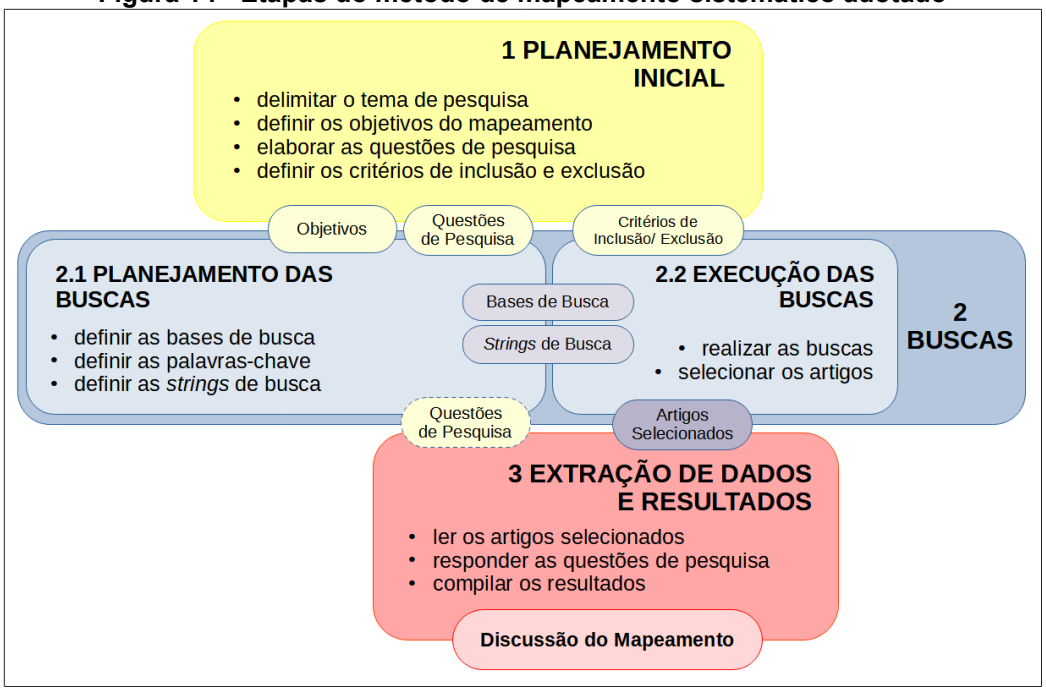
O método utilizado para realização do mapeamento sistemático foi inspirado nos protocolos desenvolvidos por Rattan, Bathia e Singh (2013) e possui as seguintes etapas: 1) planejamento inicial; 2) buscas; e 3) extração de dados e resultados. A etapa de buscas é dividida em duas subetapas: 2.1) planejamento das buscas e 2.2) execução das buscas. A Figura 14 apresenta uma visão geral do método e a sequência de etapas recomendada, enfatizando as atividades realizadas em cada etapa e os produtos de cada uma delas.

Na etapa de planejamento inicial são definidos o tema de pesquisa e os objetivos do mapeamento, além disso, são elaboradas as questões de pesquisa a serem respondidas. As questões de pesquisa devem estar alinhadas à intenção da pesquisa e guiam a extração de informações durante o mapeamento. Ainda nesta etapa, devem ser definidos os critérios de inclusão e exclusão para a seleção dos trabalhos a serem abordados no mapeamento. Também devem ser definidos o intervalo de tempo a ser considerado nas buscas, além da área e do idioma de publicação. Estes três últimos critérios possibilitam a realização de uma seleção prévia dos artigos durante a etapa de buscas, usando, para isso, recursos disponibilizados pelas bases de busca. Por serem utilizados como parâmetros durante a realização das buscas, estes critérios são chamados de critérios de busca.

Uma vez que os objetivos do mapeamento, as questões de pesquisa e os critérios de inclusão e exclusão tenham sido definidos, o método de mapeamento conduz à etapa de buscas, que, por sua vez é composta por suas subetapas. No

planejamento das buscas são realizadas as seguintes atividades: definição das bases de busca, definição das palavras-chave e definição das strings de busca. A subetapa de execução das buscas corresponde à realização das buscas nas bases e posterior seleção dos trabalhos que serão analisados.

Figura 14 - Etapas do método de mapeamento sistemático adotado



Fonte: Autoria Própria (2022)

A definição das bases de busca consiste na seleção das fontes de dados para realização das buscas. As bases elegidas devem ser as mais apropriadas, tendo em vista que todos os artigos a serem utilizados nas etapas seguintes são extraídos dessas fontes.

A definição de palavras-chave corresponde à escolha de termos pertinentes ao tema de pesquisa e aos objetivos do mapeamento. As palavras-chave que formam a base das *strings* de busca. Uma *string* de busca é formada por uma ou mais palavras-chave ligadas por meio de operações lógicos de conjunção, disjunção ou negação. Cada *string* é também um critério de busca e deve ser utilizada para a execução das buscas em conjunto com os critérios de busca já definidos.

A realização das buscas, por sua vez, refere-se à recuperação de estudos que atendem aos critérios de busca estabelecidos. Tais critérios, por sua vez, podem ter a sintaxe adaptada para atender ao formato exigido por cada uma das bases de busca escolhidas.

Uma vez que se tenha realizado as buscas, segue-se para a seleção dos trabalhos, cuja finalidade é escolher os estudos mais alinhados aos propósitos do trabalho que está sendo conduzido. Essa escolha obedece aos critérios de inclusão e exclusão estabelecidos na etapa de planejamento inicial. Nesta etapa de busca, podem ser utilizados softwares de gerenciamento de bibliografia para facilitar o controle e organização dos estudos.

Na etapa de extração de dados e resultados, procede-se à leitura dos artigos selecionados, analisando-os com base nas questões de pesquisa elaboradas na etapa de planejamento inicial. Em seguida, cada questão é respondida e são apresentadas as informações que tenham sido inferidas a partir da leitura dos artigos. Por fim, os resultados do mapeamento são compilados e discutidos, gerando-se um documento de mapeamento sistemático de literatura.

Nas seções seguintes descreve-se a aplicação do método de mapeamento sistemático no contexto deste trabalho.

4.2 Planejamento inicial

O tema geral de pesquisa para este trabalho é “Seleção de Atributos em Bases de Dados de Classificação Hierárquica Multirrótulo”. O objetivo central é identificar quais métodos de seleção de atributos são aplicados em problemas de classificação hierárquica multirrótulo. Em vista disso, foram elaboradas quatro questões de pesquisa, que são mostradas no Quadro 2, identificadas por Q_i, onde $i = 1,2,3,4$.

Como critério de inclusão, definiu-se que devem ser considerados apenas trabalhos publicados em conferências e periódicos, cuja versão completa esteja disponível nas fontes pesquisadas. Os critérios de exclusão utilizados são os seguintes: exclusão de itens duplicados, por eventualmente terem sido recuperados a partir de bases diferentes e exclusão dos estudos cujo título ou resumo não sejam compatíveis com o tema em questão. Além disso, foram adotados os seguintes critérios de busca: período de publicação de 2010 a 2022, publicações na grande área de Ciência da Computação e no idioma Inglês.

Quadro 2 - Questões de pesquisa

Q_1	Quais são as abordagens de seleção de atributos utilizadas? (Filtro, *Wrapper* ou Embutida)
Q_2	Quais foram as áreas em que a seleção de atributos foi aplicada? Qual o formato da estrutura hierárquica das classes? (Árvore ou DAG)
Q_3	Qual foi a contribuição científica dos autores nos estudos para o problema da seleção de atributos?
Q_4	O resultado obtido pelo conjunto de dados reduzido foi relevante quando comparado com o conjunto de dados formado por todos os atributos?

Fonte: Autoria Própria (2022)

4.3 Buscas

A efetuação desta etapa se inicia com o planejamento das buscas. O primeiro passo foi a definição de quais são as bases de dados a serem utilizadas no processo de busca. Com isso, foram escolhidas as bases que atendam, já como requisitos para uma primeira filtragem dos trabalhos, os seguintes critérios de inclusão e exclusão: seleção de publicações na área de Ciência da Computação, artigos na língua inglesa e definição do intervalo temporal de 2010 a 2022. Além disso, a escolha das bases de busca considerou a possibilidade de realizar consultas usando combinações de palavras-chave. O Quadro 3 apresenta a relação das bases de pesquisa e seus respectivos sites.

Quadro 3 - Definição das bases de busca

Base de busca	Site
IEEE	<https://ieeexplore.ieee.org>
Scopus	<https://www.scopus.com>
Science Direct	<https://www.sciencedirect.com>
Springer	<https://link.springer.com/>
Inderscience	< https://www.inderscience.com/>
ArXiv.org	<https://arxiv.org/>
Emerald Insight	<https://www.emeraldinsight.com/>

Fonte: Autoria Própria (2022)

Com o objetivo de obter resultados de busca precisos dentro do tema pesquisado, foi definido um conjunto de palavras-chave, em língua inglesa, relacionadas às questões de pesquisa. As palavras-chave escolhidas são as

seguintes: *Hierarchical Multi-label Classification*, *Dimensionality Reduction* e *Feature Selection*.

As strings de busca correspondem a combinações dessas palavras-chave, por meio dos operadores lógicos AND e OR. Variações desses termos foram utilizados para melhorar a qualidade dos resultados obtidos, a saber: *Multi-label Hierarchical Classification* e *Attribute Selection* são, respectivamente, variações das palavras-chave elencadas anteriormente, exceto o termo *Dimensionality Reduction*. As strings formadas, identificadas por S_i $(i = 1,\ 2)$, são apresentadas no Quadro 4.

Quadro 4 - *Strings* de busca

ID	***String* de busca**
S_1	("*Hierarchical Multi-label Classification*" OR "*Multi-label Hierarchical Classification*") AND "*Dimensionality Reduction*"
S_2	("*Hierarchical Multi-Label Classification*" OR "*Multi-Label Hierarchical Classification*") AND ("*Feature Selection*" OR "*Attribute Selection*")

Fonte: Autoria Própria (2022)

Para a realização das buscas nas bases de dados selecionadas, além das strings de busca, foram utilizados os critérios de busca definidos na etapa de planejamento inicial. O resultado das buscas, utilizando-se as strings de busca definidas e os critérios de seleção adotados, é apresentado na Tabela 2.

Tabela 2 - Total de resultados das buscas

Base de Busca	S_1	S_2	Quantidade
IEEE Xplore	02	06	08
Scopus	00	05	05
Science Direct	03	17	20
Springer	28	38	66
Inderscience	00	01	01
ArXiv.org	00	00	00
Emerald Insight	00	00	00
TOTAIS	33	67	100

Fonte: Autoria Própria (2022)

Ao final dessa atividade foram recuperados um total de 100 trabalhos e estes foram importados para o gerenciador de bibliografia *Zotero* (https://www.zotero.org/), para que fosse possível garantir um melhor gerenciamento dos resultados obtidos através dos recursos disponibilizados pela ferramenta.

O processo de seleção dos artigos foi conduzido a partir dos critérios de inclusão e exclusão definidos na etapa de planejamento inicial. Inicialmente, foram eliminadas as duplicações por meio da junção de resultados. Em seguida, procedeu-se à exclusão das publicações de livros e capítulos de livro, uma vez que os critérios de exclusão definem que serão considerados apenas artigos publicados em conferências ou periódicos. Este procedimento foi realizado de maneira automatizada por meio da ferramenta *Zotero*, resultando em 65 artigos.

Na sequência, foram descartados os artigos cujo título e o resumo não tinha relação direta com o tema deste trabalho, produzindo um total de 8 artigos selecionados para leitura e extração de dados. Os autores, título dos artigos e ano de publicação são apresentados no Quadro 5. A aplicação da etapa de extração de dados e resultados é descrita na Seção 4.4.

Quadro 5 - Descrição dos artigos selecionados

ID	Autores	Título	Ano de Publicação
1	SLAVKOV, I. *et al.*	*ReliefF for hierarchical multi-label classification*	2014
2	YAN, S.; WONG. K.	*Elucidating high-dimensional cancer hallmark annotation via enriched ontology*	2017
3	CERRI; R. *et al*.	*Multi-label Feature Selection Techniques for Hierarchical Multi-label Protein Function Prediction*	2018
4	SLAVKOV, I. *et al*.	*HMC-reliefF: Feature ranking for hierarchical multi-label classification*	2018
5	MELO, A.; PAULHEIM, H.	*Local and global feature selection for multilabel classification with binary relevance*	2019
6	HUANG, H.; LIU, H	*Feature selection for hierarchical classification via joint semantic and structural information of labels*	2020
7	ALJEDANI, N.; ALOTAIBI, R.; TAILEB, M.	*HMATC: Hierarchical multi-label Arabic text classification model using machine learning*	2021
8	SILVA, L; CERRI, R.	*Feature Selection for Hierarchical Multi-label Classification.*	2021

Fonte: Autoria Própria (2022)

Para simplificar a citação dos artigos selecionados, cada trabalho recebeu um número de identificação (ID), o qual foi utilizado como referência no restante do trabalho.

4.4 Extração de Dados e Resultados

Nesta etapa são apresentadas as respostas obtidas para as perguntas definidas no Quadro 2. Para isso, as perguntas são representadas pelos seus identificadores Q_1 a Q_4.

***Q_1**: Quais são as abordagens ou técnicas de seleção de atributos utilizadas? (Filtro, Wrapper ou Embutida)*

As abordagens e técnicas identificadas nos trabalhos analisados são apresentados no Quadro 6. Do total de trabalhos selecionados, 7 utilizam abordagem filtro, correspondendo a 87,5% do total de artigos, e 1 faz uso da abordagem *wrapper*. Nota-se uma prevalência da abordagem filtro sobre a abordagem *wrapper*. Nenhum dos trabalhos selecionados faz uso da abordagem embutida.

Quadro 6 - Abordagens e técnicas de seleção de atributos

ID	Abordagem	Técnica
1	Filtro	*HMC-ReliefF*
2	Filtro	*United Decision Tree* (UDT) *United GSS Coefficient* (UGSS) *United NGL Coefficient* (UNGL)
3	*Wrapper*	*Clus*-HMC
4	Filtro	*HMC-ReliefF*
5	Filtro	*Information Gain*
6	Filtro	*Semantic and Strutural Information (FSSS)*
7	Filtro	*Binary Relevance with Chi-Square (BR-χ^2)* *Label Powerset with Chi-Square (LP-χ^2)* *Binary Relevance with Gain Ratio (BR-GR)* *Label Powerset with Gain Ratio (LP-GR)* *Binary Relevance with Relief (BR-RF)* *Label Powerset with Relief (LP-RF)* *Binary Relevance with Information Gain (BR-IG)* *Label Powerset with Information Gain (LP-IG)*
8	Filtro	*ReliefF based on the Binary Relevance transformation (RF-BR)* *ReliefF based on the Label Powerset transformation (RF-LP)* *Information Gain based on the Binary Relevance transformation (IG-BR)* *Information Gain based on the Label Powerset transformation (IG-LP)*

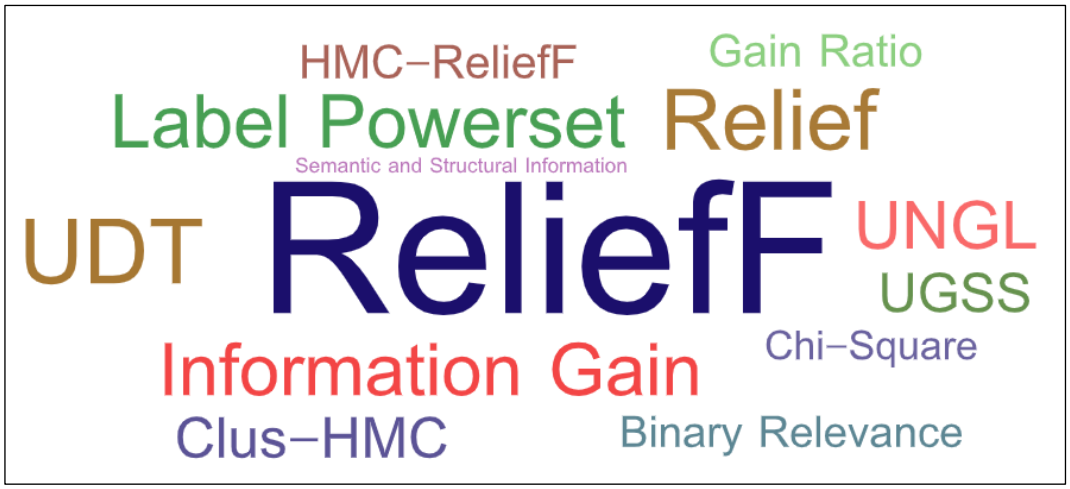
Fonte: Autoria Própria (2022)

A prevalência da abordagem filtro para a seleção de atributos pode ser justificada pelo fato de se tratar de um tipo de técnica independente de algoritmo de aprendizagem, ou seja, a seleção acontece previamente à etapa de indução do classificador. Neste caso, a seleção dos melhores atributos considera apenas as características dos próprios dados, sendo, portanto, computacionalmente eficiente.

