UNIVERSIDADE FEDERAL DE SÃO CARLOS
CAMPUS SÃO CARLOS

Luan Vinicius Moraes da Silva

Investigação de Métodos de Seleção de Atributos para Problemas de Classificação Hierárquica Multirrótulo

LUAN VINICIUS MORAES DA SILVA

Investigação de Métodos de Seleção de Atributos para Problemas de Classificação Hierárquica Multirrótulo

Trabalho de Conclusão de Curso submetido à Universidade Federal de São Carlos, como requisito necessário para obtenção do grau de Bacharel em Engenharia de Computação

São Carlos, Abril de 2023

"Desenvolvido com a permanente lembrança daqueles a quem condições pessoais e sociais não permitem desfrutar a vida por inteiro.", Catavento Museu de Ciências

Agradecimentos

"Nada realmente grande é feito sozinho". Não existe uma grande realização humana cujo mérito possa ser inteiramente pessoal, e a tarefa de concluir um curso de anos de duração e superar todos os percalços e adversidades que essa jornada implica não seria diferente. Mesmo que o progresso nesse caminho dependesse de mim, fui amparado diversas vezes, direta ou indiretamente, pelas mais diferentes entidades, sejam elas familiares, amigos, pessoas conhecidas ou desconhecidas, entidades cósmicas, instituições ou apenas o acaso.

Não citarei nominalmente todas aqui, pois corro o risco de esquecer nomes importantes. No entanto, deixo um agradecimento especial aos professores (não exclusivamente do ensino superior) que enxergam o aprendizado do estudante como critério de sucesso, à assistência estudantil da Universidade Federal de São Carlos, que torna possível que centenas de estudantes tenham condições mínimas para frequentar um ensino de ponta longe de casa e às pessoas que compartilham de forma deliberada o conhecimento.

Agradeço também meus familiares (pai e mãe), pelo esforço que fizeram para tonar esta realização possível. E aos amigos que fiz nesta jornada (do *apê* 7 e 8), que compartilharam comigo as melhores memórias que poderia ter.

Resumo

Classificação é a tarefa de atribuir exemplos de dados a classes. Na Classificação Hierárquica Multirrótulo, os exemplos podem pertencer a duas ou mais classes (rótulos) simultaneamente, onde as classes são estruturadas de forma hierárquica. A Seleção de Atributos faz parte da etapa de pré-processamento de dados e desempenha papel fundamental em tarefas de classificação para Aprendizado de Máquina, uma vez que pode reduzir de forma eficaz a dimensão do conjunto de dados, removendo atributos irrelevantes/redundantes, melhorando o desempenho preditivo do classificador. Embora muitos problemas do mundo real sejam do domínio hierárquico multirrótulo, a maioria das pesquisas relacionadas abordam a tarefa de seleção de atributos com foco em problemas monorrótulo, ou seja, de rótulo único. Em muitos trabalhos, mesmo quando a proposta aborda múltiplos rótulos, a estrutura de classes associada não é hierárquica. Portanto, neste trabalho, estudamos como a seleção de atributos pode ser empregada no contexto da Classificação Hierárquica Multirrótulo. Com esse propósito, comparamos como seletores de atributos globais conhecidos na literatura com seletores de atributos planos adaptados para estruturas hierárquicas. Os seletores de atributos globais utilizados foram Relief, Genie3 e Symbolic, e os seletores de atributos planos foram ReliefF e Information Gain. Para os seletores planos, foram adotadas estratégias para transformar o problema Hierárquico Multirrótulo em um problema multirrótulo não hierárquico, utilizando as transformações Label Powerset e Binary Relevance. Como principais resultados, os avaliadores produziram subconjuntos de atributos relevantes, aprimorando o desempenho preditivo dos classificadores enquanto reduziam a dimensionalidade do conjunto de dados original em até 75%, com destaque para os avaliadores baseados em Genie3 e Symbolic. Apesar do aprimoramento, os avaliadores planos se mostraram melhores, proporcionalmente, se comparados com os avaliadores globais.

Palavras-chave: Seleção de Atributos • Classificação Hierárquica Multirrótulo • Aprendizado de Máquina

Abstract

Classification is the task of assigning data instances to classes. In Hierarchical Multi-label Classification, instances may belong to two or more classes (labels) simultaneously, where the classes are hierarchically structured. Feature Selection is part of the data pre-processing step and plays an important role in classification tasks for Machine Learning, as it can effectively reduce the size of the dataset, removing irrelevant/redundant attributes and improving prediction performance of the classifier. Although many real-world problems are from multi-label hierarchical domain, most related research addresses the feature selection task focusing on single-label problems. In many works, even when the proposal addresses multiple labels, the associated class structure is not hierarchical. Therefore, in this work, we study how feature selection can be used in the context of Hierarchical Multi-Label Classification. For this purpose, we compare global feature selectors known in the literature with flat feature selectors adapted for hierarchical structures. The global feature selectors used were Relief, Genie3 and Symbolic, and the flat feature selectors were ReliefF and Information Gain. For flat selectors, strategies were adopted to transform the Hierarchical Multi-label problem into a non-hierarchical multi-label problem, using the Label Powerset and Binary Relevance transformations. As main results, the global evaluators produced subsets of relevant features, improving the predictive performance while reducing the original dataset by up to 75% of the original dimensionality, with emphasis on the evaluators based on the Genie3 and Symbolic set. Despite the improvement, the flat evaluators were proportionally better compared to the global evaluators.

Keywords: Feature Selection • Hierarchical Multi-label Classification • Machine Learning

Lista de abreviaturas e siglas

AM - Aprendizado de Máquina

AUPRC - *Area Under Precision-Recall Curve*

FN - *False Negative*

FS - do inglês, *Feature Selection*

FP - *False Positive*

GAD - Grafo Acíclico Direcionado

HMC - *Hierarchical Multi-label Classification*

KNN - *K Nearest Neighbor*

PCT - *Predictive Clustering Tree*

Pooled AUPRC - *Pooled Area Under Precision Recall Curve*

TP - *True Positive*

Lista de ilustrações

Lista de tabelas

Sumário

1 Introdução

Nos últimos anos, tem havido um aumento significativo na complexidade dos dados gerados e processados em diversos domínios, incluindo biologia, finanças, marketing, entretenimento digital e mídias sociais. Esses dados geralmente apresentam múltiplas classes e uma estrutura hierárquica intrínseca associada, tornando o problema pertencente ao domínio da Classificação Hierárquica Multirrótulo.

Devido à grande quantidade de atributos presentes nos dados, a Seleção de Atributos se mostra uma etapa crucial de pré-processamento dos dados e modelagem dos classificadores. Em um estudo realizado por Guyon e Elisseeff 2003, foi demonstrado que a seleção de atributos pode melhorar significativamente o desempenho de modelos de Aprendizado de Máquina em diversas tarefas de classificação, como reconhecimento de escrita manual e diagnóstico de câncer de mama. Já em um estudo mais recente realizado por Hanczar 2019, os autores mostraram que a seleção de atributos pode reduzir o custo computacional e aumentar a precisão em tarefas de classificação de imagens médicas. Além disso, a seleção de atributos pode levar a modelos mais simples e interpretáveis, permitindo a identificação dos atributos mais relevantes para a tarefa de classificação.

1.1 Problemática

Na maioria dos trabalhos de classificação encontrados na literatura, apenas uma classe é atribuída para um dado exemplo, e as classes do problema assumem uma estrutura plana (não hierárquica) [Cerri et al. 2016]. Em um estudo realizado por Arlot e Celisse 2010, os autores observaram que a maioria dos métodos de seleção de atributos é desenvolvida para dados estruturados de forma plana, em que as variáveis de entrada são independentes entre si e têm um relacionamento direto com a variável de saída (classe). No entanto, em muitas aplicações do mundo real, as classes do problema são estruturadas de forma hierárquica, ou seja, dispostas em diferentes níveis de hierarquia. Essa estrutura hierárquica pode fornecer informações adicionais e úteis para a tarefa de seleção de atributos e classificação, que podem ser exploradas para melhorar o desempenho do modelo preditivo.

