UNIVERSIDADE TECNOLÓGICA FEDERAL DO PARANÁ

DIRETORIA DE PESQUISA E PÓS-GRADUAÇÃO

PROGRAMA DE PÓS-GRADUAÇÃO EM CIÊNCIA DA COMPUTAÇÃO

ALVARO MATEUS SANTANA

UMA ABORDAGEM PARA IMPUTAÇÃO DE VALORES FALTANTES EM PROBLEMAS DE CLASSIFICAÇÃO HIERÁRQUICA MULTIRRÓTULO

PONTA GROSSA

2021

ALVARO MATEUS SANTANA

UMA ABORDAGEM PARA IMPUTAÇÃO DE VALORES FALTANTES EM PROBLEMAS DE CLASSIFICAÇÃO HIERÁRQUICA MULTIRRÓTULO

An Approach for Missing Values Imputation in Multi-Label Hierarchical Classification Problems

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

Orientadora: Prof.ª Dra. Helyane B. Borges

PONTA GROSSA

2021

Ministério da Educação
Universidade Tecnológica Federal do Paraná
Campus Ponta Grossa

UTFPR
UNIVERSIDADE TECNOLÓGICA FEDERAL DO PARANÁ

ALVARO MATEUS SANTANA

UMA ABORDAGEM PARA IMPUTAÇÃO DE VALORES FALTANTES EM PROBLEMAS 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 Dezembro de 2021

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

Prof Jose Carlos Ferreira Da Rocha, Doutorado - Universidade Estadual de Ponta Grossa (Uepg)

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

Documento gerado pelo Sistema Acadêmico da UTFPR a partir dos dados da Ata de Defesa em 07/12/2021.

AGRADECIMENTOS

Agradeço primeiramente a Deus pela vida e pela oportunidade de todo dia buscar novos desafios.

Aos meus pais, Epaminondas e Ana, que sempre me mostraram o valor da educação, além do carinho e educação que me deram durante toda a minha vida.

A minha esposa Cintia que sempre estimulou nos estudos e entendeu minha ausência em determinados momentos. Pelo companheirismo e auxílio que possibilitaram eu poder dedicar aos estudos, mesmo em um período de pandemia. Também agradeço aos meus filhos, Mateus e Melissa, que me motivam em tentar ser uma pessoa melhor.

Agradeço especialmente a minha orientadora, Professora Helyane, que teve muita paciência em esclarecer diversas dúvidas durante a orientação. Agradeço pela oportunidade de compartilhar comigo sua sabedoria e tempo.

Aos membros da banca, professora Simone, que além de participar desta etapa final foi fundamental durante o curso e também ao professor José Carlos pelos importantes conselhos na etapa de qualificação desta pesquisa.

À UTFPR, o Programa de Pós-Graduação em Ciência da Computação e a todos os professores que forneceram um ambiente agradável de aprimoramento mesmo em período de pandemia.

Aos colegas do Laboratório de Engenharia de Software e Inteligência Computacional (LESIC) pela ajuda e troca de experiências que foram muito proveitosas.

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

RESUMO

Dados faltantes são problemas comumente enfrentados por algoritmos de aprendizagem de máquina (AM) devido a diversos motivos, como por exemplo falha na inserção manual, medições incorretas de determinado sensor entre outros. Considerando isso, se torna importante usar métodos adequados para imputar dados ausentes em conjuntos de dados para tornar a aprendizagem do algoritmo mais eficiente. O problema de dados faltantes é mais desafiador quando se trata de bases de dados com classificação hierárquica multirrótulo com hierarquia estruturadas por um Grafo Acíclico Direcionado ou DAG. Este trabalho está inserido neste cenário, onde as classes estão dispostas em uma hierarquia podendo cada instância possuir mais de uma classe. Para resolver o este problema, foi criado um método de imputação de dados faltantes usando uma abordagem baseada em três tipos de regressão: linear, polinomial e múltipla. O algoritmo inicialmente verifica se há correlação entre os dados, utilizando a regressão somente caso esta correlação exista, caso contrário a abordagem de média dos valores observados é adotada. O método proposto é dividido em três etapas: verificação hierárquica multirrótulo, cálculo de correlação e aplicação do modelo. Para realização dos experimentos foram utilizadas 7 bases de dados da Ontologia Gênica com hierarquia estruturadas em formato de DAG. Os resultados mostraram que o uso da regressão apresentou a métrica baseada na área sob a curva de previsão e revocação (AUPRC) superior em 3 das bases de dados testadas quando comparadas as abordagens de não imputação de dados faltantes e média dos valores observados. Além disso, foram realizados os testes estatísticos de Friedman e Wilcoxon buscando comparar os resultados de todos os algoritmos. Os testes expõem certa diferença entre os resultados, porém mostraram que estatisticamente a diferença não é significativa.

Palavras-chave: aprendizagem de máquina; dados faltantes; classificação hierárquica multirrótulo; regressão.

ABSTRACT

Missing data are problems commonly faced by machine learning (ML) algorithms due to several reasons, such as manual insertion failure, incorrect measurements of a given sensor, among others. Taking this into consideration, it becomes essential to use appropriate methods to impute missing data into datasets and make algorithm learning more efficient. The missing data problem is more challenging when it comes to databases with multi-label hierarchical classification with hierarchy structured by a Directed Acyclic Graph or DAG. This work is part of this scenario, where classes are arranged in a hierarchy, each instance may have more than one class. To solve this problem, a method of missing data imputation is created using three types of regression-based approach: linear, polynomial and multiple. The algorithm initially checks for correlation between the data, using regression only if this correlation exists, otherwise the average approach. observed values is adopted. The proposed method is divided into three steps: multi-label hierarchical verification, correlation calculation and model application. To perform the experiments, 7 databases of the Genetic Ontology with hierarchy structured in DAG format were used. The results showed that the use of regression presented the superior area under the precision-recall curve (AUPRC) metric in 3 of the tested databases when comparing the non-imputation approaches of missing data and mean of observed values. In addition, the Friedman and Wilcoxon statistical tests were performed in order to compare the results of all algorithms. The tests show a certain difference between the results, but they showed that statistically the difference is not significant.

Keywords: Machine learning. Missing data. Hierarchical multi-label classification. Regression.

LISTA DE FIGURAS

LISTA DE GRÁFICOS

LISTA DE QUADROS

LISTA DE TABELAS

LISTA DE SIGLAS

API	*Aplication Programming Interface*
AUPRC	*Area Under the Precision Recall Curve*
CIM	*Cascaded Imputation of Missing*
DAG	*Directed Acyclic Graph*
EM	*Expected Maximization*
GO	*Gene Ontology*
JCR	*Journal Citation Reports*
KNN	*K-Nearest Neighborhood*
HMC	*Hierarchical Multi-label Classification*
HSIM	*Hierarchical Supervised Imputation Method*
MI	*Mean Imputation*
ML	*Machine Learning*
MHC-CNN	*Multi-label Hierarchical Classification - Competitive Neural Network*
MSE	*Mean Squared Error*
SJR	*Scientific Journal Rankings*
SVM	*Support Vector Machine*

SUMÁRIO

1 INTRODUÇÃO

A computação é utilizada para auxiliar atividades nas mais variadas áreas, registrando informações de diversos tipos em bases de dados. Um exemplo comum é o sistema bancário, onde os dados de cada cliente e suas movimentações bancárias são armazenadas. Os dados também podem ser obtidos por meio de sensores diversos, como por exemplo os utilizados na detecção de tráfego de veículos em que Barreto (2018) propôs o uso de técnicas de aprendizado de máquina para identificação de perfis de uso de automóveis baseado em dados automotivos. Também podem vir de câmeras utilizadas na identificação de padrões em imagens como é o exemplo de Melo (2021), que usa *deep learning* para identificação de déficit hídrico em plantas com base em imagens térmicas. Enfim, a geração de dados ocorre em diversas áreas e com o passar do tempo estas bases de dados crescem exponencialmente.

Avanços na geração e coleta de dados produzem conjuntos de dados de tamanhos massivos no comércio e em uma diversidade de disciplinas científicas. *Data warehouses* armazenam detalhes das vendas e operações de negócio, satélites orbitando na Terra enviam imagens de alta resolução de dados de sensores para a Terra, e experimentos com genomas geram dados funcionais, estruturais e sequenciais para um número cada vez maior de organismos. Devido a estes avanços, conjuntos de dados com tamanhos em *gigabytes*, *terabytes* e até mesmo em *petabytes* estão se tornando comuns (TAN *et al*., 2009).

Em Aprendizagem de Máquina (AM), uma vertente da Inteligência Artificial (IA), computadores são programados para aprender com a experiência passada. Para tal, empregam algumas técnicas, sendo uma delas a indução. A indução é um princípio no qual se obtêm conclusões genéricas a partir de um conjunto particular de exemplos (FACELI *et al*. 2011). Com isso dados massivos gerados em determinada área podem ser utilizados por um algoritmo de aprendizagem de máquina como experiência passada, realizando a indução para se obter conclusões do que se pretende. Utilizando os exemplos citados anteriormente, um algoritmo de aprendizagem de máquina como árvore de decisão, poderia ser possível criar um modelo para determinar o grau de risco de empréstimo de um cliente em um sistema bancário, ou ainda estimar o tempo de algum trajeto em determinado horário.

Duas importantes categorias dos algoritmos de aprendizagem de máquina são o aprendizado supervisionado e não supervisionado. A diferença entre ambos está relacionada à forma de como é feito o processo de aprendizagem. No aprendizado não supervisionado não existe conhecimento sobre o domínio, ou seja, há menos informação sobre os exemplos, em particular, o conjunto não é rotulado. Já no aprendizado supervisionado, uma das tarefas mais utilizada é a classificação (BORGES, 2012).

A classificação na aprendizagem supervisionada é dividida em dois tipos: plana, que é a mais convencional, e a hierárquica onde as classes se dispõem em uma hierarquia. Na classificação plana, que é a mais comum, não há relação hierárquica entre as classes. Já na classificação hierárquica, os rótulos estão dispostos em uma estrutura de hierarquia, havendo uma relação de taxonomia ou dependência, formando subclasses e superclasses (FACELI *et al*., 2011).

Além disso, a classificação hierárquica pode ser multirrótulo ou monorrótulo. Importante não confundir a classificação hierárquica monorrótulo com a multirrótulo, pois o fato de uma instância herdar as classes ancestrais pode ser confundido com esta possuir mais de um rótulo. Na classificação multirrótulo cada exemplo do conjunto de dados pode ser associado a duas ou mais classes (CERRI, 2010).

Como os dados podem ser gerados de formas variadas, muitas vezes são apresentadas bases com informações incompletas ou faltantes, ou seja, boa parte dos bancos de dados existentes é caracterizada pela imprecisão e pela incompletude, isto é, pela presença de valores ruidosos e faltantes, respectivamente (FARHANGFAR *et al*., 2004). O motivo para a ausência destes dados podem ser dos mais variados, como: equipamentos com falhas operacionais, falha na inserção manual, medições incorretas de determinado sensor entre outros (FACELI *et al*. 2011).

Tratando especificamente do caso de falta de dados, três tipos de problemas estão associados aos valores faltantes: 1) perda de eficiência; 2) complicações na manipulação e análise dos dados; e 3) discrepâncias entre os valores atribuídos aos dados faltantes e os valores reais desconhecidos (FARHANGFAR *et al*., 2007).

Técnicas de aprendizagem de máquina, como por exemplo árvore de decisão, podem gerar erro de execução quando um ou mais atributos do conjunto de treinamento não apresentam valor (FACELI *et al*. 2011).

Os dados faltantes existem desde os primórdios das atividades de coleta de dados, durante muito tempo pouca atenção foi dada a estes dados. Muitos pesquisadores utilizavam, e ainda utilizam, algoritmos que não foram desenvolvidos para tratar bases de dados com valores faltantes (GRAHAM, 2009).

Este trabalho criou um método de imputação de dados faltantes em bases de dados com classificação hierárquica multirrótulo, visando isso, inicialmente foi realizado um mapeamento sistemático da literatura dividido em dois eixos, onde no primeiro eixo de pesquisa se constatou carência em métodos de imputação desenvolvidos exclusivamente para classificação hierárquica multirrótulo. Já o segundo eixo de pesquisa mostrou que nos trabalhos relacionados a classificadores, a média ou moda foi adotada como método de imputação.

O eixo 1 do mapeamento sistemático mostrou que não há métodos de imputação desenvolvidos exclusivamente para a classificação hierárquica multirrótulo. Já o eixo 2 mostrou que os trabalhos relacionados a classificadores normalmente utilizam a média ou moda como método de imputação de dados faltantes. Ainda no eixo 2, foram verificados diversos trabalhos relacionados a classificadores em que não se deixou explícito como ocorreu a imputação dos dados faltantes.

Após o mapeamento sistemático é apresentado um método de imputação que utiliza uma técnica estatística chamada regressão, em que é possível atribuir uma relação de um atributo dependente (variável de resposta) com atributos independentes (variáveis explanatórias). O método criado utiliza as regressões linear, polinomial e múltipla e é dividida em três etapas: a) verificação hierárquica multirrótulo, b) cálculo da correlação e c) aplicação do modelo de imputação.

Aplicando o método, foram conduzidos experimentos em dados biológicos de 7 (sete) bases de dados do projeto Gene Ontology (GO), que tem hierarquia do tipo DAG. Os experimentos utilizaram três técnicas de regressão: a regressão linear, que é mais indicada quando há linearidade de dados, a regressão polinomial, onde o modelo criado busca trabalhar com dados não lineares e a regressão múltipla, onde o modelo utiliza mais de uma variável explanatória.

Para avaliar o desempenho do classificador nessas bases de dados, que tiveram seus valores imputados, foi utilizado classificador Clus-HMC (VENS et al., 2008) por meio da métrica AUPRC (*Area Under the Precision Recall Curve*). Esta medida é referente a área sob a curva de precisão e revocação, métrica comumente

utilizada em trabalhos relacionados a classificação hierárquica multirrótulo. Observou-se que em 3 (três) bases de dados, apesar da AUPRC apresentar valor superior, quando comparado com o resultado sem imputação, os testes de Friedman e Wilcoxon mostraram que não há diferença estatística significativa entre os resultados.

1.1 Descrição do problema e motivação

Um dos problemas enfrentados pelas tarefas de classificação é quando os dados apresentam atributos faltantes. A classificação de dados é o processo de encontrar um modelo que descreva as diferentes classes presentes em um conjunto de dados, ou seja, extrair informações a partir de um conjunto de dados por meio de sua categorização (CERRI, 2010). Caso o conjunto de dados não esteja completo, a acurácia da categorização ficará comprometida. Na classificação hierárquica multirrótulo esse problema é agravado devido ao aumento na complexidade de como são estruturados.

Apesar de haver aumento na complexidade da representação dos dados na classificação hierárquica multirrótulo, essa abordagem pode trazer facilitadores, pois alguns tipos de problemas podem tirar vantagem da hierarquia de classes. Por exemplo, um problema que faça busca de instâncias similares, verificando exemplos compartilham rótulos, poderiam ter essa busca maximizada havendo a verificação das classes ancestrais.

A classificação hierárquica multirrótulo busca resolver problemas em diversas áreas como classificação de textos, gêneros musicais, predição de proteínas na área de bioinformática, entre outros. Isso foi verificado no Capítulo 3, onde foi realizado um mapeamento sistemático da literatura a qual foi dividida em dois eixos: o primeiro buscou verificar quais métodos de imputação foram desenvolvidos para o cenário de classificação hierárquica multirrótulo e o segundo eixo teve como meta buscar classificadores hierárquicos multirrótulo que utilizaram em seus experimentos bases de dados estruturadas como um DAG. No segundo eixo o objetivo foi inspecionar como realizado o pré-processamento das bases de dados verificando quais métodos de imputação foram utilizados.

O mapeamento sistemático da literatura mostrou diversas abordagens para trabalhar com dados faltantes, porém se constatou carência de trabalhos

principalmente no primeiro eixo de pesquisa, que é procurar métodos desenvolvidos exclusivamente para bases de dados com classificação hierárquica multirrótulo. No segundo eixo de pesquisa foi inspecionado a fase de pré-processamento de classificadores hierárquicos multirrótulos, sendo que nos estudos onde foi possível verificar descrição do método de imputação de dados faltantes, foi utilizada a média ou moda.

Considerando que a média foi utilizada para imputar valores em diversos estudos relacionados a classificadores, este trabalho criou um método buscando contribuir para o tema incrementando a acurácia da imputação utilizando técnicas de regressão caso seja verificado que o coeficiente de correlação seja relevante. Essa contribuição busca adicionar o recurso da regressão aos métodos já utilizados por alguns trabalhos, pois caso a correlação não seja superior ao limiar definido, a média ainda será utilizada. A regressão foi adotada por ser um modelo simples, mas que pode apresentar melhor desempenho que a média dependendo da disposição dos dados.

Buuren (2018), relaciona entre seus métodos a imputação através da regressão, em que é possível atribuir o valor de uma variável em função de outra, buscando dessa forma realizar imputações mais inteligentes. Ainda sobre regressão, Neto (2002), descreve o coeficiente de correlação de Pearson, onde é possível medir a relação entre duas variáveis.

Motivado pela possibilidade de aplicar a regressão para imputação de dados faltantes no cenário de classificação de dados hierárquica multirrótulo, apresenta-se um método de imputação para auxiliar algoritmos da classificação hierárquica multirrótulo e em consequência, auxiliar o desenvolvimento das diversas áreas que utilizam a classificação hierárquica multirrótulo, em especial a bioinformática, aplicando o método criado na base de dados *Gene Ontology* (GO).

1.2 Objetivos

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

1.2.1 Objetivo Geral

Desenvolver uma abordagem para imputação de valores faltantes para classificação hierárquica multirrótulo usando regressão.

1.2.2 Objetivos Específicos

Como objetivos específicos têm-se:

- Realizar um mapeamento sistemático da literatura buscando abordagens utilizadas para tratar o problema de dados faltantes no cenário de classificação hierárquica multirrótulo;
- Analisar o desempenho dos conjuntos com os dados ausentes imputados com o método criado em um classificador hierárquico multirrótulo;
- Avaliar os resultados obtidos por meio de métrica e testes estatísticos.

1.3 Organização do trabalho

Este trabalho está organizado em seis Capítulos. O Capítulo 2 aborda o referencial teórico do trabalho, abrangendo conceitos de classificação de dados, classificação hierárquica e multirrótulo. Este capítulo também aborda noções de dados faltantes além de listar alguns métodos propostos na literatura para imputação de dados faltantes. O referencial teórico apresenta o embasamento matemático das técnicas de regressão utilizadas neste trabalho. Por fim o Capítulo 2 descreve um teste estatístico que será utilizado posteriormente para avaliar os resultados dos experimentos de cada um dos algoritmos.

O Capitulo 3 descreve o mapeamento sistemático da literatura, que é dividida em dois eixos, onde inicialmente foram levantadas as metodologias existentes para imputação de dados faltantes no cenário de classificação hierárquica multirrótulo e posteriormente são inspecionados trabalhos relacionados a classificadores hierárquicos multirrótulo buscando verificar como ocorreu a imputação de dados faltantes.

O Capítulo 4 apresenta o método para imputação de dados faltantes em bases com classificação hierárquica multirrótulo, mostrando cada uma das três etapas que o método é executado. Também é realizada uma simulação passo a passo do método em uma pequena base de dados fictícia.

O Capítulo 5 mostra as ferramentas utilizadas para o desenvolvimento do método além de ilustrar a metodologia utilizada para realizar os experimentos, que é dividida em três etapas. É comentado sobre a métrica utilizada para avaliação do algoritmo e posteriormente são apresentados os resultados de cada uma das abordagens do método. Ainda no Capítulo 5 são realizadas comparações dos resultados do método com outras duas abordagens para tratamento de dados faltantes. Por fim, os resultados comparativos são submetidos a avaliações através de teste estatístico não paramétrico

Por fim, o Capítulo 6 relata as conclusões obtidas com os resultados dos experimentos do método. Além disso, são elencados trabalhos futuros que podem estender a pesquisa.

2 REFERENCIAL TEÓRICO

Este Capítulo apresenta a base teórica dos conceitos abordados neste trabalho. Na Seção 2.1 são apresentados conceitos de classificação de dados. A Seção 2.2 aborda dados faltantes e alguns métodos comumente utilizados para a imputação de dados. A Seção 2.3 explora alguns tipos de regressão que foram a base do método deste trabalho. Por fim, a Seção 2.4 apresenta as considerações finais do capítulo.

2.1 Classificação de dados

A classificação é um processo que consiste em associar uma determinada instância (exemplo) a uma ou mais classes, dentre um conjunto de classes previamente definidas. Essa associação de um exemplo a uma determinada classe ocorre conforme as características (atributos) da instância (BORGES, 2012).

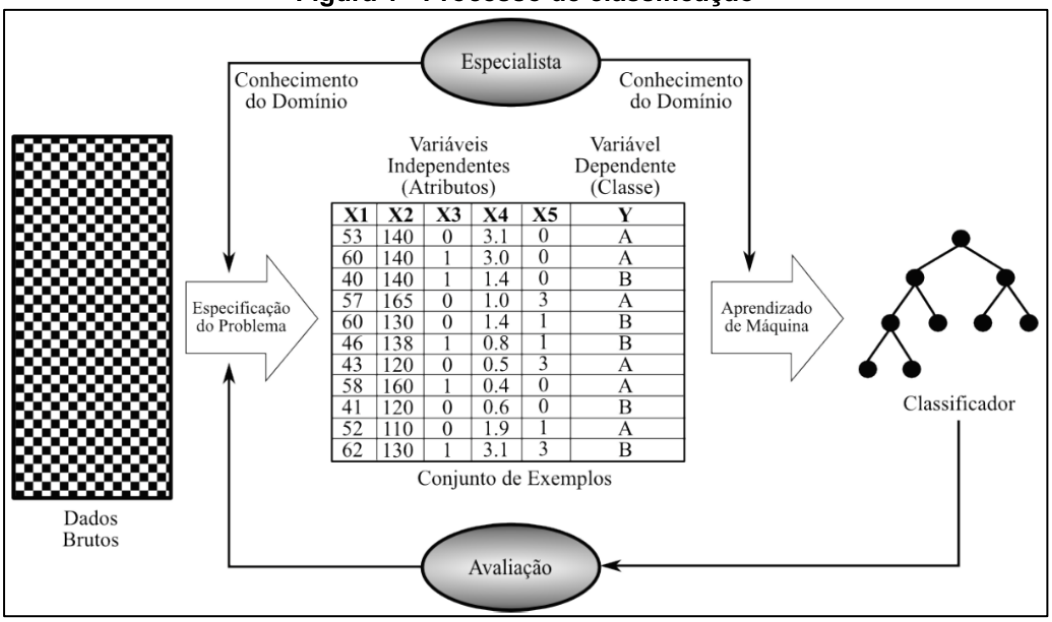
Tarefas de aprendizagem de máquina podem ser divididas em preditivas e descritivas. Nas tarefas preditivas, o objetivo é obter um modelo por meio de dados de treinamento, sendo que este modelo deverá ser capaz de prever um novo exemplo com base nos valores de seus atributos. Algoritmos utilizados nessa tarefa utilizam o paradigma de aprendizagem supervisionada, em que há o papel do rótulo para cada exemplo (FACELI *et al*., 2011).

Já nas tarefas descritivas, o objetivo é explorar ou descrever o conjunto de dados, pois não há o atributo rótulo ou classe. Algoritmos de tarefas descritivas utilizam o paradigma do aprendizado não supervisionado, ou seja, não há o papel do atributo rótulo ou classe, sendo que o objetivo é encontrar grupos de objetos semelhantes em um conjunto de dados. Tarefas supervisionadas podem ainda se diferenciar pelo tipo de rótulo, sendo que na classificação a classe é um atributo discreto e na regressão é um atributo contínuo (FACELI *et al*., 2011).

A Figura 1 ilustra a tarefa supervisionada de classificação, em que inicialmente os dados brutos são preparados em um conjunto de exemplos para que possam ser processados. Um conjunto de exemplos é composto por valores de atributos, que são as características do exemplo e pelo atributo classe.

Após o processamento dos dados, o conjunto de exemplos passará pelo algoritmo de indução para que seja feita o treinamento do classificador. O objetivo do classificador é encontrar uma função que mapeie cada exemplo com sua classe correspondente. Na figura o processamento dos dados é realizado pelo especialista, sendo que após o classificador ser treinado ele poderá passar por uma avaliação.