Além disso, há, relativamente, poucos classificadores hierárquicos multirrótulo disponíveis na literatura para que se possa utilizar as abordagens *wrapper* e embutida.

Na Figura 15, tem-se uma nuvem de palavras com os termos referentes às técnicas de seleção de atributos identificadas nos trabalhos selecionados.

Figura 15 - Nuvem de palavras das técnicas de seleção de atributos utilizadas



Fonte: Autoria Própria (2022)

No que concerne às técnicas utilizadas, identificou-se que são empregadas tanto técnicas de transformação para contextos multirrótulo quanto técnicas tradicionais da seleção, como, por exemplo, medidas de avaliação de qualidade dos atributos. Referente às técnicas de transformação, foram identificadas *Binary Relevance* e *Label Powerset*, utilizadas em 2 trabalhos de forma combinada com técnicas para ranqueamento de atributos. Das técnicas de ranqueamento de atributos, sobressai-se a técnica *ReliefF*, tendo sido adotada em 3 trabalhos, sendo que em 2 trabalhos foi utilizada de forma combinada com outras técnicas. Considerando que a medida *ReliefF* é uma variação da técnica *Relief*, pode-se dizer que 50% dos trabalhos selecionados adotaram esta técnica para a seleção de atributos. Cabe ressaltar que nos trabalhos 7 e 8 foram empregadas diversas combinando técnicas de transformação com técnicas de ranqueamento.

Q_2: Quais foram as áreas em que a seleção de atributos foi aplicada? Qual o formato da estrutura hierárquica das classes? (Árvore ou DAG)

Nota-se uma distribuição dos estudos analisados em três diferentes áreas: bioinformática, classificação de texto e processamento de imagem. A área de bioinformática foi abordada em 6 trabalhos, sendo a mais frequente entre os estudos

analisados, seguida pela área de processamento de imagem, tratada em 4 estudos. É importante observar que ambas as áreas foram abordadas simultaneamente em 3 trabalhos. A classificação de texto é abordada em 3 estudos. Estas informações são resumidas no Quadro 7, onde também é apresentada a estrutura hierárquica utilizada para organização dos dados.

Quadro 7 - Área de aplicação e tipo de hierarquia utilizada

ID	Área	Hierarquia
1	Bioinformática Processamento de imagem	DAG Árvore
2	Processamento de imagem	Árvore
3	Bioinformática	DAG
4	Bioinformática Processamento de Imagem	DAG Árvore
5	Bioinformática Classificação de texto	(DAG) Árvore Árvore
6	Bioinformática Processamento de Imagem	DAG Árvore
7	Classificação de Texto	Árvore
8	Bioinformática	Árvore

Fonte: Autoria Própria (2022)

Nota-se que a hierarquia mais abordada é a do tipo árvore, utilizada em 7 dos 8 trabalhos. A estrutura do tipo DAG é considerada em 5 trabalhos, todos da área de bioinformática. Referente ao artigo 5, embora o conjunto de dados seja hierarquicamente organizado na forma de um DAG, houve simplificação para uma estrutura do tipo árvore, exigência da ferramenta de classificação utilizada no estudo.

$\boldsymbol{Q_3}$*: Qual foi a contribuição científica dos autores nos trabalhos para o problema da seleção de atributos?*

Esta questão tem por propósito identificar se os estudos analisados propuseram a criação de um novo método para a tarefa da seleção de atributos, a adaptação de métodos existentes ou a realização de estudos experimentais com métodos existentes. O Quadro 8 resume as contribuições de cada um dos trabalhos analisados.

No trabalho 1, foi apresentado por Slavkov *et al*. (2014) um novo método de seleção de atributos para HMC (HMC-*ReliefF*), consistindo numa adaptação do algoritmo *ReliefF* para o contexto hierárquico multirrótulo. Este novo método consegue identificar os recursos mais expressivos no conjunto de dados, além de ter a

capacidade de lidar com a hierarquia de classes sem a necessidade de decompor o problema em vários problemas de classificação plana.

Quadro 8 - Tipos de contribuições dos trabalhos analisados

Contribuição	ID
Novo método de seleção de atributos	1, 2, 4 e 6
Adaptação de métodos existentes	7 e 8
Estudos experimentais com métodos existentes	3 e 5

Fonte: Autoria Própria (2022)

Yan e Wong (2017), no artigo 2, propõem uma nova abordagem para a HMC em dados textuais. Tal abordagem é composta por 3 etapas, destacando-se a etapa de representação dos recursos. Para esta etapa é sugerido um novo método de seleção de atributos, que visa selecionar os atributos mais discriminativos em relação a cada rótulo. Este método adota o aprimoramento de três técnicas existentes: IG, GSS *Coefficient* e NGL *Coefficient*.

No trabalho 4, Slavkov *et al*. (2018) realizam uma extensão do estudo conduzido por Slavkov *et al*. (2014), onde são apresentados resultados de estudos experimentais do algoritmo HMC-*ReliefF*, não sendo proposta nenhuma extensão ou modificação no método.

No artigo 6, Huang e Liu (2020) apresentam um framework baseado em rótulos de informação semântica e estrutural: *Feature Selection based on Semantic and Structural Information of Labels* (FSSS). A proposta é de um método que utilize informação semântica e estrutural dos rótulos na etapa de seleção de atributos. O procedimento consiste no cálculo de similaridade entre rótulos como regularização semântica e na extração das relações pai-filho e irmãos como regularização estrutural. Essas informações são passadas para um modelo de aprendizagem criado para a seleção de recursos.

No trabalho 7, Aljedani, Alotaibi e Taileb (2021) propõem um modelo para a HMC de textos escritos na língua árabe. O modelo proposto incorpora um método de seleção de atributos para redução da quantidade total de atributos resultantes das etapas de preparação de dados e pré-processamento. Os métodos avaliados correspondem a combinações das técnicas de BR e LP com as técnicas *Chi-Sqaure*, GR, *RelieF* e IG. Além do modelo de classificação proposto, a avaliação do impacto

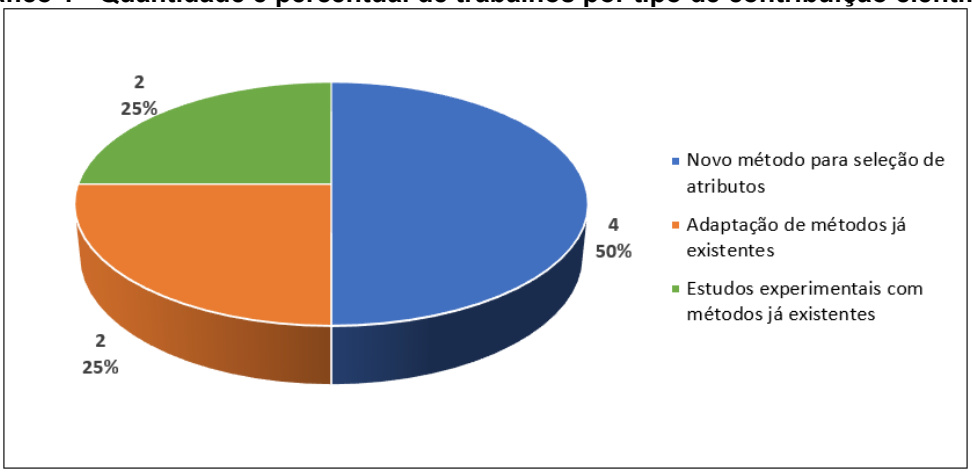
dos métodos de seleção de atributos e das dimensões do espaço de atributos sobre o modelo proposto corresponde é uma contribuição do trabalho.

No artigo 8, Silva e Cerri (2021) propõem e avaliam quatro estratégias para aplicação de métodos de seleção de atributos na HMC. Tais estratégias combinam as técnicas de BR e LP com as técnicas *ReliefF* e Ganho de Informação. Estas são utilizadas para avaliar a importância dos atributos e aquelas correspondem a técnicas de transformação multirrótulo. As quatro estratégias são aplicadas em cada nível da hierarquia de classes e cada nível é considerado como um problema multirrótulo não hierárquico. Deste modo, os atributos selecionados em cada nível são combinados para compor um novo conjunto hierárquico de dados. A principal contribuição do trabalho é a avaliação da capacidade de seleção de atributos de cada uma das estratégias propostas.

Cerri *et al*. (2018), adotando a abordagem *wrapper*, propõem, no trabalho 3, o uso do *Clus*-HMC para a seleção de atributos na HMC. Para a validação dos experimentos, fez uso dos classificadores *Hierarchical Multi-Label Classification with Local Multi-Layer Perceptron* (HMC-LMLP) e *Hierarchical Multi-Label Classification with Genetic Algorithm* (HMC-GA). Trata-se de dois classificadores não lineares baseados em RNA e algoritmos genéticos, respectivamente.

No trabalho 5, Melo e Paulheim (2019) realizam uma comparação sistemática entre as abordagens local e global de seleção de atributos para classificação hierárquica plana e multirrótulo baseada na abordagem de relevância binária.

Gráfico 1 - Quantidade e percentual de trabalhos por tipo de contribuição científica



Fonte: Autoria Própria (2022)

O Gráfico 1 resume as informações referentes às contribuições científicas dos artigos analisados. Dos estudos avaliados, quatro apresentam como principal contribuição um novo método para seleção de atributos, o que corresponde a 50% dos trabalhos analisados. Dois trabalhos propõem adaptações de métodos de seleção de atributos (25%) e outros dois realizam estudos experimentais com métodos já existentes (25%).

***Q_4**: O resultado obtido pelo conjunto de dados reduzido foi relevante quando comparado com o conjunto de dados formado por todos os atributos?*

Esta questão tem por propósito identificar se os trabalhos analisados apresentaram resultados relevantes para a tarefa de classificação hierárquica multirrótulo com a utilização de métodos de seleção de atributos. Foi analisado se os resultados obtidos para a tarefa de classificação com os dados reduzidos foram iguais ou superiores aos resultados para a mesma tarefa com os dados originais. Para os trabalhos em que não foi possível coletar esta informação, analisou-se os resultados do ponto de vista da comparação com trabalhos relacionados e cujos resultados foram citados nos artigos.

O Quadro 1 mostra, de forma sintética, as respostas à questão de pesquisa Q_4. A resposta sim indica que os experimentos indicaram que os resultados obtidos para os conjuntos de dados reduzidos foram relevantes em todos os conjuntos de dados. A resposta parcialmente revela que os resultados obtidos foram relevantes para alguns dos conjuntos de dados utilizados nos experimentos. A resposta não demonstra que os resultados não foram relevantes para nenhum dos conjuntos de dados adotados para os testes experimentais.

Quadro 9 - Resultados obtidos nos trabalhos analisados

O resultado obtido pelo conjunto de dados reduzido foi relevante quando comparado com o conjunto de dados formado por todos os atributos?	**ID**
Sim	1, 2, 4, 6 e 7
Parcialmente	3, 5 e 8
Não	-

Fonte: Autoria Própria (2022)

Os testes no trabalho 1 (SLAVKOV *et al*., 2014) foram conduzidos sobre dois conjuntos de dados em dois importantes domínios para a HMC: genômica funcional e anotação de imagem. Os resultados apresentados indicam melhor resultado na

classificação de imagem, embora demonstrem que o HCM-*ReliefF* identifica corretamente elementos de ambos os domínios.

No artigo 2 (YAN; WONG, 2017), os resultados experimentais provaram que a abordagem para a seleção de atributos realizou com sucesso a tarefa proposta, reduzindo o espaço de atributos, preservando aqueles mais informativos e filtrando os ruídos, além de diminuir a dispersão no conjunto de dados. Os resultados demonstraram uma boa taxa de redução do número de atributos, melhorando a eficácia da predição e o desempenho.

No trabalho 4 (SLAVKOV *et al*., 2018), os resultados apontam que o algoritmo realiza o processo de classificação de atributos com boa estabilidade, sendo esse resultado melhorado com o aumento do número de instâncias. Entretanto, os testes mostram que o algoritmo não é muito sensível ao tamanho da vizinhança, sendo estável já com 25 vizinhos. Além disso, os resultados dos testes experimentais mostraram que, para a maioria dos conjuntos de dados analisados, as classificações do algoritmo HCM-*ReliefF* foram melhores do que o método com o qual foi comparado (relevância binária).

No trabalho 6, Huang e Liu (2020) demonstraram, por meio de testes experimentais, que as técnicas utilizadas mostraram-se mais eficazes do que outros métodos de seleção de atributos utilizados no âmbito da classificação hierárquica.

Nos experimentos realizados no trabalho 7 (ALJEDANI; ALOTAIBI; TAILEB, 2021), cada combinação de técnicas foi utilizada para selecionar 2 mil atributos num espaço de 11 mil atributos resultante das etapas anteriores do modelo. Os resultados obtidos demonstraram que combinação BR-χ^2 obteve melhores resultados. Além disso, avaliou-se a influência do tamanho da dimensionalidade na classificação, constatando-se que a seleção de 4 mil atributos oferece melhor desempenho. O modelo proposto mostrou-se significativamente melhor que os outros modelos avaliados numa ampla gama de métricas de avaliação.

Os resultados apresentados por Cerri *et al*. (2018), no artigo 3, indicam que a seleção de atributos realizada com o algoritmo Clus-HMC mostra-se mais adequada para a classificação hierárquica multirrótulo do que a aplicação de métodos multirrótulo já conhecidos na literatura, mas que não consideram a natureza hierárquica da estrutura de classes. Os testes realizados com a abordagem HMC-LMLP, baseada em RNA, apresentaram melhores resultados quando o algoritmo foi

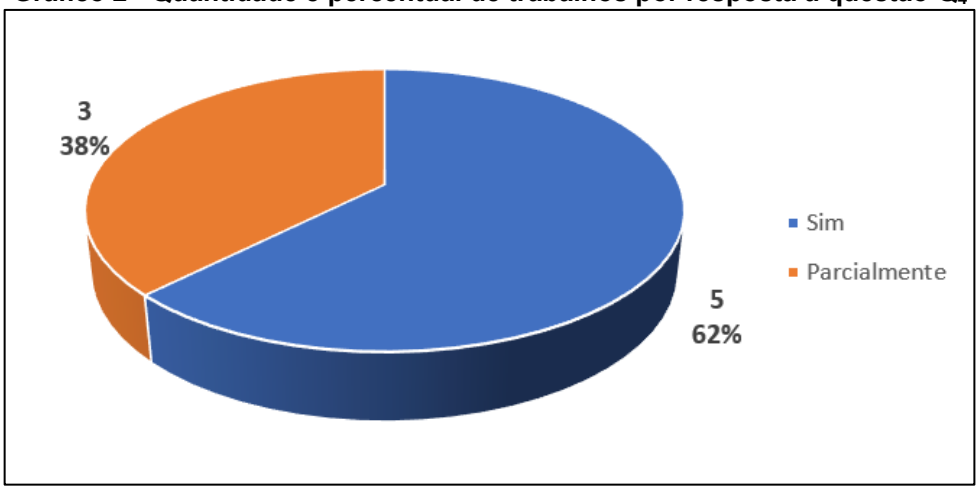
alimentado com o conjunto original de atributos. Já na abordagem genética, HMC-GA, os resultados são melhorados quando a seleção de atributos é feita com o Clus-HMC.

Os resultados comparativos em Melo e Paulheim (2019), artigo 5, demonstram que a abordagem de seleção de atributos local é melhor que a seleção de atributos globais em termos da medida de precisão da classificação, sem prejudicar o desempenho e o tempo de execução.

Os resultados dos experimentos realizados no artigo 8 (SILVA; CERRI, 2021) demonstraram que três das quatros estratégias propostas (IG-BR, RF-BR e RF-LP) conseguiram selecionar subconjuntos relevantes de atributos, de modo que com a redução do espaço de atributos a capacidade preditiva do classificador foi mantida ou melhorada.

O Gráfico 2 mostra o quantitativo e o percentual de trabalhos cujos resultados dos experimentos com os conjuntos de dados reduzidos tenham sido relevantes em relação aos conjuntos de dados originais. Em 62% dos trabalhos os resultados obtidos em testes experimentais foram considerados relevantes quando aplicado o método de seleção de atributos. Nos 38% restantes, os resultados foram parcialmente relevantes, isto é, apenas para alguns conjuntos de dados de teste foram obtidos bons resultados.

Gráfico 2 - Quantidade e percentual de trabalhos por resposta à questão Q_4



Fonte: Autoria Própria (2022)

4.5 Considerações finais do capítulo

Neste capítulo foi apresentado o estado da arte para a problema de pesquisa abordado no trabalho, o que foi feito por meio de um mapeamento sistemático de

literatura, no qual foram identificados e avaliados artigos disponíveis e relevantes no contexto da seleção de atributos para classificação hierárquica multirrótulo.

A análise dos estudos selecionados revelou que, em sua maioria, os trabalhos propõem novos métodos para seleção de atributos no contexto da classificação hierárquica e multirrótulo. Cabe destacar que a quantidade de estudos é ainda pequena. Deste modo, há espaço para a investigar a possibilidade de uso de outras técnicas para reduzir a dimensionalidade dos dados no contexto deste trabalho, de modo a melhorar o desempenho da classificação hierárquica multirrótulo.