Um exemplo de aplicação da seleção de atributos em conjuntos de dados hierárquicos é apresentado no estudo de Huang et al. 2019, em que os autores propuseram um método de seleção de atributos para classificação hierárquica baseada em árvores. Nesse trabalho, os autores exploraram a estrutura hierárquica das classes para selecionar atributos que preservam a consistência entre os diferentes níveis da hierarquia, melhorando o desempenho do modelo de classificação. Outro exemplo é o trabalho realizado por Wang et

al. 2022, que propõe um método de aprendizado por contraste que incorpora informações hierárquicas em um modelo de codificação de texto. O método utiliza uma estrutura de árvore hierárquica para representar as relações entre as classes e realiza a seleção de atributos hierárquicos usando aprendizado por contraste. O método é aplicado em uma tarefa de classificação de texto hierárquica multirrótulo, mostrando uma melhoria significativa na acurácia em comparação com outros métodos de classificação hierárquica multirrótulo.

Esses e outros estudos apresentados no Capitulo 2.3 destacam a importância da Seleção de Atributos em conjuntos de dados hierárquicos, e a necessidade de desenvolver métodos de seleção de atributos específicos para esses tipos de dados.

1.2 Objetivo

1.2.1 Objetivo Geral

Este trabalho tem por objetivo estudar como a tarefa de Seleção de Atributos pode ser empregada em problemas de Classificação Hierárquica Multirrótulo. Para isso, comparamos os resultados produzidos por três seletores de atributos globais conhecidos na literatura (Relief, Genie3 e Symbolic) com os resultados obtidos em [Silva e Cerri 2021], onde usamos os seletores de atributos planos ReliefF e Information Gain, adaptando a estrutura hierárquica multirrótulo para uma estrutura multirrótulo não hierárquica, utilizando as estratégias de transformação Label Powerset e Binary Relevance.

1.2.2 Resumo dos Resultados Obtidos

A partir de nossos experimentos, observamos que a estratégia de seleção de atributos com seletores planos produziu, proporcionalmente, melhores resultados se comparados com os seletores globais. A estratégia de abordagem por seletores planos se diferencia da abordagem por seletores globais pela forma como avalia os atributos, pelo descarte de atributos irrelevantes e pela transformação da estrutura hierárquica, o que pode explicar os resultados.

O algoritmos seletores de atributos globais produziram subconjuntos relevantes. Por ordem de melhores resultados estão, em primeiro lugar, o seletor Symbolic, produzindo 28 subconjuntos com avaliações acima da referência, 30 subconjuntos com avaliações igual a referência e 8 subconjuntos com avaliações abaixo da referência; em segundo lugar, o seletor Genie3, produzindo 26 subconjuntos com avaliações acima da referência, 30 subconjuntos com avaliações igual a referência e 10 subconjuntos com avaliações abaixo da referência; e em terceiro lugar, o seletor Relief, produzindo 23 subconjuntos com avaliações acima da referência, 10 subconjuntos com avaliações iguais a referência e 39 subconjuntos com avaliações abaixo da referência.

1.3 Organização

O restante deste documento está organizado da seguinte forma:

- Capítulo 2: Define a fundamentação teórica deste estudo, incluindo a tarefa de Seleção de Atributos de forma geral, os métodos usados neste trabalho, a tarefa de Classificação de Dados de forma geral e o problema da Classificação Hierárquica Multirrótulo, além dos trabalhos relacionados presentes na literatura;
- Capítulo 3: Apresenta e detalha a estratégia usada para a tarefa de Seleção de Atributos no contexto da Classificação Hierárquica Multirrótulo, além da metodologia com a apresentação dos conjuntos de dados, classificadores base e as medidas de avaliação;
- Capítulo 4: Apresenta e discute o resultados dos experimentos;
- Capítulo 5: Apresenta as conclusões e trabalhos futuros.

2 Fundamentação Teórica

2.1 Classificação de Dados

A classificação de dados é um problema de aprendizado supervisionado que tem como objetivo atribuir classes a exemplos com base em seus atributos. Segundo Mitchell e Learning 1997, a classificação de dados é uma tarefa fundamental em aprendizado de máquina, em que um modelo (classificador) é treinado com um conjunto de dados rotulados e, a partir desse aprendizado, é capaz de generalizar e classificar novos exemplos não vistos anteriormente. O resultado do treinamento é um modelo que pode ser utilizado para realizar predições em novos dados, tornando-se uma técnica importante em diversas áreas, como reconhecimento de padrões, análise de imagens, detecção de fraudes, entre outras.

2.1.1 Classificação Multirrótulo

A classificação multirrótulo é um problema classificação em que um modelo é treinado para classificar exemplos em várias classes ou rótulos simultaneamente. Diferentemente da classificação tradicional, em que cada exemplo é associado a apenas uma classe, a classificação multirrótulo permite que um exemplo possa ser associado a múltiplas classes simultaneamente [Tsoumakas e Vlahavas 2007].

Esse tipo de classificação é usado em muitos domínios, como processamento de linguagem natural, reconhecimento de imagens e bioinformática, onde a rotulagem de dados pode ser complexa e variável [Zhang e Zhou 2013]. O objetivo é que o modelo aprenda a prever todos os rótulos corretos para um exemplo, mesmo que haja uma grande variedade de combinações de rótulos possíveis [Read et al. 2011].

Como exemplo, suponha que estamos classificando animais em duas categorias: “mamíferos” e “animais aquáticos”, enquanto uma tartaruga seria classificada como “animal aquático” e um gato seria classificado como “mamífero”, um golfinho seria classificado tanto como “animal aquático” quanto “mamífero”, pois golfinhos pertencem ao grupo de animais aquáticos que são mamíferos.

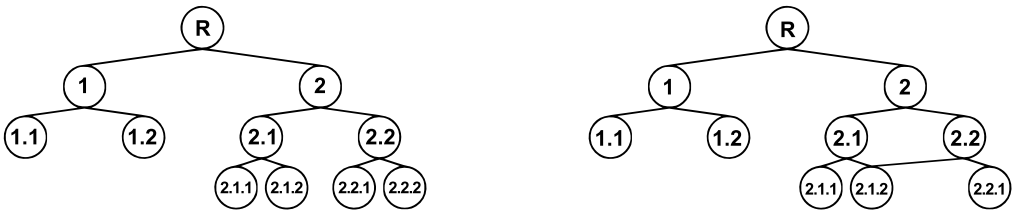
2.1.2 Classificação Hierárquica Multirrótulo

Em diversos problemas do mundo real, as classes dos problemas são organizadas em super classes e sub classes, formando uma taxonomia. Como exemplo, um complexo proteico ou organela pode ser categorizado em uma taxonomia de classe a associada com sua localização celular em uma Ontologia Genética [Consortium 2004]. Outros exemplos

de classificação podem ser encontrados na botânica ou zoologia, onde a estrutura de classificação dos seres vivos é disposta de forma hierárquica, ou ainda no campo musical, em que uma música pode ser atribuída há muitos gêneros e sub-gêneros.

Estes tipos de problemas são conhecidos na literatura de Aprendizado de Máquina como Classificação Hierárquica Multirrótulo (do inglês, *"Hierarchical Multi-label Classification"* - HMC), um caso especial da Classificação Hierárquica (do inglês *"Hierarchical Classification"* - HC), devido ao fato de um exemplo poder ser atribuído a dois ou mais caminhos na estrutura hierárquica simultaneamente. De acordo com o domínio do problema, uma estrutura hierárquica pode ser representada de duas formas, em forma de Árvore ou como um Gráfico Acíclico Direcionado (do inglês, *"Directed Acyclic Graph"* - DAG).

Em problemas hierárquicos em que a taxonomia segue uma estrutura de árvore (Figura 1a), cada nó classe tem apenas um nó pai, o que significa que cada nó tem apenas uma profundidade possível (numero de arestas entre o nó raiz e qualquer outro nó). Por outro lado, em estruturas hierárquicas no formato DAG (Figura 1b), um dado nó classe pode ter mais de um nó pai, e consequentemente múltiplos valores de profundidade. Estas características hierárquicas devem ser consideradas no desenvolvimento e avaliação do classificador hierárquico.



(a) Árvore (b) Gráfico Acíclico Direcionado

Figura 1 – Estruturas hierárquicas. Adaptado de [Silla e Freitas 2011].