Figura 1 - Processo de classificação



Fonte: Rezende (2005)

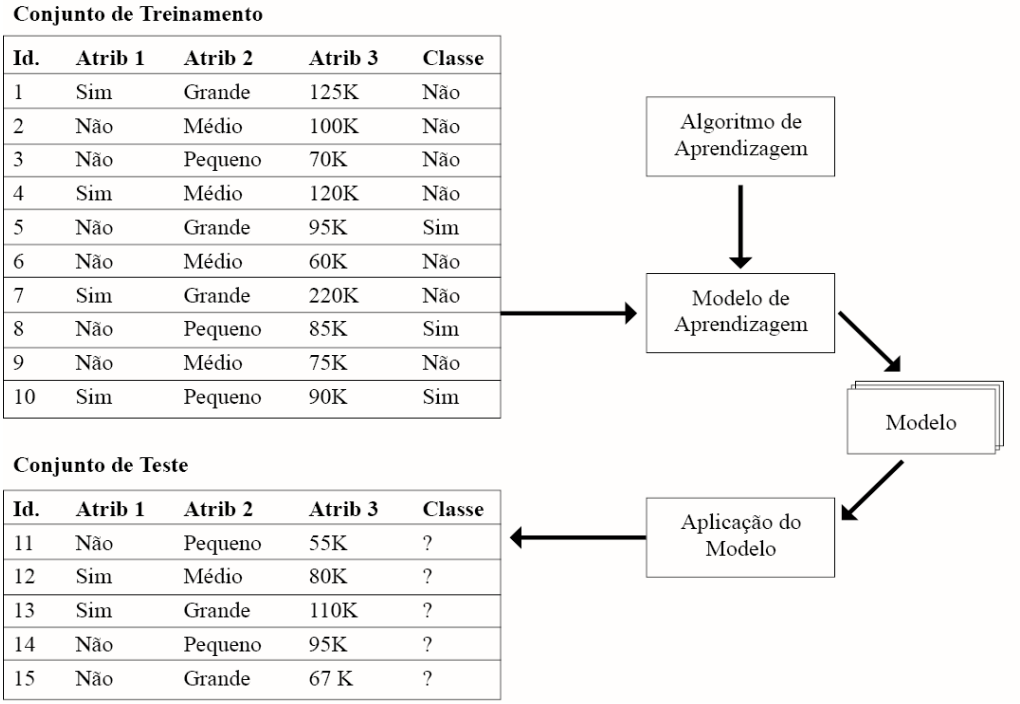
O modelo gerado pelo algoritmo de aprendizagem deve se adaptar a novos dados de entrada e prever corretamente as classes para exemplos que nunca lhe foram apresentados anteriormente (TAN *et al*., 2009), ou seja, o objetivo do classificador deve ser construir modelos com boa capacidade de generalização.

Muitas vezes é útil medir o desempenho do modelo no conjunto de teste pois esta medição fornece uma avaliação imparcial do erro de generalização. A precisão ou taxa de erro é calculada a partir do conjunto de teste e pode ser utilizada para comparar o desempenho relativo de diferentes classificadores do mesmo tipo (TAN *et al*., 2009).

A Figura 2 ilustra a abordagem geral para resolver problemas de classificação. Inicialmente há um conjunto de treinamento, que é utilizado pelo Algoritmo de Aprendizagem para criar um Modelo de Aprendizagem. Após esta etapa, o Modelo é aplicado a um conjunto de teste, onde são atribuídas classes a estes exemplos. Estas classes atribuídas podem ser comparadas as classes

originais do conjunto para se verificar se houve erro ou acerto da classificação, possibilitando avaliar o desempenho do classificador.

Figura 2 - Abordagem geral para construção de um modelo de Classificação



Fonte: Tan *et al*. (2009)

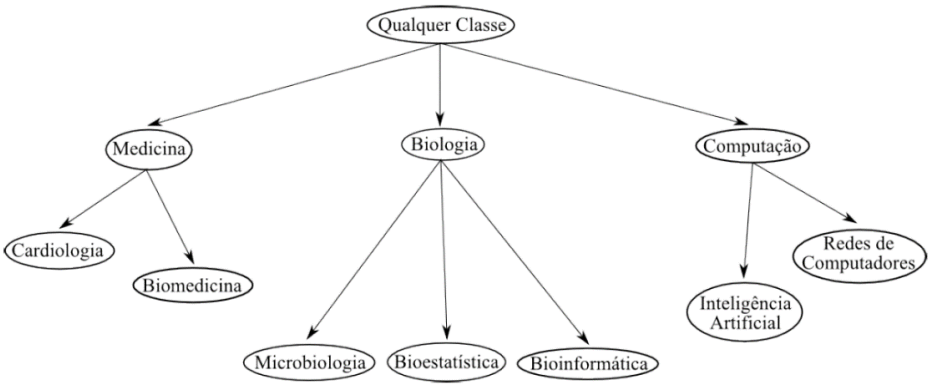
2.1.3 Classificação hierárquica

Em problemas de classificação hierárquica, deve-se ter o conjunto de dados, que será utilizado para treinamento do classificador e a hierarquia de classes que conterá as informações de ancestralidade e descendência das classes que serão preditas (BORGES, 2012).

A maioria dos problemas de classificação descritos na literatura diz respeito a problemas de classificação plana *(Flat Classification)*, em que cada exemplo é associado a uma classe pertencente a um conjunto finito de classes, não considerando relacionamentos hierárquicos. No entanto, existe um grande número de problemas em que uma ou mais classes podem ser divididas em subclasses ou agrupadas em superclasses (CERRI, 2010).

Nestas situações as classes podem ser dispostas em uma estrutura hierárquica, como uma árvore, ilustrada no exemplo da Figura 3. Este exemplo se refere a um problema de classificação hierárquica de textos científicos, em que determinada instância poderia, por exemplo, pertencer a classe Inteligência Artificial.

Figura 3 - Hierarquia de classes estruturada como árvore

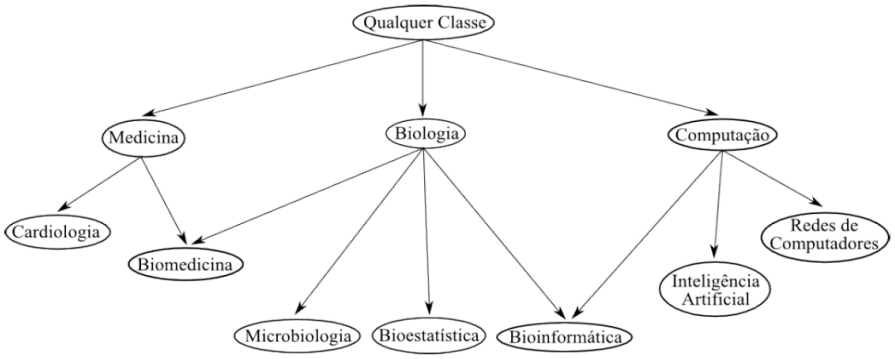


Fonte: Cerri (2010)

Ressalta-se que quanto mais profundo for o nível, normalmente a predição será mais difícil devido ao fato de as classes mais especializadas normalmente possuírem um número menor de exemplos de treinamento. O número menor de exemplos não é o único fator para a dificuldade na predição, pois em geral espera-se que uma subclasse tenha alguma semelhança com a superclasse de acordo com algum critério definido. Essa similaridade pode dificultar a separação dos dados, tornando a classificação mais difícil.

A classificação hierárquica pode ainda ser representada por um Grafo Acíclico Direcionado ou *Directed Acyclic Graph* (DAG), ilustrada na Figura 4. A principal diferença entre a estrutura em árvore e em DAG é que na árvore, cada nó possui somente um nó pai, já na DAG cada nó pode ter um ou mais nós pai.

Figura 4 - Hierarquia de classes estruturada como DAG



Fonte: Cerri (2010)

No DAG, as classes especializadas podem ser induzidas por um número maior de exemplos de treinamento (BORGES, 2012). Apesar disso, tanto para estruturas em DAG como em árvore, quanto mais especializado é o rótulo, menos

precisão haverá na predição devido à redução no número de exemplos utilizados no processo de indução (CERRI, 2010). Se o classificador não apresentar uma confiabilidade desejada em classes mais especializadas, pode ser vantajoso utilizar rótulos com níveis mais altos.

Neste contexto, o que difere a classificação hierárquica da classificação convencional, é a predição das classes na hierarquia, onde são divididas em duas categorias: predição obrigatória em nós folha e predição opcional em nós-folha (BORGES, 2012).

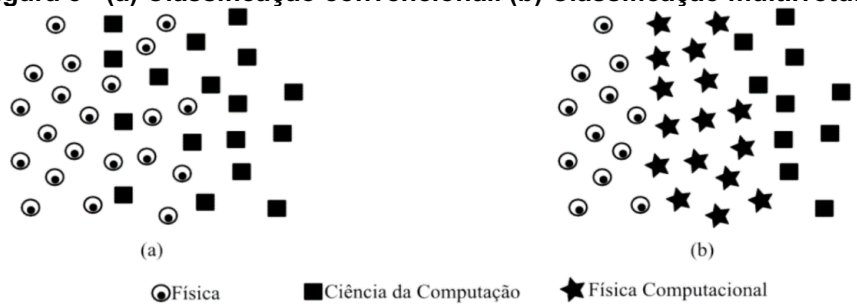
Na predição obrigatória em nós-folha, o rótulo predito em um determinado nível, estará automaticamente predizendo as classes nos níveis acima. Já na predição opcional em nós-folha os exemplos podem ser associados as classes que são representadas por qualquer nó interno da hierarquia e seus ancestrais.

2.1.4 Classificação multirrótulo

Em classificação multirrótulo, cada exemplo pode ser associado a duas ou mais classes ao mesmo tempo. Um classificador multirrótulo pode ser definido como uma função $H : x \rightarrow 2^L$ que mapeia um exemplo x em um conjunto de classes $C \in 2^L$, em que 2^L é o conjunto potência de L, ou seja, o conjunto formado por todos os subconjuntos de L (CERRI, 2010).

O estudo da classificação multirrótulo tem entre suas motivações a classificação de textos, sendo por exemplo que determinado documento pode possuir simultaneamente mais de um rótulo, pertencendo então a mais de um tema ou área. A Figura 5 ilustra a classificação convencional (plana) e multirrótulo. Neste exemplo, um determinado documento pode ter como tema física computacional, pertencendo então a classe Física e também Ciência da Computação.

Figura 5 - (a) Classificação convencional. (b) Classificação multirrótulo.

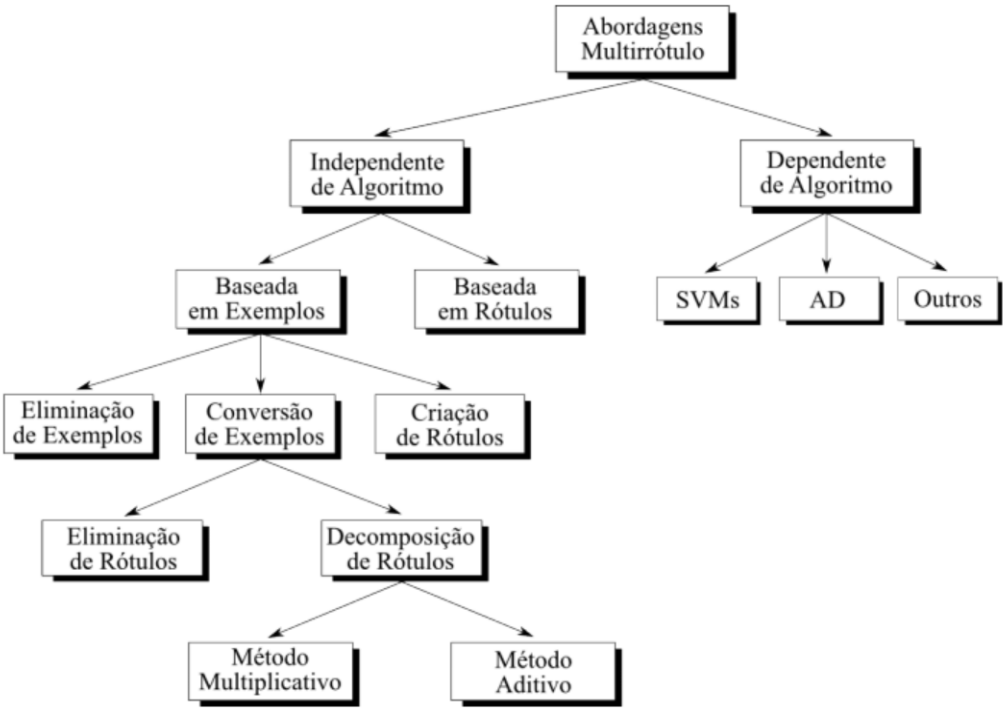


Fonte: Cerri (2010)

Freitas e Carvalho (2007) propõe duas formas de abordar os métodos de classificação multirrótulo, que são: dependente e independente do algoritmo. A abordagem independente do algoritmo transforma o problema multirrótulo em monorrótulo. Já a forma dependente do algoritmo é baseada na criação de algoritmos específicos para o problema multirrótulo, podendo gerar melhores resultados se comparados a independente de algoritmo.

A Figura 6 ilustra a visão das abordagens propostas para tratar problemas de classificação multirrótulo. A abordagem independente de algoritmo utiliza algoritmos tradicionais de classificação para tratar problemas multirrótulo, transformando o problema multirrótulo original em um conjunto de problemas simples-rótulo. Já a abordagem dependente de algoritmo adapta algoritmos para tratar o problema multirrótulo, como SVMs e árvores de decisão, ou podem ser especificamente desenvolvidos para classificação multirrótulo (CERRI, 2010).

Figura 6 - Abordagens para classificação multirrótulo



Fonte: Faceli *et al*. (2011)

2.1.5 Classificação hierárquica multirrótulo

Para Blockeel et al. (2006), o problema de classificação hierárquica multirrótulo pode ser definido da seguinte maneira:

Dado: um espaço de exemplos X e uma hierarquia de classes $(C, \leq_h)$, no qual C é um conjunto de classes e $\leq_h$ é uma ordem parcial estruturada como árvore enraizada, representando o relacionamento de superclasse (para todo $c1, c2 \in C: c1 \leq_h c2$ se e somente se $c1$ é uma superclasse de $c2$); um conjunto T de exemplos (xi, Yi) com $xi \in X \; e \; Yi \subseteq C$ tal que $c \in Yi \rightarrow \forall c' \leq_h c : c' \in Yi$; e algum critério de qualidade q tipicamente recompensa modelos de alta acurácia preditiva e baixa complexidade).

Encontrar: uma função $f: X \rightarrow 2^C$ (na qual 2^C é o conjunto potência de C), tal que $c \in f(x) \rightarrow \forall c' \leq_h c : c' \in f(x) e \; f$ maximiza q.

Há problemas de classificação que possuem características tanto de classificação multirrótulo, quanto de classificação hierárquica, se chamando de problemas de Classificação Hierárquica Multirrótulo, onde uma instância pertence a múltiplas classes ao mesmo tempo que estão organizadas de forma hierárquica (CERRI, 2010).

Neste tipo de classificação cada instância irá pertencer a um ou mais rótulos, sendo que estas classes estarão dispostas em uma estrutura de hierarquia. Este tipo de problema é comum em problemas de classificação de textos (ESULI *et al*. 2006) e predição e funções de proteínas (BORGES 2012). A hierarquia dos rótulos pode ser organizada em uma estrutura de árvore ou DAG que foram ilustradas nas Figuras 3 e 4, respectivamente.

Não deve haver confusão entre problemas de classificação hierárquica simples-rótulo e problemas de classificação hierárquica multirrótulo. Pode-se pensar em um problema hierárquico simples-rótulo como sendo naturalmente um problema multirrótulo, devido ao fato de que um ramo da hierarquia contém mais de uma classe. Quando se atribui a um exemplo às classes “Esportes/Coletivos/Futebol”, está se atribuindo um ramo da hierarquia que contém três classes. Entretanto, um problema hierárquico é considerado multirrótulo quando são atribuídos mais de um ramo da hierarquia a um exemplo (CERRI, 2010). Caso os dados possam seguir mais de um caminho na hierarquia, tem-se um problema de classificação hierárquica com múltiplos rótulos.

Um exemplo de trabalho em problemas de classificação hierárquica multirrótulo para classificação de textos é o método proposto por Esuli *et al*. (2006), em que é utilizado um algoritmo chamado *TreeBoost.MH*, que consiste de uma

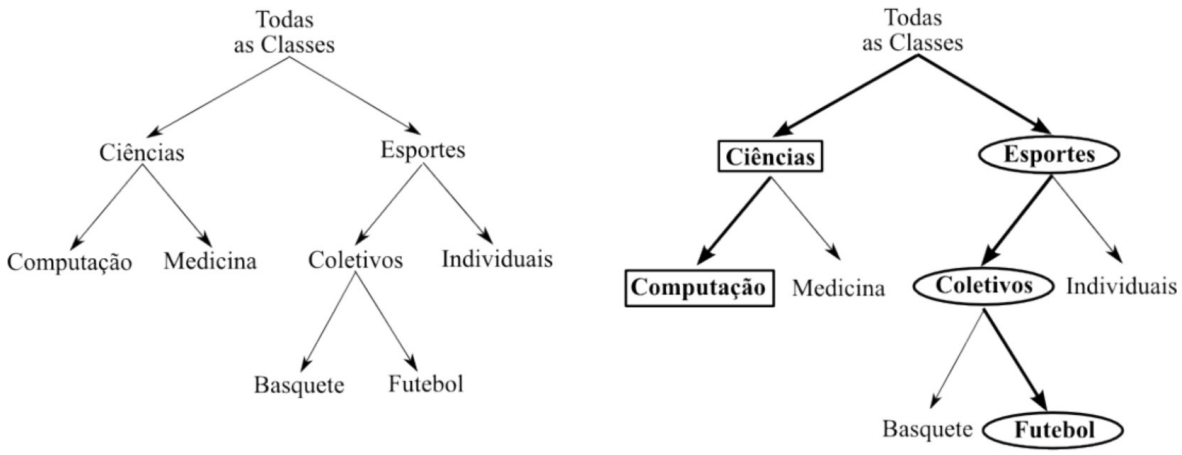
variação hierárquica do algoritmo *AdaBoost.MH* (SCHAPIRE; SINGER, 1999), membro da família de *boosting* de algoritmos de aprendizado (FREUND; SCHAPIRE, 1999). O algoritmo faz uma seleção de atributos e uma seleção de exemplos de treino negativos localmente, considerando a topologia do esquema de classificação. A distribuição dos pesos do algoritmo *boosting* atualiza a cada iteração também é atualizada localmente. O *TreeBoost.MH* é definido pelos autores como um algoritmo recursivo que utiliza o *AdaBoost.MH* como base, e é chamado recursivamente na estrutura da árvore.

Borges (2012), propõe um método chamado MHC-CNN *(Multilabel Hierarchical Classification using a Competitive Neural Network)* para resolução do problema de classificação hierárquica multirrótulo de funções proteínas. O algoritmo utiliza a abordagem de classificação global baseado em Rede Neural Competitiva. Esse método se baseia no aprendizado competitivo, os neurônios artificiais competem entre si para serem ativados, existindo um vencedor que terá seus pesos atualizados junto a seus vizinhos. O trabalho é pioneiro no uso de redes neurais para tratar problemas hierárquicos estruturados em DAG e que utilizam abordagem de de classificação hierárquica global. Nos experimentos houve comparação com os algoritmos Clus-HMC e Clus-HC, usando a medida de avaliação AUPRC.

Vens *et al*. (2008), apresenta abordagens que utilizam árvores de decisão (baseadas no conceito de *Predictive Clustering Trees*) em problemas de classificação hierárquica multirrótulo. As abordagens são utilizadas e comparadas em bases de dados relacionadas à genômica funcional. Os autores comparam o desempenho de induzir uma única árvore de decisão, que faz predições para todas as classes da hierarquia de uma única vez (chamada de HMC), com outras duas abordagens, que fazem o aprendizado de uma árvore de decisão para cada classe. A primeira, chamada de SC, define a tarefa de classificação simples-rótulo independente para cada classe, ignorando os relacionamentos hierárquicos entre as classes. A segunda, chamada HSC, explora os relacionamentos para induzir a árvore de decisão para cada classe. Os experimentos que a abordagem HMC supera as abordagens HSC e SC com relação a acurácia, tamanho do modelo e tempo de indução. Foram avaliados 24 conjuntos de dados da área genômica funcional. A avaliação do desempenho foi realizada também com a métrica baseada na área sobre a curva de precisão e revocação (AUPRC).

A Figura 7 ilustra a classificação hierárquica multirrótulo, determinado texto poderia abordar assuntos relacionados à ciência da computação e esportes coletivos, sendo então pertencentes as classes “Ciência/Computação” e “Esportes/Coletivos/Futebol”.

Figura 7 - Problema hierárquico multirrótulo estruturado como árvore. A esquerda a hierarquia de classes, a direita predições gerando uma subárvore

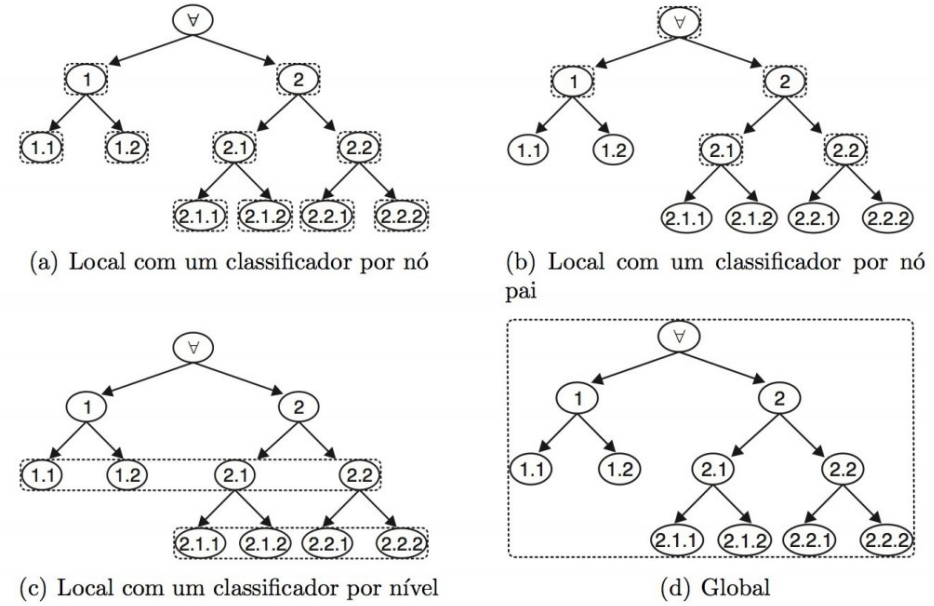


Fonte: Cerri (2010)

Freitas e Carvalho (2007), divide as abordagens para classificação hierárquica em classificadores locais e globais, também chamada de *Big-Bang.* Os classificadores locais se dividem ainda em classificação por nó, classificação por nível e classificação por nó pai. A Figura 8 ilustra as abordagens para classificação hierárquica proposta.

A Figura 8 (a), ilustra a abordagem local com um classificador por nó, que consiste em treinar um classificador binário local para cada nó da hierárquica de classes. A Figura 8 (b), a classificação local com um classificador por nó pai, onde a técnica treina cada nó pai da hierarquia de classes em um classificador multirrótulo. A Figura 8 (c) mostra a abordagem local com um classificador por nível, onde é criado um classificador para cada nível da hierarquia. Por fim a Figura 8 (d) mostra a abordagem global ou *Big-Bang*, que consiste em construir um único modelo de classificação levando em consideração a hierarquia de classes de todo o conjunto de treinamento.

Figura 8 - Abordagens para classificação hierárquica



Fonte: Freitas e Carvalho (2007)

A técnica de classificação hierárquica local para cada nó consiste em treinar um classificador binário para cada nó da hierarquia de classes, ou seja, consiste em usar *M* classificadores independentes para cada classe, onde *M* é o número total de nós na hierarquia de classes (BORGES, 2012). Segundo Blockeel *et al*. (2006), essa estratégia apresenta algumas desvantagens como por exemplo o número elevado de iterações, as relações hierárquicas não são consideradas entre outros.