Verificou-se, ainda que, em sua maioria, os trabalhos adotaram a abordagem filtro univariado, sendo utilizadas diferentes técnicas de ordenação, como IG, GR, *Chi-Square* e *Relief*. Além disso, identificou-se que a utilização de técnicas de transformação para conjuntos de dados multirrótulo também podem apresentar bons resultados para a seleção de atributos em contextos de classificação hierárquica multirrótulo, destacando-se as técnicas BR e LP.

Observou-se que em nenhum dos trabalhos revisados houve a utilização do *Fisher Score* como medida para avaliar qualidade dos atributos. Deste modo, uma vez que o FS é comumente utilizado para a seleção de atributos em contextos de classificação tradicional, sendo considerada uma técnica de baixo custo computacional e de implementação simples, este trabalho investigou a possibilidade de estender a definição de tal métrica na tarefa de seleção de atributos para a HMC. Como resultado de tal investigação, um novo método para seleção de atributos foi proposto. Este método é apresentado no capítulo 0.

5 MÉTODO PARA SELEÇÃO DE ATRIBUTOS: *FEATURE SELECTION BASED ON FISHER SCORE FOR HIERARCHICAL MULTI-LABEL CLASSIFICATION* (FSF-HMC)

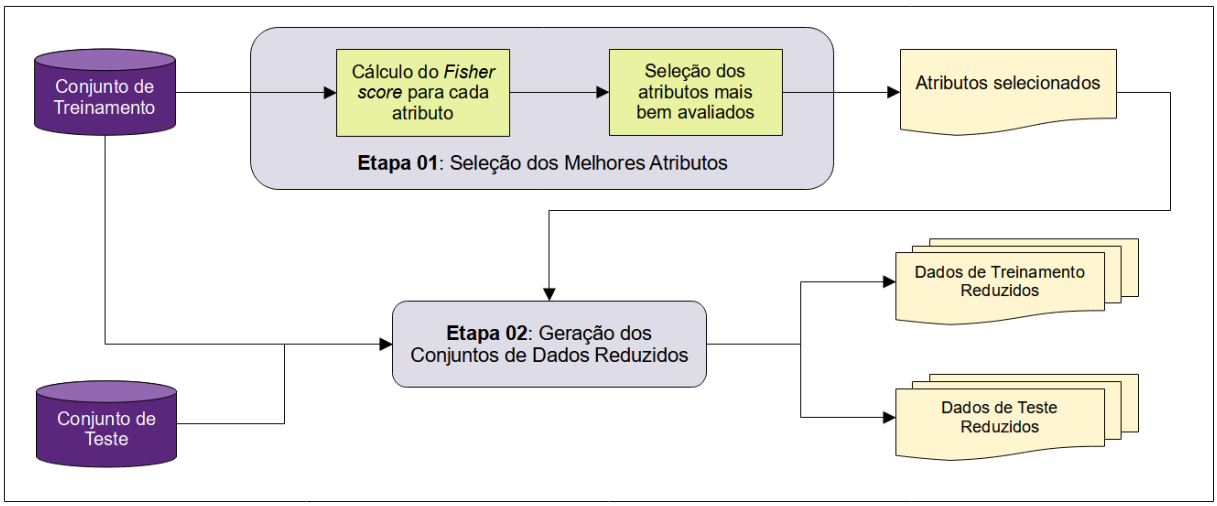
Este Capítulo apresenta o método proposto para seleção de atributos no contexto da classificação hierárquica multirrótulo. A Seção 5.1 é dedicada à descrição geral do método proposto. A Seção 5.2 mostra uma simulação do método. Por fim, na Seção 5.3 são apresentadas as considerações finais do capítulo.

5.1 Visão geral do método FSF-HMC

O *Feature Selection based on Fisher score for Hierarchical Multi-Label Classification* (FSF-HMC) é um método filtro univariado para seleção de atributos em bases de dados de HMC, que se enquadra na categoria técnicas de ordenação. Portanto, avalia individualmente cada atributo através de uma medida para determinar quão bom um atributo é. Para isso, o método FSF-HMC adota a métrica *Fisher Score*. Quanto maior é o valor da medida, melhor é o atributo, o que significa dizer que quanto melhor for o atributo, maior é a sua capacidade de discriminação entre diferentes classes. No contexto deste trabalho, o cálculo da medida FS foi adaptado para considerar a organização hierárquica das classes no cálculo do valor. As adaptações realizadas são detalhadas adiante.

A Figura 16 apresenta uma visão geral do método FSF-HMC. O método possui duas etapas: seleção dos melhores atributos e geração dos conjuntos de dados reduzidos. A etapa seleção de atributos é composta por duas atividades, que são executadas sobre o conjunto de dados de treinamento original: o cálculo da medida FS para cada atributo e a seleção dos atributos mais bem avaliados. Na etapa geração dos conjuntos de dados reduzidos, são produzidos novos conjuntos de treinamento e teste a partir dos conjuntos de dados originais e da lista de atributos selecionados que foi gerada na etapa anterior.

Figura 16 - Etapas do método FSF-HMC para seleção de atributos



Fonte: Autoria Própria (2022)

A seguir, é apresentada uma descrição formal do cenário de aplicação do método FSF-HMC e, na sequência, o próprio método é definido por meio de tal formalismo. Seja $D = \{E_1, E_2, \dots, E_n\}$ uma base de dados de treinamento com n instâncias, conforme ilustrado na Figura 4 e cujas classes estão organizadas hierarquicamente, segundo foi definido na Seção 2.3 e na Seção 2.4, onde $E_i = (X_i, Y_i)$, com $i = 1, \dots n$. Cada instância E_i é composta por dois conjuntos, X_i e Y_i, que correspondem, respectivamente, aos valores dos atributos e aos rótulos de classe aos quais a instância está associada, de modo que $X_i = (x_{i1}, x_{i2}, \dots, x_{im})$ e $Y_i \subseteq Y = \{y_j \mid j = 1, \dots, q\}$, onde m é quantidade de atributos na base de dados, q é a quantidade de rótulos de classe possíveis e $Y \subseteq L = \{y_1, y_2, \dots, y_q\}$, em que L é o conjunto de todos os possíveis rótulos de classes. O conjunto de atributos na base de dados é $X = \{x_1, x_2, \dots x_m\}$, onde um atributo x_k é definido por $x_k = (x_{1k}, x_{2k}, \dots, x_{nk})^T$ e x_{ik} $(i = 1, 2, \dots, n)$ é o valor que o atributo x_k assume na i-ésima instância.

Dado um conjunto de atributos X, onde $|X| = m$, o método FSF-HMC tem por objetivo encontrar o menor subconjunto $X^R \subseteq X$, de dimensão $d < m$, que satisfaça ao seguinte critério de avaliação: o valor do FS de cada atributo em X^R é maior do que o valor médio de FS no conjunto X. A partir do subconjunto de atributos X^R e dos conjuntos de treinamento (D) e teste originais (T), são produzidos os conjuntos de treinamento e teste reduzidos, indicados por D^R e T^R, respectivamente. As etapas deste procedimento são descritas na Subseção 5.1.1 e na Subseção 5.1.2.

5.1.1 Seleção dos melhores atributos

A primeira etapa do método FSF-HMC tem por objetivo gerar o conjunto X^R a partir do conjunto de treinamento D. Para isto, são realizadas duas atividades, conforme mostrado na Figura 16. A primeira delas consiste em calcular, para cada atributo $x_k \in X$, um valor de avaliação $FS(x_k)$ por meio da medida FS. A outra atividade corresponde à geração do conjunto $X^R \subseteq X$ com os d atributos mais bem avaliados.

O cálculo do FS para cada atributo x_k da base de dados D deve considerar a natureza da tarefa de classificação como hierárquica e multirrótulo. Como uma instância E_i está associada a um conjunto de classes $Y_i \subseteq Y$, é preciso considerar cada uma das classes $y_j \in Y_i$ no processo de cálculo da medida de avaliação de cada um dos atributos. Para lidar com essa particularidade, o método FSF-HMC trata o problema multirrótulo como um problema monorrótulo. Para tanto, cada instância E_i é considerada em r vezes no processo de cálculo da medida *Fisher Score*, onde $r = |Y_i|$. Por exemplo, na base de dados fictícia da Tabela 1, a instância E_1 está associada ao conjunto de classes $Y_1 = CC@EE@II = \{CC, EE, II\}$. Neste caso, E_1 é considerada três vezes no processo, uma vez para cada classe a que está associada. Este procedimento confere ao método FSF-HMC a capacidade de lidar tanto com problemas monorrótulo quanto com problemas multirrótulo.

Além disso, como o cálculo do FS para cada atributo x_k requer a média dos valores desse atributo em relação a cada classe a que ele está associado, e como as classes são organizadas segundo uma hierarquia, a aplicação da Equação exige algumas adaptações para considerar a estrutura hierárquica das classes. A estrutura hierárquica da j-ésima classe é dada pelo conjunto $H_j = \{y_H | \, y_H = y_j \text{ ou } y_H \text{ é } ancestral\ de\ y_j\}$. Neste conjunto, o nível hierárquico da j-ésima classe é 0. O nível hierárquico de cada ancestral direto da classe y_j é igual a 1 e assim sucessivamente, adicionando-se 1 para cada novo nível da hierarquia. Por exemplo, para a base de dados fictícia da Tabela 1, estrutura hierárquica da classe $y_8 = HH$, de acordo com a Figura 10, é $H_8 = \{HH(\text{nível } 0), DD(\text{nível } 1), AA(\text{nível } 2)\}$. O método FSF-HMC é capaz de lidar tanto com hierarquias de classes estruturadas como árvore quanto com hierarquias organizadas como DAG e adota a política do menor caminho para calcular os níveis hierárquicos das classes.

Pela Equação (22), o cálculo de FS para o atributo x_k em relação à classe y_j, requer que sejam determinados o número de exemplos da classe y_j (n_j), a média dos valores de x_k em relação à classe y_j (μ_k^j), a média dos valores de x_k em relação a todo o conjunto D (μ_k) e $\sum_{j=1}^{q} S_t^j(x_k)$, conforme definido na Seção 3.3.5. A primeira adaptação necessária para se considerar a hierarquia de classes no cálculo do *Fisher Score* diz respeito ao valor de n_j e a segunda refere-se ao cálculo de μ_k^j, conforme detalhado nos parágrafos seguintes.

O valor de n_j na Equação (22) é dado por $n_j = |D_j|$, onde é o conjunto de todas as instâncias de dados associadas à classe y_j e às classes ancestrais de y_j. Por exemplo, na base de dados fictícia da Tabela 1, considerando a instância E_3 e a classe associada $y_{10} = JJ$, o conjunto de instâncias associadas à classe $y_{10} = JJ$ e às suas classes ancestrais é dado por $D_{10} = \{E_1, E_2, E_3, E_4, E_5\}$, pois $H_{10} = \{JJ(\text{nível } 0), FF(\text{nível } 1), GG(\text{nível } 1), BB(\text{nível } 2), CC(\text{nível } 2)\}$. As instâncias E_2 e E_4 estão associadas diretamente à classe BB e as instâncias E_1 e E_5 estão associadas diretamente à classe CC. Neste exemplo, não há instâncias associadas diretamente às classes FF e GG.

O valor de μ_k^j, que na Equação (22) é a média aritmética simples dos valores do atributo x_k em relação à j-ésima classe, foi adaptado para ser a média ponderada dos valores de x_k em D_j, calculada pela Equação (24). Isto indica que os valores do atributo x_k em relação às classes associadas a ele, em cada nível da hierarquia, colaboram para o cômputo do valor de sua medida FS. Quanto maior é o nível hierárquico da classe, menor é a contribuição do valor do atributo quando associado a ela.

$$\mu_k^j = \frac{\sum_{i=1}^{n_j} p_i \cdot x_{ik}^j}{\sum_{i=1}^{n_j} p_i} \tag{24}$$

onde,

- p_i é o peso para a i-ésima instância no conjunto D_j e $p_i = 1 - \frac{h_i}{max_h + 1}$, onde h_i corresponde ao nível hierárquico da classe à qual o exemplo está

associado em H_j e max_h é o maior nível hierárquico no conjunto de classes.

- x_{ik}^{j} é o valor do k-ésimo atributo para a i-ésima instância na j-ésima classe.

A média dos valores do atributo x_k em relação a todo o conjunto de dados D é a média aritmética simples dada pela Equação (25), onde $|D|$ é a quantidade de exemplos no conjunto D e x_{ik} é o valor do k-ésimo atributo para a i-ésima instância.

$$\mu_k = \frac{\sum_{i=1}^{|D|} x_{ik}}{|D|} \tag{25}$$

O valor de $\sum_{j=1}^{q} S_t^j(x_k)$, conforme já mencionado, é calculado como definido na Subseção 3.3.5, onde $S_t^j(x_k) = \sum_{k=1}^{n_j}\left(x_{ik}^j - \mu_k^j\right)^2$ é a matriz de dispersão da classe do k-ésimo atributo em relação à j-ésima classe e x_{ik}^{j} é o valor do k-ésimo atributo para a i-ésima instância na j-ésima classe.

O Algoritmo 1 descreve o passo a passo do cálculo do *Fisher Score* para um atributo x_k no conjunto de dados de treinamento D. As entradas do algoritmo são o atributo x_k, o conjunto de dados de treinamento D, o conjunto dos rótulos de classe associados às instâncias em D e a hierarquia de classes H. O conjunto dos rótulos de classe associados às instâncias em D, dado por $L_D \subseteq L$, corresponde às classes que possuem pelo menos uma instância associada. A saída do algoritmo é o valor da medida FS para o atributo x_k, isto é $FS(x_k)$.

Para o cálculo de FS do atributo x_k, inicialmente é calculada a média dos valores deste atributo para todo o conjunto de dados D conforme Equação (25) (linha 1). Em seguida, para cada classe $y_j \in L_D$, determina-se o conjunto H_j, isto é, sua estrutura hierárquica composta pela classe y_j, cujo nível é 0, e por suas classes ascendentes com seus respectivos níveis (linha 4). Na sequência, define-se o conjunto D_j de instâncias da classe y_j e de suas classes ascendentes (linha 5) e calcula-se a quantidade de exemplos n_j (linha 6). Posteriormente, é calculada a média ponderada dos valores de x_k em D_j, conforme Equação (24) (linha 7) e o valor de $S_t^j(x_k)$ de acordo com a Equação (23) (linha 8). Adiante, é calculada a soma dos

valores de $S_t^j(x_k)$ (linha 9). Por fim, conforme Equação (22), o valor de FS para o atributo x_k é calculado (linha 10) e, em seguida, retornado.

Algoritmo 1. *FisherScore*

Entrada: x_k, D, L_D, H
Saída: $FS(x_k)$
1: calcular μ_k conforme Equação 25
2: calcular max_h para o conjunto de classes L_D
3: **para cada** y_j em L_D **faça**
4: determinar o conjunto H_j da estrutura hierárquica para a classe y_j
5: determinar o conjunto D_j de instâncias da classe y_j
6: fazer $n_j = |H_j|$
7: calcular μ_k^j conforme Equação 24
8: calcular $S_t^j(x_k)$ conforme Equação 23
9: calcular $\sum_{j=1}^{q} S_t^k(x_k)$
10: calcular $FS(x_k)$ conforme Equação 22
11: **retorna** $FS(x_k)$

Após o cálculo de FS para cada atributo x_k em D, procede-se à seleção dos atributos mais bem avaliados para compor o conjunto $X^R \subseteq X$. O Algoritmo 2 estabelece o passo a passo para a seleção dos melhores atributos por meio do método FS-MHC, onde a entrada é o conjunto D de dados de treinamento original, o conjunto X dos atributos em D, o conjunto L_D de classes em D e a hierarquia das classes H. A saída X^R corresponde ao conjunto de atributos reduzido.

Algoritmo 2. *SelecionaAtributos*

Entrada: D, X, L_D, H
Saída: X^R
1: **para cada** x_k em X **faça**
2: calcular FisherScore(x_i, T, C, H) e adicionar ao conjunto F
3: calcular a média (mfs) dos valores de Fisher Score em F
4: **para cada** $F(x_k)$ em F **faça**
5: **se** $FS(x_k) > mfs$ **então**
6: adicionar x_k a X^R
7: **senão**
8: ignorar x_k
9: **retorna** X^R

Inicialmente, para cada atributo $x_k \in X$, o valor de seu *Fisher Score* é calculado e adicionado a um conjunto F (linhas 1 e 2). O método FSF-HMC seleciona os atributos cujo valor do *Fisher Score* seja superior à média dos valores de *Fisher*

Score calculados para cada atributo x_k. Neste caso, calcula-se a média aritmética dos valores de *Fisher Score* em F e, em seguida, adicionam-se no conjunto X^R apenas os atributos cujo *score* seja superior à média calculada (linhas 3 a 8).

5.1.2 Geração dos conjuntos de dados reduzidos

A segunda etapa do método FSF-HMC tem por objetivo produzir os conjuntos de dados de treinamento e de teste reduzidos, D^R e T^R, respectivamente. O Algoritmo 3 descreve o procedimento para geração desses conjuntos. A entrada é o conjunto de atributos selecionados X^R, gerado na etapa anterior, e os conjuntos de dados de treinamento e teste originais, D e T, respectivamente. E a saída são os conjuntos D^R e T^R.