Definição: Considerando $\mathbf{X}$ o espaço de exemplos, um problema hierárquico multirrótulo consiste de encontrar uma função (classificador) f para mapear cada exemplo $\mathbf{x}_i \in \mathbf{X}$ para um conjunto de classes $C_i \in C$, sendo C o conjunto de todas as classes do problema. A função f deve respeitar as restrições hierárquicas, e otimizar um critério de qualidade [Cerri et al. 2016].

A restrição hierárquica estabelece que quando uma classe é predita, todas as suas super-classes devem também serem preditas. Como exemplo, na Figura 1a, um exemplo classificado na classe 2.1.1 deve também ser classificado nas classes 2.1 e 2, simultaneamente.

A seleção de Atributos desempenha um papel importante em problemas de classificação, já que pode efetivamente reduzir a dimensionalidade do conjunto de dados removendo atributos irrelevantes/redundantes, e aprimorando o desempenho preditivo. Ainda que existam muitos problemas do mundo real hierárquicos e multirrótulo, a maioria dos

métodos de seleção de atributos na literatura lida com problemas de rótulos únicos (monorrótulo). Mesmo a maioria dos trabalhos direcionados para problemas multirrótulo são focados em cenários não hierárquicos. Assim, neste trabalho investigamos a aplicação de métodos de seleção de atributos em problemas de Classificação Hierárquica Multirrótulo focando em estruturas no formato árvore e DAG.

2.2 Seleção de Atributos

A Seleção de Atributos (do inglês, *"Feature Selection"* - FS) objetiva encontrar um número mínimo de atributos que descreva o conjunto de dados tão bem quanto o conjunto original de atributos. FS é uma etapa de pré-processamento de dados, desempenhando um importante papel em tarefas de Classificação no campo de Aprendizado de Máquina, uma vez que pode reduzir o espaço de características removendo efetivamente atributos redundantes ou irrelevantes, reduzindo o tempo de treino enquanto aumenta ou mantém o desempenho preditivo [Peralta et al. 2017, Wei et al. 2017].

A importância da Seleção de Atributos já foi comprovada na literatura: em um estudo realizado por Huang et al. 2019, os autores propuseram um método de seleção de atributos para classificação hierárquica baseado em árvores. Nesse trabalho, os autores exploraram a estrutura hierárquica dos dados para selecionar atributos que preservam a consistência entre os diferentes níveis da hierarquia, melhorando o desempenho do modelo de classificação.

Outro exemplo é o trabalho realizado por Wang et al. 2022, em que os autores prepuseram um método de aprendizado por contraste que incorpora informações hierárquicas em um modelo de codificação de texto. O método utiliza uma estrutura de árvore hierárquica para representar as relações entre as classes e realiza a seleção de atributos hierárquicos usando aprendizado por contraste. O método é aplicado em uma tarefa de classificação de texto hierárquica multirrótulo, mostrando uma melhoria significativa na acurácia em comparação com outros métodos de classificação hierárquica multirrótulo.

Há três abordagens principais para seleção de atributos: filtro, wrapper e embedded. A abordagem de filtro utiliza uma função de avaliação para medir a relevância dos atributos independentemente do modelo de aprendizado de máquina. Essa função de avaliação é aplicada a cada atributo, e um limiar de corte é utilizado para selecionar os atributos mais relevantes. Os atributos selecionados são, então, utilizados como entrada para o modelo de aprendizado de máquina.

Existem diversas funções de avaliação disponíveis, tais como correlação, informação mútua, teste estatístico, análise de componentes principais (PCA), entre outras. A escolha da função de avaliação dependerá do problema específico e das características do conjunto de dados.

A abordagem wrapper utiliza um modelo de aprendizado de máquina para avaliar a relevância dos atributos. Essa abordagem é mais custosa computacionalmente, pois envolve o treinamento de um modelo para cada subconjunto de atributos selecionados.

A abordagem wrapper utiliza uma estratégia de busca para selecionar os atributos mais relevantes. A estratégia de busca pode ser exaustiva ou heurística, como busca sequencial contínua ou para frente (forward), busca sequencial para trás (backward), ou busca baseada em algoritmos genéticos. A busca é feita através da seleção de um subconjunto de atributos, treinamento do modelo de aprendizado de máquina e avaliação do desempenho do modelo. A estratégia de busca é repetida até que um conjunto satisfatório de atributos seja encontrado.

A abordagem embedded utiliza um modelo de aprendizado de máquina que incorpora a seleção de atributos no processo de treinamento. Essa abordagem é mais eficiente em termos computacionais do que o wrapper, pois a seleção de atributos é incorporada no modelo de aprendizado de máquina.

Um exemplo de abordagem embedded é a Regressão Logística L1, que utiliza uma função de custo L1 para penalizar a magnitude dos coeficientes associados aos atributos. A penalidade L1 força alguns coeficientes a zero, eliminando assim os atributos menos relevantes. Outro exemplo é a Árvore de Decisão, que seleciona automaticamente os atributos mais relevantes ao dividir o conjunto de dados.

Em suma, a Seleção de Atributos é uma tarefa comprovadamente importante no campo de Aprendizado de Máquina. Uma vez objeto de estudo deste trabalho, descrevemos a seguir os algoritmos seletores usados neste trabalho para a tarefa de seleção de atributos.

2.2.1 Relief

A família Relief de algoritmos de seleção de atributos não usa qualquer modelo preditivo. Seus algoritmos podem lidar com várias tarefas de predição, incluindo classificação multirrótulo [Spolaôr et al. 2016], regressão [Robnik-Šikonja e Kononenko 2003] e classificação hierárquica multirrótulo [Petkovic, Dzeroski e Kocev 2020]. A ideia principal do Relief é estimar a qualidade do atributo de acordo com o quão bem seus valores distinguem entre exemplos que estão perto uns dos outros. Isto é, o atributo x_i é relevante se as diferenças no espaço de classes entre dois exemplos vizinhos são notáveis se e apenas se a diferença nos valores dos atributos de x_i entre os dois exemplos são notáveis.

De acordo com Petković, Džeroski e Kocev 2022, se $r = (x^1, y^1) \in D_{train}$ é aleatoriamente selecionado, e $n = (x^2, y^2)$ é um dos seus k vizinhos, então as importâncias computadas $importance_{Relief}(x_i)$ do algoritmo Relief são estimadas pela Equação 2.1:

$$P_1 - P_2 = P(x_i^1 \neq x_i^2 \mid y^1 \neq y^2) - P(x_i^1 \neq x_i^2 \mid y^1 = y^2) \tag{2.1}$$

em que as probabilidades são modeladas pela distância entre r e n no subespaço apropriado. Para o espaço descritivo X estendido pelos domínios X_i dos atributos x_i, temos (Equação 2.2):

$$d_x(x^1, x^2) = \frac{1}{F}\sum_{i=1}^{F} d_i(x^1, x^2); d_i(x^1, x^2) = \begin{cases} 1[x_i^1 \neq x_i^2] & : X_i \nsubseteq \mathbb{R} \\ \frac{|x_i^1 - x_i^2|}{\max_x x_i - \min_x x_i} & : X_i \subseteq \mathbb{R} \end{cases} \tag{2.2}$$

em que 1 denota a função indicador. A definição da distância do espaço de classes d_y depende do domínio da função. No caso da classificação e problemas de regressão multirrótulo, as parte categórica e numérica da definição d_i na Equação 2.2 são aplicadas. De forma semelhante, em tarefas de regressão multirrótulo, d_y é o análogo de d_x.

Para os problemas de Classificação Multirrótulo e Hierárquica Multirrótulo, há mais de uma opção para a definição de distância da classe [Petković, Kocev e Džeroski 2018]. Neste trabalho é usada a distância de Hamming entre os dois conjuntos. Seja $S \subseteq \mathbb{L}$ representado como um vetor binário s, a distância de Hamming é dada pela Equação 2.3:

$$d_y(s^1, s^2) = \gamma \sum_{i=1}^{|\mathbb{L}|} \alpha_i 1[s_i^1 \neq s_i^2] \tag{2.3}$$

em que os pesos α_i são definidos com base na hierarquia, γ é o fator normalização que assegura que d_y mapeie para $[0, 1]$. Essa fator e igual a $\frac{1}{|\mathbb{L}|}$ em problemas de regressão multirrótulo e depende dos dados em problemas de classificação hierárquica multirrótulo [Petkovic, Dzeroski e Kocev 2020].