Na classificação hierárquica local em nós pais para cada um dos nós, devem-se distinguir as subclasses contidas em seus nós filhos, usando classificadores multiclasse ou combinações de classificadores binários. Esta estratégia não costuma ser empregada em hierarquias do tipo DAG, uma vez que pode haver muita redundância nos conjuntos de treinamento de diferentes classificadores (Freitas; Carvalho, 2007).

A abordagem de classificação hierárquica local por nível consiste em criar um classificador multirrótulo para cada nível da hierarquia. Esta técnica pode ser usada em estruturas em árvore e em DAG (BORGES, 2012).

Os classificadores globais considerar a hierárquica como um todo na etapa de treinamento e não possuem a modularidade característica das abordagens locais. O modelo final obtido nessa abordagem, também denominada de *Big-Bang*, pode ser menor do que aquele formado por múltiplos classificadores locais. Contudo, esse classificador tende a ser mais complexo do que os individuais induzidos na

abordagem local. Logo, na abordagem *Big-Bang*, geralmente um único modelo é induzido. Na etapa de teste, esse modelo pode ser empregado potencialmente na predição de qualquer classe da hierarquia (FACELI *et al*., 2011).

O estudo acerca da classificação hierárquica multirrótulo possui vertentes em diversas áreas como bioinformática, classificação de textos, processamento de imagem entre outros. Inclusive não há consenso sobre qual algoritmo utilizar para classificação hierárquica multirrótulo.

Como este trabalho irá se focar na imputação de dados faltantes em bases de dados da Gene Ontology (GO), será tratado um algoritmo que também foi testado sob esta base de dados, que é Clus-HMC (VENS *et al*, 2008).

2.1.6 Medida baseada na Curva de Previsão e Revocação

Vens *et. al.* (2008) propôs uma medida baseada na análise de curvas de precisão e revocação (curvas PR). Tal medida funciona da seguinte maneira: um conjunto de limiares entre 0 e 1 são selecionados. Cada limiar corresponde a um ponto no espaço da curva PR e variando esses limiares obtêm-se a curva PR

Para um determinado limiar, um ponto de precisão e revocação no espaço da curva PR se dá por meio da Equação (1) e Equação (2), respectivamente.

$$\overline{Prec} = \frac{\sum_i VP_i}{\sum_i VP_i + \sum_i FP_i} \quad (1)$$

$$\overline{Rev} = \frac{\sum_i VP_i}{\sum_i VP_i + \sum_i FN_i} \quad (2)$$

Onde o i varia de 1 até o número de classes do problema, enquanto o número de Verdadeiros Positivos, Falsos Positivos e Falsos negativos são representados por VP, FP e FN, respectivamente. Os valores de AUPRC variam entre 0 e 1 e quanto mais próximo de 1, melhor (VENS *et al.*, 2008).

2.2 Imputação de dados faltantes

Conjuntos de dados podem apresentar dificuldades relacionadas à qualidade, exemplos mais frequentes dessas dificuldades são dados ruidosos, inconsistentes, redundantes ou incompletos (FACELI *et al*., 2011). A Figura 9 ilustra

um conjunto de dados onde alguns atributos apresentam dados ausentes, representados pelo caractere “?”. No exemplo, caso fosse preciso aplicar determinado algoritmo de aprendizagem, como Rede Neural Artificial, seria necessário realizar alguma abordagem para tratar estes dados ausentes. Isto é necessário pois estes dados serão utilizados como entradas para algum neurônio artificial.

Este tratamento do dado faltante pode ser através de métodos simples, como por exemplo, preenchimento manual e eliminação de valores ausentes. Também é possível realizar abordagens mais complexas como emprego de algoritmos de aprendizagem de máquina que lidam diretamente com os valores ausentes como por exemplo, indutores de árvores de decisão ou aplicação de algum método ou heurística para automaticamente definir valores atributos com valores ausentes (FACELI *et al*. 2011).

> Dados ausentes estão lá, gostemos ou não. Nas ciências sociais, é quase inevitável que alguns entrevistados se recusem a participar ou a responder a certas perguntas. Em estudos médicos, o desgaste de pacientes é muito comum. A teoria, metodologia e software para lidar com problemas de dados incompletos foram amplamente expandidos e refinados nas últimas décadas (BUUREN, 2018, p. 7).

Figura 9 - Conjunto e dados apresentando dados faltantes

Id.	Atrib 1	Atrib 2	Atrib 3	Classe
1	Sim	Grande	125	Não
2	?	Médio	100	Não
3	Não	Pequeno	70	Não
4	Sim	Médio	?	Não
5	Não	Grande	95	Sim
6	Não	?	60	Não
7	Sim	Grande	?	Não
8	Não	Pequeno	85	Sim
9	Não	Médio	75	Não
10	Sim	Pequeno	90	Sim

Fonte: Adaptado de Tan *et al*. (2009)

A imputação dos dados faltantes é necessária pois algumas técnicas de aprendizagem de máquina podem gerar erro de execução caso um ou mais atributos do conjunto de treinamento não apresentarem valores. Segundo Faceli *et al*. (2011),

a ausência de valores em alguns atributos pode ter diferentes causas como por exemplo:

- O atributo não foi considerado importante quando foi inserido.
- Desconhecimento do valor do atributo na ocasião do preenchimento dos valores do exemplo.
- Distração na hora do preenchimento
- Falta da obrigatoriedade de inserção de um valor para o atributo.
- Inexistência do valor em alguns objetos.
- Problema no equipamento que realizou a coleta dos dados.

Um aspecto importante na análise dos dados faltantes é compreender o motivo pelo qual surgiram a falta destes dados. Rubin (1976) classificou os problemas de dados faltantes em três categorias. Em sua abordagem, cada ponto de dados tem alguma probabilidade de estar faltando. O processo que governa essas probabilidades é chamado de mecanismo de dados perdidos ou mecanismo de resposta. O modelo para o processo é chamado de modelo de dados ausentes ou modelo de resposta (BUUREN, 2018).

As três categorias propostas por Rubin (1976) são:

1. Se a probabilidade de estar faltando é a mesma para todos os casos, então se diz que os dados estão faltando completamente ao acaso ou *Missing Completely at Random* (MCAR) (BUUREN, 2018). Isso implica efetivamente que as causas dos dados ausentes não estão relacionadas aos dados. Consequentemente, pode-se ignorar muitas das complexidades que surgem devido à falta de dados, além da óbvia perda de informações. Um exemplo é quando se tem uma amostra aleatória de uma população, em que cada membro tem a mesma chance de ser incluído na amostra. Os dados (não observados) de membros da população que não foram incluídos na amostra são MCAR .
2. Se a probabilidade de estar faltando é a mesma apenas dentro dos grupos definidos pelos dados observados, então os dados estão faltando aleatoriamente ou *Missing at Random* (MAR) (BUUREN, 2018). Um exemplo é um estudo retrospectivo de memória, em que é questionado a idosos eventos de sua infância, onde possivelmente alguns terão muita dificuldade e

não lembrarão destes acontecimentos. Neste caso a ausência é relacionada a idade avançada e não aos eventos em si.

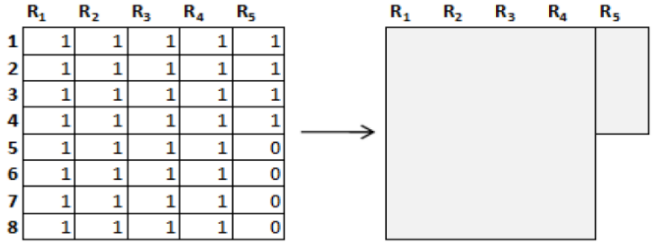
3. Se MCAR e MAR não forem válidos, então diz-se que dados ausentes gerados não ao acaso ou *Missing Not at Random* (MNAR) (BUUREN, 2018). MNAR significa que a probabilidade de haver dados ausentes varia por motivos que se desconhece. Por exemplo, o mecanismo de uma balança de pesagem pode se desgastar com o tempo, produzindo mais dados perdidos à medida que o tempo passa, mas pode-se deixar de notar isso. Se os objetos mais pesados forem medidos posteriormente, obtém-se uma distribuição das medições que será distorcida. MNAR inclui a possibilidade de que a escala produza mais valores ausentes para os objetos mais pesados (como acima), uma situação que pode ser difícil de reconhecer e controlar. MNAR é o caso mais complexo entre as três categorias propostas por Rubin (1976).

A distinção de Rubin (1976) é importante para entender por que alguns métodos funcionam e outros não. Sua teoria estabelece as condições sob as quais um método de dados ausentes pode fornecer inferências estatísticas válidas (BUUREN, 2018).

A escolha do procedimento mais adequado para tratar os dados faltantes, também passa pela detecção do padrão dos dados faltantes. Através do padrão, é possível identificar se há ou não consistência no modo pelo qual os dados não foram observados (LITTLE; RUBIN, 2002).

Por exemplo, a Figura 10 exibe uma matriz 8x5 e o padrão de dados faltantes que se pode detectar. Esse padrão é chamado de Univariado (SILVA, 2012). É um padrão consistente, único e detectável e pode acontecer, por exemplo, se alguns participantes de uma pesquisa não respondem a um item do questionário. Neste padrão é apresentada a falta isoladamente em uma variável, o que é comum em estudos experimentais

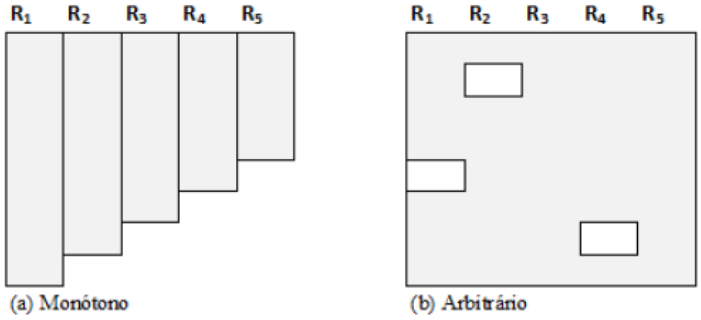
Figura 10 - Padrão de dados faltantes Univariado



Fonte: Veroneze (2011)

A Figura 11 apresenta dois exemplos de padrões de dados faltantes, em que uma matriz de dados tem o padrão Monótono e outra Arbitrário (SILVA, 2012). O Padrão monótono geralmente ocorre em pesquisas clínicas, onde os indivíduos participantes da pesquisa em algum momento não podem continuar no estudo devido à alguns fatores, por exemplo, reação de alguma droga. Este tipo de padrão de dados em falta é característico de experimentos longitudinais, sendo as variáveis medidas ao longo do tempo. Por fim no padrão arbitrário, que acontece quando alguns itens não são respondidos de forma aleatória.

Figura 11 - Padrões de dados faltantes monótono e arbitrário



Fonte: Veroneze (2011)

2.2.1 Métodos de imputação

Buuren (2018) realizou um estudo de técnicas para imputação de dados faltantes, onde são relacionados os seguintes métodos: exclusão dos dados com atributos faltantes, imputação pela média ou moda, imputação por meio da regressão de dados, da última observação realizada ou LOCF (*Last observation carried forward*), imputação pelo método do indicador e imputação múltipla. Além destes o autor cita a existência de outros métodos baseados em análise de probabilidade e procedimentos de ponderação como por exemplo redes bayesianas.

Uma forma de abordar o problema de dados faltantes é remover as instâncias que possuam falta de dados. A exclusão dos dados pode ser dividida em duas abordagens: exclusão *listwise* (ou análise de caso completo) e exclusão *pairwise* (exclusão por pares).

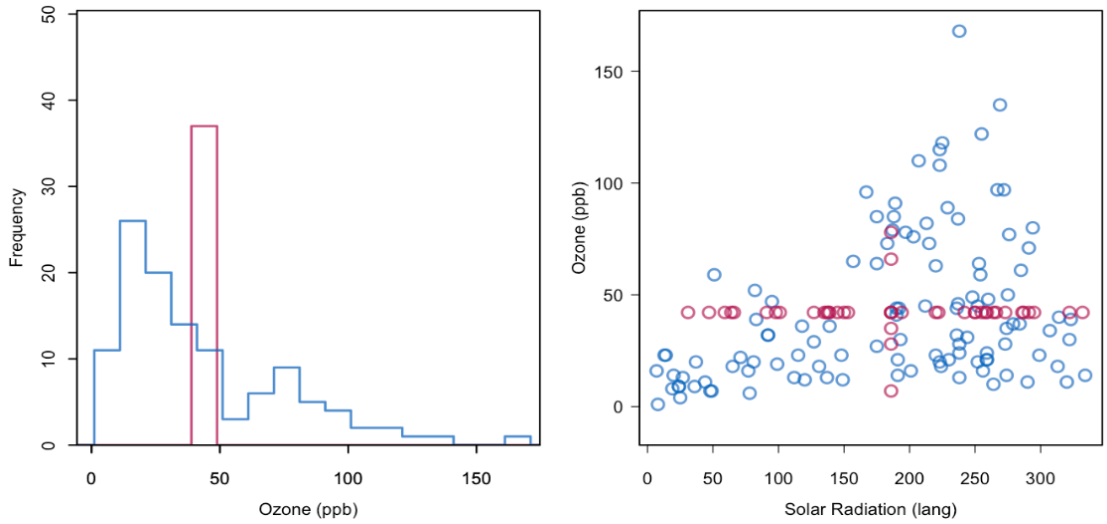
Na abordagem de exclusão *listwise* são eliminados todos os casos com um ou mais valores ausentes. Já a exclusão *pairwise* também conhecida como análise de caso disponível tenta remediar o problema da perda excessiva de dados da

exclusão *listwise*, sendo que o método analisa os exemplos em que as variáveis de interesse estão presentes, excluindo os casos em que não tem respostas completas quando está sendo feita a associação entre duas variáveis (BUUREN, 2018). Uma forma de análise da associação é realizar o cálculo das médias em todos os dados observados. Assim, a média da variável X é baseado em todos os casos com dados observados em X, a média da variável Y usa todos os casos com observados Y, e assim por diante.

Outro método para a realização da imputação de dados faltantes é substituí-los pela média, podendo ser utilizada a moda para dados categóricos. A imputação média é uma solução rápida e simples, no entanto irá subestimar a variância e pode enviesar qualquer estimativa quando os dados não são *MCAR* (BUUREN, 2018).

O Gráfico 1 ilustra este método utilizando a dispersão de dados, sendo que a esquerda há a frequência dos dados e a direita o valor dos dados. Os pontos azuis indicam os dados observados e os vermelhos os dados imputados.

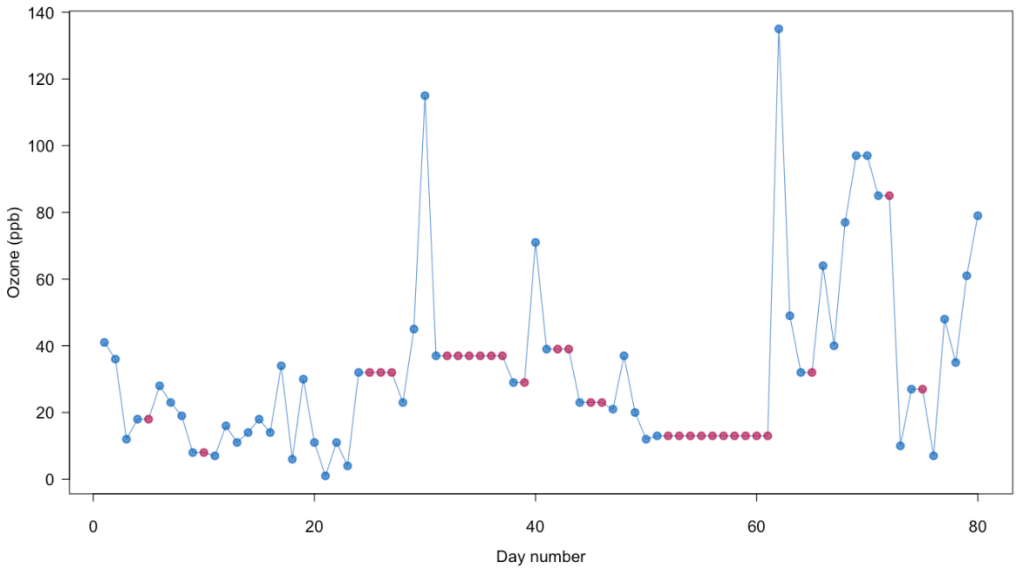
Gráfico 1 - Imputação pela média



Fonte: Buuren (2018)

O método de última observação realizada ou LOCF (*Last observation carried forward*) tem como objetivo tomar o valor observado anteriormente como um substituto para os dados ausentes. O Gráfico 2 ilustra a esta abordagem utilizando o gráfico de dispersão de dados. Os pontos em vermelho indicam as imputações e os azuis dados existentes.

Gráfico 2 - Imputação pela última observação realizada



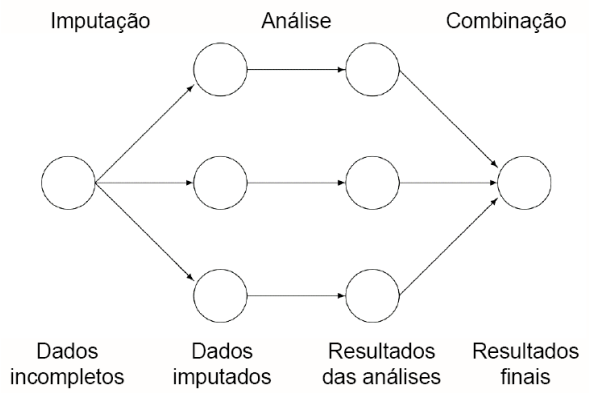
Fonte: Buuren (2018)

Para explicar o método indicador, suponha que se quer utilizar uma regressão, mas faltam valores em uma das variáveis explicativas. O método do indicador substitui cada valor em falta por um zero e prolonga-se o modelo de regressão pelo indicador de resposta. O procedimento é aplicado a cada variável incompleta. Este método é popular em saúde pública e epidemiologia. Uma vantagem é que o método do indicador conserva o conjunto de dados completo. O método do indicador pode ter seus usos em situações particulares, mas falha como método genérico para lidar com dados ausentes (BUUREN, 2018).

2.2.2 Imputação múltipla

A imputação múltipla cria várias versões completas dos dados, substituindo os valores ausentes por dados plausíveis. Considera-se valores plausíveis quando os valores estejam de acordo com determinada restrição definida. Esses valores são extraídos de uma distribuição modelada especificamente para cada entrada ausente. Por exemplo, a Figura 12 exemplifica a imputação utilizando três conjuntos. Os três conjuntos de dados imputados são idênticos para as entradas de dados observados, mas diferem nos valores imputados. A magnitude dessa diferença reflete na incerteza sobre qual valor imputar (BUUREN, 2018).

Figura 12 - Etapas da imputação múltipla



Fonte: Adaptado de Buuren (2018)

Imputação Múltipla é uma técnica estatística desenvolvida para tirar vantagem da flexibilidade em cálculos para tratar dados faltantes. Com isso, cada valor faltante é substituído por dois ou mais valores imputados, ao invés de apenas um valor, a fim de representar a incerteza sobre qual valor imputar (RUBIN, 1987).

A primeira etapa começa com a análise dos dados observados e incompletos. A imputação múltipla cria várias versões completas dos dados, substituindo os valores ausentes por valores de dados plausíveis. Esses valores são extraídos de uma distribuição modelada especificamente para cada valor ausente.

Na segunda etapa, as $x \geq 2$ bases de dados geradas são usadas em análises convencionais para testar os modelos estatísticos de interesse para o estudo. Cada uma das x bases de dados completas é analisada individualmente. Logicamente, como são realizadas x análises, serão geradas x estimativas de cada um dos parâmetros de interesse (VERONOZE, 2011). Ainda na segunda etapa é necessário agregar os resultados das análises feitas nas $x \geq 2$ bases de dados, gerando estimativas globais para os parâmetros de interesse e para o erro padrão.

Por fim, na última etapa são reunidas estimativas de parâmetros para estimar sua variância. A variância combina a variância de amostragem convencional (variância dentro da imputação) e a variância extra (variância entre imputação). Sob as condições apropriadas, as estimativas combinadas são imparciais e têm as propriedades estatísticas corretas (BUUREN, 2018). As condições apropriadas normalmente são validadas por meio de simulações a avaliações estatísticas dos resultados.

Para BUUREN (2018), a imputação múltipla é agora aceita como o melhor método para lidar com dados incompletos em muitos campos. A imputação múltipla foi desenvolvida por Donald B. Rubin na década de 1970. Na época Rubin observou que a imputação de um valor (imputação única) para o valor ausente não poderia ser correta em geral. Ele precisava de um modelo para relacionar os dados não observados aos dados observados e notou que, mesmo para um determinado modelo, os valores imputados não podiam ser calculados com certeza. Sua solução foi simples: criar várias imputações que refletem a incerteza dos dados ausentes (BUUREN, 2018).

2.3 Regressão

A regressão é uma técnica preditiva de análise de dados em que a variável alvo a ser avaliada é continua. Neste método, dado uma variável (ou um conjunto de variáveis) explicativas é possível estimar a variável alvo. Os atributos explicativos de uma tarefa de regressão podem ser discretos ou contínuos. Exemplos de aplicações de regressão incluem a previsão de um índice de bolsa de valores, a previsão da quantidade de precipitação em uma região baseada nas características dos ventos, a projeção do total de vendas de uma empresa baseada na quantidade gasta em publicidade e a avaliação da idade de um fóssil de acordo com a quantidade de carbono-14 presente no material orgânico (TAN *et. al.*, 2006).

Regressão também pode ser definida como uma tarefa de aprender uma função f que mapeie cada conjunto de atributos x em uma saída de valores contínuos y. O objetivo da regressão é encontrar uma função alvo que possa ajustar os dados de entrada com um erro mínimo (TAN *et. al*., 2006).

A imputação por regressão incorpora o conhecimento de outras variáveis, tentando produzir imputações plausíveis. A primeira etapa envolve a construção de um modelo a partir dos dados existentes. As previsões para os casos incompletos são então calculadas sob o modelo (BUUREN, 2018).

2.3.1 Regressão linear

Muitas vezes a posição dos pontos em um diagrama de dispersão sugere a existência de uma relação funcional entre duas variáveis. Surge então o problema

de se determinar uma função que exprima esse relacionamento. Essa relação funcional corresponderia a linha de regressão (NETO, 2002).

Existem diversos métodos para obtenção da reta desejada. O mais simples de todos, pode se chamar de método do ajuste visual, consiste simplesmente em traçar diretamente a reta com auxílio de uma régia, no diagrama de dispersão, procurando fazer da melhor forma possível, com que essa reta passe por entre os pontos (NETO, 2002). Esse procedimento entretanto, somente será razoável se a correlação linear for muito forte, caso contrário levará a resultados subjetivos.

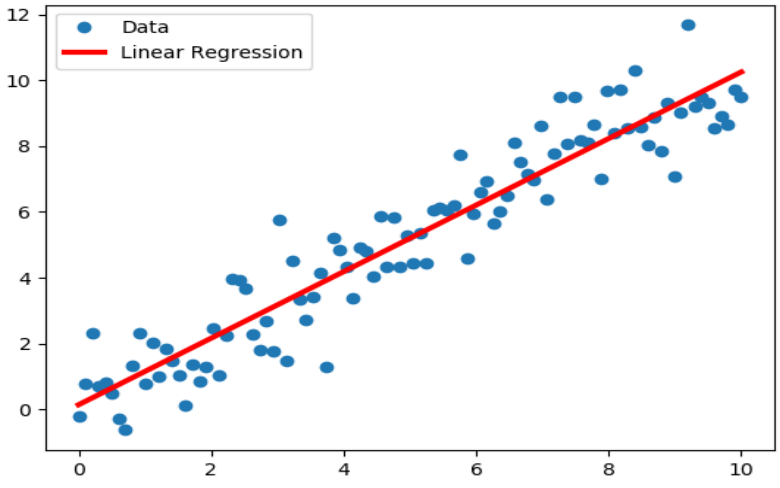
Por outro lado, a aplicação do princípio de máxima verossimilhança, leva nas condições admitidas, ao chamado método dos mínimos quadrados, segundo o qual a reta a ser adotada deverá ser aquela que torna mínima a soma dos quadrados das distâncias da rede aos pontos experimentais (NETO, 2002).