Algoritmo 3. *GeraConjuntosReduzidos*

Entrada: $\boldsymbol{D, T, X^R}$
Saída: $\boldsymbol{D^R, T^R}$
1: **para cada** $\boldsymbol{x_k}$ em $\boldsymbol{X^R}$ **faça**
2: adicionar $\boldsymbol{D(x_k)}$ a $\boldsymbol{D^R}$
3: adicionar $\boldsymbol{T(x_k)}$ a $\boldsymbol{T^R}$
4: **retorna** $\boldsymbol{D^R, X^R}$

Nas linhas 2 e 3 $D(x_k)$ e $T(x_4)$ correspondem à inserção do atributo $x_k = (x_{1k}, x_{2k}, \dots, x_{nk})^T$ nos conjuntos D^R e T^R, onde n é o número de instâncias nos conjuntos de dados originais.

5.2 Aplicação do método FSF-HMC numa base de dados fictícia

Para ilustrar a aplicação do método FSF-HMC, seja considerada a base de dados fictícia da Tabela 1, cuja hierarquia de classes H está representada na Figura 10. Inicialmente, é possível definir os seguintes conjuntos: D, o conjunto das instâncias de treinamento; X, o conjunto de atributos; e L_D, o conjunto das classes associadas às instâncias em D.

$$D = \{E_1, E_2, E_3, E_4, E_5, E_6\}$$
$$X = \{x_1, x_2, x_3, x_4, x_5, x_6, x_7, x_8, x_9\}$$

$$L_D = \{BB, CC, EE, HH, II, JJ\}$$

O conjunto H, conforme apresentado na Subseção 2.4.1, pode ser representado da seguinte forma:

$$H = \{RAIZ/AA, RAIZ/BB, RAIZ/CC, AA/DD, AA/EE, AA/II, BB/EE, BB/FF, CC/GG, DD/HH, DD/II, FF/JJ, GG/JJ, HH/KK, II/KK, JJ/KK\}$$

É possível, ainda, detalhar os elementos dos conjuntos D e X. Para simplificar, é apresentado apenas um exemplo, visto que os dados podem ser acessados na Tabela 1, onde os elementos do conjunto D correspondem às linhas e os elementos do conjunto X, às colunas.

$$E_1 = (0{,}45; 0{,}42; 0{,}56; 0{,}92; 0{,}95; 0{,}34; 0{,}21; 0{,}88; 0{,}16; CC@EE@II)$$
$$x_1 = (0{,}45; 0{,}84; 0{,}27; 0{,}85; 0{,}26; 0{,}36)^T$$

A primeira etapa do método, de acordo com o Algoritmo 2, consiste em, dados os conjuntos D, X, L_D e H, calcular o FS para cada um dos atributos $x_k \in X$. A seguir, é apresentado o passo a passo desse cálculo para o atributo x_1, conforme o Algoritmo 1.

Inicialmente, é calculada a média aritmética μ_1 dos valores de x_1 para todo o conjunto de dados D. Neste caso,

$$x_1 = \{0{,}45;\ 0{,}84;\ 0{,}27;\ 0{,}85;\ 0{,}26; 0{,}36\}$$
$$\mu_1 = 0{,}505$$

Em seguida, é calculado o nível hierárquico máximo max_h para as classes em L_D, de acordo a hierarquia de classes H, ilustrada na Figura 10. Neste caso, $max\_h = 2$, relativo às classes HH e JJ.

Na sequência, para cada uma das classes y_j em $L_D = \{BB, CC, EE, HH, II, JJ\}$, determina-se sua hierarquia (conjunto H_j). Depois disso, calcula-se o subconjunto D_j de exemplos cujos rótulos de classe contenham y_j (e suas classes ascendentes) e o tamanho do subconjunto D_j (n_j). Posteriormente, calcula-se a média ponderada dos

valores de x_1 em D_j (μ_1^j) conforme Equação (24). Na sequência, computa-se o valor de $n_j \cdot \left(\mu_1^j - \mu_1\right)^2$ e o valor de $S_t^j(x_1)$ pela Equação (23).

Para a classe $y_1 = BB$, a estrutura hierárquica, obtida a partir do conjunto H, é a seguinte:

$$H_1 = \{BB\ (nível\ 0), RAIZ\}$$

Neste caso, como o conjunto H_1 possui apenas a classe $BB\ (nível\ 0)$ e a $RAIZ$ (a $RAIZ$ é desconsiderada para fins da seleção de atributos), o conjunto das instâncias relacionadas à classe BB e a seus ancestrais é dado por:

$$D_1 = \{E_2, E_4\}$$

Como D_1 possui dois elementos, $n_1 = 2$. Pela Tabela 1, os valores do atributo x_1 para os exemplos E_2 e E_4 são $0{,}84$ e $0{,}85$, respectivamente. Com isso, a média ponderada dos valores do atributo x_1 considerando a classe BB, calculada pela Equação (24), é

$$\mu_1^1 = \frac{1 \cdot 0{,}84 + 1 \cdot 0{,}85}{1 + 1} = \frac{1{,}69}{2} = 0{,}845$$

O peso p_i para o valor do atributo x_1 no i-ésimo exemplo é dado por $p_i = 1 - \frac{h_i}{max_h + 1}$, onde h_i corresponde ao nível hierárquico da classe em relação à classe BB. Neste caso, como há dois exemplos e ambos são exemplos da classe BB, cujo nível hierárquico é 0, tem-se que

$$p_1 = 1 - \frac{h_1}{max\_h + 1} = 1 - \frac{0}{3} = 1$$

$$p_2 = 1 - \frac{h_2}{max_h + 1} = 1 - \frac{0}{3} = 1$$

Uma vez que se tenha calculado μ_1^1, computa-se para o atributo x_1, em relação à classe BB, o valor de $n_1 \cdot (\mu_1^1 - \mu_1)^2$, conforme requisitado pela Equação (22). Neste sentido,

$$n_1 \cdot (\mu_1^1 - \mu_1)^2 = 2 \cdot (0{,}845 - 0{,}505)^2 = 0{,}290$$

Por fim, utiliza-se a Equação (23) para calcular $S_t^1(x_1)$ para o atributo x_1 considerando a classe BB. Assim,

$$S_t^1(x_1) = \sum_{k=1}^{2}(x_{1k}^1 - \mu_1^1)^2 = (0{,}84 - 0{,}845)^2 + (0{,}85 - 0{,}845)^2 = 0{,}00005.$$

O mesmo procedimento é realizado para a classe $y_2 = CC$. Neste caso,

$$H_2 = \{CC\ (nível\ 0), RAIZ\}$$
$$D_2 = \{E_1, E_1\}$$
$$n_2 = 2$$

Os valores do atributo x_1 para a classe CC, conforme a Tabela 1, são, respectivamente, $0{,}45$ e $0{,}26$. Neste caso, a média ponderada dos valores do atributo x_1 em relação à classe CC é dada por

$$\mu_1^2 = \frac{1 \cdot 0{,}45 + 1 \cdot 0{,}26}{1 + 1} = \frac{0{,}71}{2} = 0{,}355$$

Os pesos para estes valores são iguais a 1, dado que os exemplos são da classe CC, cujo nível hierárquico é 0.

Tendo sido calculado o valor de μ_1^2, tem-se que

$$n_2 \cdot (\mu_1^2 - \mu_1)^2 = 2 \cdot (0{,}355 - 0{,}505)^2 = 0{,}045$$
$$S_t^2(x_1) = \sum_{k=1}^{2}(x_{1k}^2 - \mu_1^2)^2 = (0{,}45 - 0{,}355)^2 + (0{,}26 - 0{,}355)^2 = 0{,}01805.$$

Para a classe $y_3 = EE$, tem-se que sua estrutura hierárquica, conforme apresentado no conjunto H (ou Figura 10) é dada por

$$H_3 = \{EE\ (nível\ 0), AA\ (nível\ 1), BB(nível\ 1)\}$$

Neste caso, o conjunto de exemplos D_3 contém as instâncias da classe EE e as instâncias de suas classes ancestrais presentes na base de dados da Tabela 1. Assim,

$$D_3 = \{E_1, E_2, E_4\}$$

cujos valores para o atributo x_1 são, respectivamente, $0{,}45$, $0{,}84$ e $0{,}85$. O valor de n_3 é igual a 3, dado que D_3 possui três elementos. O cálculo da média ponderada dos valores de x_1 para a classe EE é dado por

$$\mu_1^3 = \frac{1 \cdot 0{,}45 + 0{,}667 \cdot 0{,}84 + 0{,}667 \cdot 0{,}85}{1 + 0{,}667 + 0{,}667} = \frac{1{,}577}{2{,}333} = 0{,}6757$$

Os pesos calculados para cada um dos valores consideram o nível de hierarquia da classe ao qual o exemplo está associado. Neste caso,

$$p_1 = 1 - \frac{h_1}{max_h + 1} = 1 - \frac{0}{3} = 1$$

$$p_2 = 1 - \frac{h_2}{max_h + 1} = 1 - \frac{1}{3} = 1 - 0{,}333 = 0{,}667$$

$$p_3 = 1 - \frac{h_3}{max_h + 1} = 1 - \frac{1}{3} = 1 - 0{,}333 = 0{,}667$$

O primeiro valor corresponde ao exemplo E_1, cuja classe é EE e seu nível hierárquico é 0. Os demais valores correspondem aos exemplos E_2 e E_3, cuja classe é BB e seu nível hierárquico é 1.

Uma vez que se tenha calculado o valor de μ_1^3, tem-se:

$$n_3 \cdot (\mu_1^3 - \mu_1)^2 = 3 \cdot (0{,}6757 - 0{,}505)^2 = 0{,}0874$$

$S_t^3(x_1) = \sum_{k=1}^{3}(x_{1k}^3 - \mu_1^3)^2 = (0{,}45 - 0{,}6757)^2 + (0{,}84 - 0{,}6757)^2 + (0{,}85 - 0{,}6757)^2 = 0{,}1083$.

Para a classe $y_4 = HH$, a estrutura hierárquica é dada por

$$H_4 = \{HH\ (nível\ 0), DD\ (nível\ 1), AA(nível\ 2)\}$$

Como não há exemplos na base de dados da Tabela 1 para as classes ancestrais da classe HH, o conjunto D_4 resume-se aos exemplos da própria classe HH, isto é,

$$D_4 = \{E_2, E_6\}$$

Os valores de x_1 para os exemplos considerados são, respectivamente, $0{,}84$ e 0,36. Além disso, $n_4 = 2$. A média ponderada dos valores de x_1 para a classe HH é dada por

$$\mu_1^4 = \frac{1 \cdot 0{,}84 + 1 \cdot 0{,}36}{2} = \frac{1{,}2}{2} = 0{,}6000$$

Além disso,

$$n_4 \cdot (\mu_1^4 - \mu_1)^2 = 2 \cdot (0{,}6000 - 0{,}505)^2 = 0{,}0181$$
$$S_t^4(x_1) = \sum_{k=1}^{2}(x_{1k}^4 - \mu_1^4)^2 = (0{,}84 - 0{,}6)^2 + (0{,}36 - 0{,}6)^2 = 0{,}1152.$$

A classe $y_5 = II$, possui, conforme mostrado na Figura 10, a seguinte estrutura hierárquica:

$$H_5 = \{II\ (nível\ 0), DD\ (nível\ 1), FF\ (nível\ 1), AA\ (nível\ 2), \quad BB(nível\ 2)\}$$

Neste caso, de acordo com a Tabela 1,

$$D_5 = \{E_1, E_2, E_4\}$$

cujos valores para x_1 são, respectivamente, $0{,}45$, $0{,}84$ e $0{,}85$. Visto que todos os exemplos são relacionados à classe II, cujo nível hierárquico é 0, os pesos para cada uma das instâncias são

$$p_1 = 1 - \frac{h_1}{max_h + 1} = 1 - \frac{0}{3} = 1$$

$$p_2 = 1 - \frac{h_2}{max_h + 1} = 1 - \frac{0}{3} = 1$$

$$p_3 = 1 - \frac{h_3}{max_h + 1} = 1 - \frac{0}{3} = 1$$

A média ponderada dos valores de x_1 para a classe II é dada por

$$\mu_1^5 = \frac{1 \cdot 0{,}45 + 1 \cdot 0{,}84 + 1 \cdot 0{,}85}{1 + 1 + 1} = \frac{2{,}14}{3} = 0{,}7133$$

Como $n_5 = 3$, tem-se que

$$n_5 \cdot (\mu_1^5 - \mu_1)^2 = 3 \cdot (0{,}7133 - 0{,}505)^2 = 0{,}1302$$

$S_t^5(x_1) = \sum_{k=1}^{3}\left(x_{1k}^5 - \mu_1^5\right)^2 = (0{,}45 - 0{,}7133)^2 + (0{,}84 - 0{,}7133)^2 + (0{,}85 - 0{,}7133)^2 = 0{,}1041.$

Para a classe $y_6 = JJ$, tem-se a seguinte estrutura hierárquica, conforme a Figura 10:

$$H_6 = \{JJ\ (nível\ 0), FF\ (nível\ 1), GG(nível\ 1), BB(nível\ 2), CC(nível\ 2)\}$$

Neste caso,

$$D_6 = \{E_1, E_2, E_3, E_4, E_5, E_6\}$$

cujos valores de x_1 são, respectivamente, $0{,}45$, $0{,}84$, $0{,}27$, $0{,}85$, $0{,}26$ e $0{,}36$. Tem-se, ainda, que $n_6 = 6$.

A média ponderada dos valores de x_1 é dada por

$$\mu_1^6 = \frac{0{,}333 \cdot 0{,}45 + 0{,}333 \cdot 0{,}84 + 1 \cdot 0{,}27 + 0{,}333 \cdot 0{,}85 + 0{,}333 \cdot 0{,}26 + 1 \cdot 0{,}36}{0{,}333 + 0{,}333 + 1 + 0{,}333 + 0{,}333 + 01}$$

$$= \frac{1{,}430}{3{,}333} = 0{,}4290$$

O cálculo dos pesos considera o nível hierárquico das classes relacionadas aos exemplos. Neste caso, os exemplos E_3 e E_6 estão associados à classe JJ cujo nível hierárquico é 0. Os exemplos E_2 e E_4 estão associadas à classe BB cujo nível é 2 e os exemplos E_1 e E_5 pertencem à classe CC, cujo nível é 2. Assim,

$$p_1 = 1 - \frac{h_1}{max_h + 1} = 1 - \frac{2}{3} = 1 - 0{,}667 = 0{,}333$$

$$p_2 = 1 - \frac{h_2}{max_h + 1} = 1 - \frac{2}{3} = 1 - 0{,}667 = 0{,}333$$

$$p_3 = 1 - \frac{h_3}{max_h + 1} = 1 - \frac{0}{3} = 1$$

$$p_4 = 1 - \frac{h_4}{max_h + 1} = 1 - \frac{2}{3} = 1 - 0{,}667 = 0{,}333$$

$$p_5 = 1 - \frac{h_5}{max_h + 1} = 1 - \frac{2}{3} = 1 - 0{,}667 = 0{,}333$$

$$p_6 = 1 - \frac{h_6}{max_h + 1} = 1 - \frac{0}{3} = 1$$

Por fim, tem-se que

$$n_6 \cdot (\mu_1^6 - \mu_1)^2 = 6 \cdot (0{,}4290 - 0{,}505)^2 = 0{,}0173$$

$S_t^6(x_1) = \sum_{k=1}^{6}(x_{1k}^6 - \mu_1^6)^2 = (0{,}27 - 0{,}4290)^2 + (0{,}36 - 0{,}4290)^2 + (0{,}84 - 0{,}4290)^2 + (0{,}85 - 0{,}4290)^2 + (0{,}45 - 0{,}4290)^2 + (0{,}26 - 0{,}4290)^2 = 0{,}4052.$

Uma vez que todos os cálculos referentes a cada classe $c_k \in L_D$ tenham sido realizados, computa-se o valor de $\sum_{j=1}^{q} S_t^j(x_1)$, isto é

$\sum_{j=1}^{6} S_t^j(x_1) = S_t^1(x_1) + S_t^2(x_1) + S_t^3(x_1) + S_t^4(x_1) + S_t^5(x_1) + S_t^6(x_1) = 0{,}00005 + 0{,}01805 + 0{,}1083 + 0{,}1152 + 0{,}1041 + 0{,}4052 = 0{,}7509.$

Tendo sido calculados todos os elementos necessários, calcula-se o valor da medida FS para o atributo x_1 utilizando-se a Equação (22).

$$FS(x_1) = \frac{\sum_{j=1}^{q} n_j \left(\mu_1^j - \mu_1\right)^2}{\sum_{j=1}^{q} S_t^j(x_1)} = \frac{0{,}29 + 0{,}045 + 0{,}0874 + 0{,}0181 + 0{,}1302 + 0{,}0173}{0{,}7509}$$

$$= \frac{0{,}588}{0{,}7509} = 0{,}7831$$

Iterativamente, o processo é realizado para todos os demais atributos. O resultado é mostrado na Tabela 3.