Para estimar as probabilidades condicionais da Equação 2.1, elas são primeiro expressadas na forma incondicional, por exemplo, $P_1 = P(x_i^1 \neq x_i^2 \wedge y^1 \neq y^2)/P(y^1 \neq y^2)$. Por fim, o numerador é modelado como o produto $d_i d_y$, eo denominador como d_y.

2.2.2 Symbolic

Symbolic Ranking é um algoritmo de seleção de atributos que se baseia em uma abordagem simbólica, onde os atributos são considerados como variáveis simbólicas e a seleção é realizada com base em um ranking dessas variáveis [Liu e Motoda 1998, Liu et al. 2010]. O algoritmo é útil para a seleção de atributos em problemas de classificação com um grande número de atributos, pois é capaz de selecionar os atributos mais relevantes de forma eficiente. No entanto, o algoritmo pode não funcionar bem em casos em que os atributos têm interações complexas entre si ou em que a distribuição de classes é desbalanceada.

De acordo com Petkovic, Dzeroski e Kocev 2020, a pontuação simbólica é calculada com base na frequência de cada valor do atributo em cada classe. Essa frequência é usada

para computar o ganho de informação simbólico, que é uma medida de quanta informação um atributo fornece sobre a classe. Quanto maior o ganho de informação simbólico, maior a pontuação simbólica do atributo (Equação 2.4). Como os atributos mais próximos da raiz são tidos como mais importantes que os atributos mais distantes, a ocorrência dos atributos é ponderada pelo número de amostras $e(N)$ que são atribuídas a um nó N.

$$importance_{SYMB}(x_i) = \frac{1}{|FO|} \sum TR \in FO \sum N \in TR(x_i) e(N)/|D_{train}| \tag{2.4}$$

Na Equação 2.4, TR representa uma Árvore, $N \in TR$ denota um nó, FO denota uma floresta (conjunto de Árvores), o conjunto de todos os nós internos de uma árvore em que o atributo x_i aparece como parte do teste é representado por $TR(x_i)$, e D_{train} representa o conjunto de dados original.

2.2.3 Genie3

Genie3 (Gene Network Inference with Ensemble of trees 3) é um algoritmo de seleção de atributos baseado em comitês (ensemble) de árvores de decisão aleatórias. Cada árvore é construída a partir de uma amostra aleatória dos dados de entrada e de um subconjunto aleatório de atributos, e a importância de cada atributo é determinada em cada árvore, medindo a redução média da impureza. Essas medidas de importância são usadas para calcular um ranking de atributos, em que os atributos mais importantes são aqueles que aparecem com mais frequência nas diferentes árvores do conjunto. O algoritmo também permite a identificação de relações de dependência entre os atributos, por meio da análise das correlações entre as medidas de importância.

O Genie3 é frequentemente utilizado em problemas de inferência de redes de genes a partir de dados de expressão gênica [Huynh-Thu et al. 2010, Maduranga et al. 2013]. Além de ser capaz de lidar com grandes conjuntos de dados, ele é capaz de capturar relações não-lineares e interações entre os atributos.

Ainda sobre o trabalho de Petkovic, Dzeroski e Kocev 2020, a importância do algoritmo seletor de atributos é definida pela Equação 2.5:

$$importance_{GENIE3}(x_i) = \frac{1}{|FO|} \sum TR \in FO \sum N \in TR(x_i) h^* \tag{2.5}$$

em que h^* é o valor heurístico da função redutora de variância. Na equação, h^* é proporcional a $e(N) = |E|$, portanto atribuindo maior importância aos atributos frequentemente próximos do nó raiz da árvore.

2.3 Trabalhos Relacionados

2.3.1 Classificação Hierárquica Multirrótulo

A seleção de atributos voltada para problemas de Classificação Hierárquica Multirrótulo tem ganhado atenção da literatura de Aprendizado de Máquina. Por isso, esta seção apresenta algumas propostas de trabalho de Seleção de Atributos em problemas Hierárquicos e Não Hierárquicos Multirrótulo.

No trabalho de Amazal, Ramdani e Kissi 2020, os autores abordaram a tarefa de Seleção de Atributos para problemas Multirrótulo não Hierárquicos propondo uma abordagem com uso do Chi-Square ponderado, intitulada *Distributed Category Term Frequency Based on Chi-square* (CTF-CHI), usando um classificador Multinomial Nive Bayes (MNB) para avaliar a eficiência do subconjunto de atributos selecionado. Os autores aplicaram a Seleção de Atributos transformando o problema Multirrótulo não Hierárquico em um problema monorrótulo.

Petković, Kocev e Džeroski 2020 investigou dois grupos de seletores de atributos para a tarefa de *multi-target regression* (MTR), através do estudo dos ranqueadores de características (*Symbolic*, *Genie3*, e *Random Forest scores*) baseado nos agrupadores (*bagging*, *random forest* e *extra trees*) de árvores de agrupamento preditivo, e um score derivado do métodos RReliefF. Problemas MTR consistem de múltiplas variáveis alvos contínuas, isto é, pertencentes ao domínio dos números reais, onde o objetivo é aprender um modelo para prever todas as variáveis simultaneamente para um dado exemplo.

Slavkov et al. 2018 endereçou o ranqueamento de características no contexto da Classificação Hierárquica Multirrótulo focando no estimador de importância ReliefF, uma continuação de seu trabalho anterior [Slavkov et al. 2013]. O autor testou sua proposta em cinco conjuntos de dados do campo biológico e de imagens, e obteve melhores resultados quando comparado com o algoritmo de pontuação de características baseado em *Binary Relevance*.

Ainda relacionado ao ranqueamento de atributos, Petkovic, Dzeroski e Kocev 2020 estendeu seu trabalho anterior *Multi-target regression* [Petković, Kocev e Džeroski 2020] para o contexto da Classificação Hierárquica Multirrótulo. Os autores aplicaram um grupo de abordagens de pontuadores de atributos baseados nas funções avaliadoras *Symbolic*, *Genie3* e *Random Forest*, combinadas com as árvores de agrupamento preditivo *Bagging*, *Random Forest* e *Extra Trees*. Os autores avaliaram suas abordagens em 30 conjuntos de dados reais multirrótulo estruturados de forma hierárquica através de um modelo KNN que considera a importância do atributo na função distância. Os resultados obtidos superaram o método seletor de atributos HMC-Relief, e mostraram que os ranqueadores *Symbolic* e *Genie3* levam a subconjunto de atributos relevantes.

3 Proposta e Metodologia

3.1 Proposta

Uma vez aplicado um método de seleção de atributos, faz-se necessário estabelecer um critério de escolha dos melhores atributos por meio da definição de um limiar. A definição de um limiar para a seleção de atributos pode ser um desafio, pois pode afetar significativamente o desempenho do modelo de classificação. Se o limiar for muito alto, pode resultar na seleção de um número muito pequeno de atributos, o que pode levar a perda de informação e a um desempenho inferior do modelo. Por outro lado, se o limiar for muito baixo, pode resultar na seleção de muitos atributos, incluindo atributos irrelevantes ou redundantes, o que pode levar a um modelo de aprendizado de máquina com alta variância e baixa capacidade de generalização.

Outro desafio na seleção de atributos é que diferentes conjuntos de dados e problemas de aprendizado podem exigir diferentes limiares. O limiar ideal para um conjunto de dados pode não ser o mesmo para outro conjunto de dados. Além disso, a seleção de atributos é muitas vezes uma tarefa exploratória, que requer experimentação e ajuste do limiar para encontrar o melhor conjunto de atributos para um determinado problema.

Em resumo, os principais desafios em estipular um limiar para a tarefa de seleção de atributos são encontrar um equilíbrio entre o número de atributos selecionados e o desempenho do modelo de aprendizado de máquina, ajustar o limiar para diferentes conjuntos de dados e problemas de aprendizado de máquina, e lidar com a natureza exploratória da seleção de atributos.