O modelo matemático da regressão linear pode ser ilustrado na Equação (3) onde é fornecido uma relação linear entre duas variáveis.

$$y = a + b(x) \tag{3}$$

Onde y é a variável dependente, a denota a intersecção no eixo y, b a inclinação da reta e x a variável independente. Utilizando o modelo é possível estabelecer uma linha em um gráfico de dispersão conforme ilustrado no Gráfico 5, sendo que a tarefa do modelo é encontrar uma curva parametrizada que corresponda aproximadamente a um conjunto de dados.

Gráfico 3 - Dispersão de dados com linha de regressão



Fonte: Autoria própria (2021)

A Equação (4) e a Equação (5) descrevem o cálculo para obtenção dos valores de a e b, em que $\bar{y}$ é a média dos valores de y e $\bar{x}$ representa a média dos valores de x.

$$a = \bar{y} - b\bar{x} \tag{4}$$

$$b = \frac{n\sum xy - \sum x \sum y}{n\sum x^2 - (\sum x)^2} \tag{5}$$

Existem diversas abordagens para obtenção dos valores de a e b. A abordagem utilizada neste trabalho é a que foi relatada nas Equações (4) e (5) é a mais comum e é chamada método dos mínimos quadrados.

2.3.2 Regressão polinomial

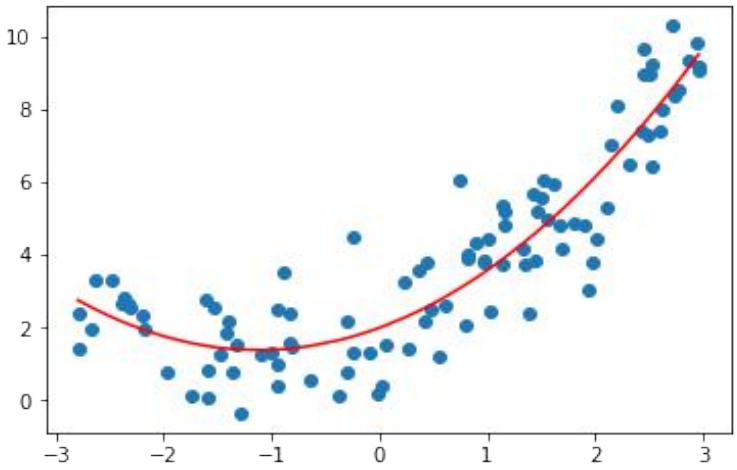
O mesmo princípio dos mínimos quadrados visto para a regressão linear poderá ser aplicado se admitirmos que a função de regressão é um polinômio de grau $k > 1$. A diferença está em que teremos $k + 1$ coeficientes a estimar (NETO, 2002).

Tomando as derivadas parciais em relação às $k + 1$ estimativas, chega-se a um sistema de $k + 1$ equações com $k + 1$ incógnitas o qual, resolvido, fornece a solução do problema (NETO, 2002). Assim no caso de se admitir que a função de regressão seja uma parábola na forma da Equação (6), onde é caracterizado como regressão polimomial de grau 2.

$$y = a + b(x)^2 \tag{6}$$

O Gráfico 6 ilustra o modelo delimitado por um polinômio quadrático. Observando o gráfico, é possível verificar que a criação do modelo utiliza uma parábola ao invés de uma reta. Isto pode diminuir a quantidade de erros do modelo quando comparados à regressão linear simples, pois muitas vezes os dados não estarão dispostos de forma linear como mostrado no Gráfico 5.

Gráfico 4 - Dispersão de dados com linha de regressão polinomial



Fonte: Autoria própria (2021)

2.3.3 Regressão múltipla

A ideia da correlação entre duas variáveis pode ser estendida para o caso de várias variáveis (NETO, 2002), ou seja, a regressão múltipla pode ser usada no intuito de melhorar o modelo desenvolvido para explicar o comportamento das variáveis do banco de dados que estão sendo estudadas.

Em regressão múltipla, a variável determinada é aquela que tenha correlação significativa com a variável a ser prevista. A variável está no centro das análises e deve ser identificado o seu impacto coletivo, assim como a contribuição de cada variável separada para o efeito geral da variável preditora.

A regressão linear múltipla é uma técnica multivariada cuja finalidade principal é obter uma relação matemática entre uma das variáveis estudadas (variável dependente ou resposta) e o restante das variáveis que descrevem o sistema (variáveis independentes ou explicativas). Sua principal aplicação, após encontrar a relação matemática, é produzir valores para a variável dependente quando se têm as variáveis independentes (cálculo dos valores preditos). Ou seja, ela também pode ser usada na predição de resultados, por meio da regra estatística dos mínimos quadrados (NETO, 2002).

2.3.4 Correlação de Pearson

O coeficiente de correlação r, ou correlação de Pearson (PEARSON; HERON, 1913), mostra a relação entre as variáveis dependente e independente.

Valores positivos indicam variação no mesmo sentido das variáveis. Valores negativos indicam variáveis inversamente relacionadas. Um valor nulo significa que não existe relação. Quanto mais próximo r estiver de 1, maior será a correlação (NETO, 2002). A Equação (7) mostra o cálculo para obtenção do valor de r.

$$r = \frac{n\sum xy - \sum x \sum y}{\sqrt{n(\sum x^2) - (\sum x)^2}\sqrt{n(\sum y^2) - (\sum y)^2}} \quad (7)$$

Onde n é a quantidade de linhas da amostra e x e y são as variáveis dependente e independente.

Além da correlação, o coeficiente de determinação R^2 pode ser utilizado para indicar a intensidade da correlação. Valores próximos a um são desejáveis e indicam que as variáveis dependente e independente estão fortemente correlacionadas (NETO, 2002). Em outras palavras, indica quanto o modelo foi capaz de explicar os dados originais. O coeficiente de determinação R^2 é obtido com o resultado da fórmula da Equação (7) sendo elevada ao quadrado.

Para Cohen (1988), valores entre 0.10 e 0.29 podem ser considerados pequenos; escores entre 0.30 e 0.49 podem ser considerados como médios; e valores entre 0.50 e 1 podem ser interpretados como grandes. Já para Mukaka (2012), valores entre 0.00 e 0.30 são insignificantes, de 0.30 a 0.50 são considerados correlação baixa, 0.50 a 0.70 correlação moderada, 0.70 a 0.90 correlação alta e acima de 0.90 correlação muito alta.

2.5 Testes estatísticos para validação dos resultados

Testes estatísticos são importantes para pesquisas que tem como objetivo comparar algoritmos, sendo que existem diversos tipos de testes para auxiliar as pesquisas. O testes estatísticos fornecem respaldo às pesquisas para que estas tenham aceitabilidade no meio científico. Os testes podem ser divididos em paramétricos e não-paramétricos, sendo que a diferença entre eles se refer ao tipo de variável estudada.

Devido ao não conhecimento da distribuição dos dados em experimentos realizados para problemas hierárquicos é indicado o uso de testes não-paramétricos, como o teste de Friedman (BORGES, 2012).

2.5.1 Teste de Friedman

O teste de Friedman é um teste não paramétrico (FRIEDMAN, 1940) recomendado para comparar o desempenho de vários algoritmos sobre diferentes bases de dados.

A ideia deste teste é que os algoritmos sejam organizados por postos de acordo com o desempenho, atribuindo 1 ao primeiro colocado, 2 ao segundo e assim sucessivamente. Após isso calcula-se a média dos postos obtidos pelos algoritmos sobre todos os conjuntos de dados usados nos experimentos. Posteriormente ao cálculo da média dos postos médios dos classificadores é necessário verificar a diferença estatística existente entre eles. Sendo que este cálculo é feito através da aplicação da Equação (8).

$$X_F^2 = \frac{12n}{k(k+1)}\left[\sum_j r_j^2 - \frac{k(k+1)^2}{4}\right] \tag{8}$$

em que n é a quantidade de base de dados e k é a quantidade de algoritmos e rj é o posto médio para o $j - esimo$ algoritmo.

De acordo com os autores Iman e Davenport (1980) o teste de Friedman é considerado muito conservador, e por isso sugerem a utilização da estatística F_F, definida na Equação (9), distribuída de acordo com a tabela F de Snedecor com $k - 1$ e $(k-1) * (n-1)$ graus de liberade.

$$F_F = \frac{(n-1)X_F^2}{n(k-1) - X_F^2} \tag{9}$$

2.5.2 Teste de Wilcoxon

O teste de Wilcoxon (1945) é utilizado para comparação de dois grupos relacionados onde a variável é ordinal. O teste classifica em postos a diferença entre os algoritmos e em seguida soma as diferenças positivas e negativas.

Na sequência calcula-se o valor de T que representa o menor das somas de postos com mesmo sinal (DESMAR, 2006).

Em seguida determina-se o valor de n que é o total das diferenças com sinal. Se $n \leq 25$ os valores críticos de T são tabelados, onde n representa o número de bases de dados avaliadas, descontados o número de empates. Já para valores de $N \geq 25$, utiliza-se a estatística z definida na Equação (10).

$$z = \frac{T - \frac{1}{4}n(n+1)}{\sqrt{\frac{1}{24}n(n+1)(2n+1)}} \tag{10}$$

A hipótese nula assume que a diferença de desempenho entre algoritmos não é significativa. Com nível de confiança $\alpha = 0{,}05$, a hipótese nula não pode ser rejeitada se $-1{,}96 \leq z \leq 1{,}96$.

2.6 Considerações finais do capítulo

Este Capítulo apresentou conceitos da classificação, descrevendo suas duas divisões: tarefas descritivas e preditivas. A tarefa preditiva, que utiliza o paradigma de aprendizagem supervisionada foi mais explorada, pois este trabalho utiliza este tipo de classificação. Foi ilustrado um fluxo de classificação com aprendizagem supervisionada, onde há o papel do atributo rótulo ou classe.

Também foram abordados conceitos de classificação hierárquica, onde o papel do rótulo está disposto em uma hierarquia de classes. Foram delimitadas as diferenças de representações de hierarquia, distinguindo árvore e DAG (*Directed Acyclic Graph*). Foram expostos conceitos de classificação multirrótulo, onde um exemplo pode possuir uma ou mais classes, além de descrever abordagens para resolução de problemas de classificação multirrótulo propostas na literatura.

Foi apresentada a regressão linear simples, onde o modelo é criado a partir da definição de uma linha em um gráfico de dispersão de dados. Esta abordagem de regressão é mais efetiva caso os dados estejam dispostos de forma linear, fato que nem sempre acontece em um cenário real.

Também foi abordada a regressão polinomial, onde ao invés do modelo ser definido por uma linha no gráfico de dispersão de dados, uma parábola é utilizada. Dessa forma, mesmo caso os dados sejam lineares, a curva da parábola irá minimizar os erros do modelo.

A regressão múltipla foi apresentada, sendo a terceira abordagem para utilização de regressão que será utilizada no método apresentado no Capítulo 4.

Foi descrita a correlação de Pearson, onde é possível estabelecer o quanto uma variável pode explicar outra dentro de um conjunto de dados. Este conceito é fundamental para verificar se é viável a utilização da regressão, já que caso a correlação seja muito pequena, outros métodos serão mais efetivos para criação do modelo de imputação.

Por fim foram apresentados os testes estatísticos de Friedman e Wilcoxon que são importantes para comparar resultados de experimentos como os que serão realizados utilizando o método deste trabalho.

3 MAPEAMENTO SISTEMÁTICO DA LITERATURA

Neste Capítulo é apresentado um mapeamento sistemático da literatura sobre a imputação de valores faltantes em bases de dados hierárquica multirrótulo. Na Seção 3.1 é descrito como foi organizada a pesquisa, que é dividida em dois eixos. A Seção 3.2 descreve a metodologia adotada para realização do estudo. Na Seção 3.3 foram levantadas questões a serem respondidas durante a pesquisa bibliográfica. A Seção 3.4 descreve as bases de pesquisa utilizadas, além dos termos para realização das buscas. A Seção 3.5 descreve a realização das buscas e os resultados encontrados em cada base de dados. Na Seção 3.6 são apresentados os procedimentos de filtragens e na Seção 3.7 os critérios de ordenação das publicações e documentos. Posteriormente na Seção 3.8, é apresentado os resultados obtidos com o mapeamento sistemático. Por fim, na Seção 3.9 é descrita as considerações do capítulo.

3.1 Organização da pesquisa

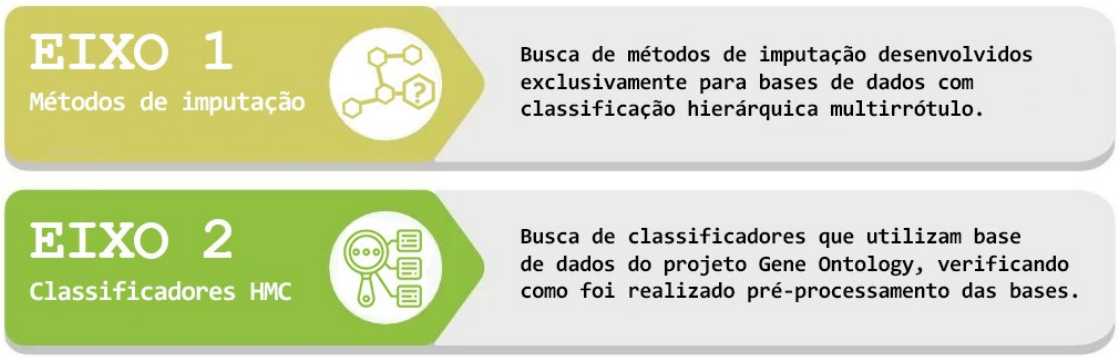
O mapeamento sistemático da literatura foi dividido em dois eixos, sendo o primeiro voltado a busca por trabalhos relacionados à imputação de dados faltantes em bases de dados com classificação hierárquica multirrótulo. No eixo 1 o objetivo principal foi elencar métodos de imputação que a literatura propôs para resolver problemas neste cenário específico de classificação hierárquica multirrótulo.

O eixo 2 da revisão eixo de pesquisa foram buscados trabalhos relacionados a classificadores hierárquicos multirrótulo que utilizam estrutura de dados do tipo DAG e apresentem experimentos que utilizam bases de dados do projeto *Gene Ontology*. O objetivo do eixo 2 da pesquisa foi verificar como foram conduzidos os experimentos destes trabalhos, pois muitas bases de dados do projeto *Gene Ontology* apresentam dados faltantes. Ou seja, verificando como ocorreu o pré-processamento das bases de dados dos experimentos destes trabalhos, é possível verificar quais métodos de imputação foram utilizados.

Buscando oferecer melhor compreensão dos dois eixos de pesquisa, a Figura 13 ilustra ambas as abordagens e seus objetivos. No eixo 1 foi realizada busca por trabalhos que descrevam métodos de imputação para dados faltantes em bases de dados com classificação hierárquica multirrótulo. Já o eixo 2 da pesquisa

busca verificar como os trabalhos relacionados a classificadores hierárquicos multirrótulo realizaram a imputação de dados faltantes.

Figura 13 - Os dois eixos de pesquisa do mapeamento sistemático da literatura



Fonte: Autoria própria (2021)

3.2 Descrição do método de mapeamento sistemático

Para elaboração do levantamento bibliográfico de ambos os eixos de pesquisa, foi utilizado um método adaptado da literatura proposto por Pagani, Kovaleski e Resende (2015) chamado *Methodi Ordinatio*. Neste método os trabalhos são classificados em um *ranking* de importância, utilizando critérios de avaliação e validação. Para calcular a ordem de importância são levados em conta fator de impacto, ano da publicação e número de citações. A Figura 14 apresenta as etapas do método.

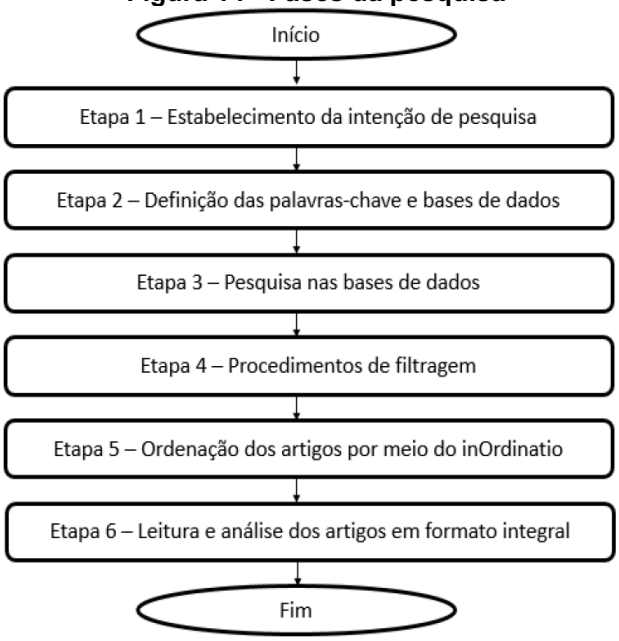
Durante a primeira etapa deve-se definir a intenção de pesquisa, juntamente com o levantamento de questões que devem ser respondidas pelo mapeamento sistemático da literatura. Ainda nesta etapa são definidas outras características da pesquisa como por exemplo o período das publicações. Neste trabalho, foram considerados estudos realizados entre os anos de 2009 e 2021.

Posteriormente são definidas palavras-chave e são realizadas pesquisas em diversos repositórios de artigos científicos. Ainda nesta etapa, deve-se realizar uma busca preliminar com palavras-chave relacionadas ao tema, buscando validar o resultado de busca obtido e verificar a necessidade de alteração e/ou combinação delas. Por fim, deve-se definir qual o gerenciador de bibliografia que será utilizado.

Sequencialmente é realizada a pesquisa nas bases de dados definidas anteriormente alimentando o gerenciador bibliográfico com os resultados. Logo após,

são executadas as etapas iniciais de filtragem. Neste mapeamento sistemático foi realizado uma adaptação das fases de pesquisa propostas por Pagani, Kovaleski e Resende (2015). A adaptação buscou redução da quantidade de etapas de forma a simplificar a representação. Por exemplo, originalmente são propostas as etapas 2 e 3, que descrevem a pesquisa preliminar de palavras-chave e definição e combinação de palavras-chave e bases de pesquisa, sucessivamente. Já na adaptação, estas duas etapas estão abarcadas na etapa 2, definição das palavras-chave e bases de dados.

Figura 14 - Fases da pesquisa



Fonte: Adaptado de Pagani, Kovaleski e Resende (2015)

Após a filtragem, é necessário realizar a identificação do ano, número de citações e fator de impacto de cada publicação, utilizando os índices JCR (Journal Citation Reports) e SJR (Scientific Journal Rankings). A ordenação dos resultados é realizada na quinta etapa, e a equação deste cálculo é mostrada na Equação (11).

$$InOrdinatio = \frac{F_i}{1000} + \alpha * [10 - (Aa - Ap)] + C_i \quad (11)$$

Em que F_i representa o Fator de impacto, Aa o ano atual em que o mapeamento sistemático está sendo realizado, Ap o ano da publicação e C_i o número de citações do artigo. O α é o número entre 1 e 10 que representa quão importante um estudo novo é em relação a um mais velho, portanto quanto maior o parâmetro, mas bem colocado ficarão os estudos mais novos.

Por fim, com a busca dos artigos selecionados é realizada a leitura dos documentos mais relevantes buscando responder aos questionamentos levantados.

3.3 Questões de pesquisa

Os dois eixos de pesquisa utilizam o mesmo método de pesquisa, porém cada um possui questões de pesquisa específicas. O Quadro 1 lista as questões de pesquisa para o eixo 1, onde o objetivo é verificar métodos de imputação desenvolvidos para base de dados com classificação hierárquica multirrótulo. Já o eixo 2 da pesquisa busca inspecionar como ocorreu o pré-processamento de dados em trabalhos relacionados a classificadores hierárquicos multirrótulo que utilizam estruturas do tipo DAG e realizaram experimentos em bases de dados do projeto *Gene Ontology*. As questões do eixo 2 estão listadas no Quadro 2.

Quadro 1 - Questões levantadas no eixo 1 da pesquisa

Item	Descrição
P1	Quais os métodos e técnicas de Imputação de Valores em base de dados utilizadas na Classificação Hierárquica Multirrótulo?
P2	Quais foram as áreas em que a Imputação de Valores foi realizada? Qual o formato da estrutura hierárquica das classes?
P3	Que tipo de base de dados foram utilizadas? Quais as características destas bases de dados?
P4	Qual foi a contribuição científica dos autores no trabalho? Ocorreu a criação de novo método ou aprimoramento de método existente?
P5	Os resultados do método foram comparados a outras abordagens?

Fonte: Autoria própria (2021)

Quadro 2 - Questões levantadas no eixo 2 da pesquisa

Item	Descrição
P1	O autor relatou se realizou imputação de dados faltantes nos experimentos realizados?
P2	Qual método utilizado para imputação dos dados faltantes?

Fonte: Autoria própria (2021)

3.4 Seleção das bases de pesquisa e termos de busca

Após a definição da intenção e questões de pesquisa, posteriormente, são definidas as bases de dados que serão utilizadas, sendo que no Quadro 3 são definidos estes repositórios, que foram utilizadas para ambos os eixos de pesquisa.

Quadro 3 - Definição das bases de pesquisa

Base de Pesquisa	Site
ACM DL	<https://dl.acm.org>
IEEE	<https://ieeexplore.ieee.org>
EmeraldInsight	<https://www.inderscience.com>
ScienceDirect	<https://www.sciencedirect.com>
Scopus	<https://www.scopus.com>
Springer	<https://link.springer.com>

Fonte: Autoria própria (2021)

Foram definidas palavras-chave em buscas preliminares. Nesta etapa se buscou maximizar os resultados relevantes para o tema da pesquisa. O Quadro 4 lista as palavras-chaves que foram definidas para ambos os eixos de pesquisa.

Quadro 4 - Definição das palavras chave da pesquisa

Termo na Língua Inglesa	Termo na Língua Portuguesa
Imputation of Missing data	Imputação de dado ausente
Missing value imputation	Imputação de valor ausente
Hierarchical multi-label classification	Classificação hierárquica multirrótulo
Hierarchical Classification	Classificação hierárquica
Multi-label Classification	Classificação multirrótulo
Gene ontology	Ontologia Genética
Bioinformatics	Bionformática

Fonte: Autoria própria (2021)

Com a definição das palavras-chave foi possível a combinação destas para formação das *strings* de busca em conjunto com os operadores lógicos AND e OR. As *strings* do eixo 1 de pesquisa estão descritas no Quadro 5. Os termos foram definidos buscando relacionar publicações referenciadas a imputação de dados faltantes e classificação hierárquica multirrótulo.

Já o Quadro 6 mostra a única *string* utilizada no eixo 2 da pesquisa. É possível observar a referência aos termos “*gene ontology*” ou “*bioinformatics*”, que foram definidos para buscar trabalhos relacionados a bioinformática ou ontologia genética, temas que foram amplamente abordados em publicações referentes a classificadores hierárquicos multirrótulo como nos trabalhos de Borges (2012) e Vens (2008). Também foram utilizados os termos "*Hierarchical multi-label*

classification" ou "*Multi-label hierarchical classification*" de forma a maximizar os resultados do eixo 2 da pesquisa.

Quadro 5 - Definição das strings de busca do eixo 1 de pesquisa

ID	***Strings* de Busca**
S1	*"Imputation of missing data" OR "missing value imputation" AND "Hierarchical Classification"*
S2	*"Imputation of missing data" OR "missing value imputation" AND "Multi-label Classification"*
S3	*"Imputation of missing data" OR "missing value imputation" AND "Hierarchical Multi-label Classification"*

Fonte: Autoria própria (2021)

Quadro 6 - A string de busca do eixo 2 da pesquisa

ID	***Strings* de Busca**
S4	*("Hierarchical multi-label classification" OR "Multi-label hierarchical classification") AND ("gene ontology" OR "bioinformatics")*

Fonte: Autoria própria (2021)

3.5 Realização das buscas

Concluída as etapas anteriores, foi possível a realização das pesquisas nas bases de dados citadas, sendo que o resultado do eixo 1 da pesquisa é apresentado na Tabela 1, nas colunas "S1", "S2" e "S3". O eixo 2 da pesquisa tem os resultados apresentados também na Tabela 1 na coluna "S4".

Tabela 1 - Resultado das buscas

Base de pesquisa	Eixo 1			Eixo2
	S1	S2	S3	S4
ACM DL	0	0	0	8
IEEE	0	0	0	7
EmeraldInsight	77	2	1	0
ScienceDirect	2	6	1	20
Scopus	1	1	0	27
Springer	8	9	1	78
Total	88	18	3	140

Fonte: Autoria própria (2021)

3.6 Procedimentos de filtragem

Após a realização da pesquisa nas bases de dados, foi realizada filtragem dos resultados seguindo etapas abaixo:

1) Eliminação de resultados duplicados.
2) Eliminação de todas as publicações de livros e capítulos de livros.