Tabela 3 - Resultados do cálculo do *Fisher Score* para os atributos

x_1	x_2	x_3	x_4	x_5	x_6	x_7	x_8	x_9
0,7831	0,5976	0,0051	0,5501	0,4905	0,2181	0,2660	0,0673	0,2458

Fonte: Autoria própria (2022)

Uma vez que os valores de FS para cada atributo tenham sido calculados, a seleção dos mais bem avaliados se dá a partir da média dos valores de FS calculada sobre todos os atributos. Neste caso, a média calculada a partir dos resultados na Tabela 3 é $0{,}3582$. Este valor é utilizado como limiar para seleção dos atributos, ou seja, aqueles atributos cujo valor de FS for superior ao limiar são selecionados. Portanto, o resultado da seleção de atributos para o contexto apresentado é: x_1, x_2, x_4, x_5.

O conjunto de dados reduzido gerado a partir da lista de atributos selecionados é apresentado na Tabela 4.

Tabela 4 - Base de dados reduzida

Exemplos	x_1	x_2	x_4	x_5	**Classe**
E_1	0,45	0,42	0,92	0,95	CC@EE@II
E_2	0,84	0,37	0,35	0,47	BB@HH@II
E_3	0,27	0,49	0,41	0,83	JJ
E_4	0,85	0,11	0,89	0,69	BB@II
E_5	0,26	0,82	0,95	0,74	CC
E_6	0,36	0,76	0,04	0,11	HH@JJ

Fonte: Autoria Própria (2022)

5.3 Considerações finais do capítulo

Neste capítulo foi apresentado o método FSF-HMC, que realiza a seleção de atributos com base na avaliação da qualidade dos atributos por meio da métrica FS, que foi adaptada para considerar a hierarquia das classes no procedimento de cálculo da pontuação individual para cada atributo.

Uma das vantagens do método FSF-HMC é que ele utiliza a abordagem filtro e, por isso, é capaz de reduzir a dimensão dos atributos de uma base de dados de classificação hierárquica multirrótulo de forma que o conjunto de dados reduzido possa ser utilizado por qualquer classificador. Além disso, o método proporciona, ao algoritmo de aprendizagem, mais rapidez no processamento dos conjuntos de dados.

Uma desvantagem do método proposto é que, pelo fato de selecionar apenas os atributos que estejam acima da média dos valores de FS calculados, a quantidade de atributos selecionados tende a ser fixa para cada base de dados em que o método seja aplicado. Neste caso, é possível que algum atributo relevante possa não estar contido no conjunto de atributos reduzido. Outro ponto que pode se mostrar como uma desvantagem é que a utilização apenas do FS não considera a correlação entre os atributos, o que pode levar o método a desconsiderar a redundância entre os atributos.

O método proposto foi avaliado através da medida AUPRC produzida por dois classificadores hierárquicos multirrótulo, Clus-HMC e MHC-CNN. Os experimentos realizados e os respectivos resultados são relatados e discutidos no Capítulo 6.

6 EXPERIMENTOS E RESULTADOS

Neste Capítulo são apresentados os experimentos e resultados preliminares do método proposto. Para isso, a Seção 6.1 apresenta a metodologia de avaliação adotada para os experimentos e as ferramantas utilizadas. A Seção 6.2 descreve as bases de dados utilizadas. A Seção 6.3 detalha os experimentos realizados e os resultados obtidos, além de estabelcer uma breve comparação dos resultados. A Seção 6.4 mostra uma análise comparativa do método proposto com um outro método de seleção de atributos. Por fim, a seção 6.4 trata das considerações finais do capítulo.

6.1 Metodologia de avaliação e ferramentas utilizadas

A metodologia para avaliação e validação estatística dos resultados produzidos pela aplicação do método FSF-HMC é ilustrada na Figura 17.

Inicialmente, o método é aplicado sobre um conjunto de dados para HMC e, em seguida, os conjuntos de dados reduzidos são utilizados para a avaliação da tarefa de classificação. Para isso, o conjunto D^T é utilizado para induzir um classificador hierárquico multirrótulo f, que, por sua vez, é aplicado sobre T^R. Neste caso, avalia-se a precisão do classificador na classificação das instâncias de treinamento em suas respectivas classes associadas.

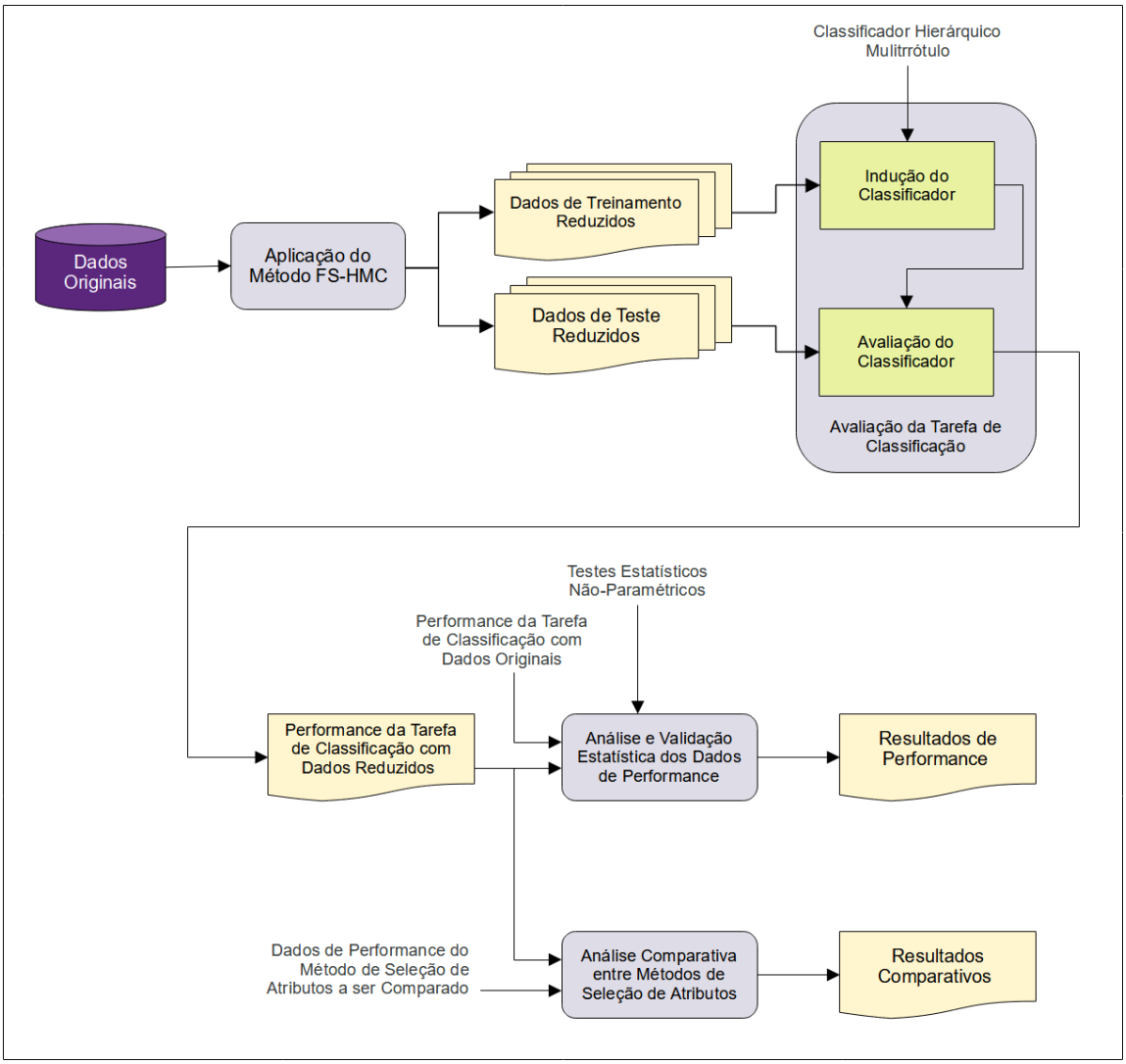
O algoritmo 4, apresenta o passo a passo básico da aplicação do método FSF-HMC, onde as entradas são os conjuntos de dados de treinamento e teste originais, D e T, respectivamente, e a hierarquia de classes H. A saída produzida é a performance P do classificador hierárquico multirrótulo para a tarefa de classificação com os conjuntos de dados reduzidos, D^R e T^R.

Algoritmo 4. *FSF-HMC*

Entrada: $\boldsymbol{D}, \boldsymbol{T}, \boldsymbol{H}$
Saída: $\boldsymbol{P}$
1: obter a partir de $\boldsymbol{D}$ os conjuntos $\boldsymbol{X}$ e $\boldsymbol{L_D}$
2: $\boldsymbol{X^R}$ = SelecionaAtributos($\boldsymbol{D}, \boldsymbol{X}, \boldsymbol{L_D}, \boldsymbol{H}$)
3: $\boldsymbol{D^R}, \boldsymbol{T^R}$ = GeraConjuntosReduzidos($\boldsymbol{D}, \boldsymbol{T}, \boldsymbol{X^R}$)
4: $\boldsymbol{f}$ = InduzClassificadorHM($\boldsymbol{D^R}$)
5: $\boldsymbol{P}$ = $\boldsymbol{f}$($\boldsymbol{T^R}$)
6: **retorna** $\boldsymbol{P}$

As linhas 1, 2 e 3 correspondem à aplicação do método FSF-HMC. A linha 1 indica que os conjuntos de atributos (X) e de classes associadas às instâncias (L_D) devem ser extraídos a partir do conjunto de dados de treinamento original. Na linha 2, tem-se uma chamada à função `SelecionaAtributos`, que corresponde à execução da primeira etapa do método, conforme apresentado na Subseção 5.1.1. Na linha 3 é executada a função `GeraConjuntosReduzidos`, que se refere à realização da segunda etapa do método, conforme exposto na Subseção 5.1.2. As linhas 4 e 5 indicam as atividades referentes à avaliação da tarefa de classificação.

Figura 17 - Metodologia de avaliação do FSF-HMC



Fonte: Autoria Própria (2022)

Os dados de performance da tarefa de classificação para um classificador hierárquico multirrótulo específico são, então, comparados com a performance da tarefa de classificação realizada sobre o conjunto de dados original. Aplicam-se testes

estatísticos não-paramétricos sobre os dados de performance coletados, a fim de validar os resultados obtidos, gerando-se os resultados de performance para o método.

Por fim, os dados de performance são também comparados com dados de performance produzidos por outro método de seleção de atributos, produzindo-se um resultado comparativo entre os métodos.

As bases de dados utilizadas nos experimentos, são descritas na Seção 6.2. Os algoritmos de classificação adotados para a tarefa de classificação foram o Clus-HMC e o MHC-CNN, descritos na Subseção 2.4.1. O desempenho desses classificadores foi mensurado em termos da medida AUPRC, descritas na Subseção 2.4.2. Estes resultados, obtidos na etapa de classificação hierárquica multirrótulo, foram avaliados utilizando o teste de hipótese de Wilcoxon (1945), que é aplicado para comparar dois grupos relacionados, sendo a variável é de mensuração ordinal. O teste classifica em postos a diferença entre os algoritmos sobre cada base usada para avaliação de desempenho (BORGES, 2012). A adoção de testes não-paramétricos justifica-se pela dificuldade de se conhecer a distribuição dos dados.

Os dados de performance gerados pela aplicação do método FSF-HMC são comparados com os dados de performance produzidos por outro método de seleção de atributos no contexto da HMC: o método *Feature Selection based on Wrapper approach for Hierarchical Multi-label Classification* (FSW-HMC) (ALMEIDA, 2018). Trata-se de um método para seleção de atributos que adota a abordagem *wrapper* e utiliza como estratégia de busca uma combinação de AG com SIA. O método foi testado com 10 bases de dados da GO, sendo adotada a medida de desempenho AUPRC e o classificador Clus-HMC. Os resultados dos experimentos demonstraram um ganho nas bases com todas as classes superior a 63,4% em termos da redução do número de atributos. Além disso, quanto ao desempenho do classificador, concluiu-se que não houve diferença estatisticamente significativa na medida AUPRC produzida pelo classificador quando submetido aos conjuntos de dados original e reduzido.

Escolheu-se o método FSW-HMC para realizar essa análise comparativa com o FSF-HMC porque os testes experimentais para ambos foram realizados com o mesmo conjunto de dados, obtidos a partir do trabalho de Borges (2012) e pela facilidade de acesso aos dados de performance.

Para a implementação dos algoritmos do método FSF-HMC propostos neste trabalho, optou-se por utilizar a linguagem de programação *Python*, em sua versão 3.8.5, e a biblioteca Pandas (PANDAS, 2022). Pandas é uma biblioteca *open source* e de uso gratuito para manipulação e análise de dados amplamente utilizada, que, de modo particular, permite trabalhar com dados tabulares e matrizes de forma simples e com poucos comandos. As principais estruturas de dados disponibilizadas pela biblioteca Pandas são as *Series* e os *DataFrames* (PANDAS, 2022). Além disso, é possível utilizar as diversas funcionalidades de indexação oferecidas pela biblioteca, para manipular, alterar a formatação, filtrar ou agregar subconjuntos específicos de dados (MCKINNEY, 2018).

6.2 Bases de dados

Os dados utilizados nos experimentos são oriundos do projeto *Gene Ontolgy* (GO). Trata-se de dados biológicos da área genômica funcional cuja organização hierárquica é na forma de um DAG. Vens *et al.* (2008), Borges (2012), Cerri *et al.* (2018), Almeida (2018), Melo e Paulheim (2019) e Siqueira (2019) desenvolveram trabalhos com esses mesmos dados. Na Tabela 5 são apresentadas as principais características dessas bases de dados.

Tabela 5 - Características das bases de dados GO

Base de Dados	Quant. Amostras	Quant. Atributos	Quant. Classes	Quant. Min/Max Classes por Amostra	Quant. Min/Max Amostras por Classe
Cellcycle	3751	77	4126	03/28	0/785
Church	3749	27	4126	03/28	0/786
Derisi	3719	63	4120	03/28	0/781
Eisen	2418	79	3574	03/28	0/492
Expr	3773	551	4132	03/28	0/789
Gasch1	3758	173	4126	03/28	0/786
Gasch2	3773	52	4132	03/28	0/789
Pheno	1586	69	3128	03/21	0/388
Seq	3900	478	4134	03/28	0/791
Spo	3697	80	4120	03/28	0/775

Fonte: Borges (2012, p. 80)

De modo particular, no trabalho de Borges (2012) esses dados já passaram por uma etapa de pré-processamento, que incluiu processos de transformação de valores, imputação de valores e normalização dos dados, além da separação em dois conjuntos, um para treinamento e outro para teste, sendo 2/3 de amostra para treinamento e 1/3 para teste.

6.3 Experimentos e resultados

Os experimentos realizados seguiram a metodologia de aplicação e avaliação apresentada na Seção 6.1, seguindo as duas etapas definidas: aplicação do método FSF-HMC e análise de validação estatística dos dados de performance. A metodologia é repetida para cada um dos conjuntos de dados utilizados.

As bases de dados de treinamento foram utilizadas pelo método FSF-HMC na execução da etapa de seleção de atributos. Na etapa seguinte, geração dos conjuntos de dados reduzidos, o método gerou conjuntos de treinamento e testes reduzidos com base na lista de atributos selecionados na etapa anterior. Neste sentido, cada conjunto de dados ao passar pelo método FSF-HMC gerou uma base de dados reduzida somente com os atributos cujo valor de FS seja superior à média dos valores de FS calculada para todos os atributos.

A Tabela 6 mostra a quantidade de atributos geradas para cada um dos conjuntos de dados utilizados e o percentual de redução alcançado em cada conjunto de dados, calculado pela Equação (26). Este percentual indica o ganho obtido com o processo de seleção de atributos na base de dados original.

$$PR = 100 - \frac{TS * 100}{TA} \tag{26}$$

onde TS é o número de atributos no subconjunto reduzido e TA é o número de atributos na base original. Este ganho foi superior a 53% para as 10 bases de dados analisadas, podendo chegar a 82,50% na base Spo. Com base nestes resultados, pode-se afirmar que o método FSF-HMC foi capaz de promover uma redução do número de atributos acima de 50% em todas as bases de dados. Este resultado é significativo do ponto de vista do desempenho para a tarefa de classificação, visto que tal tarefa pode ser realizada com um conjunto de dados de menor dimensão.

Tabela 6 - Percentual de redução de atributos após aplicação do método FSF-HMC

Base de Dados	TA	TS	PR (%)
Cellcycle	77	29	62,34
Church	27	7	74,07
Derisi	63	29	53,97
Eisen	79	30	62,03
Expr	551	154	72,05
Gasch1	173	65	62,43
Gasch2	52	21	59,62
Pheno	69	22	68,12
Seq	478	143	70,08
Spo	80	14	82,50

Fonte: Autoria Própria (2022)

Os resultados de avaliação do método FSF-HMC em termos da medida AUPRC são apresentados a seguir. Primeiro são apresentados os resultados obtidos e avaliados no classificador Clus-HMC e, posteriormente, no MHC-CNN.