Com isto em mente, definimos um conjunto de limiares específicos para cada conjunto de dados, com base na abordagem de partição por quartis. Para isso, com os atributos selecionados e ordenados, da menor pontuação para a maior, selecionamos três limiares com base na distribuição da pontuação dos atributos, isto é, o primeiro limiar é o Q1, que corresponde aos atributos que obtiveram pontuação acima do primeiro quartil. O segundo limiar é o Q2, que corresponde aos atributos que pontuaram acima do segundo quartil (ou Mediana), e por fim, o terceiro limiar corresponde aos atributos que pontuaram acima do terceiro quartil.

Como resultado final, são gerados três subconjuntos, Q1, Q2 e Q3, com os top 75% dos atributos selecionados, 50% selecionados e 25% selecionados, respectivamente. A distribuição por quartis pode ser vista na representação de uma caixa (diagrama) boxplot, exemplificada na Figura 2.

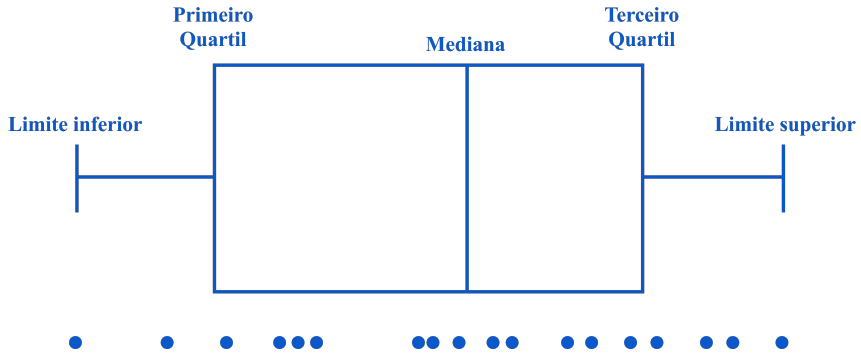


Figura 2 – Partições por quartis em um diagrama Boxplot.

3.2 Metodologia

Esta seção apresenta os conjuntos de dados utilizados em nossos experimentos, o classificador base usado para validar os experimentos e as medidas de avaliação.

3.2.1 Conjunto de Dados

Usamos 24 conjuntos de dados hierárquicos multirrótulo reais compostos de funções proteicas, estruturados em forma de Árvore (FunCat) e DAG (GO). Estes conjuntos de dados são frequentemente usados para avaliar classificadores hierárquicos multirrótulo [Nakano, Lietaert e Vens 2019], e são disponibilizados publicamente[1]. Estes conjuntos de dados são particionados em conjuntos de treino, validação e teste. A Tabela 1 apresenta as principais características dos conjuntos de dados usados.

Muitos métodos na literatura, tais como [Clare 2003, Vens et al. 2008, Cerri et al. 2016, Wehrmann, Cerri e Barros 2018], apresentaram o resultado de suas propostas nestes conjuntos de dados, treinando seus modelos em 2/3 de cada conjunto e testando nos 1/3 restantes. Neste trabalho, usamos essas mesmas partições de dados, unindo conjunto de treino e validação, e testando no conjunto de teste remanescente.

3.2.2 Classificador Base

Clus-HMC [Vens et al. 2008] é um algoritmo que constrói classificadores baseados em árvores de decisão. É baseado no conceito de Predictive Clustering Trees (PCTs), em que o nó raiz corresponde ao cluster (conjunto) contendo todos os dados de treino, que por sua vez é recursivamente particionado em clusters menores enquanto a árvore cresce até os nós folhas. PCTs são construídas de maneira a minimizar a variância dentro de cada cluster.

Para aplicar PCT à tarefa de Classificação Hierárquica Multirrótulo, primeiro, os rótulos dos exemplos são representados como vetores binários, isto é, o i-ésimo compo-

[1] <https://itec.kuleuven-kulak.be/?page_id=5236>

Tabela 1 – Conjuntos de dados: Atributos ($|A|$); Exemplos ($|X|$); Tipo de Atributo (Tipo): Numérico (Quanti), Qualitativo (Quali) ou ambos (Misto).

No.	Conjunto de Dados	$\|A\|$	Type	$\|X\|$		
				Treino	Validação	Teste
D1	Cellcycle_Fun	77	Quanti	1628	848	1281
D2	Cellcycle_GO	77	Quanti	1620	844	1270
D3	Church_Fun	27	Misto	1630	844	1281
D4	Church_GO	27	Misto	1622	840	1269
D5	Derisi_Fun	63	Quanti	1608	842	1275
D6	Derisi_GO	63	Quanti	1589	832	1249
D7	Eisen_Fun	79	Quanti	1058	529	837
D8	Eisen_GO	79	Quanti	1053	526	831
D9	Expr_Fun	551	Misto	1639	849	1291
D10	Expr_GO	551	Misto	1631	845	1280
D11	Gasch1_Fun	173	Quanti	1634	846	1284
D12	Gasch1_GO	173	Quanti	1626	842	1273
D13	Gasch2_Fun	52	Quanti	1639	849	1291
D14	Gasch2_GO	52	Quanti	1631	845	1280
D15	Hom_Fun	47034	Quali	1669	870	1315
D16	Hom_GO	47034	Quali	1656	863	1301
D17	Pheno_Fun	69	Quali	656	353	582
D18	Pheno_GO	69	Quali	650	350	576
D19	Seq_Fun	478	Misto	1701	879	1339
D20	Seq_GO	478	Misto	1686	832	1324
D21	Spo_Fun	80	Misto	1600	837	1266
D22	Spo_GO	80	Misto	1591	832	1250
D23	Struc_Fun	19628	Quali	4806	2515	3819
D24	Struc_GO	19628	Quali	1653	855	1298

nente do vetor é sinalizado com 1 se o exemplo pertencer à classe c_i, e 0 caso contrário. Para analisar a variância no contexto hierárquico multirrótulo, o Clus-HMC considera a similaridade em níveis hierárquicos mais altos como mais importante que a similaridade em níveis mais baixos. Essa importância é aplicada ponderando as classes no cálculo da distância Euclidiana entre os vetores binários. Neste trabalho, usamos a implementação do Clus-HMC provida dentro do framework Clus[2]. Para maiores detalhes sobre o Clus-HMC, sugerimos a leitura do artigo de Vens et al. 2008.

3.2.3 Medidas de Avaliação

Para avaliar nossos resultados, usamos medidas baseadas em curvas Precisão-Revocação (Precision-Recall) (PR). Devido ao Clus-HMC produzir como saída valores numéricos que podem ser interpretados com as probabilidades dos exemplos pertencerem às classes, podemos usar diferentes valores de limiar em um intervalo [0.0,1.0] a fim de produzir pontos PR, gerando uma curva de precisão em função de sua revocação. Neste trabalho, usamos uma variante da curva PR, definida a seguir:

- **Área Agrupada sob as curvas PR** ou *Pooled Area Under Precision Recall Curve* (*Pooled AUPRC*) é uma medida comumente adotada na literatura para tarefas hierárquicas multirrótulo. Devido ao fato de dados hierárquicos serem muito desba-

[2] <https://dtai.cs.kuleuven.be/clus/>

lanceados, a quantidade de classificações negativas para as classes menos frequentes tende a ser maior que a quantidade de classificações positivas, com a possibilidade de levar a falsos negativos. Portanto, apenas a medida AUPRC não é recomendada [Nakano, Lietaert e Vens 2019]. *Pooled AUPRC* corresponde à área ponderada sob as curvas de precisão-revocação. Essa curva é gerada tomando a precisão e a revocação ponderadas de cada classe, para diferentes limiares. Os limiares variam no intervalo $[0.0, 1.0]$ e são incrementados em passos de 0.02. Em resumo, quanto mais próximo de 1.0, melhor o modelo considerado.