3) Separação dos resultados nas seguintes categorias: artigos de periódicos e artigos de conferências.

4) Exclusão de resultados que não apresentam relação com o tema proposto. Para realizar esta etapa foi considerado o título, resumo e palavras-chave.

Na Tabela 2 é ilustrado o resultado das etapas 1 e 2. Para esta etapa e para as demais descritas acima foi criado um repositório no gerenciador de bibliografia *Zotero*[1] e realizado os procedimentos de filtragem utilizando os recursos desta ferramenta. Nas Tabelas 2, 3 e 4 há informações de ambos os eixos de pesquisa.

Tabela 2 - Aplicação das etapas de filtragem 1 e 2 no processo de filtragem

	Busca inicial (eixo 1)	Filtragem 1 e 2 (eixo 1)	Busca inicial (eixo 2)	Filtragem 1 e 2 (eixo 2)
Total	109	73	140	106

Fonte: Autoria própria (2021)

Na terceira etapa as publicações são separadas por categoria, pois será realizada avaliação de forma diferente para publicações em periódico e em conferências. A Tabela 3 relaciona a divisão por tipo de publicação.

Tabela 3 - Aplicação das etapas 3 no processo de filtragem

Tipo de publicação	Filtragem 3 (eixo 1)	Filtragem 3 (eixo 2)
Periódicos	67	59
Conferências	6	47
Total	73	106

Fonte: Autoria própria (2021)

Na quarta etapa foi realizada a leitura do título, palavras-chave e resumo dos trabalhos, identificando os que possuem relação com tema da revisão sistemática. O eixo 2 da pesquisa possui uma particularidade, que foi verificado se os autores explicitaram como realizaram a imputação de dados faltantes, caso tenha sido necessário.

Tabela 4 - Aplicação da etapa 4 no processo de filtragem

Tipo de publicação	Filtragem 4 (eixo 1)	Filtragem 4 (eixo 2)
Periódicos	3	3
Conferências	2	6
Total	5	9

Fonte: Autoria própria (2021)

[1] https://www.zotero.org/

3.7 Critérios de ordenação

Os trabalhos resultantes da etapa de filtragem foram ordenados utilizando o *Methodi Ordinatio* de Pagani, Kovaleski e Resende (2015). Para o cálculo foi utilizado como Fator de impacto (Fi) os valores dos índices SJR, número de citações (Ci) obtido por meio do *Google Scholar* (GOOGLE SCHOLAR, 2021) e o valor de $\alpha = 10$.

O Quadro 7 e 8 mostram, os trabalhos dos eixos 1 e 2 da pesquisa, organizados conforme o ranking de *Methodi Ordinatio* de Pagani, Kovaleski e Resende (2015), sucessivamente, de acordo com seu número de identificação (ID).

Quadro 7 - Artigos de periódicos selecionados do eixo 1 da pesquisa

ID	Autores	Periódico	Título
1	Ma, Q.; Lee, W.; Fu, T.; Gu, Y. e Yu, G.	*Data Mining and Knowledge Discovery*	*MIDIA: exploring denoising autoencoders for missing data imputation*
2	Benahed, L. e Houichi, L.	*Environmental Monitoring and Assessment*	*The effect of simple imputations based on four variants of PCA methods on the quantiles of annual rainfall data*
3	Celton, M.; Malpertyu, A.; Lelandais, G. e de Brevern, A.	*BMC Genomics*	*Comparative analysis of missing value imputation methods to improve clustering and interpretation of microarray experiments*

Fonte: Autoria própria (2021)

Quadro 8 - Artigos de periódicos selecionados do eixo 2 da pesquisa

ID	Autores	Periódico	Título
1	Cerri, R.; Barros, R. C. e Carvalho, A. C.	*BMC Bioinformatics*	*Reduction strategies for hierarchical multi-label classification in protein function prediction*
2	Feng, S.; Fu, P. e Zheng, W.	*Biotechnology & Biotechnological Equipment*	*A hierarchical multi-label classification method based on neural networks for gene function prediction*
3	Feng, S.; Fu, P. e Zheng, W.	*Review of Scientific Instruments*	*A postprocessing method in the HMC framework for predicting gene function based on biological instrumental data*

Fonte: Autoria própria (2021)

Como o *Methodi Oridinatio proposto* por Pagani, Kovaleski e Resende (2015), foi criado somente para ordenação de artigos, foi realizada uma adaptação para ordenar as publicações em conferências, sendo que foi retirado o Fi (Fator de Impacto) do cálculo. O Quadro 9 e o Quadro 10 relacionam em ordem de importância

das publicações em conferências com seu número de identificação (ID), qual foi utilizado também como referência no restante do trabalho. É possível observar que a publicação de Galvão e Merschmann (2016) consta entre os resultados de ambos os eixos de pesquisa.

Quadro 9 - Artigos de conferências selecionados no eixo 1 da pesquisa

ID	Autores	Conferência	Título
4	Read, J.; Bifet, A.; e Abdessalem, T.	*Advances in Knowledge Discovery and Data Mining (22nd Pacific-Asia Conference, 2018)*	*Scalable Model-Based Cascaded Imputation of Missing Data*
5	Galvão, L. e Merschmann, L.	*Discovery Science (19th International Conference, 2016)*	*HSIM: A supervised imputation method for hierarchical classification scenario*

Fonte: Autoria própria (2021)

Quadro 10 - Artigos de conferências selecionados no eixo 2 da pesquisa

ID	Autores	Periódico	Título
1	Wehrmann, J; Cerri, R. e Barros, R.	*International Conference on Machine Learning*	*Hierarchical Multi-Label Classification Networks*
2	Cerri, R.; Barros, R. C. e De Carvalho, A. C. P. L. F.	*Proceedings of the International Joint Conference on Neural Networks (IJCNN 2015)*	*Hierarchical classification of Gene Ontology-based protein functions with neural networks*
3	Galvão, L. e Merschmann, L.	*Discovery Science (19th International Conference, DS 2016)*	*HSIM: A Supervised Imputation Method for Hierarchical Classification Scenario*
4	Fabris, F.; Freitas, A. A.	*IEEE Symposium on Computational Intelligence and Data Mining (CIDM)*	*Dependency network methods for Hierarchical Multi-label Classification of gene functions*
5	Barros, R. C.; Cerri, R.; Freitas, A. A. e De Caravalho, A. C. P. L. F.	*Machine Learning and Knowledge Discovery in Databases (European Conference, PKDD 2013)*	*Probabilistic Clustering for Hierarchical Multi-Label Classification of Protein Functions*
6	Borges, H. B. e Nievola, J. C.	*The 2012 International Joint Conference on Neural Networks (IJCNN)*	*Multi-Label Hierarchical Classification using a Competitive Neural Network for protein function prediction*

Fonte: Autoria própria (2021)

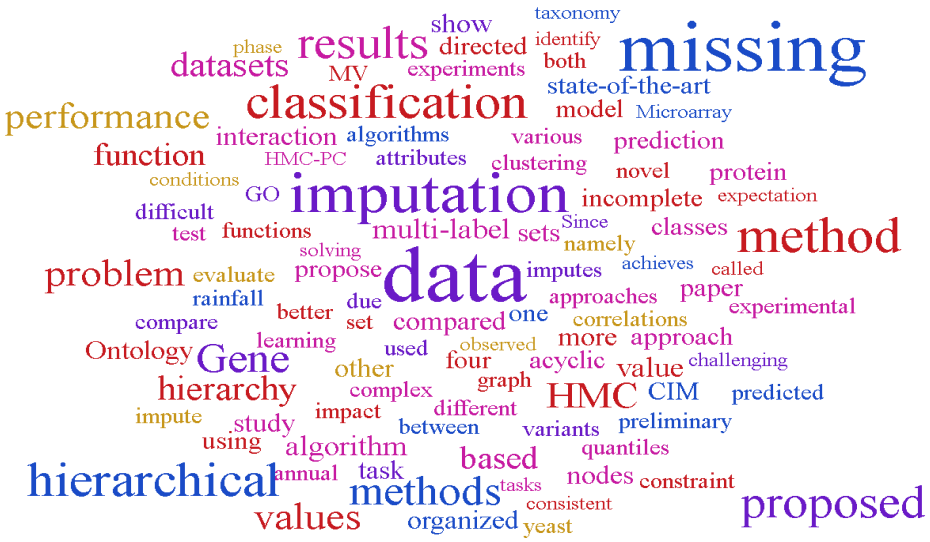
3.8 Resultados

Na etapa 6 do *Methodi Ordinatio* foi realizada a leitura completa dos artigos após as etapas de filtragens. Esta leitura busca responder as questões levantadas anteriormente para ambos os eixos de pesquisa.

A Figura 15 ilustra uma nuvem de palavras que é composta pelas palavras que mais apareceram nos resumos das publicações que foram resultados da pesquisa. É possível verificar que o termo que mais aparece nos resumos é *data*, seguido das palavras *missing* e *imputation.* Todas as palavras listadas são termos

na língua inglesa. Isso mostra que se levar em conta os resumos dos trabalhos, a pesquisa obteve sucesso na busca por trabalhos relacionados a imputação de dados faltantes ou do inglês *imputation missing data.*

Figura 15 - Nuvem de palavras resultante dos resumos das publicações



Fonte: Autoria própria (2021)

3.8.1 Eixo 1 da pesquisa – métodos de imputação

P1: Quais os métodos e técnicas de Imputação de Valores em base de dados utilizadas na Classificação Hierárquica Multirrótulo?

No eixo 1 da pesquisa desta revisão sistemática da literatura, não foram encontradas abordagens desenvolvidas exclusivamente para tarefas de imputação de dados faltantes no cenário de Classificação Hierárquica Multirrótulo.

Foi possível observar trabalhos onde foram propostas e testadas diversas abordagens para tratar os dados faltantes, mas nenhuma foi desenvolvida exclusivamente para tratar a classificação hierárquica multirrótulo. Estas técnicas são relatadas no Quadro 11.

Quadro 11 - Técnicas de imputação encontradas nos trabalhos

(continua)

ID	Método proposto
1	É proposto o método *Missing Data Imputation denosing* (MIDIA) ou imputação de dados ausentes com autoencoder de eliminação de ruído.
2	Não é proposto nenhum método, sendo realizado estudo dos efeitos da imputação utilizando o procedimento matemático para análise exploratória chamada *Principal Component Analysis* ou análise de componentes principais.

Quadro 12 - Técnicas de imputação encontradas nos trabalhos

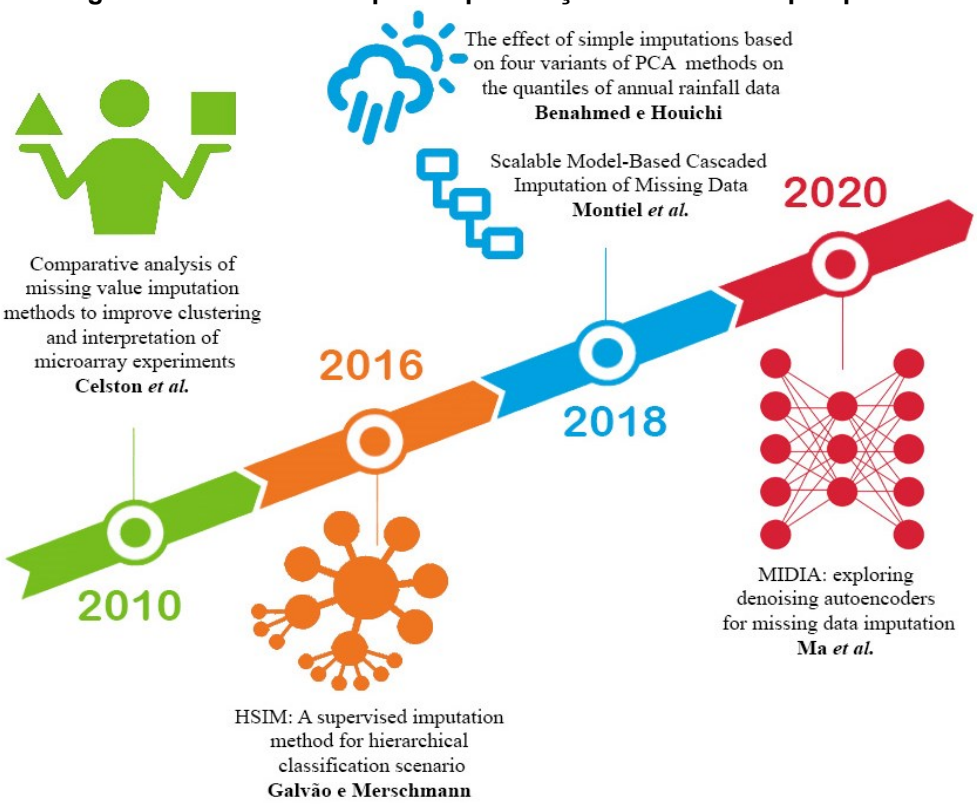
(conclusão

3	Não é proposto nenhum método, sendo realizado estudo comparativo dos métodos de imputação para melhorar o agrupamento e interpretação de experimentos em *microarray*.Os métodos comparados são baseados em *Expected Maximization*.
4	É proposto o método CIM (*Cascaded Imputation of Missing*).
5	É proposto o método HSIM (*Hierarchical Supervised Imputation Method*) ou Método de imputação hierárquica supervisionada.

Fonte: Autoria própria (2021)

A Figura 16 busca ilustrar a linha do tempo das publicações selecionadas no eixo 1 de pesquisa. Iniciando no ano de 2010, quando Celston *et al* (2016). realiza um estudo comparativo de métodos de imputação. Passando pelo ano de 2016 quando Galvão e Merschmann (2016) propõem um método de imputação para bases de dados com classificação hierárquica utilizando bases de dados da bioinformática. Em 2018 temos dois trabalhos de Benahmed e Houichi (2018) que propõem método de imputação para dados de precipitações de chuva e Montiel *et al* (2018). propõem um modelo de imputação em cascata. Por fim em 2020, Ma *et al*. (2010) utilizam autoencoders para realizar a imputação.

Figura 16 - Linha do tempo das publicações do eixo 1 da pesquisa



Fonte: Autoria própria (2021)

Apesar de não ser possível encontrar métodos desenvolvidos exclusivamente para o cenário de classificação hierárquica multirrótulo, os trabalhos foram selecionados pois se entendeu que alguns contribuíram com esta pesquisa seja por trabalhar com dados com classificação hierárquica ou por apresentarem métodos de imputação de dados.

Ma *et al*. (2020), propõe o método *Missing Data Imputation denoising Autoencoder* (MIDIA) ou imputação de dados ausentes com *autoencoder* de eliminação de ruído.

Benahmed e Houichi (2018), realizam estudo acerca dos efeitos dos métodos de imputação em dados anuais de precipitação de chuva.

Celston *et al*. (2010), realiza um estudo comparativo de métodos de imputação de dados faltantes para melhorar o agrupamento e interpretação de experimentos em microarray ou microarranjos.

Montiel *et al*. (2018) propuseram um método para trabalhar em base de dados multirrótulo utilizando classificação binária, porém este método desconsidera a hierarquia das classes.

Galvão e Merschmann (2016) propuseram um método para imputação considerando a classificação hierárquica e concluem o estudo incluindo como trabalho futuro a adição do cenário de classificação multirrótulo.

P2: Quais foram as áreas em que a Imputação de Valores foi realizada? Qual o formato da estrutura hierárquica das classes?

O Quadro 12 apresenta a distribuição das áreas das bases de dados em que as publicações foram norteadas e sua respectiva a hierarquia de classes. Somente a publicação de Galvão e Merschmann (2016) foi concebida sobre hierarquia DAG ou árvore. Nas demais publicações não é previsto hierarquia nas classes.

Quadro 13 - Área de aplicação e hierarquia das classes

ID	Área	Hierarquia
1	Computação / Social	Inexistente
2	Meteorologia / séries temporais	Inexistente
3	Bioinformática	Inexistente
4	Demografia / Música / Texto / Bioinformática	Inexistente
5	Bioinformática	DAG / *Tree*

Fonte: Autoria própria (2021)

P3: Que tipo de base de dados foi utilizada? Quais as características?

Ma *et al*. (2020), realizou experimentos sob três bases de dados: *Air Quality*, *Adult* e *Car*. A base *Air Quality* contém a resposta de um sensor de gás implantado em uma área poluída significativa da Itália. Ele contém 8.991 observações, onde cada observação tem oito atributos. *Adult* é um conjunto de dados de informações do censo que contém cerca de 32.000 observações com 14 atributos. A base *Car* contém 1782 observações em que cada observação tem seis atributos.

No trabalho Benahmed e Houichi (2018) foram utilizados dados de precipitações de chuva de 30 estações meteorológicas da Argélia considerando 69 anos (1936-2005). O estudo é conduzido para quatro diferentes percentagens de valores ausentes: 10%, 20%, 30% e 40%.

Celston *et al*. (2010) utiliza 5 conjuntos de dados da bioinformática. O percentual de dados faltantes das bases de dados é de 0,8%, 3%, 7,6%, 11,4%.

Os experimentos de Montiel *et al*. (2018) são realizados em 10 conjuntos de dados de diversos domínios.

Bases de dados com classificação hierárquica multirrótulo foram encontradas somente nos experimentos do método HSIM de Galvão e Merschmann (2016), onde foram realizadas análises sob 8 bases de dados da bioinformática relacionadas a funções genéticas. Originalmente estas bases de dados possuíam classificação multirrótulo, porém como o foco do método proposto foi a utilização da imputação em base de dados com classificação hierárquica monorrótulo, estas bases de dados foram convertidas, escolhendo somente uma classe para cada instância, sendo esta a mais frequente no conjunto de dados original. As bases utilizadas no trabalho de Galvão e Merschmann (2016) são as mesmas deste trabalho, sendo que posteriormente são detalhadas demais características como quantidade de dados faltantes, atributos e classes.

P4: Qual foi a contribuição científica dos autores no trabalho? Ocorreu a criação de novo método ou aprimoramento de método existente?

Ma *et al*. (2020), propõe o método *Missing Data Imputation denoising Autoencoder* (MIDIA) ou imputação de dados ausentes com *autoencoder* de eliminação de ruído. *Autoencoders* são redes neurais onde inicialmente o algoritmo codifica um registro de dados em um vetor latente de baixa dimensão que por sua vez é decodificado de volta para o registro de dados original a fim de incorporar as

propriedades inerentes e as correlações (geralmente não lineares) entre os atributos em um vetor latente. O *denoising Autoencoder* (*Autoencoder* de redução de ruído) utiliza uma entrada parcialmente corrompida e treina um modelo que recupera a entrada original não corrompida.

O MIDIA foi motivado pelas deficiências observadas na aplicação do *denosing Autoencoder* ao problema de imputação de dados faltantes. Sendo que dado um conjunto de dados, o MIDIA visa capturar as correlações ocultas entre os dados faltantes e não faltantes, e em seguida estimar valores a serem imputados.

Benahmed e Houichi (2018) realizaram estudo acerca dos efeitos dos métodos de imputação em dados anuais de precipitação de chuva. Foram utilizados quatro métodos de imputação para realização dos experimentos, que são *Expected-Maximization*, *Probabilistic PCA*, *Regularized PCA* e *Singular Value Decomposition* PCA. O método de *Expected-Maximization* teve desempenho superior aos demais.

Celston *et al*. (2010) realizaram um estudo comparativo de métodos de imputação de dados faltantes para melhorar o agrupamento e interpretação de experimentos em *microarray* ou microarranjos. São realizadas mais de 6.000.000 simulações para avaliar a qualidade de 12 métodos de imputação em cinco conjuntos de dados biológicos diferentes. O autor destaca um método chamado *EM_array* que é baseado em *Expected Maximization*. Este mesmo método também obteve os melhores resultados nos experimentos junto a *LSI_array*, *LSI_Combined* e *LSI_Adaptative*, sendo estes também baseados em *Expected Maximization*. Ou seja, os resultados dos experimentos conduzidos pelos autores determinaram que métodos baseados em *Expected Maximization* obtém melhores resultados no cenário de *microarray*.

Montiel *et al*. (2018), propuseram um modelo em cascata de imputação aplicando algoritmo de regressão linear de forma iterativa. O CIM (*Cascaded Imputation of Missing*) itera através das colunas com valores ausentes, realizando a imputação de forma incremental. O método realiza a imputação de dados numéricos e nominais e mitiga o impacto dos dados MAR e MCAR na classificação binária e de várias classes. Nos experimentos do autor, o CIM é comparado a outros 4 técnicas de imputação: Imputação constante, média/moda, *Expectation-Maximization* e *KNN*.

O método *Hierarchical Supervised Imputation Method* (HSIM), proposto por Galvão e Merschmann (2016), prevê a imputação média de valores considerando sua hierarquia. O algoritmo calcula a média dos valores existentes nas demais

instâncias com mesma classe da instância que possui valor faltante e caso não encontre, calcula a média dos valores existentes nas instâncias com classes ascendentes ou descendentes. O algoritmo trabalha com dados categóricos, neste caso substituindo a média pela moda. Os dados em que o algoritmo é executado são organizados em uma estrutura de árvore e para cada instância haverá somente classe, sendo, portanto, um método para classificação hierárquica monorrótulo.

Os experimentos de Galvão e Merschmann (2016) foram realizados em conjuntos de dados incompletos para avaliar o efeito do método de imputação proposto no desempenho do classificador hierárquico global *Global-Model Naive Bayes* (GMNB), uma extensão do classificador tradicional *Naive Bayes* para problemas de classificação hierárquica. Os resultados apresentados pelo HSIM utilizando o GMNB foram comparados a utilização das técnicas de *Mean Imputation* (MI), *Expected Maximization* (EM) e *K-Nearest Neighborhood* (KNN) todos também utilizando o classificador GMNB. Na maioria dos conjuntos de dados, o classificador GMNB alcançou um desempenho preditivo mais alto quando o conjunto de dados foi pré-processado usando o HSIM.

P4: Os resultados do método foram comparados a outras abordagens?

Ma *et al.* (2020) compara seu método os seguintes algoritmos: média, *KNN*, *Kernel, GBKII, Hot-deck*, Regressão linear multivariada, SVM, árvore de decisão, *Low-rank Matrix Recovery*, *Bayesian* PCA, dAE e MIDA.

Nos experimentos do estudo de Montiel *et al.* (2018), o CIM é comparado a outras 4 técnicas de imputação: Imputação constante, Média / moda, *Expectation-Maximization* e KNN.

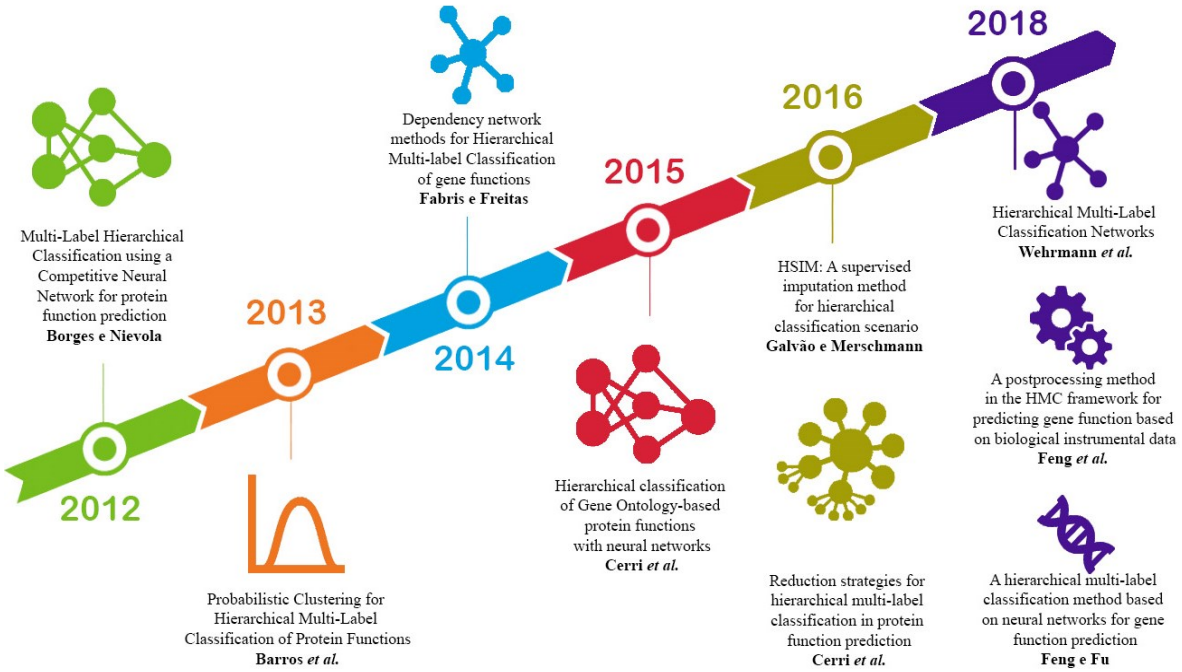
O método Hierarchical Supervised Imputation Method (HSIM), proposto por Galvão e Merschmann (2016) foram comparados a utilização das técnicas de *Mean Imputation* (MI), *Expected Maximization* (EM) e *K-Nearest Neighborhood* (KNN) todos também utilizando o classificador GMNB. Na maioria dos conjuntos de dados, o classificador GMNB alcançou um desempenho preditivo mais alto quando o conjunto de dados foi pré-processado usando o HSIM proposto.