6.3.1 Resultados para o classificador Clus-HMC

A Tabela 7 apresenta os resultados obtidos, em relação à medida AUPRC, para os conjuntos de dados no classificador Clus-HMC, tendo sido utilizados os parâmetros padrões do algoritmo, conforme definido por Vens et al. (2008). Estes resultados foram obtidos e relatados por Borges (2012). Ainda na Tabela 7 são apresentados os dados obtidos para o conjunto de dados reduzidos, obtido a partir da aplicação do método FSF-HMC, onde se pode verificar que a maior diferença obtida da medida AUPRC da base dados completa com o subconjunto reduzido foi igual a 0,014, para a base Expr. Para as demais bases analisadas essa diferença foi inferior a 0,005.

Esses resultados foram analisados estatisticamente por meio do teste Wilcoxon para duas amostras pareada, com nível de significância $\alpha = 0{,}05$. Por se tratar de uma amostra pequena ($N < 25$), os valores críticos são tabelados. Neste trabalho foi utilizada a tabela disponível na página *Real Statistics Using Excel*[1]. O valor calculado para W foi 8. Neste caso, para $N = 6$, pois no teste não são considerados

[1] https://www.real-statistics.com/statistics-tables/wilcoxon-signed-ranks-table/

postos para diferenças nulas, e $\alpha = 0{,}05$, o valor de W crítico é 0. Como $8 > 0$, a hipótese nula não pode ser rejeitada, indicando que não há diferença estatisticamente significativa no desempenho do classificador quando é utilizado o conjunto de dados original e o conjunto de dados reduzido.

Tabela 7 - Medida AUPRC para o classificador Clus-HMC

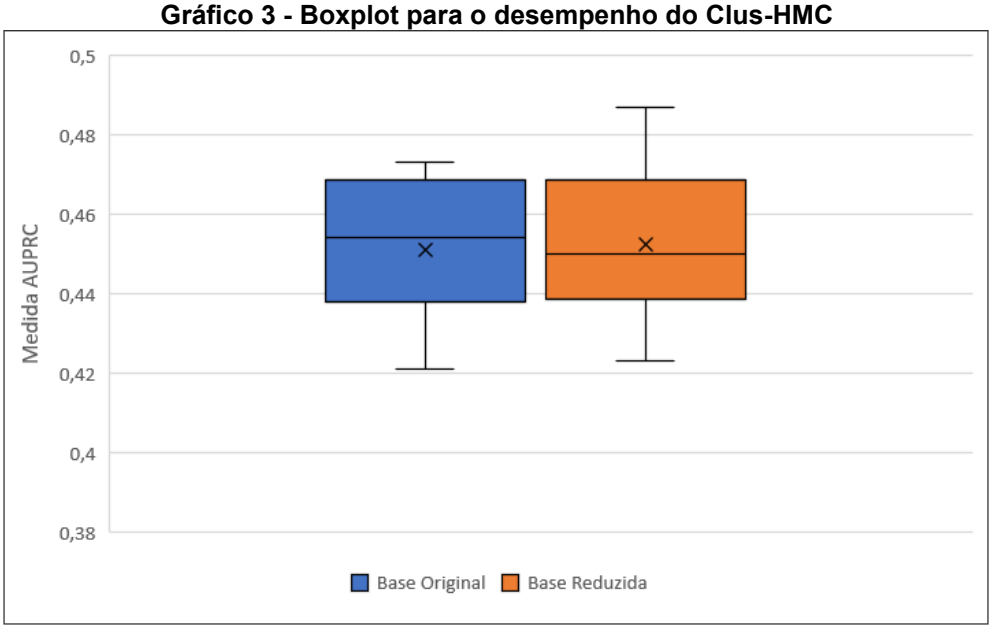
Base de Dados	AUPRC para a Base de Dados Original	AUPRC para a Base de Dados Reduzida	Diferença (Red. – Orig.)
Cellcycle	0,439	0,442	0,003
Church	0,452	0,449	-0,003
Derisi	0,438	0,438	0
Eisen	0,456	0,456	0
Expr	0,473	0,487	0,014
Gasch1	0,456	0,451	-0,005
Gasch2	0,438	0,439	0,001
Pheno	0,421	0,423	0,002
Seq	0,470	0,470	0
Spo	0,468	0,468	0

Fonte: Autoria Própria (2022)

O Gráfico 3 é um *boxplot* para o desempenho do algoritmo Clus-HMC com o conjunto de dados original e com o conjunto de dados reduzido. Observa-se uma grande semelhança na distribuição dos dados de desempenho, com intervalo de variação entre o menor e o maior valor de AUPRC próximos. Pode-se dizer, portanto, que os dados estão sobrepostos. Além disso, os valores da mediana e da média para os dois experimentos, são muito similares. A média do valor de AUPRC apurado para o conjunto de dados original foi de 0,454 e para o conjunto de dados reduzido foi de 0,450. Os valores das medianas foram 0,451 e 0,452, respectivamente, para o conjunto de dados original e para o conjunto de dados reduzido.

Nota-se, por meio dos valores das médias e medianas e do Gráfico 3, que as distribuições de desempenho aferidas para o classificador Clus-HMC, tanto com os dados originais quanto com os dados reduzidos, são aproximadamente simétricas, pois a mediana é próxima da média. Com base nessas análises, pode-se concluir que se trata de populações similares, o que sugere que a hipótese nula não possa seja rejeitada.

Gráfico 3 - Boxplot para o desempenho do Clus-HMC



Fonte: Autoria Própria (2022)

6.3.2 Resultados para o classificador MHC-CNN

Para a avaliação realizada com uso do classificador MHC-CNN, os parâmetros utilizados foram descritos em Borges (2012) e a condição de parada foi definida para 1000 épocas. A Tabela 8 apresenta os resultados obtidos em termos da medida AUPRC para o MHC-CNN considerando 1000 épocas. Os valores para o conjunto de dados original foram relatados por Borges (2012).

Tabela 8 – Medida AUPRC para o classificador HMC-CNN

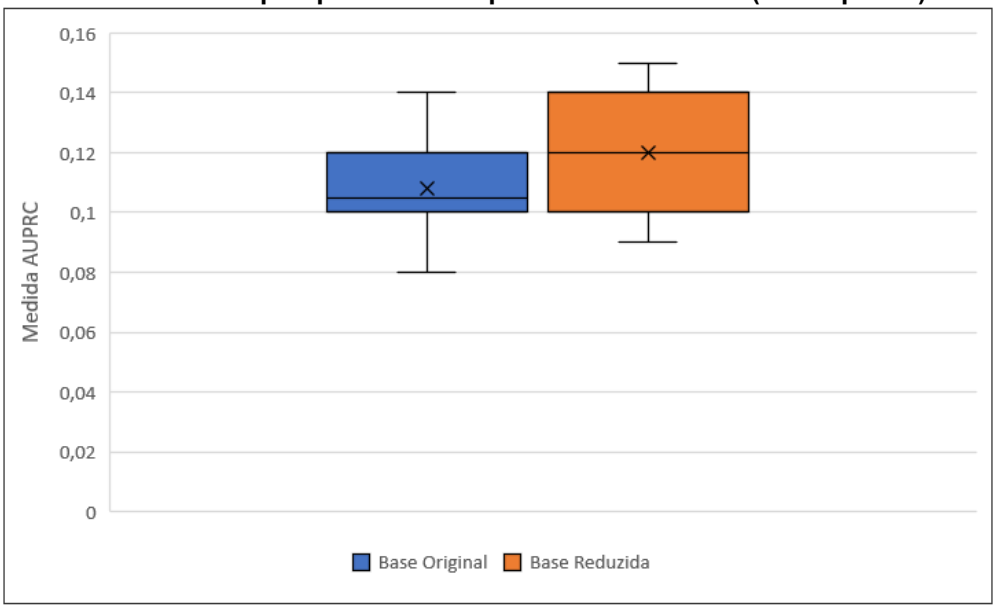
Base de Dados	AUPRC para a Base de Dados Original	AUPRC para a Base de Dados Reduzida	Diferença (Red. – Orig.)
Cellcycle	0,10	0,10	0
Church	0,12	0,11	-0,01
Derisi	0,11	0,15	0,04
Eisen	0,14	0,14	0
Expr	0,08	0,13	0,05
Gasch1	0,11	0,13	0,02
Gasch2	0,10	0,09	-0,01
Pheno	0,10	0,14	0,04
Seq	0,10	0,10	0
Spo	0,12	0,11	-0,01

Fonte: Autoria Própria (2022)

Aplicando o teste Wilcoxon com os resultados obtidos pelo algoritmo MHC-CNN para a base de dados original e a base de dados reduzida, com nível de significância $\alpha = 0{,}05$, obteve-se $W = 6$. O valor crítico é obtido na tabela específica para $N = 7$, sendo seu valor igual a 2. Como $6 > 2$, não se pode rejeitar a hipótese nula. Portanto, o desempenho do classificador MHC-CNN não apresentou diferenças estatisticamente significativas quanto ao desempenho com os conjuntos de dados original e reduzido.

O Gráfico 4 é um *boxplot* para o desempenho do algoritmo MHC-CNN com o conjunto de dados original e com o conjunto de dados reduzido. Como se pode observar, as distribuições de desempenho considerando os conjuntos de dados original e reduzido são bem semelhantes. Os valores da mediana e da média para os dois experimentos, são similares. A média do valor de AUPRC apurado para o conjunto de dados original foi de 0,108 e para o conjunto de dados reduzido foi de 0,112. Os valores das medianas foram 0,105 e 0,120, respectivamente, para o conjunto de dados original e para o conjunto de dados reduzido. Com base nessas análises, pode-se concluir que se trata de populações similares, o que sugere que a hipótese nula não pode seja rejeitada.

Gráfico 4 - Boxplot para o desempenho do MHC-CNN (1000 épocas)



Fonte: Autoria Própria (2022)

Os resultados obtidos pelos testes experimentais podem ser utilizados para comparar o método FSF-HMC com outros métodos de seleção de atributos. Neste

trabalho, foi feita a comparação do FSF-HMC com o método FSW-HMC. A Seção 6.4 apresenta a análise comparativa dos dois métodos.

6.4 Análise comparativa com o método FSW-HMC

A análise comparativa dos métodos FSF-HMC e FSW-HMC foi feita em termos do número de atributos selecionados (NAS), do percentual de redução do número de atributos na base de dados originais (PR) e dos resultados da classificação em termos da medida AUPRC para o classificador Clus-HMC. Nos experimentos conduzidos por Almeida (2018) para avaliação do método FSW-HMC, foram gerados 10 subconjuntos de dados reduzidos para cada conjunto de dados originais, sendo considerado o subconjunto com melhor medida AUPRC. Esses dados são mostrados na Tabela 9.

Tabela 9 - Comparativo entre os métodos FSW-HMC e FSF-HMC

Base de Dados	FSW-HMC			FSF-HMC		
	N_{AS}	PR	AUPRC	N_{AS}	PR	AUPRC
Cellcycle	1	98,7	0,432	29	62,34	0,442
Church	7	74,1	0,446	7	74,07	0,449
Derisi	4	93,7	0,436	29	53,97	0,438
Eisen	8	89,9	0,448	30	62,03	0,456
Expr	16	97,1	0,459	154	72,05	0,487
Gasch1	54	68,8	0,458	65	62,43	0,451
Gasch2	19	63,5	0,435	21	59,62	0,439
Pheno	1	98,6	0,422	22	68,12	0,423
Seq	46	90,4	0,470	143	70,08	0,470
Spo	13	83,8	0,463	14	82,50	0,468

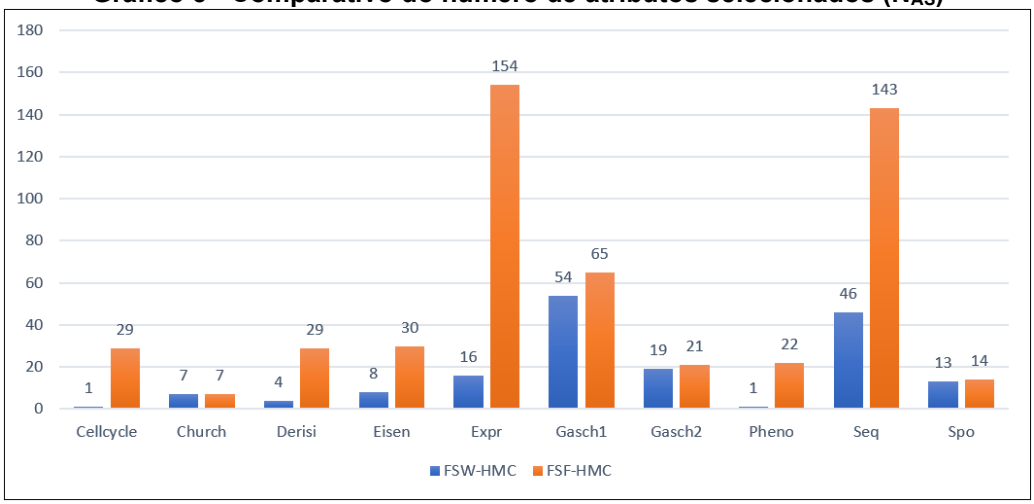
Fonte: Autoria Própria (2022)

O Gráfico 5 mostra a comparação dos métodos FSW-HMC e FSF-HMC quanto ao número de atributos selecionados.

Como se pode observar, exceto para a base de dados Church, onde ambos os métodos selecionaram o mesmo número de atributos, o método FSW-HMC conseguiu selecionar um número inferior de atributos nas bases de dados originais. Aplicando o teste Wilcoxon, com nível de significância $\alpha = 0{,}05$, obteve-se $W = 0$. O valor crítico é obtido na tabela específica para $N = 9$, sendo seu valor igual a 5. Como

$0 < 5$, a hipótese nula deve ser rejeitada, indicando que há diferença estatística significativa em favor do método FSW-HMC referente ao número de atributos selecionados.

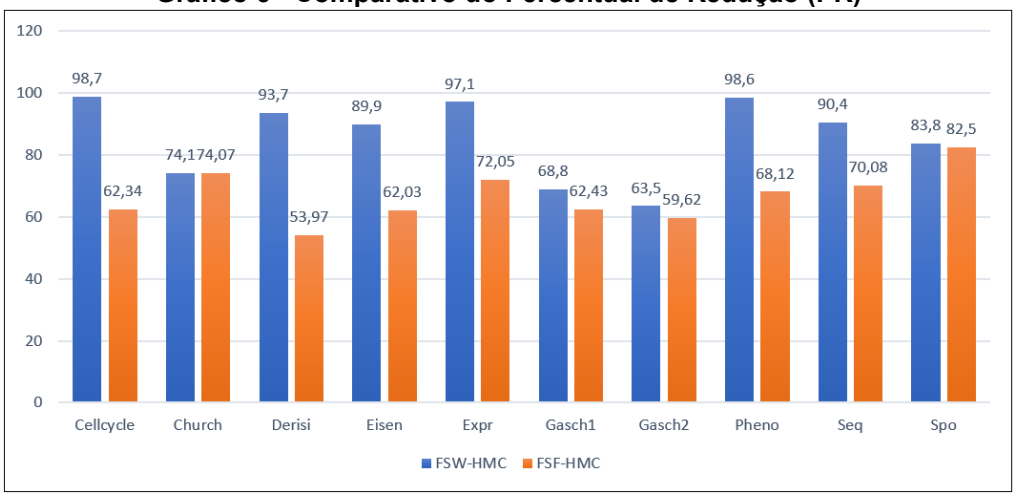
Gráfico 5 - Comparativo do número de atributos selecionados (N_{AS})



Fonte: Autoria Própria (2022)

O Gráfico 6 mostra a comparação dos métodos FSW-HMC e FSF-HMC quanto ao percentual de redução do número de atributos em relação à base de dados original. Como se pode observar, o método FSW-HMC conseguiu um melhor percentual de redução em todas as bases de dados testadas. Aplicando o teste Wilcoxon, com nível de significância $\alpha = 0{,}05$, obteve-se $W = 0$. O valor crítico é obtido na tabela específica para $N = 10$, sendo seu valor igual a 8. Como $0 < 8$, a hipótese nula deve ser rejeitada, indicando que há diferença estatística significativa em favor do método FSW-HMC referente ao percentual de redução obtido.