Os cálculos da Precisão e Revocação são apresentados nas Equações 3.1 e 3.2, respectivamente. Nas equações, TP refere-se à quantidade de verdadeiros positivos, FP refere-se à quantidade de falsos positivos, e FN à quantidade de falsos negativos. A precisão visa identificar a proporção de classificações positivas que estão de fato corretas, enquanto a revocação busca identificar a proporção de verdadeiros positivos classificados corretamente.

$$Precisão = \frac{TP}{TP + FP} \tag{3.1}$$

$$Revocação = \frac{TP}{TP + FN} \tag{3.2}$$

4 Resultados

A Tabela 2 apresenta os resultados da tarefa de classificação com diferentes números de atributos selecionados, como descrito na estratégia de partição por quartis. Executamos o classificador Clus-HMC com os melhores valores de hiperparâmetros conforme sugerido por Vens et al. 2008[1].

Os experimentos foram realizados em um Sistema Operacional Windows 10 Pro 64 bits, 16Gb de memória RAM, 512Gb de HD e processador AMD Ryzen 5 4600H com 12 núcleos.

A Tabela 2 contem a seguinte notação:

- **Referência:** valor resultado da classificação sem a aplicação de qualquer método de seleção de atributos;
- $|A|$**:** Espacialidade do subconjunto, ou, quantidade de atributos resultante da seleção;
- **Q1:** O subconjunto de dados resultado da seleção dos top 75% atributos (valor referente aos atributos posicionados acima do primeiro quartil);
- **Q2:** O subconjunto de dados resultado da seleção dos top 50% atributos (valor referente aos atributos posicionados acima da mediana);
- **Q3:** O subconjunto de dados resultado da seleção dos top 25% atributos (valor referente aos atributos posicionados acima do terceiro quartil);
- Valores não computados são representados com traços (-);
- As pontuações acima da referência, e portanto os melhores valores, são destacadas em negrito.

Tabela 2 – Resultados: *Pooled Area Under Precision Recall Curve.*

Conjunto de Dados	Referência	Partição	|A|	Relief	Symbolic	Genie3
Cellcycle Fun	0,194	Q1	59	**0,195**	**0,2**	0,194
		Q2	40	**0,199**	**0,198**	**0,195**
		Q3	21	**0,196**	**0,197**	**0,199**

[1] <https://dtai.cs.kuleuven.be/clus/hmcdatasets/ftests.txt>

Tabela 2 - continuação da página anterior.

Conjunto de Dados	Referência	Partição	|A|	Relief	Symbolic	Genie3
Cellcycle GO	0,363	Q1	59	**0,364**	0,363	**0,365**
		Q2	40	**0,367**	**0,364**	**0,364**
		Q3	21	0,358	0,363	0,363
Church Fun	0,187	Q1	22	0,187	0,187	0,187
		Q2	15	0,184	0,187	0,186
		Q3	8	**0,194**	**0,19**	**0,19**
Church GO	0,358	Q1	22	**0,361**	0,358	0,358
		Q2	15	**0,36**	0,358	0,358
		Q3	8	0,353	**0,36**	**0,36**
Derisi Fun	0,198	Q1	49	**0,199**	0,198	0,198
		Q2	33	**0,201**	0,198	0,198
		Q3	17	0,184	0,198	0,198
Derisi GO	0,362	Q1	49	0,362	0,362	0,362
		Q2	33	0,362	0,362	0,362
		Q3	17	0,362	0,361	0,359
Eisen Fun	0,232	Q1	61	0,22	0,23	0,23
		Q2	41	0,227	**0,233**	0,231
		Q3	21	0,214	**0,24**	**0,241**
Eisen GO	0,383	Q1	61	**0,388**	**0,384**	0,382
		Q2	41	0,381	**0,385**	**0,384**
		Q3	21	0,377	0,38	0,381
Expr Fun	0,221	Q1	415	0,218	0,221	0,221
		Q2	277	0,216	0,22	0,22
		Q3	139	0,215	0,221	0,221
Expr GO	0,38	Q1	415	0,379	0,379	0,379
		Q2	277	0,38	**0,382**	0,38
		Q3	139	0,374	**0,384**	**0,383**
Gasch1 Fun	0,219	Q1	131	0,218	0,219	0,219
		Q2	88	0,214	0,219	0,219
		Q3	46	0,214	0,218	0,217
Gasch1 GO	0,379	Q1	131	0,379	**0,38**	**0,38**
		Q2	88	**0,381**	**0,381**	**0,381**
		Q3	46	**0,382**	**0,383**	**0,382**
Gasch2 Fun	0,208	Q1	40	**0,209**	**0,21**	**0,211**
		Q2	27	0,203	**0,21**	**0,21**
		Q3	14	0,202	**0,214**	**0,212**
Gasch2 GO	0,377	Q1	40	0,375	0,377	0,377

Tabela 2 - continuação da página anterior.

Conjunto de Dados	Referência	Partição	|A|	Relief	Symbolic	Genie3
		Q2	27	0,373	**0,378**	0,377
		Q3	14	0,366	0,377	**0,378**
		Q1	35277	**0,315**	**0,302**	**0,302**
Hom Fun	0,3	Q2	23518	**0,303**	**0,303**	**0,303**
		Q3	11760	0,278	**0,308**	**0,308**
		Q1	35277	0,237	-	-
Hom GO	0,244	Q2	23518	0,228	-	-
		Q3	11760	0,211	-	-
		Q1	53	**0,18**	0,173	0,173
Pheno Fun	0,173	Q2	36	**0,18**	0,173	0,173
		Q3	19	**0,18**	0,173	0,173
		Q1	53	0,346	0,346	0,346
Pheno GO	0,346	Q2	36	0,346	0,346	0,346
		Q3	19	0,346	0,346	0,346
		Q1	360	0,224	0,233	0,233
Seq Fun	0,233	Q2	240	0,223	**0,236**	**0,235**
		Q3	121	0,208	**0,235**	**0,242**
		Q1	360	**0,409**	0,408	0,408
Seq GO	0,408	Q2	240	0,398	0,408	0,408
		Q3	121	0,394	0,408	0,408
		Q1	61	0,202	**0,203**	**0,203**
Spo Fun	0,202	Q2	41	0,195	**0,203**	**0,203**
		Q3	21	0,189	0,201	0,201
		Q1	61	0,35	0,365	0,365
Spo GO	0,365	Q2	41	0,36	0,365	0,364
		Q3	21	0,363	0,364	**0,366**
		Q1	14722	**0,207**	0,181	0,181
Struc Fun	0,181	Q2	9815	**0,203**	**0,216**	**0,216**
		Q3	4908	**0,195**	**0,207**	**0,207**
		Q1	14722	0,148	-	-
Struc GO	0,156	Q2	9815	0,148	-	-
		Q3	4908	0,147	-	-

De acordo com os resultados dos experimentos registrados da Tabela 2, o seletor de atributos Relief produziu 23 subconjuntos (dentre os 72) com pontuações *Pooled AUPRC* acima da referência, o seletor Symbolic produziu 28 subconjuntos (dentre os 66) com pontuações *Pooled AUPRC* acima da referência, e o seletor Genie3 produziu 26 subconjuntos

(dentre os 66) com pontuações *Pooled AUPRC* acima da referência.

O seletor Relief produziu 10 subconjuntos (dentre os 72) com pontuação *Pooled AUPRC* igual à referência enquanto reduziu o espaço de atributos original em até 75%. O seletor Symbolic produziu 30 subconjuntos (dentre os 66) com pontuação *Pooled AUPRC* igual à referência enquanto reduziu o espaço de atributos original em até 75%. O seletor Genie3 produziu 30 subconjuntos (dentre os 66) com pontuação *Pooled AUPRC* igual à referência enquanto reduziu o espaço de atributos original em até 75%.

A Tabela 3 apresenta os resultados obtidos em um trabalho que publicamos na conferência International Symposium on Intelligent Data Analysis (IDA 2021) em Porto, Portugal [Silva e Cerri 2021]. Naquele trabalho estudamos como a seleção de atributos pode ser empregada no contexto da classificação hierárquica multirrótulo com foco em estruturas baseadas em árvore (FunCat), fazendo a transformação da estrutura hierárquica em n estruturas planas, sendo n o número de níveis da árvore. Em seguida empregamos um seletor de atributos para problemas multirrótulo planos em cada um dos n níveis, e reconstruímos a hierarquia com os atributos selecionados em cada nível, seguindo como critério de limiar a mesma abordagem por quartis descrita no Capitulo 3.