3.8.2 Eixo 2 da pesquisa – pré-processamento dos classificadores

Como mencionado anteriormente, foram selecionadas somente publicações onde os autores explicitaram como realizaram a imputação de dados faltantes. Além

disso os experimentos foram executados em bases de dados da bioinformática. A Figura 17 busca ilustrar a linha do tempo das publicações selecionadas no eixo 2 de pesquisa.

Figura 17 - Linha do tempo das publicações do eixo 2 da pesquisa



Fonte: Autoria própria (2021)

Wehrmann *et al.* (2018), norteou sua pesquisa em desenvolver uma arquitetura de rede neural para classificação hierárquica multirrótulo capaz de otimizar perdas locais e globais para descobrir relações hierárquicas. No detalhamento dos experimentos realizados foi relatado que todos os dados faltantes foram substituídos pela média ou moda.

Feng e Fu (2018), propõe um método de classificação hierárquica multirrótulo baseada em redes neurais para as bases de dados do projeto Gene Ontology. Os autores descrevem que os dados ausentes foram substituídos pela média dos valores observados.

Cerri *et al.* (2016), estendeu um de seus trabalhos anteriores Cerri *et al.* (2015) em que foram utilizadas redes neurais em um classificador hierárquico multirrótulo para resolver o problema de predição genética. O autor utilizou neste trabalho a substituição dos dados ausentes pela média ou moda.

Cerri *et al.* (2015), propõe um classificador hierárquico multirrótulo utilizando redes neurais para resolver o problema de predição genética. O autor substituiu os dados ausentes numéricos pela média dos valores observados e utilizou a moda para atributos nominais.

A estudo de Feng *et al.* (2018) desenvolve um método pós-processamento no cenário de classificação hierárquica multirrótulo. O método de revisa os resultados preliminares e garante que as previsões são consistentes com a restrição de hierarquia. Nos experimentos deste estudo os dados faltantes foram substituídos pela média dos valores observados.

Fabris e Freitas (2014), propõe dois novos algoritmos usando um modelo gráfico probabilístico baseado em redes de dependência para resolver o problema de classificação hierárquica multirrótulo para de funções gênicas. Neste trabalho os autores relataram que descartaram dois conjuntos de dados devido a estes possuírem muitos (mais de 50%) de valores ausentes.

Barros *et al.* (2013), realiza proposta de classificador hierárquico multirrótulo utilizando agrupamento probabilístico. O autor também substitui os valores ausentes pela média dos valores do respectivo atributo.

O artigo de Borges e Nievola (2012), apresenta um algoritmo de classificação hierárquica utilizando a abordagem global, denominado MHC-CNN (*Multilabel Hierarchical Classification using Competitive Neural Network*). O método foi testado em alguns conjuntos de dados da área de bioinformática. Os autores descrevem que o critério usado para imputação de valores de atributos ausentes foi calcular a média aritmética das classes ancestrais mais próximas da classe à qual pertence a amostra que possui o atributo faltante.

3.9 Considerações do capítulo

Neste Capítulo foram apresentados os trabalhos encontrados por meio da revisão sistemática sobre imputação de dados faltantes em bases de dados de classificação hierárquica multirrótulo. A revisão foi dividida em dois eixos. O eixo 1 teve como objetivo verificar métodos de imputação para bases de dados com classificação hierárquica multirrótulo. O eixo 2 pesquisou trabalhos relacionados a

classificadores hierárquicos multirrótulo que utilizam estruturas do tipo DAG e tenham apresentado experimentos com dados do projeto *Gene Ontology*.

Ressalta-se que no eixo 1 da pesquisa, não foi possível verificar método de imputação desenvolvido exclusivamente para o cenário de classificação hierárquica multirrótulo utilizando o método proposto de revisão sistemática da literatura. Foi encontrado método para a classificação hierárquica e outro para classificação multirrótulo, mas, utilizando a metodologia descrita, não foi possível verificar estudo englobando ambos os cenários.

O eixo 2 teve como objetivo identificar trabalhos na literatura relacionados a problemas classificação hierárquica que utilizassem bases de dados com dados faltantes. Foi possível listar nove trabalhos relacionados a classificadores que explicitaram como a imputação de dados faltantes, sendo que nestas publicações o resultado foi unânime: utilização da média ou moda dos valores existentes para definir os valores ausentes.

Dessa forma, foi constatado que há carência na descrição detalhada dos métodos de imputação utilizados no cenário de classificação hierárquica multirrótulo e quando houve descrição a média ou moda foi utilizada. Este resultado norteou a criação do método de imputação do Capítulo 4.

4 IMPUTAÇÃO BASEADA NA REGRESSÃO PARA CLASSIFICAÇÃO HIERÁRQUICA MULTIRRÓTULO (IR-HMC)

Este Capítulo apresenta o método para imputação de dados faltantes em bases de dados com classificação hierárquica multirrótulo utilizando regressão. A Seção 4.1 mostra a descrição do método IR-HMC (*Imputation based on Regression for Hierarchical Multi-label Classification*) ou Imputação baseada na regressão para classificação hierárquica multirrótulo. Nesta seção são mostradas as etapas a serem seguidas pelo método, além de serem apresentados os fluxos geral e específicos do IR-HMC, são definidos os formatos e parâmetros de cada uma das três etapas. A Seção 4.2 apresenta um exemplo do método que realiza a imputação de dados faltantes em uma base de dados fictícia. Por fim, a Seção 4.3 relata as considerações finais do Capítulo.

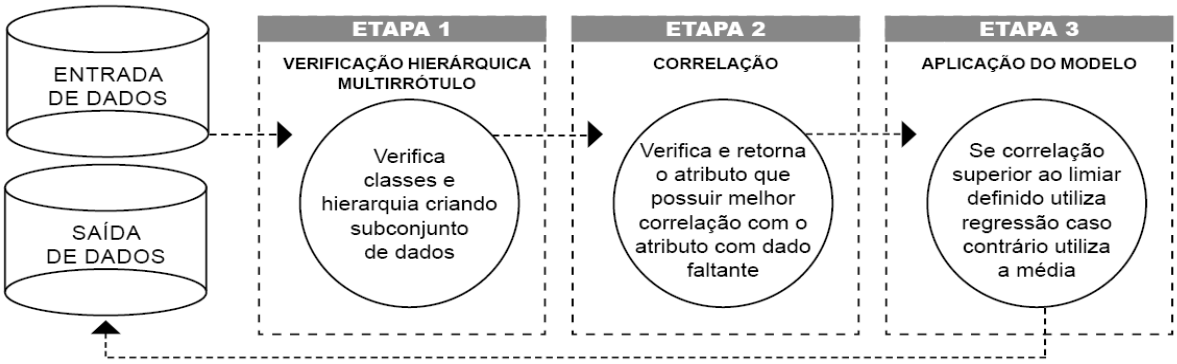
4.1 Descrição do método

O método IR-HMC, utiliza regressão caso exista correlação superior ao limiar definido e na hipótese do limiar ser inferior, a regressão é substituída pela média ou moda dos valores observados.

O método trabalha com três tipos de regressão: linear, múltipla e polinomial. Na regressão polinomial é possível parametrizar o grau do polinômio que seja deseja utilizar. Na hipótese da utilização da regressão múltipla serão utilizados todos os atributos que apresentarem correlação superior ao limiar definido.

O IR-HMC é ilustrado na Figura 18 e se divide em 3 etapas: Verificação hierárquica multirrótulo, Correlação, Aplicação do Modelo. As etapas são repetidas de forma iterativa até que todos os dados faltantes sejam imputados.

Figura 18 - Fases do método de imputação



Fonte: Autoria própria (2021)

A seguir serão detalhadas cada uma das etapas, além da entrada e saída de dados do método, onde também serão inseridos ilustrações e pseudocódigos para melhor compreensão do que ocorre dentro de cada etapa.

4.1.1 Entrada de Dados

O método tem como entrada de dados a base de dados hierárquica multirrótulo. Cada nível da hierarquia de classes é definido em uma linha do arquivo da entrada de dados, sendo que é utilizado o separador “/” para dividir a classe pai e filho, sucessivamente. Por exemplo, a linha “GO0003774/GO0003824” define que há classe GO0003774 é ancestral da classe GO0003824. Ainda dentro da entrada de dados, uma instância que possuir mais de uma classe deverá ser dividida pelo caractere “@”. Por exemplo uma instância que possua em como classe o conteúdo “GO0003774@GO0003824” pertencerá a ambas as classes.

4.1.2 Etapa 1 - Verificação hierárquica multirrótulo

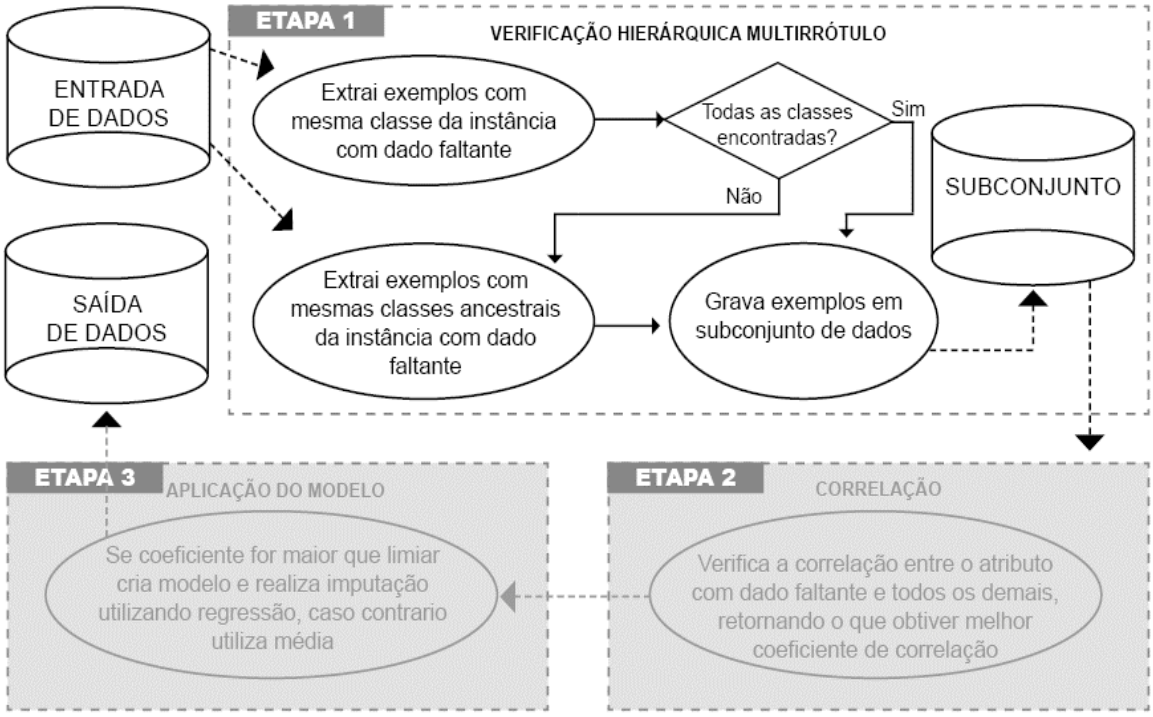
Na Etapa 1, verificação hierárquica multirrótulo, ilustrada na Figura 19, são analisadas as instâncias de cada atributo que possuir dados faltantes bem como os seus rótulos (classes). Para cada instância, que possui valor faltante, são verificados seus rótulos. Na sequência são examinados os rótulos de todas as instâncias da base de dados que possuem os mesmos rótulos da instância que está sendo analisada. Estes dados são copiados para um subconjunto de dados, que será utilizado posteriormente na segunda e terceira etapa. Por se tratar de problemas de classificação hierárquica, em que há relação de ancestralidade entre as classes da hierarquia, há ainda a necessidade de preservar esta relação. Sendo assim, caso não existam outras instâncias com os mesmos rótulos que a instância com dado faltante, é realizada análise da hierarquia das classes, extraindo exemplos com as mesmas classes ancestrais da instância com dado faltante. Ou seja, as instâncias das classes ancestrais do exemplo com dado faltante também irão compor o subconjunto de dados que que será utilizado na segunda e terceira etapa. Como o método também trata estruturas do tipo DAG, caso a classe do exemplo com dado faltante possuir mais de um ancestral, serão considerados todos os exemplos que possuírem esse rótulo.

Caso o subconjunto resultante possua dados faltantes será utilizada a média dos valores observados como modelo de imputação. Importante ressaltar que a

imputação destes valores se refere somente ao subconjunto gerado. Posteriormente todos os dados serão imputados utilizando o método IR-HMC. Outra característica do método é que como ele foi desenvolvido para as bases de dados da bioinformática, o IR-HMC considera que sempre haverá outras instâncias compartilhando rótulos, mesmo que não estejam no mesmo nível hierárquico.

A imputação pela média nos subconjuntos foi necessária, pois em testes preliminares em bases de dados com grande quantidade de dados faltantes, observou-se que o uso do IR-HMC nos subconjuntos faria com que em determinado momento o algoritmo não pudesse extrair um subconjunto de dados que compartilhasse rótulos, pois este subconjunto não existiria.

Figura 19 - Detalhamento da primeira etapa do método



Fonte: Autoria própria (2021)

Buscando melhor ilustrar a primeira etapa, o Quadro 14 mostra os passos do procedimento de verificação hierárquica multirrótulo em forma de pseudocódigo. Os parâmetros deste procedimento são: base de dados e vetor com os rótulos da amostra com dado faltante, conforme ilustra a linha 1. Na linha 2 é criada uma lista em que será inserido o subconjunto de dados que será utilizado na segunda etapa do método. Nas linhas 3 e 4 são ilustrados os laços de repetições, mostrando a iteração entre os rótulos do exemplo com dado faltante e todas as linhas da base de

dados, ou seja, cada rótulo do exemplo com dado faltante será comparado com os rótulos da base dados, extraindo os exemplos que irão compor o subconjunto de dados. Na linha 5 é verificado se o rótulo do exemplo com dado faltante está presente em outra instância da base de dados. Caso positivo, o item é inserido no subconjunto de dados, caso negativo, é inserido na base de dados os exemplos que possuam o mesmo rótulo ascendente do exemplo. Por fim, na linha 9 é retornado o subconjunto.

Quadro 14 - Passos do procedimento de verificação hierárquica multirrótulo

```
1  procedimento verificacao_hmc (entrada_dados, rotulos_dado_faltante)
2     subconjunto[ ] ← lista
3     para i = 1 até quantidade_rotulos faça
4        para j = 1 até numero_linhas_dados faça
5           se entrada_dados [ j ].classe contem rotulos_dado_faltante [ i ]
6              subconjunto.insere (dados [ j ].classe)
7           senão
8              subconjunto.insere (verifica_rotulos_ascendentes (rotulos_dado_faltante [ i ]))
9     retorne subconjunto
```

Fonte: Autoria própria (2021)

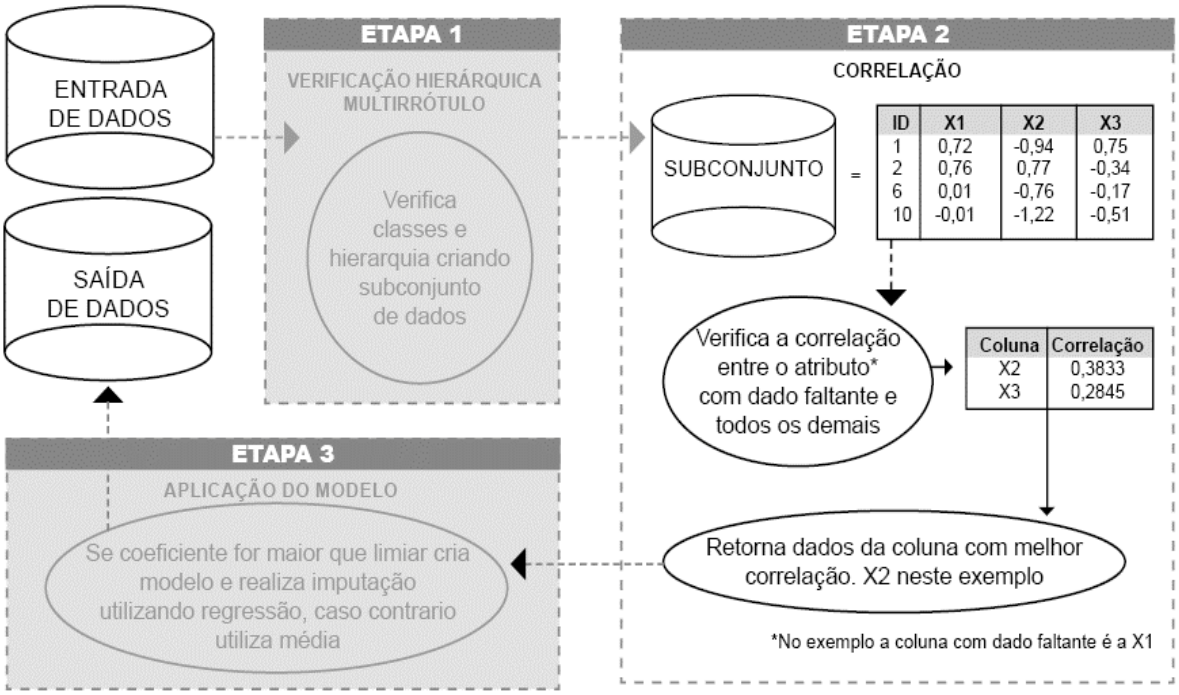
Realizado os procedimentos da Etapa 1, têm-se um subconjunto de dados somente com exemplos que compartilham rótulos da instância com dado faltante ou que possuam rótulos que estejam na hierarquia da classe da instância com dado faltante. Além disso, entende-se que esse subconjunto de dados resultante apresentará maior grau de similaridade com a instância que possui dado faltante se comparado com todo o conjunto de dados, pois fez-se essa verificação de suas classes e hierarquia de classes.

4.1.3 Etapa 2 - Correlação

Na Etapa 2, correlação, ilustrada na Figura 20, é realizada a verificação da correlação do atributo que possui dado faltante com todos os demais atributos do subconjunto, gerado na Etapa 1. Nesse caso é atribuído um coeficiente entre 0 e 1 para cada atributo deste subconjunto. O objetivo desta etapa é estimar o grau de explicabilidade do atributo faltante analisando todos os atributos do subconjunto gerado na Etapa 1. Após essa análise é encontrado o atributo que possui melhor correlação com o atributo com dado faltante, obtida por meio da aplicação da

Equação (7) apresentada no Capítulo 4. Ainda nesta etapa, caso o método esteja utilizando a abordagem de regressão múltipla, serão retornados todos os atributos que apresentarem coeficiente de correlação a superior ao limiar definido. Na Figura 20, é possível observar a etapa 1 já foi cumprida, sendo na ilustração também é possível verificar o subconjunto gerado e o grau de correlação do atributo X1 com os atributos X2 e X3, que são 0,3833 e 0,2845, respectivamente. Ao final o coeficiente de correlação mais elevado é retornado para a terceira etapa do método.

Figura 20 - Detalhamento da verificação da correlação



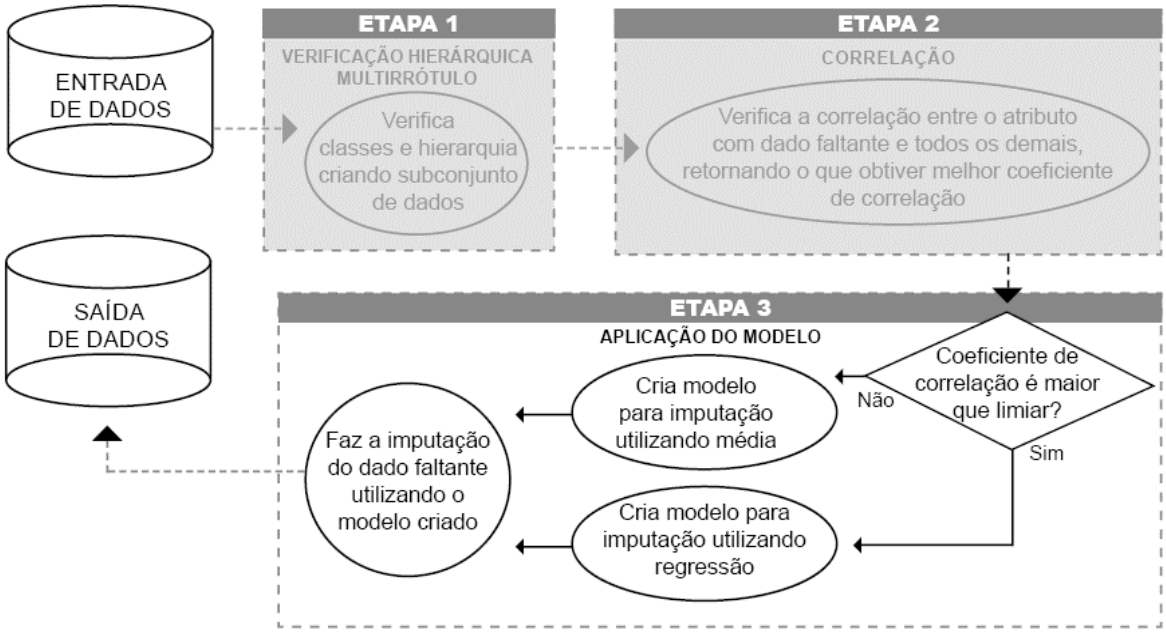
Fonte: Autoria própria (2021)

4.1.4 Etapa 3 – Aplicação do modelo

Nesta etapa é verificado se o atributo que possui melhor correlação com a coluna com dado faltante tem valor superior ao limiar definido. Caso esse valor for superior, um modelo de imputação com regressão será utilizado, caso contrário a regressão não será adotada, sendo que a média dos valores observados será a abordagem para imputar os dados faltantes. Esta etapa é ilustrada na Figura 21 em que inicialmente é feito uma verificação do coeficiente de correlação. Caso este valor seja superior ao limiar definido, será utilizado o modelo com regressão, caso contrário o modelo usado é a média.

Ao final desta etapa, se existirem mais dados faltantes as Etapa 1, 2 e 3 serão repetidas.

Figura 21 - Detalhamento da aplicação do modelo de imputação



Fonte: Autoria própria (2021)

4.1.5 Saída de Dados

Após a imputação de todos os valores faltantes tem-se a base de dados hierárquica multirrótulo com os valores imputados. O formato da base de dados é o mesmo da entrada, sendo preservados todas os atributos existentes e classes

4.1.6 Pseudocódigo geral do IR-HMC

O Quadro 13 apresenta o pseudocódigo geral do método de imputação IR-HMC. As linhas 1 e 2 ilustram os laços de repetições para haver iteração nas linhas, que correspondem as instâncias da base de dados, e as colunas que são os atributos da base de dados. A linha 3 verifica se há dado faltante. A linha 4 descreve Etapa 1 do método, em que é realizada a verificação dos rótulos da instância e relação de ancestralidade, demonstrado anteriormente no Quadro 14. A linha 5 mostra a Etapa 2 no qual é verificado o atributo com melhor correlação. Na linha 6 é verificado se o coeficiente de correlação encontrado é superior ao limiar definido, caso positivo na linha 7 é imputado o valor estimado utilizando a regressão linear, caso negativo na linha 9 é utilizada a média.