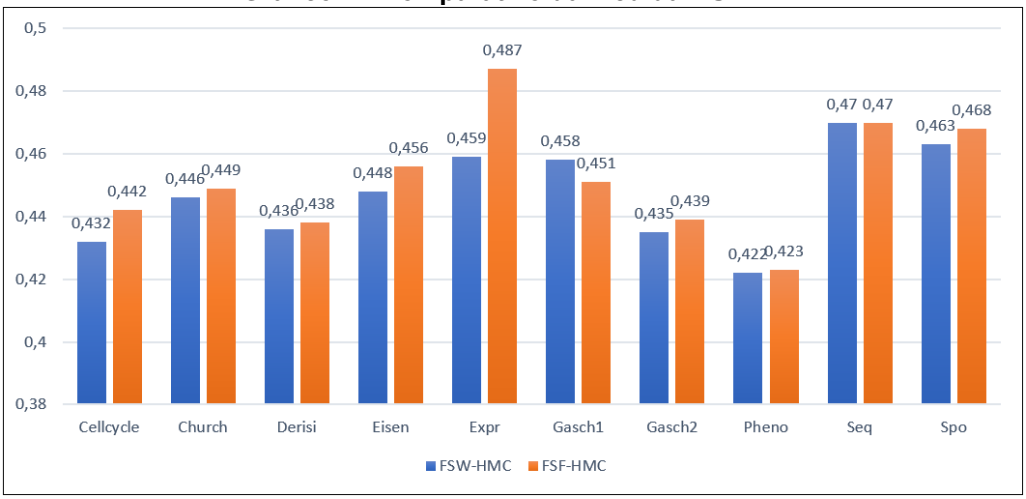
Gráfico 6 - Comparativo do Percentual de Redução (PR)



Fonte: Autoria Própria (2022)

O Gráfico 7 mostra a comparação dos métodos FSW-HMC e FSF-HMC referente à medida AUPRC gerada pelo classificador Clus-HMC. Como se pode observar, exceto para as bases Gasch1 e Seq, o método FSF-HMC conseguiu um melhor valor para a medida AUPRC. Porém, aplicando o teste Wilcoxon, com nível de significância $\alpha = 0{,}05$, obteve-se $W = 6$. O valor crítico é obtido na tabela específica para $N = 9$, sendo seu valor igual a 5. Como $6 > 5$, a hipótese nula não deve ser rejeitada, indicando que não há diferença estatística significativa quanto ao desempenho do classificador para os conjuntos de dados reduzidos gerados pelos métodos comparados.

Gráfico 7 - Comparativo da medida AUPRC



Fonte: Autoria Própria (2022)

Apesar das significativas diferenças quanto ao número de atributos selecionados e ao percentual de redução da atributos na base de dados, os dois métodos apresentaram resultados estatisticamente iguais em relação à tarefa de classificação com o classificador Clus-HMC. Deve-se ressaltar que o FSW-HMC adota a abordagem *wrapper* e, por ser dependente de algoritmo, utiliza a medida de precisão do algoritmo de aprendizado também como critério de parada. Neste sentido, é necessário um número maior de iterações até que se alcance o conjunto reduzido com melhor desempenho. Isto leva o método FSW-HMC a selecionar o melhor dentre os possíveis subconjuntos de atributos.

De forma diferente, o método FSF-HMC, por usar a abordagem filtro univariado com uma medida de avaliação para ranquear os atributos de modo independente do algoritmo de classificação e por estabelecer um critério de seleção baseado no valor dessa medida, tem como limitação que a quantidade de atributos

selecionados é fixada pelo valor da medida, não sendo, portanto, possível testar outros subconjuntos de atributos sem que seja alterado o critério.

Dado que o método FSW-HMC usa a abordagem *wrapper*, seu custo computacional tende a ser maior do que o método FSF-HMC, pois este, por ser independente de algoritmo, aproveita-se do menor custo computacional proporcionado pela abordagem filtro. O que é considerado uma vantagem do FSF-HMC sobre o FWS-HMC, visto que não há diferença entre esses métodos no que se refere à tarefa de classificação.

6.5 Considerações finais do capítulo

Neste capítulo foram apresentados a metodologia utilizada nos experimentos e os resultados obtidos para o método de seleção de atributos proposto em 10 bases de dados da GO. Foi realizada uma análise da tarefa de seleção de atributos e uma comparação do desempenho de dois classificadores hierárquicos multirrótulo, Clus-HMC e MHC-CNN, em termos da medida de AUPRC obtida nesses classificadores com os conjuntos de dados originais e com os conjuntos de dados reduzidos obtidos com a aplicação do método FSF-HMC. Além disso, foi feita uma análise comparativa do FSF-HMC com outro método de seleção de atributos, FSW-HMC.

Quanto à capacidade de seleção de atributos, o método avaliado proporcionou um ganho em todas as bases, em termos do percentual de redução, que ficou acima de 53%, tendo chegado a 82,5% na base de dados Spo. Esse percentual de redução é uma função da quantidade de atributos na base de dados original e o valor médio do *Fisher Score* calculado para cada um dos atributos na base original, visto que o método estabelece que são selecionados apenas os atributos cujo valor da medida *Fisher Score* seja maior do que o valor médio obtido para a base original.

Referente ao desempenho do classificador com o conjunto de dados reduzidos. Tanto no Clus-HMC quanto no HMC-CNN a avaliação estatística demonstrou que os resultados obtidos são equivalentes às medidas AUPRC obtidas com os conjuntos de dados originais.

Por fim, apesar de se ter demonstrado que o método proposto apresentou resultados similares aos obtidos pelo Clus-HMC e HMC-CNN sem a seleção de atributos, os valores obtidos de AUPRC foi para um conjunto de atributos de tamanhos

inferiores. Neste caso, pode-se falar em ganho computacional no tempo de processamento das bases de dados reduzidas, além de ser gerado um modelo de classificação mais simples.

7 CONCLUSÃO

O tratamento de dados com alta dimensionalidade é uma tarefa desafiadora para a aprendizagem de máquina. De modo particular, em bases de dados de classificação hierárquica multirrótulo, a performance de um algoritmo de aprendizagem pode ser prejudicada tanto em tempo de execução quanto em precisão na classificação de exemplos ainda não rotulados.

A seleção de atributos é uma das técnicas que podem ser utilizadas para a redução da dimensionalidade em bases de dados. Trata-se de uma tarefa importante, pois permite a exclusão de atributos redundantes e/ou irrelevantes, gerando-se um subconjunto de atributos de menor dimensão, sem, com isso, perder o significado dos dados. A seleção de atributos pode contribuir, portanto, para melhorar o desempenho dos algoritmos de aprendizagem para a tarefa de classificação.

O método FSF-HMC, desenvolvido neste trabalho, adota a abordagem filtro para a seleção de atributos, sendo um método do tipo filtro univariado de ordenação de atributos. Isto quer dizer que o método utiliza apenas informações dos próprios dados para medir seu nível de qualidade. Neste sentido, o método incorpora as principais vantagens da abordagem filtro, como eficiência, menor custo computacional que as demais abordagens, além de ser independente de algoritmo.

A medida de avaliação da qualidade de atributos adotada foi o *Fisher Score*. Trata-se de uma estratégia para calcular uma pontuação individual para cada atributo na base de dados. Esta estratégia apresenta bons resultados em bases de alta dimensionalidade, porém não foram encontrados trabalhos relacionados que utilizem tal métrica para a tarefa de seleção de atributos.

Para os experimentos foram utilizados 10 bases de dados da GO, em que as classes estão dispostas hierarquicamente em forma de DAG. Cada base de dados foi submetida ao método FSF-HMC, sendo gerado um conjunto de dados reduzido para cada uma delas. Nos subconjuntos gerados observou-se um percentual de redução superior a 53% em relação à base completa. O maior percentual de redução verificado foi de 82,5%.

A fim de verificar a capacidade preditiva do método em relação à redução alcançada, foram utilizados dois classificadores hierárquicos multirrótulo, Clus-HMC e MHC-CNN, e a medida de avaliação de desempenho AUPRC. Esses classificadores foram testados com cada uma das bases de dados reduzidas e os desempenhos

apurados foram comparados com os desempenhos gerados quando utilizados os conjuntos de dados originais.

Após análise dos resultados obtidos com as bases de dados reduzidas, verificou-se que a diferença obtida da medida AUPRC foi inferior a 0,08 considerando os cenários de testes de ambos os classificadores. Os resultados obtidos foram validados estatisticamente pelo teste de Wilcoxon e pelo teste de Friedman (este apenas para o classificador MHC-CNN) e verificou-se que os resultados apresentados por ambos os classificadores para as bases de dados reduzida e original são estatisticamente equivalentes.

Quando comparado com outro método de seleção de atributos, FSW-HMC, observou-se que o desempenho do FSF-HMC em relação ao número de atributos selecionados, por consequência, ao percentual de redução, foi inferior. Porém, deve-se destacar que os dois métodos adotam abordagens de seleção de atributos diferentes. O método FSW-HMC utiliza a abordagem *wrapper*, que, em geral, possui desempenho melhor. Por fazer uso da abordagem filtro univariado, o método FSF-HMC, não consegue lidar com a correlação entre os atributos e entre as classes, o que pode levar o método a selecionar atributos redundantes. Quanto à tarefa de classificação, verificou-se que estatisticamente os dois métodos apresentaram desempenho equivalente. Uma vantagem do método FSF-HMC sobre o método FSW-HMC é conseguir produzir igual desempenho na tarefa de classificação a um custo computacional inferior.

No que se refere à seleção de atributos para bases de dados de classificação hierárquica multirrótulo, conforme mapeamento sistemático apresentado no Capítulo 4, alguns métodos têm sido propostos para esta tarefa. Em geral, esses métodos adotam a abordagem filtro e são aplicados para hierarquias estruturadas como árvores. Nenhum dos métodos citados no mapeamento adota o *Fisher Score* como medida de avaliação da qualidade dos atributos.

São contribuições deste trabalho o panorama do estado da arte para o problema da seleção de atributos em bases de dados de classificação hierárquica multirrótulo, a investigação e adaptação do cálculo da medida *Fisher Score* para considerar a hierarquia de classes e a apresentação de um novo método para seleção de atributos que pode ser utilizado tanto para bases hierárquicas monorrótulo como multirrótulo e cujas classes estão organizadas hierarquicamente como DAG ou como árvore.

7.1 Trabalhos futuros

Diversas extensões podem ser exploradas para ampliação e aprimoramento deste trabalho. Uma das propostas é aplicar o método FSF-HMC em outras bases de dados fora do domínio da bioinformática, para fins de comparação de desempenho. Outra proposta é utilizar outros classificadores hierárquicos multirrótulo e outras medidas de avaliação de desempenho, a fim de validar os resultados obtidos.

Também como trabalho futuro propõe-se um estudo comparativo do método FSF-HMC com outros métodos de seleção de atributos que adotem a abordagem filtro, visando a comparação de desempenho, de modo particular os estudos identificados no mapeamento sistemático de literatura realizado, especialmente o trabalho 8. Por fim, um desafio é estender o método para que seja capaz de lidar com a correlação entre os atributos. Desse modo, será possível identificar e excluir atributos redundantes do subconjunto reduzido gerado pelo método.

REFERÊNCIAS

ALJEDANI N, ALOTAIBI R, TAILEB M. HMATC: Hierarchical multi-label arabic text classification model using machine learning. **Egyptian Informatics Journal**. vol. 22. num. 3 p. 225-37, 2021.

ALMEIDA, T. B. **Seleção de atributos usando abordagem wrapper para classificação hierárquica multirrótulo**. 2018. Dissertação (Mestrado em Ciência da Computação) – Programa de Pós-Graduação em Ciência da Computação. Universidade Tecnológica Federal do Paraná, Ponta Grossa, 2018.

ALVES, R. T. **Um Sistema Imunológico Artificial para Classificação Hierárquica e Multi-Label de Funções de Proteínas**. 2010. Tese (Doutorado em Engenharia Elétrica e Informática Industrial) - Programa de Pós-graduação em Engenharia Elétrica e Informática Industrial. Universidade Tecnológica Federal do Paraná (UTFPR), Curitiba, 2010.

ANSARI, M.; SHAHANE, N. M. A Review on Multi-label Classification. **International Journal of Research and Analytical Reviews**, vol. 06, n. 01, p. 816-819, jan/mar. 2019.

AZHAGUSUNDARI, B.; THANAMANI, A. S. Feature Selection based on Information Gain. **International Journal of Innovative Technology and Exploring Engineering** (IJITEE), vol. 2, n. 2, p. 18-21, 2013.

BLOCKEEL, H.; *et al*. Hierarchical multi-classification. *In*: ACM SIGKDD INTERNATIONAL CONFERENCE ON KNOWLEDGE DISCOVERY AND DATA MINING. 8, 2002, Edmonton. **Proceedings [...]** New York: Association for Computing Machinery, 2002, p. 21-35.

BLOCKELL, H. *et al*. Decision trees for hierarchical multilabel classification: A case study in functional genomics. *In*: EUROPEAN CONFERENCE ON PRINCIPLES AND PRACTICE OF KNOWLEDGE DISCOVERY IN DATABASES. 10, 2006, Edinburgh. **Proceedings [...]**, Edinburgh: DPLP, 2006, p. 18-29.

BLUM, A. L.; LANGLEY, P. Selection of relevant features and examples in machine learning. **Artificial intelligence**, Elsevier, v. 97, n.1, p.245–271,1997.

BORGES, H. B. **Classificador hierárquico multirrótulo usando uma rede neural competitiva**. 2012. Tese (Doutorado em Informática) – Programa de Pós-Graduação em Informática. Pontifícia Universidade Católica do Paraná, Curitiba, 2012.

BOYD, K. *et al*. Unachievable region in precision-recall space and its effect on empirical evaluation. *In*: INTERNATIONAL CONFERENCE ON MACHINE LEARNING. 29, 2012, Edinburgh. **Proceedings [...]** Madison: Omnipress, 2012, p. 349-368.

BOZ, Olcay. Feature subset selection by using sorted feature relevance. *In*: INTERNATIONAL CONFERENCE ON MACHINE LEARNING AND APPLICATIONS. 2002, Las Vegas **Proceedings [...]** Las Vegas: CSREA Press, 2002, p. 147-153.

ICMLA. [S.l.: s.n.], 2002. p.147–153.

CARVALHO, A.C.P.L.F.; FREITAS, A. A. A Tutorial on Multi-label Classification Techniques. *In*: ABRAHAM, A.; HASSANIEN, A. E.; SNÁŠEL, V. (org**). Foundations of Computational Intelligence**, vol. 05. Studies in Computational Intelligence, v. 205. Berlin: Springer, 2009, p. 177-195.

CERRI, R. **Técnicas de classificação hierárquica multirrótulo**. 2010. Dissertação (Mestrado em Ciência da Computação) – Programa de Pós-Graduação em Ciência da Computação e Matemática. Universidade de São Paulo, São Paulo, 2010.

CERRI, R.; BARROS, R. C.; CARVALHO, A. C. P. L. F. A Genetic Algorithm for Hierarchical Multi-Label Classification. *In:* ANNUAL ACM SYMPOSIUM ON APPLIED COMPUTING. 27, 2012, New York. **Proceedings [...]** New York: Association for Computing Machinery, 2012, p. 250–255.

CERRI, R.; *et al*. Multi-label Feature Selection Techniques for Hierarchical Multi-label Protein Function Prediction. *In*: INTERNATIONAL JOINT CONFERENCE ON NEURAL NETWORKS. 2018, Rio de Janeiro. **Proceedings [...]** Rio de Janeiro: IEEE, 2018, p. 1-7.

CHANDRASHEKAR, G.; SAHIN, F. A survey on feature selection methods. **Computers and Electrical Engineering**, vol. 14, p. 16-28, 2014.

CHEN, J.; HU, J.; ZHANG, G. Feature Selection Based on Gain Ratio in Hybrid Incomplete Information Systems. *In*: INTERNATIONAL CONFERENCE ON INTELLIGENT SYSTEMS AND KNOWLEDGE ENGINEERING. 16, 2021, Chengdu. **Proceedings [...]** Chegdu: IEEE, 2021, p. 728-735.

CHERMAN, E.A.; MONARD, M. C.; METZ, J. Multi-label Problem Transformation Methods: a Case Study. **CLEI Electronic Journal**, vol. 14, n. 1, p. 1-10, 2011.

COSTA, E. P. **Investigação de Técnica de Classificação Hierárquica para Problemas em Bioinformática**. 2008. Dissertação (Mestrado em Ciências) – Instituto de Ciências Matemáticas e de Computação da Universidade de São Paulo, São Carlos, 2008.

COSTA JUNIOR, J. D. *et al.* Label Powerset for Multi-label Data Streams Classification with Concept Drift. *In***:** SYMPOSIUM ON KNOWLEDGE DISCOVERY, MINING AND LEARNING. 5, 2017, Uberlândia. **Proceedings [...]** Uberlândia: SBC, 2017, p. 97-104.

COVÕES, T. F. **Seleção de atributos via agrupamento**. 2010. Dissertação (Mestrado em Ciências) – Instituto de Ciências Matemáticas e de Computação da Universidade de São Paulo, São Carlos, 2010.

DASH, M.; LIU, H. Feature selection for classification. **Intelligent Data Analysis**, Elsevier, v. 1, n. 1, p. 131-156, 1997.

DASH, M.; LIU, H.; MOTODA, H. Consistency Based Feature Selection. *In*: PACIFIC-ASIA CONFERENCE ON KNOWLEDGE DISCOVERY AND DATA MINING. 4, 2000, Delhi. **Proceedings [...]** Delhi: ACM, 2000, p. 98-109.

DIMITROVSKI, I. *et al*. Hierarchical annotation of medical images. **Pattern Recognition**, Elsevier, v. 44, n. 10-11, 2011, p. 2436-2449.