A Tabela 3 apresenta os resultados de [Silva e Cerri 2021], utilizando mesma notação da Tabela 2. Os seguintes seletores foram avaliados:

- **RF-BR**: Seletor ReliefF baseado na estratégia de transformação Binary-Relevance;
- **RF-LP**: Seletor ReliefF baseado na estratégia de transformação Label-Powerset;
- **IG-BR**: Seletor Information Gain baseado na estratégia de transformação Binary-Relevance;
- **IG-LP**: Seletor Information Gain baseado na estratégia de transformação Label-Powerset.

Tabela 3 – Resultados: *Pooled Area Under Precision Recall Curve [Silva e Cerri 2021].*

Conjunto de Dados	Referência	Partição	IG-BR	RF-BR	IG-LP	RF-LP
Cellcycle Fun	0,183	Q1	**0,184**	0,183	0,173	0,179
		Q2	**0,190**	**0,189**	0,172	0,179
		Q3	**0,183**	0,182	0,173	**0,191**
Church Fun	0,183	Q1	0,183	**0,186**	0,178	0,183
		Q2	**0,187**	**0,188**	0,177	**0,185**
		Q3	**0,191**	0,175	0,177	0,179
Derisi Fun	0,157	Q1	**0,164**	**0,162**	-	0,156
		Q2	**0,167**	**0,157**	-	0,152

Tabela 3 - continuação da página anterior.

Conjunto de Dados	Referência	Partição	IG-BR	RF-BR	IG-LP	RF-LP
		Q3	**0,175**	**0,158**	-	**0,167**
Eisen Fun	0,205	Q1	0,201	0,205	0,200	0,205
		Q2	**0,210**	0,204	0,188	**0,213**
		Q3	**0,221**	**0,214**	0,188	**0,232**
Expr Fun	0,171	Q1	**0,174**	**0,185**	**0,198**	**0,185**
		Q2	**0,184**	**0,186**	**0,209**	**0,183**
		Q3	**0,195**	**0,200**	**0,210**	**0,202**
Gasch1 Fun	0,195	Q1	**0,199**	**0,201**	**0,213**	0,192
		Q2	0,194	**0,197**	**0,212**	0,193
		Q3	**0,200**	0,189	**0,19**	0,192
Gasch2 Fun	0,180	Q1	**0,186**	**0,185**	0,175	**0,180**
		Q2	0,178	**0,186**	0,169	**0,188**
		Q3	**0,186**	**0,190**	0,169	**0,185**
Pheno Fun	0,170	Q1	0,170	0,170	0,170	0,170
		Q2	0,170	0,170	0,170	0,170
		Q3	**0,171**	**0,172**	**0,172**	**0,172**
Seq Fun	0,207	Q1	0,205	0,206	0,207	0,207
		Q2	0,202	**0,209**	0,203	0,201
		Q3	0,198	0,198	0,195	0,184
Spo Fun	0,182	Q1	**0,185**	0,182	0,173	0,182
		Q2	**0,191**	**0,188**	0,173	**0,189**
		Q3	**0,199**	**0,197**	0,173	**0,197**

Podemos observar, por ordem de melhores resultados, que a estratégia $IG-BR$ produziu aprimoramento preditivo em 21 dos 30 conjuntos de dados, a estratégia $RF-BR$ produziu aprimoramento preditivo em 19 dos 30 conjuntos de dados, a estratégia $RF-LP$ produziu aprimoramento preditivo em 13 dos 30 conjunto de dados, e por fim, a estratégia $IG-LP$ produziu aprimoramento preditivo em 7 dos 27 conjuntos de dados.

É importante enfatizar que os seletores de atributos propostos em [Silva e Cerri 2021] e os usados neste trabalho são diferentes, isto é, os seletores usados aqui levam em consideração a estrutura hierárquica e ranqueiam os atributos, enquanto os propostos em [Silva e Cerri 2021] são direcionados para estruturas planas podendo atribuir uma pontuação nula para um atributo, significando que o atributo não tem informação relevante sobre o conjunto de dados.

Outra distinção entre [Silva e Cerri 2021] e o trabalho atual está na forma de selecionar os atributos. No trabalho atual os seletores ranqueiam os atributos e subpartições

são selecionadas. Por isso, na Tabela 2, a subpartição Q1 do conjunto de dados Cellcycle Fun tem 59 atributos como resultado dos três seletores. Já no trabalho de Silva e Cerri 2021, os seletores podem produzir avaliações nulas de um atributo. Estes atributos são descartados, gerando como resultado uma subpartição com número de atributos diferentes para cada seletor. Isso pode ser observado na Figura 3, em que a estratégia **IG-BR** selecionou 90% dos atributos do conjunto original na subpartição Q1, enquanto na mesma subpartição, a estratégia **RF-BR** selecionou quase 100% dos atributos.

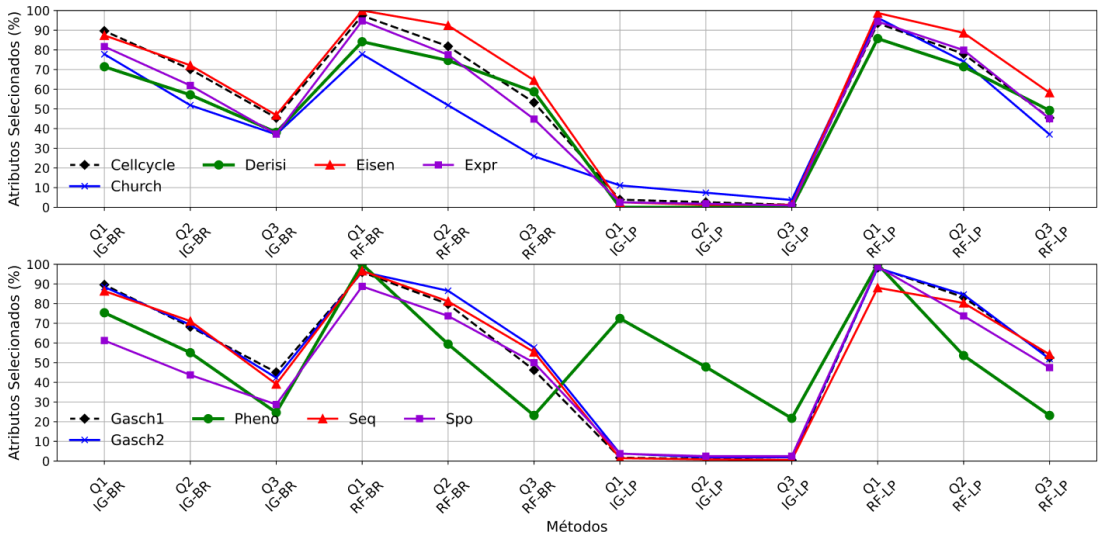


Figura 3 – Porcentagem de atributos selecionados em função do método e partição. Adaptado de [Silva e Cerri 2021]

5 Conclusão

Neste trabalho, investigamos como a seleção de atributos pode ser empregada no contexto da Classificação Hierárquica Multirrótulo, através da comparação de seletores de atributos globais conhecidos na literatura com seletores de atributos planos adaptados para estruturas hierárquicas. Os seletores de atributos globais investigados foram Relief, Genie3 e Symbolic, e os seletores de atributos planos foram ReliefF e o Information Gain adotando como estratégia de adaptação a transformação do problema Hierárquico Multirrótulo para multirrótulo não hierárquico, com os transformadores Label Powerset e Binary Relevance.

A partir de nossos experimentos, observamos que a estratégia de seleção de atributos com seletores planos produziu, proporcionalmente, melhores resultados se comparado com os seletores globais. A estratégia de abordagem por seletores planos se diferencia da abordagem por seletores globais pela forma como avalia os atributos, pelo descarte de atributos irrelevantes e pela transformação da estrutura hierárquica, o que pode explicar os resultados.