Quadro 15 – Verificação hierárquica multirrótulo

```
1  para i = 1 até numero_colunas_dados faça
2     para j = 1 até numero_linhas_dados faça
3        se dados[ i ] [ j ] == Nulo então
4           subconjunto[ ][ ] ← verificacao_hmc (entrada_dados, rotulos_dado_faltante, hierarquia)
5           atributo_melhor_correlacao ← verifica_correlacao (subconjunto, coluna_dado_faltante)
6           se atributo_melhor_correlacao.correlacao > limiar então
7              entrada_dados[ i ] [ j ] ← regressao_linear (atributo_melhor_correlacao, subconjunto)
8           senão
9              entrada_dados[ i ] [ j ] ← média(subconjunto)
```

Fonte: Autoria própria (2021)

Nota-se que o IR-HMC realiza uma varredura em cada célula da base de dados buscando dados faltantes, buscando a criação de um subconjunto para verificar se a correlação é superior ao limiar definido. A criação do subconjunto (linha 4) remete a uma outra função descrita no pseudocódigo do Quadro 14.

4.2 Exemplo do método

Para ilustrar o funcionamento do método IR-MHC, o processo será exemplificado utilizando uma base de dados fictícia com classificação hierárquica multirrótulo apresentada na Tabela 5. Já a hierarquia das classes desta base é mostrada na Figura 22. Para a simulação serão seguidas todas as etapas do método descritas na seção anterior.

Neste exemplo será abordado somente a regressão linear simples, porém nos posteriormente nos experimentos do método serão realizados testes com a utilização das abordagens de regressão linear, regressão polinomial e regressão múltipla. Na hipótese de utilização da regressão polinomial um dos parâmetros adicionais do método será o grau do polinômio que deseja.

Como trata-se de um exemplo fictício com fins didáticos, o limiar de correlação definido foi de 0,2 para que a regressão seja utilizada em todas as instâncias com dados faltantes, ou seja, neste exemplo a média não será adotada em nenhuma instância. Vale ressaltar que em uma base de dados real deve ser adotado algum critério para definição do limiar de correlação.

Tabela 5 - Base de dados fictícia

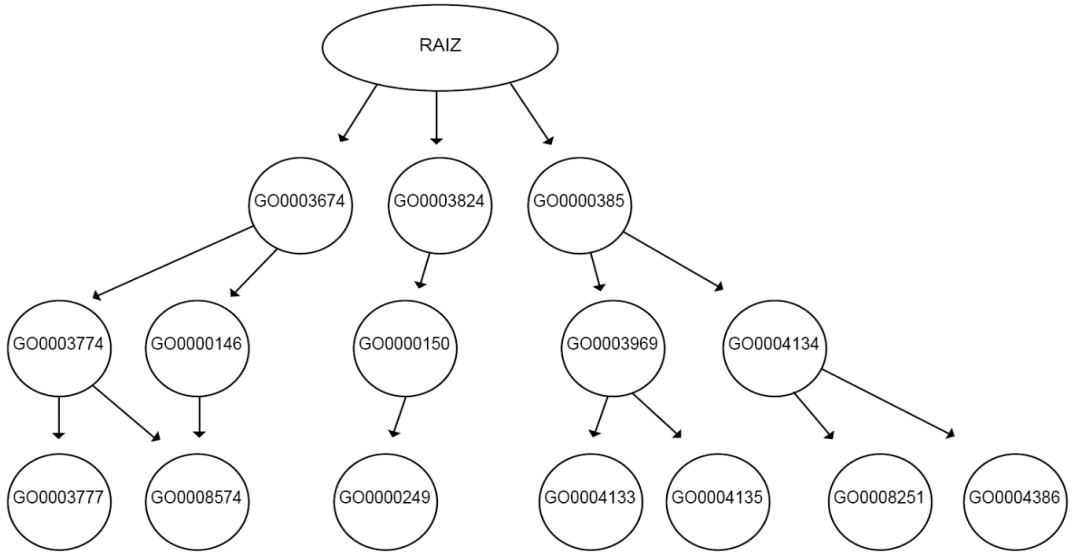
ID	X1	X2	X3	Classe
1	0,72	-0,94	0,75	GO0003774@GO0003824@GO0004134
2	0,76	0,77	-0,34	GO0003824@GO0003969
3	?	0,66	-0,04	GO0003824@GO0003774@GO0004134
4	-0,74	-0,84	-0,2	GO0000150@GO0000249
5	-0,36	-0,58	?	GO0000146@GO0004133@GO0003777
6	0,01	-0,76	-0,17	GO0003824@GO0004135
7	-0,45	-0,84	0,56	GO0003674@GO0000150
8	0,36	-1,03	-0,84	GO0000146@GO0008251
9	0,03	?	-0,42	GO0004133@GO0008574
10	-0,01	-1,22	-0,51	GO0003774@GO0004386

Fonte: Autoria própria (2021)

Cada linha da Tabela 6 representa um exemplo, sendo que os dados faltantes são representados pelo caractere “?”. A primeira coluna “ID” representa o identificador sequencial, e as colunas X1 a X3 representam os atributos (características) da base de dados. Como trata-se de um conjunto de dados que já passou por uma etapa de classificação, a última coluna representa o atributo classe ou rótulo de cada instância. Como se trata de uma base de dados multirrótulo, cada classe é separada pelo caractere “@”.

A Figura 22 ilustra a hierarquia das classes do conjunto de dados que será utilizado na simulação. Importante observar que se trata de uma estrutura hierárquica do tipo DAG.

Figura 22 - Hierarquia das classes



Fonte: Autoria própria (2021)

Entrada de Dados

A base de dados fictícia é formada por 10 (dez) instâncias e 3 (três) atributos. Além disso, cada instância possui uma variação na quantidade de classes entre 2 (dois) a 3 (três) rótulos, além disso a hierarquia é composta por 15 (quinze) classes.

Etapa 1 – Verificação hierárquica multirrótulo

O método é iniciado pela primeira coluna, analisando o atributo X1. Nota-se que há um dado faltante nesta coluna com número de identificação (ID) igual a 3. Na Etapa 1 é preciso verificar quais outras instâncias compartilham rótulo com a instância que possui dado faltante.

A instância com número de identificação 3 possui três rótulos sendo GO0003774, GO0003824 e GO0004134. Verificando o primeiro rótulo GO0003774, é possível observar que as instâncias com números de identificação 1 e 10 compartilham o mesmo rótulo. Analisando o rótulo GO0003824 é verificado que as instâncias 1, 2 e 6 compartilham esta classe.

Por fim, a classe GO0004134 é compartilhada pela instância com número de identificação 1. Eliminando os rótulos em duplicidade restarão os exemplos com números de identificação 1, 2 e 6, sendo que este será o subconjunto que será utilizado na etapa 2 do método.

Neste exemplo não é necessário verificar as classes ancestrais da hierarquia de classes, pois foi possível encontrar para cada uma das três classes outras instâncias que também a compartilha do mesmo rótulo. A Tabela 6 ilustra o subconjunto de dados após a conclusão da primeira etapa do método.

Tabela 6 - Subconjunto gerado na verificação dos rótulos da instância 3

ID	X1	X2	X3	Classe
1	0,72	-0,94	0,75	GO0003774@GO0003824@GO0004134
2	0,76	0,77	-0,34	GO0003824@GO0003969
6	0,01	-0,76	-0,17	GO0003824@GO0004135
10	-0,01	-1.22	-0,51	GO0003774@GO0004386

Fonte: Autoria própria (2021)

Etapa 2 - Correlação

Na Etapa 2 é verificado qual dos dois atributos, X2 ou X3, possui melhor correlação com o atributo X1 ou seja, qual desses atributos tem maior grau de explicabilidade com o atributo X1.

Para isso, é necessário aplicar o cálculo de r ilustrado no Capítulo 4, Equação (7) e obter o coeficiente de correlação R^2. Todo o cálculo é realizado utilizando o subconjunto gerado na primeira etapa. Aplicando fórmula têm-se os valores de 0.3833 para o atributo X2 e 0.2845 para o atributo X3. Dessa forma, nota-se que o atributo X2 possui o melhor coeficiente de correlação e, portanto, explica melhor as variáveis do atributo X1.

Etapa 3 – Aplicação do modelo

Na Etapa 3 é criado o modelo de regressão linear entre os atributos X1 e X2. Este cálculo também é definido no Capítulo 4, Equação (5). Já os cálculos para definição da inclinação da reta de regressão e intersecção no eixo y são demonstradas no mesmo capítulo nas Equações (4) e (5), respectivamente.

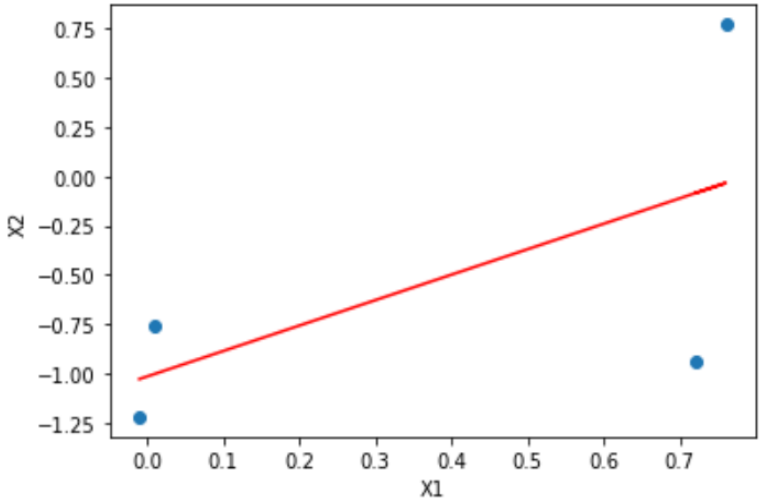
Realizados os cálculos, o modelo de imputação do exemplo é definido na Equação (12).

$$y = -1{,}0153 + 1{,}2914(x) \qquad (12)$$

Substituindo o x na equação do modelo pelo valor 0.66, presente no atributo X2 no exemplo com dado faltante, obtêm-se o resultado de -0.1629, sendo este o valor estimado para ser imputado. Realizada a imputação, a terceira e última etapa do método é completada.

O Gráfico 7 ilustra a dispersão de dados de X1 e X2 e a linha de regressão linear do exemplo proposto.

Gráfico 5 - Modelo de regressão para a imputação da instância ID 3



Fonte: Autoria própria (2021)

A Tabela 7 apresenta a base de dados após ser realizado o procedimento de imputação na primeira coluna (X1) para o atributo com número de identificação 3, sendo que este valor está destacado.

Tabela 7 - Conjunto de dados após imputação da primeira coluna

ID	X1	X2	X3	Classe
1	0,72	-0,94	0,75	GO0003774@GO0003824@GO0004134
2	0,76	0,77	-0,34	GO0003824@GO0003969
3	**-0,16**	0,66	-0,04	GO0003824@GO0003774@GO0004134
4	-0,74	-0,84	-0,2	GO0000150@GO0000249
5	-0,36	-0,58	?	GO0000146@GO0004133@GO0003777
6	0,01	-0,76	-0,17	GO0003824@GO0004135
7	-0,45	-0,84	0,56	GO0003674@GO0000150
8	0,36	-1,03	-0,84	GO0000146@GO0008251
9	0,03	?	-0,42	GO0004133@GO0008574
10	-0,01	-1,22	-0,51	GO0003774@GO0004386

Fonte: Autoria própria (2021)

Na próxima coluna, atributo X2, é possível verificar um dado faltante no exemplo com número de identificação 9. Sendo assim, inicia-se a execução das Etapas 1, 2 e 3.

Etapa 1 - Verificação hierárquica multirrótulo

A instância com número de identificação 9 possui dois rótulos, sendo que o rótulo GO0004133 é compartilhado pelo exemplo com identificação 5. Já o rótulo GO0008574 não é encontrado em nenhum outro exemplo, sendo necessário verificar os ancestrais na hierarquia das classes.

A classe GO0008574 possui duas classes ancestrais na hierarquia: GO0003774 e GO0000146. Essas classes são rótulos das instâncias com números de identificação 1, 3, 10, 5 e 8. Neste exemplo, devido a estrutura das classes ser do tipo DAG, o subconjunto irá conter exemplos que possuem as duas classes ancestrais. O subconjunto gerado após a verificação dos rótulos e hierarquia é apresentado na Tabela 8. É possível observar que este subconjunto já possui o valor imputado na iteração anterior. Além disso, na instância com identificação número 5, originalmente tem-se um dado faltante na coluna X3, sendo que para estes casos onde o subconjunto apresenta um dado faltante, o valor será substituído pela média dos valores observados, que neste caso possui valor igual a -0,16.

A imputação deste valor será utilizada somente no subconjunto gerado que será utilizado para criação do modelo de imputação da instância 9. Na iteração da coluna X3 será estimado um valor utilizando o método proposto. Dessa forma, tem-se o subconjunto de dados após aplicação da primeira etapa do método.

Tabela 8 - Subconjunto gerado na verificação dos rótulos da instância 9

ID	X1	X2	X3	Classe
5	-0,36	-0,58	-0,16	GO0000146@GO0004133@GO0003777
1	0,72	-0,94	0,75	GO0003774@GO0003824@GO0004134
3	-0,16	0,66	-0,04	GO0003824@GO0003774@GO0004134
8	0,36	-1,03	-0,84	GO0000146@GO0008251
10	-0,01	-1,22	-0,51	GO0003774@GO0004386

Fonte: Autoria própria (2021)

Etapa 2 - Correlação

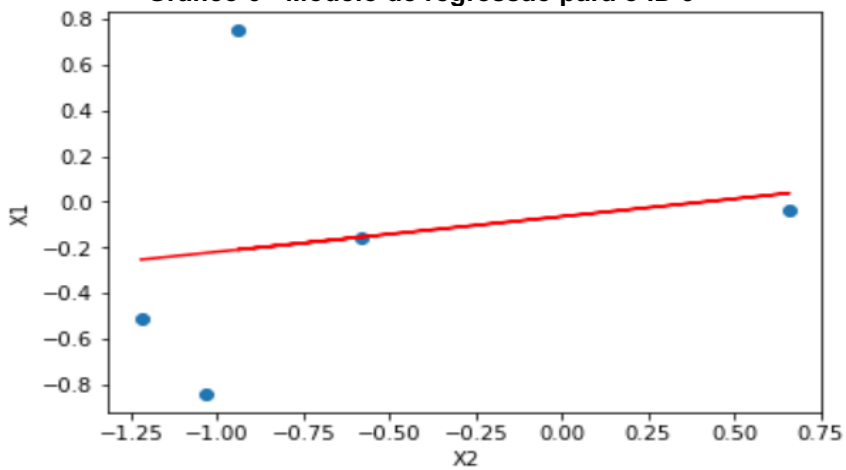
Na Etapa 2 é calculado as correlações entre os atributos, sendo que o coeficiente de correlação entre o atributo X2 e o atributo X1 tem o valor de 0.2064 enquanto o coeficiente entre o atributo X2 e o atributo X3 é de 0.0380. Logo, é possível concluir que neste caso o atributo X2 é melhor explicado pelo atributo X1.

Etapa 3 – Aplicação do modelo

Seguindo para a Etapa 3 é criado o modelo de imputação, definido na Equação (13) e a linha de regressão linear, ilustrada no Gráfico 8. Aplicando o modelo, o valor a ser imputado é de -0.06 (com o arredondamento).

$$y = -0{,}0637 + 0{,}1546(x) \tag{13}$$

Gráfico 6 - Modelo de regressão para o ID 9



Fonte: Autoria própria (2021)

A Tabela 9 mostra a base de dados após ser realizado o procedimento de imputação de valores coluna X2 para o atributo com número de identificação 9.

Tabela 9 - Conjunto de dados após imputação da segunda coluna

ID	X1	X2	X3	Classe
1	0,72	-0,94	0,75	GO0003774@GO0003824@GO0004134
2	0,76	0,77	-0,34	GO0003824@GO0003969
3	**-0,16**	0,66	-0,04	GO0003824@GO0003774@GO0004134
4	-0,74	-0,84	-0,2	GO0000150@GO0000249
5	-0,36	-0,58	?	GO0000146@GO0004133@GO0003777
6	0,01	-0,76	-0,17	GO0003824@GO0004135
7	-0,45	-0,84	0,56	GO0003674@GO0000150
8	0,36	-1,03	-0,84	GO0000146@GO0008251
9	0,03	**-0,06**	-0,42	GO0004133@GO0008574
10	-0,01	-1,22	-0,51	GO0003774@GO0004386

Fonte: Autoria própria (2021)

Analisando dado faltante na base de dados, observa-se que há um valor faltante no atributo X3. Assim, é necessário a realização da Etapa 1 novamente.

Etapa 1 – Verificação hierárquica multirrótulo

O próximo dado faltante se encontra no exemplo com número de identificação 5, na terceira coluna no atributo X3. Os rótulos deste exemplo são GO0000146, GO0004133 e GO0003777. Observando o conjunto de dados, é possível encontrar as instâncias com números de identificação 8 e 9 que compartilham os rótulos GO0000146 e GO0004133. Já no que se refere ao o rótulo GO0003777 não é possível encontrar outras instâncias que compartilham dessa classe. Sendo assim, é necessário verificar as classes ancestrais da hierarquia de classes. Neste caso, o rótulo ascendente mais próximo a GO0003777 é o rótulo GO0003774. Esse rótulo é encontrado nas instâncias com número de identificação 1, 3 e 10. A Tabela 10 mostra o subconjunto após a conclusão da primeira etapa.

Tabela 10 - Subconjunto gerado na verificação dos rótulos da instância 5

ID	X1	X2	X3	Classe
8	0,36	-1,03	-0,84	GO0000146@GO0008251
1	0,72	-0,94	0,75	GO0003774@GO0003824@GO0004134
3	-0,16	0,66	-0,04	GO0003824@GO0003774@GO0004134
9	0,03	-0,6	-0,42	GO0004133@GO0008574
10	-0,01	-1,22	-0,51	GO0003774@GO0004386

Fonte: Autoria própria (2021)

Etapa 2 - Correlação

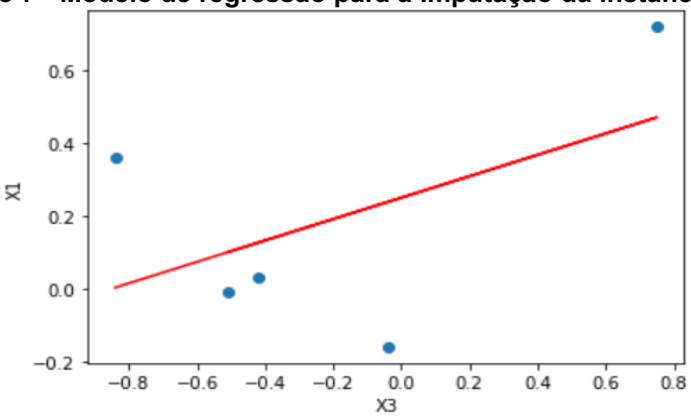
Na Etapa 2 é calculado os coeficientes de correlação, sendo que entre o atributo X3 e o atributo X1 o valor é de 0.2574, já entre o atributo X3 e o atributo X2 é de 0.0352, sendo possível concluir que o atributo X3 é melhor explicado por X1.

Etapa 3 – Aplicação do modelo

A linha de regressão linear realizada com estes dados é ilustrada na Gráfico 9 e o modelo de imputação é definido na Equação (14). Aplicando o modelo, o valor a ser imputado é de 0.14, concluindo então a terceira etapa do IR-HMC.

$$y = 0{,}2503 + 0{,}2941(x) \quad (14)$$

Gráfico 7 - Modelo de regressão para a imputação da instância ID 5



Fonte: Autoria própria (2021)

A Tabela 11 apresenta a base de dados após a conclusão da terceira etapa e conclusão da execução do método de imputação.

Tabela 11 - Conjunto de dados após imputação da terceira coluna

ID	X1	X2	X3	Classe
1	0,72	-0,94	0,75	GO0003774@GO0003824@GO0004134
2	0,76	0,77	-0,34	GO0003824@GO0003969
3	**-0,16**	0,66	-0,04	GO0003824@GO0003774@GO0004134
4	-0,74	-0,84	-0,2	GO0000150@GO0000249
5	-0,36	-0,58	**0,14**	GO0000146@GO0004133@GO0003777
6	0,01	-0,76	-0,17	GO0003824@GO0004135
7	-0,45	-0,84	0,56	GO0003674@GO0000150
8	0,36	-1,03	-0,84	GO0000146@GO0008251
9	0,03	**-0,6**	-0,42	GO0004133@GO0008574
10	-0,01	-1,22	-0,51	GO0003774@GO0004386

Fonte: Autoria própria (2021)

4.3 Considerações do capítulo

Neste Capítulo foi apresentado o método de imputação de dados faltantes para bases de dados com classificação hierárquica multirrótulo utilizando regressão. Foi relatado que o método inicialmente verifica qual atributo possui maior coeficiente de correlação com o atributo que possui dado faltante, para então utilizar a técnica de regressão, sendo que o método pode utilizar a regressão linear, polinomial ou múltipla.

Foram apresentadas as três etapas do método que são: verificação hierárquica multirrótulo, correlação e aplicação do modelo. Buscando melhor entendimento foram detalhadas as etapas de verificação hierárquica multirrótulo e aplicação do modelo através de pseudocódigos. Foi informado que a correlação pode ser calculada através da Equação (7) definida no Capítulo 3.

Foi apresentada uma base de dados fictícia com dados faltantes e sua hierarquia de classes do tipo DAG, onde buscando melhor entendimento do método, foi realizada simulação passo a passo. Durante a simulação foi detalhada cada uma das três etapas, apresentando o valor dos subconjuntos gerados, sendo que ao final a simulação retornou uma base de dados sem dados faltantes.

5 EXPERIMENTOS E RESULTADOS

Neste Capítulo são apresentados os experimentos realizados e os resultados obtidos com o método IR-HMC. A Seção 5.1 descreve as ferramentas utilizadas na implementação do método. A Seção 5.2 descreve as bases de dados utilizadas nos experimentos. A Seção 5.3 ilustra a metodologia usada para na aplicação dos experimentos. Na Seção 5.4 são apresentados os resultados obtidos nos experimentos realizados. Por Fim a seção 5.5 relata as considerações finais do Capítulo.

5.1 Ferramentas utilizadas

Para implementação do método proposto neste trabalho foram utilizadas as bibliotecas *Pandas e Sklearn,* abordadas por Sarkar (2018), que simplificam o manuseio de dados e aplicação de técnicas de regressão em linguagem *Python* versão 3.9.6.

Pandas oferece estruturas e operações para manipular tabelas numéricas e séries temporais. Através desta biblioteca é possível realizar operações de junção e ordenação de conjunto de dados de forma facilitada.

Sklearn é uma biblioteca de aprendizado de máquina que abrange uma grande quantidade de algoritmos e entre eles provê a criação de modelos de regressão.

O código fonte da implementação do método e também as bases de dados utilizadas nos experimentos estão disponíveis em repositório público no endereço http://dainf.pg.utfpr.edu.br/lesic/site/produto/130.

5.2 Bases de dados

Na realização dos experimentos foram utilizadas bases de dados do projeto (GO) Gene Ontology (2021), que possuem estrutura hierárquica entre classes que são organizadas em DAG. Além disso estas bases de dados também foram utilizadas em estudos de classificadores hierárquicos multirrótulo como os propostos por Vens, *et. al*, (2008), Borges (2012) e Cerri (2010).

Para a realização dos experimentos são utilizadas sete bases de dados Na Tabela 12 é possível observar as características das bases de dados utilizadas nos experimentos. As bases de dados com ID 1, 3, 4, 5, 6 e 7 possuem atributos numéricos faltantes. A base de dados com ID 2 possui atributos numéricos e um categórico faltante.