DIMITROVSKI, I. *et al.* Hierarchical classification of diatom images using ensembles of predictive clustering trees. **Ecological Informatics**, Elsevier, v. 7, n. 1, p. 19-29, 2012.

DUDA, R. O.; HART, P. E.; STORK, D. G. **Pattern Classification**. 2 ed. Wiley-Interscience, 2001.

FACELI, K. *et al*. **Inteligência Artificial:** uma abordagem de aprendizagem de máquina. Rio de Janeiro: LTC, 2011.

FENG, S.; ZHAO, C.; FU, P. A deep neural network based hierarchical multi-label classification method. **Review of Scientific Instruments**. AIP Publishing, v. 91, n. 2, p. 024103, 2020.

FREITAS, A. A.; CARVALHO, A. C. P. F. A Tutorial on Hierarchical Classification with Applications in Bioinformatics. *In*: TANIAR, D. Research and Trends in Data Mining Technologies and Applications. **Advances in Data Warehousing and Mining**. IGI Publishing, Cap.7, p. 179-209, 2007.

GALAR, M. *et al.* An overview of ensemble methods for binary classifiers in multi-class problems: experimental study on one-vs-one and one-vs-all schemes. **Pattern Recognition**, Elsevier, v. 44, n. 8, 2011, p. 1761–1776.

GALAR, M. *et al.* A review on ensembles for the class imbalance problem: bagging, boosting, and hybrid-based approaches**. IEEE Transactions on Systems, Man, and Cybernetics**, Part C (Applications and Reviews), v. 42, n. 4, 2012, p. 463–484.

GARGIULO, F. *et al*. Deep neural network for hierarchical extreme multi-label text classification. **Applied Soft Computing**, Elsevier, v. 79, p. 125-138, 2019.

GEVERT, V. G. *et al.* Modelos de Regressão Logística, Redes Neurais e Support Vector Machine (SVMs) na Análise de Crédito a Pessoas Jurídicas. **Revista Ciências Exatas e Naturais**, vol.12, n. 2, 2010.

GHODSI, A. **Dimensionality reduction a short tutorial**. Department of Statistics and Actuarial Science, University of Waterloo, Ontario, Canada, p. 38, 2006.

GOLDSCHMIDT, R.; PASSOS, E.; BEZERRA, E. **Data Mining:** conceitos, técnicas, orientações e aplicações. 2 ed. Rio de Janeiro: Elsevier, 2015.

GOPIKA, N.; KOWSHALAYA, M. Correlation Based Feature Selection Algorithm for Machine Learning. *In*: INTERNATIONAL CONFERENCE ON COMMUNICATION AND ELECTRONICS SYSTEMS. 3, 2018, Calcutta. **Proceedings […]** Calcultta: IEEE, 2018, p. 692-695.

GU, Q.; LI, Z.; HAN, J. Generalized Fisher Score for feature selection. **arXiv preprint arXiv:1202.3725**, 2012.

GÜNES, S.; POLAT, K.; YOSUNKAYA, S. Multi-class f-score feature selection approach to classification of obstructive sleep apnea syndrome. **Expert Systems with Applications**, Elsevier, v. 37, n. 2, p. 998-1004, 2010.

HAN, J.; KAMBER, M.; PEI, J. **Data Mining**: Concepts and Techniques. 3 ed. Waltham: Elsevier, 2012.

HERRERA, F. *et al.* Multilabel Classification. *In*: HERRERA, F. *et al.* **Multilabel Classification**: Problems Analysis, Metrics and Techniques. Cham: Springer International Publishing Switzerland, 2016, p. 17-31.

HOQUE, N.; BHATTACHARYYA, D. K.; KALITA, J. K. MIFS-ND: a mutual information-based feature selection method. **Expert Systems with Applications**, vol. 41, n. 14, p. 6371–6385, 2014.

HUANG, H.; LIU, H. Feature selection for hierarchical classification via joint semantic and structural information of labels. **Knowledge-Based Systems**, v. 195, 105655, mai. 2020.

HUANG, J. *et al.* Multi-Label Learning via Feature and Label Space Dimension Reduction. **IEEE Access**, v. 8, p. 20289-20303, 2020.

JAIN, A. K.; DUIN, R. P. W.; MAO, J. Statistical pattern recognition: a review. **IEEE Transactions on Pattern Analysis and Machine Intelligence**, IEEE, v. 22, n. 1, p. 4-37, 2000.

JIN, C. *et al.* Chi-square Statistics Feature Selection Based on Term Frequency and Distribution for Text Categorization. **IETE Journal of Research**, vol. 61, n. 4, p. 351-362, 2015.

JOVIĆ, A.; BRKIĆ, K.; BOGUNOVIĆ, N. A review of feature selection methods with applications. *In*: INTERNATIONAL CONVENTION ON INFORMATION AND COMMUNICATION TECHNOLOGY, ELECTRONICS AND MICROELECTRONICS. 38, 2015, Opatija. **Proceedings [...]** Opatija: IEEE, 2015, p. 1200-1205.

KOHAVI, R.; JOHN, G. H.; Wrappers for feature subset selection. **Artificial Intelligence**, vol. 97, p. 273-324, 1997.

KOHONEN, T. The self-organizing map. **Proceedings of the IEEE**, IEEE, v. 78, n. 9, p. 1464-1480, set. 1990.

KRUSKAL, J. B. Multidimensional scaling by optimizing goodness of fit to a nonmetric hypothesis. **Psychometrika**, Springer, v. 29, n. 1, p. 1-27, 1964.

KUDO, M.; SKLASNKY, J. A comparative evaluation of medium and large-scale feature selectors for pattern classifiers. **Kybernetika**, vol. 34, n. 4, p. 429-434, 1998.

KUMAR, V.; MINZ, S. Feature selection: a literature review. **Smart Computer Review**, v. 4, n. 3, p. 211-229, 2014.

LAZAR, C. *et al.* A Survey on Filter Techniques for Feature Selection in Gene Expression Microarray Analysis. **IEEE/ACM Transactions on Computational Biology and Bioinformatics**, vol. 9, n. 4, p. 1106-1119, 2012

LEE, Huei Diana. **Seleção de atributos importantes para a extração de conhecimento em bases de dados**. 2005. Tese (Doutorado em Ciências) - Programa de Pós-graduação em Ciências – Ciência da Computação e Matemática. Instituto de Ciências Matemáticas e de Computação (ICMC-USP), São Carlos, 2005.

LIN, J.; GUNOPULOS, D. Dimensionality reduction by random projection and latent semantic indexing. Disponível em: <https://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.386.8494&rep=rep1&type=pdf.> Acesso em: 25 nov. 2021.

LIU, H.; MOTODA, H. **Computational methods of feature selection**. [S.l.]: CRC Press, 2008, 411p.

LORENA, L. H. N.; CARVALHO, A. C. P. L. F.; LORENA, A. C. Filter Feature Selection for One-Class Classification. **Journal of Intelligent and Robotic Systems**, vol. 80, p. 227-243, 2015.

MAATEN, L. V. D.; POSTMA, E., HERIK, J. V. D. Dimensionality reduction: a comparative. **Journal of Machine Learning Research**. vol. 10, p. 66-71, 2009.

MANIKANDAN, G., ABIRAMI, S. A Survey on Feature Selection and Extraction Techniques for High-Dimensional Microarray Datasets. *In:* ANOUNCIA, S. M.; WIIL, U. K. (org). **Knowledge Computing and its Applications**: Knowledge Computing in Specific Domains. vol. II. Singapore: Springer, 2018. p. 311-333.

MCKINNEY, W. **Python para Análise de Dados**: Tratamento de Dados com Pandas, NumPy e IPython. Rio de Janeiro: Novatec, 2018.

MELO, A.; PAULHEIM, H. Local and global feature selection for multilabel classification with binary relevance. **Artificial Intelligence Review**, Springer, v. 51, n. 1, p. 33-60. 2019.

MICHALAK, K.; KWASNICKA, H. Correlation based feature selection method. **International Journal of Bio-Inspired Computation**, vol. 2, n. 5, p. 319-332, 2010.

MITCHELL, T. M. **Machine Learning**. McGraw-Hill Higher Education, 1997.

MONARD, A. C.; BARANAUSKAS, J. A. Conceitos sobre Aprendizagem de Máquina. In: REZENDE, S. O. **Sistemas Inteligentes**: Fundamentos e Aplicações. Barueri: Manole, 2005, p. 89-114.

NAGPAL, A.; GAUR, D. A New Proposed Feature Subset Selection Algorithm Based on Maximization of Gain Ratio. In: KUMAR, N.; BHATNAGAR, V. (org). **Big Data Analytics**. Cham: Springer International Publishing, 2015, p.181-197.

O CADERNO. Intérprete: Pe. Fábio de Melo. Compositor: Toquinho. *In*: VIDA. Intérprete: Pe. Fábio de Melo. Rio de Janeiro: Som Livre, 2008. 1 CD, faixa 6.

OMUYA, E. O.; OKEYO, G. O.; KIMWELE, M. W. Feature Selection for Classification using Principal Component Analysis and Information Gain. **Expert Systems with Applications**, vol. 174, n. 114765, 2021.

PANDAS. Pandas Documentation, 2022. Disponível em: < https://pandas.pydata.org/docs/> Acesso em: 20 de jun. de 2022.

PEREIRA, R. B. *et al.* Information Gain Feature Selection for Multi-Label Classification. **Journal of Information and Data Management**, vol. 6, n. 1, p. 48-58, 2015.

PÉREZ-ORTIZ, M. *et al.* Fisher Score-Based Feature Selection for Ordinal Classification: A Social Survey on Subjective Well-Being. *In*: INTERNATIONAL CONFERENCE ON HYBRID ARTIFICIAL INTELLIGENCE SYSTEMS. 11, 2016, Seville. **Proceedings [...]** Seville: Springer, 2016, p. 597-608.

PRABOWO, F. A.; IBROHIM, M. O.; BUDI, I. Hierarchical Multi-label Classification to Identify Hate Speech and Abusive Language on Indonesian Twitter. *In*: INTERNATIONAL CONFERENCE ON INFORMATION TECHNOLOGY, COMPUTER AND ELECTRICAL ENGINEERING. 6, 2019, Semarang. **Proceedings [...]**, Semarang: IEEE, 2019, p. 1-5.

PRIYADARSINI, R. P.; VALARMATHI, M. L.; SIVAKUMARI, S. Gain Ratio based feature selection method for privacy preservation. **ICTACT Journal of Soft Computing**, vol. 1, n. 1, p. 201-205, 2011.

PUTRI, N. K.; RUSTAM, Z.; SARWINDA, D. Learning Vector Quantization for Diabetes Data Classification with Chi-Square Feature Selection. **IOP Conference Series: Materials Science and Engineering**, vol. 546, n. 5, p. 052059, 2019.

QUINLAN, J. R. Induction of decision trees. **Machine learning**, v. 1, n. 1, p. 81–106, 1986.

RATTAN, D.; BATHIA, R.; SINGH, M. Software clone detection: a systematic review. **Information and Software Technology**, [S.l.], v. 55, n. 7, p. 1165-1199, 2013.

REMESEIRO, B.; BOLON-CANEDO, V. A review of feature selection methods in medical applications. **Computers in Biology and Medicine**, vol.112, n. 103375, 2019.

RUSSELL, S.; NORVIG, P. **Inteligência Artificial**. 3 ed. Rio de Janeiro: Elsevier, 2013.

SHIN, K; XU, X. M. Consistency-Based Feature Selection. *In*: INTERNATIONAL CONFERENCE ON KNOWLEDGE-BASED AND INTELLIGENT INFORMATION AND ENGINEERING SYSTEMS. 13, 2009, Santiago. **Proceedings [...]** Santiago: Springer-Verlag, 2009, p. 342-350.

SIBLINI, W.; KUNTZ, P.; MEYER, F. A Review on Dimensionality Reduction for Multi-label Classification. **IEEE Transactions on Knowledge and Data Engineering**, [s. l.], v. 14, n. 8, p. 1-20, set. 2019.

SIKONJA, M. R.; KONONENKO, I. Theoretical and empirical analysis of Relief and ReliefF. **Machine Learning**, vol. 53, p. 23-69, 2003.

SILLA JUNIOR, C.; FREITAS, A. A survey of hierarchical classification across different application domains. **Data Mining and Knowledge Discovery**, Kluwer Academic Publishers, v. 22, n. 1-2, p. 31-72, abr. 2010.

SILVA, L. V. M.; CERRI, R. Feature Selection for Hierarchical Multi-label Classification. *In:* ABREU, P.H. *et al.* (org) **Advances in Intelligent Data Analysis XIX. IDA 2021. Lecture Notes in Computer Science**, vol 12695, p. 196-208 Springer, Cham, 2021.

SINGH, B. *et al.* Optimization of feature selection method for high dimensional data using fisher score and minimum spanning tree. *In*: ANNUAL IEEE INDIA CONFERENCE. 11, 2014, Pune. **Proceedings [...]** Pune: IEEE, 2014, p. 1-6.

SIQUEIRA, R. F. **Redução de dimensionalidade em bases de dados de classificação hierárquica multirrótulo usando *autoencoders***. 2019. Dissertação (Mestrado em Ciência da Computação) – Programa de Pós-Graduação em Ciência da Computação. Universidade Tecnológica Federal do Paraná, Ponta Grossa, Paraná, 2019.

SLAVKOV, I. *et al*. ReliefF for hierarchical multi-label classification**. Lecture Notes in Computer Science**, LNAI, v. 8399, p. 148-161, 2014.

SLAVKOV, I. *et al*. HMC-ReliefF: Feature Ranking for Hierarchical Multi-label Classification. **Computer Science and Information Systems**, v. 15, n. 1, p. 187-209, 2018.

SPOLAÔR, N. *et al.* A Comparison of Multi-label Feature Selection Methods using the Problem Transformation Approach**. Electronic Notes in Theoretical Computer Science**, v. 292, p. 135-151, 2013a.

SPOLAÔR, N. *et al.* ReliefF for Multi-label Feature Selection. *In*: BRAZILIAN CONFERENCE ON INTELLIGENT SYSTEMS. 2, 2013, Fortaleza. **Proceedings [...]** Fortaleza: IEEE, 2013b, p. 6-11.

STOJANOVA, D. *et al.* Using PPI network autocorrelation in hierarchical multi-label classification trees for gene function prediction. **BMC Bioninformatics**, v. 14, n. 285, 2013.

SUN, A.; LIM, E. Hierarchical Text Classification and Evaluation. *In*: INTERNATIONAL CONFERENCE ON DATA MINING. 1, 2001, San Jose. **Proceedings [...]**, California: IEEE, 2001, p. 521-528.

SUN, L.; *et al*. Feature selection using Fisher score and multilabel neighborhood rough sets for multilabel classification. **Information Sciences**, Elsevier, v. 578, p. 887-912, 2021.

TAN, P.; STEINBACH, M.; KUMAR, V. **Introdução ao data mining:** mineração de dados. Rio de Janeiro: Ciência Moderna, 2009.

TANG, J.; ALELYANI, S.; LIU, H. Feature selection for classification: a review. *In*: AGGARWALL, C. C (org.). **Data Classification**: Algorithms and Applications. Minneapolis: CRC Press, 2014. p. 37-64.

TSOUMAKAS, G.; KATAKIS, I. Multi-Label Classification: An Overview. *In*: ERICKSON, J. **Database Technologies**: Concepts, Methodologies, Tools, and Applications. Hershey, PA: IGI Global, 2009, p.309-319.

URBANOWICZ, R. J. *et al*. Relief-based feature selection: Introduction and review. **Journal of Biomedical Informatics**, vol. 85, p. 189-203, 2018.

VAN DER MAATEN, L.; POSTMA, E.; VAN DEN HERIK, J. Dimensionality reduction: a comparative. **Journal of Machine Learning Research**, v. 10, n. 66-71, p. 13, 2009.

VENKATESH, B; ANURADHA, J. A Review of Feature Selection and Its Methods. **Cybernetics and Information Technologies**, vol. 19, n. 1, p. 3-26, 2019.

VENS, C. *et al*. Decision trees for hierarchical multi-label classification. **Machine Learning**, Springer, v. 73, n. 2, p. 185, 2008.

WITTEN I. H.; FRANK, E. **Data mining: Practical machine learning tools and techniques**. San Francisco CA, USA: Morgan Kaufmann, 2011.

YAN, S.; WONG, K.-C. Elucidating high-dimensional cancer hallmark annotation via enriched ontology. **Journal of Biomedical Informatics**, Elsevier, v.73, p. 84-94, 2017.

YU. L.; LIU, H. Feature Selection for High-Dimensional Data: A Fast Correlation-Based Filter Solution. *In*: INTERNATIONAL CONFERENCE ON MACHINE LEARNING. 20, 2003, Washington D.C. **Proceedings [...]** Washington D.C.: AAAI Press, 2003, p. 856-863.

ZHAI, Y. *et al.* A Chi-Square Statistics Based Feature Selection Method in Text Classification *In*: IEEE INTERNATIONAL CONFERENCE ON SOFTWARE ENGINEERING AND SERVICE SCIENCE. 9, 2018, Beijing. **Proceedings [...]**, Beijing: IEEE ICSESS, 2018, p. 160-163.