Os algoritmos seletores de atributos globais produziram subconjuntos relevantes. Por ordem de melhores resultados estão em primeiro lugar, o seletor Symbolic, produzindo 28 subconjuntos com avaliações acima da referência, 30 subconjuntos com avaliações iguais à referência, e 8 subconjuntos com avaliações abaixo da referência; em segundo lugar, o seletor Genie3, produzindo 26 subconjuntos com avaliações acima da referência, 30 subconjuntos com avaliações iguais à referência, e 10 subconjuntos com avaliações abaixo da referência; e em terceiro lugar, o seletor Relief, produzindo 23 subconjuntos com avaliações acima da referência, 10 subconjuntos com avaliações iguais à referência, e 39 subconjuntos com avaliações abaixo da referência.

Como pode ser constatado deste trabalho, a tarefa de Seleção de Atributos no contexto da Classificação Hierárquica Multirrótulo ainda precisa ser muito investigada. Genie3 e Symbolic são seletores baseados em conjuntos, enquanto Relief é um método de seleção estatística. De acordo com os resultados apresentados, os seletores em conjuntos se sobressaíram. Assim, este trabalho nos direciona para a exploração dos métodos baseados em conjunto e como suas características podem ser compartilhadas com os métodos puramente estatísticos.

Referências

AMAZAL, H.; RAMDANI, M.; KISSI, M. Towards a feature selection for multi-label text classification in big data. In: SPRINGER. *International Conference on Smart Applications and Data Analysis*. [S.l.], 2020. p. 187–199. Citado na página 29.

ARLOT, S.; CELISSE, A. A survey of cross-validation procedures for model selection. 2010. Citado na página 19.

CERRI, R. et al. Reduction strategies for hierarchical multi-label classification in protein function prediction. *BMC bioinformatics*, Springer, v. 17, n. 1, p. 373, 2016. Citado 3 vezes nas páginas 19, 24 e 32.

CLARE, A. Machine learning and data mining for yeast functional genomics. *Aberystwyth: The University of Wales.(Doctor of Philosophy)*, 2003. Citado na página 32.

CONSORTIUM, G. O. The gene ontology (go) database and informatics resource. *Nucleic Acids Research*, v. 32, n. suppl_1, p. D258–D261, 2004. ISSN 0305-1048. Citado na página 23.

GUYON, I.; ELISSEEFF, A. An introduction to variable and feature selection. *Journal of machine learning research*, v. 3, n. Mar, p. 1157–1182, 2003. Citado na página 19.

HANCZAR, B. Performance visualization spaces for classification with rejection option. *Pattern Recognition*, Elsevier, v. 96, p. 106984, 2019. Citado na página 19.

HUANG, W. et al. Hierarchical multi-label text classification: An attention-based recurrent network approach. In: *Proceedings of the 28th ACM international conference on information and knowledge management*. [S.l.: s.n.], 2019. p. 1051–1060. Citado 2 vezes nas páginas 19 e 25.

HUYNH-THU, V. A. et al. Inferring regulatory networks from expression data using tree-based methods. *PloS one*, Public Library of Science San Francisco, USA, v. 5, n. 9, p. e12776, 2010. Citado na página 28.

LIU, H.; MOTODA, H. Feature transformation and subset selection. *IEEE Intell Syst Their Appl*, Citeseer, v. 13, n. 2, p. 26–28, 1998. Citado na página 27.

LIU, H. et al. Feature selection: An ever evolving frontier in data mining. In: PMLR. *Feature selection in data mining*. [S.l.], 2010. p. 4–13. Citado na página 27.

MADURANGA, D. et al. Inferring gene regulatory networks from time-series expressions using random forests ensemble. In: SPRINGER. *Pattern Recognition in Bioinformatics: 8th IAPR International Conference, PRIB 2013, Nice, France, June 17-20, 2013. Proceedings 8*. [S.l.], 2013. p. 13–22. Citado na página 28.

MITCHELL, T. M.; LEARNING, M. Mcgraw-hill science/engineering/math, march 1997. *Chap. II*, 1997. Citado na página 23.

NAKANO, F. K.; LIETAERT, M.; VENS, C. Machine learning for discovering missing or wrong protein function annotations. *BMC bioinformatics*, Springer, v. 20, n. 1, p. 485, 2019. Citado 2 vezes nas páginas 32 e 34.

PERALTA, D. et al. Distributed incremental fingerprint identification with reduced database penetration rate using a hierarchical classification based on feature fusion and selection. *Knowledge-Based Systems*, Elsevier, v. 126, p. 91–103, 2017. Citado na página 25.

PETKOVIC, M.; DZEROSKI, S.; KOCEV, D. Feature ranking for hierarchical multi-label classification with tree ensemble methods. *Acta Polytechnica Hungarica*, v. 17, n. 10, p. 129–148, 2020. Citado 4 vezes nas páginas 26, 27, 28 e 29.

PETKOVIĆ, M.; DŽEROSKI, S.; KOCEV, D. Feature ranking for semi-supervised learning. *Machine Learning*, Springer, p. 1–30, 2022. Citado na página 26.

PETKOVIĆ, M.; KOCEV, D.; DŽEROSKI, S. Feature ranking with relief for multi-label classification: Does distance matter? In: SPRINGER. *Discovery Science: 21st International Conference, DS 2018, Limassol, Cyprus, October 29–31, 2018, Proceedings*. [S.l.], 2018. p. 51–65. Citado na página 27.

PETKOVIĆ, M.; KOCEV, D.; DŽEROSKI, S. Feature ranking for multi-target regression. *Machine Learning*, Springer, v. 109, n. 6, p. 1179–1204, 2020. Citado na página 29.

READ, J. et al. Classifier chains for multi-label classification. *Machine learning*, Springer, v. 85, n. 3, p. 333, 2011. Citado na página 23.

ROBNIK-ŠIKONJA, M.; KONONENKO, I. Theoretical and empirical analysis of relieff and rrelieff. *Machine learning*, Springer, v. 53, n. 1-2, p. 23–69, 2003. Citado na página 26.

SILLA, C. N.; FREITAS, A. A. A survey of hierarchical classification across different application domains. *Data Mining and Knowledge Discovery*, Springer, v. 22, n. 1-2, p. 31–72, 2011. Citado 2 vezes nas páginas 13 e 24.

SILVA, L. V. da; CERRI, R. Feature selection for hierarchical multi-label classification. In: SPRINGER. *Advances in Intelligent Data Analysis XIX: 19th International Symposium on Intelligent Data Analysis, IDA 2021, Porto, Portugal, April 26–28, 2021, Proceedings 19*. [S.l.], 2021. p. 196–208. Citado 6 vezes nas páginas 13, 15, 20, 38, 39 e 40.

SLAVKOV, I. et al. Relieff for hierarchical multi-label classification. In: SPRINGER. *International Workshop on New Frontiers in Mining Complex Patterns*. [S.l.], 2013. p. 148–161. Citado na página 29.

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. Citado na página 29.

SPOLAÔR, N. et al. A systematic review of multi-label feature selection and a new method based on label construction. *Neurocomputing*, Elsevier, v. 180, p. 3–15, 2016. Citado na página 26.

TSOUMAKAS, G.; VLAHAVAS, I. Random k-labelsets: An ensemble method for multilabel classification. In: SPRINGER. *European conference on machine learning*. [S.l.], 2007. p. 406–417. Citado na página 23.

VENS, C. et al. Decision trees for hierarchical multi-label classification. *Machine learning*, Springer, v. 73, n. 2, p. 185, 2008. Citado 3 vezes nas páginas 32, 33 e 35.

WANG, Z. et al. Incorporating hierarchy into text encoder: a contrastive learning approach for hierarchical text classification. *arXiv preprint arXiv:2203.03825*, 2022. Citado 2 vezes nas páginas 19 e 25.

WEHRMANN, J.; CERRI, R.; BARROS, R. Hierarchical multi-label classification networks. In: DY, J.; KRAUSE, A. (Ed.). [S.l.: s.n.], 2018. (Proceedings of Machine Learning Research, v. 80), p. 5075–5084. Citado na página 32.

WEI, L. et al. A novel hierarchical selective ensemble classifier with bioinformatics application. *Artificial intelligence in medicine*, Elsevier, v. 83, p. 82–90, 2017. Citado na página 25.

ZHANG, M.-L.; ZHOU, Z.-H. A review on multi-label learning algorithms. *IEEE transactions on knowledge and data engineering*, IEEE, v. 26, n. 8, p. 1819–1837, 2013. Citado na página 23.