Tabela 12 - Características das bases utilizadas nos experimentos preliminares

ID	Nome	Amostras	Atributos	Classes	Atributos com dados faltantes	Amostras com dados faltantes
1	Cellcycle	3751	77	4125	16136	3507
2	Church	3749	27	4125	9751	2312
3	Eisen	2418	79	3573	3686	1823
4	Expr	3773	551	4131	185383	3773
5	Gasch1	3758	173	4125	18067	3245
6	Gasch2	3773	52	4131	2854	994
7	Seq	3900	478	4133	32	26

Fonte: Adaptado de Borges (2012)

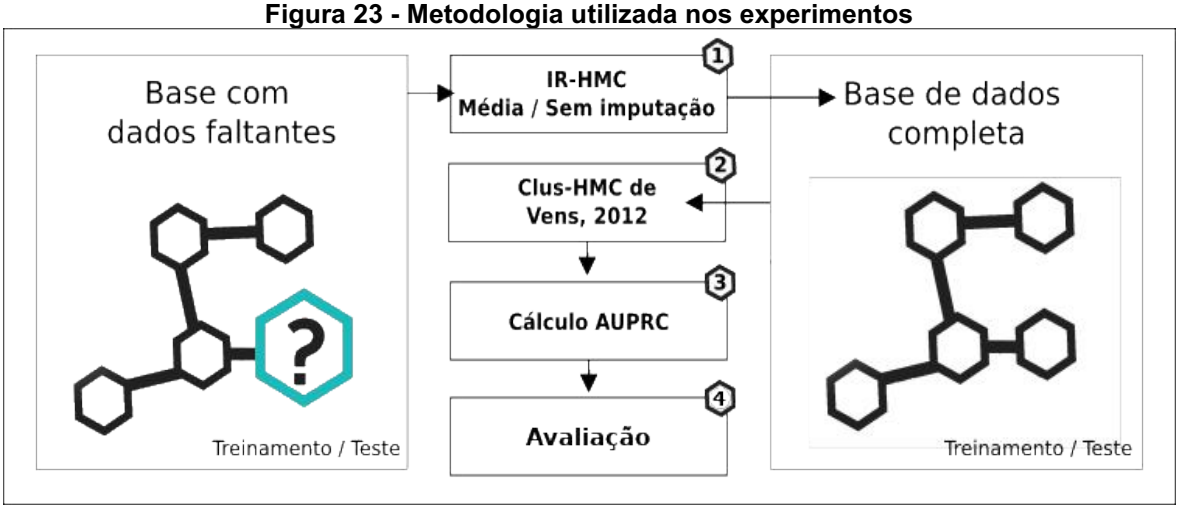
As bases de ID 1, 2, 3, 4, 6 e 7 são dados biológicos de expressão gênica obtidos pela técnica de microarranjo (BORGES, 2012).

Esses conjuntos de dados de Borges (2012), que também são utilizadas nestes experimentos, foram divididos em conjuntos de treinamento e teste, sendo 2/3 das amostras para o conjunto de treinamento e 1/3 para o teste.

5.3 Metodologia utilizada nos experimentos

Os experimentos foram realizados em quatro etapas distintas e são ilustradas na Figura 23. Inicialmente é utilizado a abordagem para tratar dados faltantes seguido do uso do classificador Clus-HMC proposto por Vens *et al* (2008) e verificação da acurácia da classificação por meio da medida baseada na curva de precisão e revocação (AUPRC). Por fim, é realizado avaliação dos resultados utilizando os testes estatísticos estatístico de Friedman (FRIEDMAN,1940) e Wilcoxon (WILCOXON, 1945).

Figura 23 - Metodologia utilizada nos experimentos



Fonte: Autoria própria (2021)

Na etapa 1, onde é utilizado o método IR-HMC foi adotado o limitar de correlação igual a 0,3. Este valor foi adotado, pois de acordo com Cohen (1988) e Mukaka (2012), correlações abaixo de 0,3 são pequenas ou desprezíveis. Por isso nos experimentos, o método proposto utilizará a regressão somente quando o coeficiente de correlação for superior a 0.3.

Para avaliar o desempenho preditivo do método de imputação será utilizado o classificador hierárquico multirrótulo, proposto por *Vens et al.* (2008), chamado Clus-HMC.

Além disso, os resultados do método IR-HMC serão comparados ao uso da média proposto no trabalho de Borges (2012) e o uso do classificador sem imputar os dados faltantes.

O classificador proposto por Vens *et al.* (2008) foi adotado para utilização neste trabalho, pois este foi amplamente citado nos trabalhos relacionados ao eixo 2 de pesquisa da revisão sistemática da literatura. Nos 9 trabalhos selecionados no eixo 2 de pesquisa, em 8 o trabalho proposto por Vens *et al.* (2008) consta como bibliografia.

Os trabalhos apresentados na literatura de classificação hierárquica multirrótulo demonstram o desempenho da classificação, por meio de algumas medidas de avaliação definidas para essa área. Para o experimento se optou pela medida baseada na curva de precisão e revocação (AUPRC), que estão disponíveis no algoritmo Clus-HMC de Vens (2008) *et al*. e MHC-CNN de Borges (2012).

Vens *et al*. (2008) propõe uma medida baseada na análise de curvas de precisão e revocação (curvas PR). Nessa medida é escolhido um conjunto de limiares entre 0 e 1, sendo que para cada limiar corresponde a um ponto no espaço da curva PR e variando esses limiares obtêm-se a curva PR. Esses limiares podem ser interpretados com a probabilidade de associação de cada classe a uma instância. Uma curva PR reflete a precisão de um classificador como uma função de sua revocação. Quanto mais perto de 1 for o valor da AUPRC, melhor é o classificador (CERRI, 2010).

Cerri (2010), afirma que de acordo com os autores da área, a utilização das medidas de precisão e revocação torna a avaliação mais adequada para problemas de classificação hierárquica multirrótulo, porque, nesse tipo de problema, geralmente classes individuais possuem poucos exemplos positivos. Um exemplo disso pode ser extraído do problema de classificação funcional de genes. Nesse problema, apenas poucos genes possuem uma função ativa particular, o que implica que, para a maioria das classes, o número de exemplos negativos supera o número de exemplos positivos. Nesse caso, é mais interessante saber o número de exemplos positivos para um conjunto de classes do que os negativos.

Vens *et al.* (2008) argumenta que optou pelo uso do AUPRC pois em conjuntos de dados HMC, geralmente ocorre que as classes individuais têm poucas instâncias positivas. Por exemplo, na genética funcional, normalmente apenas alguns genes têm uma função particular. Isso implica que, para a maioria das classes, o número de instâncias negativas excede em muito o número de instâncias positivas. Por isso ele afirma estar mais interessado em reconhecer as instâncias positivas (que uma instância tem um determinado rótulo), em vez de prever corretamente as negativas (que uma instância não tem um rótulo específico). Além disso, como optou-se pelo uso do classificador Clus-HMC de Vens *et al.* (2008), entendeu-se ser coerente a utilização da métrica proposta pelo autor.

5.3 Resultados

Conforme ilustrado na Figura 23, a metodologia utilizada nos experimentos tem como métrica de avaliação o valor do AUPRC de cada algoritmo em todas as sete bases, sendo que após isso estes valores são avaliados estatisticamente. A

Tabela 13 apresenta os resultados do AUPRC do método proposto neste trabalho. Os dados estão destacados em negrito para demonstrar os valores de AUPRC mais elevados em cada uma das bases de dados. Conforme descrito na metodologia, é definido que na fase 1 é utilizada uma abordagem para tratar os dados faltantes, sendo que o método propõe o uso da regressão linear, regressão polinomial (grau 2 a 5) e regressão múltipla, relacionados nas colunas da Tabela 13, respectivamente.

Tabela 13 – Resultados dos experimentos

Base de dados	Regressão linear	Regressão polinomial (grau 2)	Regressão polinomial (grau 3)	Regressão polinomial (grau 4)	Regressão polinomial (grau 5)	Regressão múltipla
Cellcycle	0,4394	0,4397	0,4399	**0,4422**	0,4382	0,4395
Church	0,4366	**0,4463**	0,4432	0,4462	0,4411	0,4364
Eisen	0,4536	0,4536	0,4536	0,4536	0,4536	**0,4592**
Expr	0,4539	0,4618	**0,4652**	0,4619	0,4611	0,4507
Gasch1	0,4550	0,4558	0,4484	0,4509	0,4567	**0,4600**
Gasch2	0,4356	0,4385	0,4389	0,4360	0,4360	**0,4397**
Seq	0,4678	**0,4683**	**0,4683**	**0,4683**	**0,4683**	0,4682

Fonte: Autoria própria (2021)

Observa-se na Tabela 13, que a regressão polinomial grau 2 apresentou valor de AUPRC mais elevado nas bases de dados Church, a regressão polinomial grau 3 apresentou AUPRC mais elevado na base de dados Expr, já a regressão polinomial grau 4 apresentou AUPRC mais alto na base de dados Cellcycle e a regressão múltipla apresentou AUPRC mais elevado em 3 conjuntos: Eisen, Gash1 e Gash2. Por fim, no conjunto Seq a regressão polinomial de grau 2 a 5 apresentou o mesmo valor de AUPRC. A regressão linear simples não apresentou valor mais elevado de AUPRC em nenhuma das bases de dados.

O método IR-HMC realiza imputação dos dados faltantes utilizando regressão somente caso coeficiente de correlação seja superior ao limiar definido, sendo que nestes experimentos este valor é 0,3. Por isso nem sempre a regressão foi utilizada para realizar a imputação dos dados. Buscando melhor entendimento dos resultados do método foi extraída a informação de qual foi o percentual de vezes que cada método de regressão foi utilizado. Estes números estão descritos na Tabela 14 onde é possível verificar por exemplo que a base de dados Gash1 foi a que mais vezes a regressão foi utilizada, já a base de dados Church foi onde a regressão foi utilizada em um número menor de vezes.

Tabela 14 – Percentual dos dados faltantes em que regressão foi utilizada

Base de dados	Regressão linear	Regressão polinomial (grau 2)	Regressão polinomial (grau 3)	Regressão polinomial (grau 4)	Regressão polinomial (grau 5)	Regressão múltipla
Cellcycle	39,06%	12,01%	11,99%	12,01%	11,99%	27,79%
Church	6,11%	0,35%	0,25%	0,25%	0,25%	3,80%
Eisen	74,55%	43,90%	43,98%	43,92%	43,84%	43,71%
Expr	59,17%	47,81%	47,74%	47,74%	47,16%	59,17%
Gasch1	79,75%	64,38%	64,00%	63,84%	62,93%	79,75%
Gasch2	64,33%	46,11%	46,06%	46,06%	45,99%	41,15%
Seq	25,00%	9,38%	9,38%	9,38%	9,38%	25,00%

Fonte: Autoria própria (2021)

5.3.1 Comparação dos resultados obtidos

A Tabela 15 ilustra a comparação das abordagens de imputação de dados com utilização da média e a não imputação com o método IR-HMC. A primeira coluna se refere a base de dados, a segunda coluna é referente ao valor do AUPRC obtido com a utilização da média proposta por Borges (2012), a terceira coluna é referente aos dados do AUPRC para não imputação.

Os dados da quarta a nova coluna são advindos dos resultados dos experimentos realizados com o método IR-HMC, onde foram utilizadas as técnicas de regressão linear, regressão polinomial (grau 2 a 5) e regressão múltipla, respectivamente.

Os valores mais elevados de cada uma das bases de dados estão destacados em negrito. É possível observar que a média apresentou AUPRC superior em duas bases de dados (Expr e Seq) e a regressão em outros três conjuntos (Eisen, Cellcycle e Gash1). Além disso, observa-se que o uso do classificador Clus-HMC sem imputação apresentou AUPRC superior nas bases de dados restantes (Church e Gash2).

Tabela 15 – Comparação dos resultados dos experimentos

Base de dados	Média*	Sem input	Reg. linear	Reg. pol. (grau 2)	Reg. pol. (grau 3)	Reg. pol. (grau 4)	Reg. pol. (grau 5)	Reg. múltipla
Cellcycle	0,4375	0,4377	0,4394	0,4397	0,4399	**0,4422**	0,4382	0,4395
Church	0,4420	**0,4505**	0,4366	0,4463	0,4432	0,4462	0,4411	0,4364
Eisen	0,4541	0,4570	0,4536	0,4536	0,4536	0,4536	0,4536	**0,4592**
Expr	**0,4713**	0,4483	0,4539	0,4618	0,4652	0,4619	0,4611	0,4507
Gasch1	0,4546	0,4581	0,4550	0,4558	0,4484	0,4509	0,4567	**0,4600**
Gasch2	0,4367	**0,4411**	0,4356	0,4385	0,4389	0,4360	0,4360	0,4397
Seq	**0,4684**	0,4641	0,4678	0,4683	0,4683	0,4683	0,4683	0,4682

* Média definida por Borges (2012)

Fonte: Autoria própria (2021)

Conforme mencionado anteriormente, na média foi utilizada abordagem definida Borges (2012), onde o critério usado para imputar valores ausentes foi calcular a média aritmética de todos os ancestrais mais próximos da classe à qual pertence a amostra. Nas amostras multirrótulo é feita a média aritmética também sobre a quantidade de rótulos da amostra.

Os resultados obtidos foram avaliados utilizando o teste de hipótese de Friedman (FRIEDMAN,1940). Este teste é indicado para comparar o desempenho de vários algoritmos em diferentes bases de dados. O teste é sugerido por não ser paramétrico, haja vista a dificuldade de se conhecer a distribuição dos dados.

O teste de Friedman foi realizado sob os dados apresentados na Tabela 15, onde são comparados os resultados do método IR-HMC com a média e a não imputação considerando as sete bases de dados.

A Tabela 16 ilustra o cálculo dos postos médios necessários para realização do teste de Friedman. Cada valor corresponde ao posto obtido considerando os resultados apresentados na Tabela 15, onde o primeiro colocado tem o posto 8, o segundo o posto 7 e assim por diante. Ao final é possível verificar a soma dos postos, onde a regressão polinomial grau 2 apresentou soma dos postos com valor mais elevado, seguido da regressão polinomial grau 3, média e não imputação. Os postos mais altos de cada base de dados estão destacados em negrito.

Tabela 16 - Postos médios dos resultados obtidos no teste de Friedman

Base de dados	Média *	Sem input	Reg. linear	Reg. pol. (grau 2)	Reg. pol. (grau 3)	Reg. pol. (grau 4)	Reg. pol. (grau 5)	Reg. múltipla
Cellcycle	1	2	4	6	7	**8**	3	5
Church	4	**8**	2	7	5	6	3	1
Eisen	6	7	3	3	3	3	3	**8**
Expr	**8**	1	3	5	7	6	4	2
Gasch1	3	7	4	5	1	2	6	**8**
Gasch2	4	**8**	1	5	6	2,5	2,5	7
Seq	**8**	1	2	5,5	5,5	5,5	5,5	3
Soma dos postos	**34**	**34**	**19**	**36,5**	**34,5**	**33**	**27**	**34**

* Média definida por Borges (2012)

Fonte: Autoria própria (2021)

Realizado o cálculo dos postos médios e aplicando nestes dados a Equação (8) têm-se $X_F^2 = 1{,}16$. Usando a estatística F_F para correção de X_F^2, conforme a Equação (9), têm-se $F_F = 0{,}14$

Para os graus de liberdade $k-1\ e\ (k-1)*(n-1)$ em que $k=8\ e\ n=7$, encontra-se o seguinte valor tabelado na distribuição de F de Snedecor $F(6{,}42)=2{,}34$. Como $F_F\ < F(6{,}42)$, logo a hipótese nula não pode ser rejeitada, indicando que não há diferença estatística entre os resultados dos algoritmos comparados.

Considerando que o teste de Friedman mostrou que a hipótese nula não pode ser rejeitada, foram realizados outros dois testes utilizando a abordagem de Wilcoxon (1945).

Diferente do teste de Friedman, onde todas as abordagens são comparadas de uma só vez, no teste de Wilcoxon a comparação ocorre em dois grupos. Inicialmente foram realizados o teste comparando o melhor resultado do método IR-HMC com a não imputação. Os dados deste teste são descritos na Tabela 17, sendo além dos valores do AUPRC de cada uma das duas abordagens há a informação do sinal, posição positiva e posição negativa além da soma dos postos. Sinais positivos indicam que o IR-HMC obteve valor mais elevado de AUPRC, sinais negativos indicam que a não imputação obteve valor mais elevado de AUPRC. As colunas referentes as posições (negativa e positiva) indicam o posto de cada uma das abordagens, sendo que quando maior o posto maior a diferença. Ao final há a informação da soma dos postos positivos que é igual a 22 e a soma dos postos negativos igual é 6. Logo a menor soma é $T=6$.

Tabela 17 - Comparativo com a não imputação no teste de Wilcoxon

Base de dados	Sem imputação	Melhor resultado do IR-HMC	Sinal	Posição positiva	Posição negativa
Cellcycle	0,4377	0,4422	+	**6**	
Church	0,4505	0,4462	-		**5**
Eisen	0,4570	0,4592	+	**3**	
Expr	0,4483	0,4619	+	**7**	
Gasch1	0,4581	0,4600	+	**2**	
Gasch2	0,4411	0,4397	-		**1**
Seq	0,4641	0,4683	+	**4**	
Soma dos postos				**22**	**6**

Fonte: Autoria própria (2021)

Considerando que $N=7$, onde N é o número de bases de dados, é verificado o valor tabelado para valores críticos com confiança $\alpha=0{,}05$, chegando ao valor 2. Como $T>2$, a hipótese nula não pode ser rejeitada, indicando que não há diferença estatística significativa entre os dois resultados.

A Tabela 18 mostra os resultados do teste de Wilcoxon considerando a comparação do melhor resultado do IR-HMC com a média proposta por Borges (2012). Na tabela é possível verificar que a menor soma de postos é igual a 8, sendo $T = 8$. Como $T > 2$, onde 2 é o valor tabelado para valores críticos com confiança $\alpha = 0{,}05$, a hipótese nula não pode ser rejeitada, indicando que não há diferença estatística significativa entre os dois resultados.

Tabela 18 - Comparativo com a média no teste de Wilcoxon

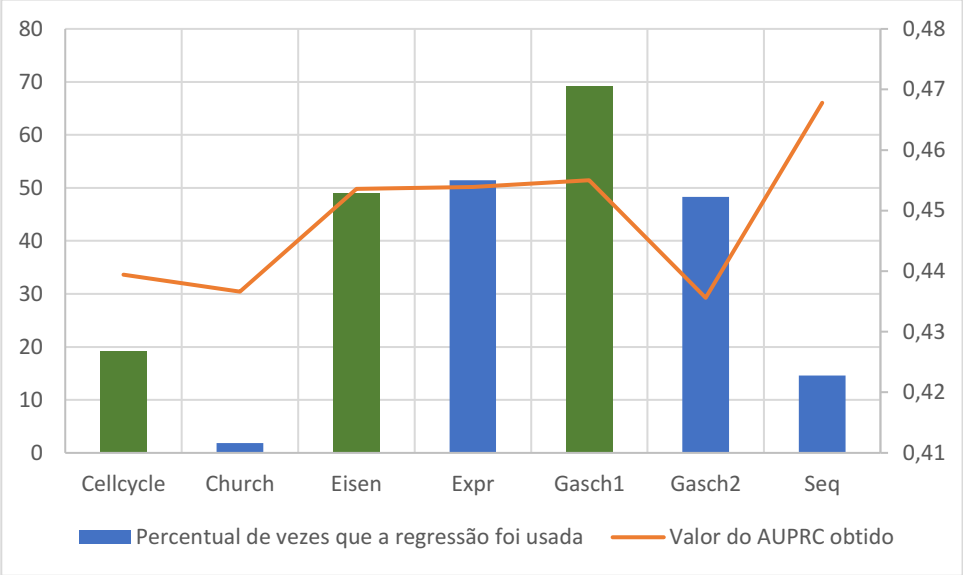
Base de dados	Média (Borges 2012)	Melhor resultado do IR-HMC	Sinal	Posição positiva	Posição negativa
Cellcycle	0,4375	0,4422	+	**4**	
Church	0,4420	0,4462	+	**3**	
Eisen	0,4541	0,4592	+	**5**	
Expr	0,4713	0,4619	-		**7**
Gasch1	0,4546	0,4600	+	**6**	
Gasch2	0,4367	0,4397	+	**2**	
Seq	0,4684	0,4683	-		**1**
	Soma dos postos			**20**	**8**

Fonte: Autoria própria (2021)

Buscando realizar uma análise dos valores do AUPRC definidos na Tabela 12, o Gráfico 8 ilustra os valores da métrica em cada base de dados, relacionando estes dados com o percentual de vezes em que a regressão foi utilizada, descrito na Tabela 14. As barras do gráfico indicam o valor do percentual de vezes em que a regressão foi utilizada em média, a linha laranja indica o valor do AUPRC obtido com a regressão. As barras verdes presentes nas bases de dados Cellcycle, Eisen e Gash1 indicam que a regressão obteve o valor de AUPRC mais elevado.

Esta análise buscou estabelecer uma relação entre o uso da regressão e o valor do AUPRC obtido, porém considerando somente estes dados não é possível chegar a uma conclusão. Apesar disso é interessante observar alguns dados como: a regressão foi mais utilizada na base de dados Gash1, sendo que a abordagem também apresentou o valor de AUPRC mais elevado neste conjunto (0,455); na base de dados Church, onde a regressão foi menos utilizada, a regressão apresentou AUPRC inferior as demais abordagens.

Gráfico 8 - Percentual de uso da regressão x AUPRC



Fonte: Autoria própria (2021)

5.4 Considerações finais

Este capítulo apresentou as ferramentas utilizadas no desenvolvimento do método proposto, abordando linguagem de programação e bibliotecas utilizadas. Foi apresentada a metodologia utilizada nos experimentos que é dividida em três etapas. Para avaliação da acurácia do método de imputação foi avaliado o desempenho do classificador Clus-HMC em relação medida AUPRC obtida para cada um dos algoritmos.

Foi mostrado os resultados do método proposto, indicando o AUPRC de cada algoritmo. Além disso foi realizada uma comparação do desempenho do classificador Clus-HMC considerando o pré-processamento das bases de dados utilizando o método proposto e as abordagens com média e não imputação. Como mencionado anteriormente, a métrica utilizada para verificação do classificador foi o AUPRC.

Os métodos de imputação utilizando regressão polinomial ou múltipla apresentaram AUPRC superior em três das sete bases de dados em que foram realizados os experimentos quando comparados a média ou a não imputação de dados faltantes.

O teste de Friedman mostrou que a soma dos postos médios das abordagens utilizando o IR-HMC apresentaram valores superiores as demais

abordagens, além disso o teste de Wilcoxon mostrou que o melhor resultado do método proposto apresentou posições positivas com valores superiores as das abordagens de não imputação e média.

Apesar dos resultados dos testes de Friedman e Wilcoxon mostrarem a existência de certa diferença entre os métodos, o cálculo destas diferenças mostrou que em ambos os testes a hipótese nula não pôde ser rejeitada, indicando que a diferença existente não é suficientemente significativa estatisticamente.

Também foi realizada uma tentativa de análise dos resultados, buscando estabelecer uma relação entre o uso da regressão e o valor do AUPRC obtido, porém somente com os dados atuais não foi possível chegar a uma conclusão.

6 CONCLUSÃO

Técnicas para lidar com dados faltantes foram expandidas nas últimas décadas, porém a revisão sistemática da literatura deste trabalho mostrou que há carência em estudos direcionados para cenários específicos como a da classificação hierárquica multirrótulo. Além disso, os trabalhos relacionados a classificadores como os propostos por Vens *et al*, (2008) Cerri (2010) e Borges (2012) têm foco nas etapas de classificação de dados não havendo detalhes das etapas de pré-processamento de dados incluindo atividades de imputação de dados faltantes.

O referencial teórico e a revisão sistemática da literatura mostraram alguns métodos de imputação e também serviram para ressaltar a importância da necessidade de realizar a imputação de dados faltantes de forma a melhorar a acurácia dos classificadores, já que dados faltantes estão presentes em diversos tipos de bases de dados.

Neste sentido, este trabalho se propõe em dar um passo adicional na otimização de técnicas de imputação de dados faltantes para bases de dados com classificação hierárquica multirrótulo, utilizando inclusive como parte do método algumas abordagens já utilizadas em estudos relacionados a classificadores como a média, isso considerando casos em que o uso da regressão pode não ser o mais adequado levando em conta o coeficiente de correlação encontrado.

Nos experimentos foram utilizadas 7 bases de dados do projeto *Gene Ontology* em que as classes estão dispostas em uma estrutura de DAG. Para avaliação do método de imputação, foi utilizado o classificador Clus-HMC.

Para avaliar a relevância dos dados estimados pelo método de imputação proposto, os valores da métrica AUPRC obtidos pelo classificador Clus-HMC após a imputação de dados foram comparados a outras duas abordagens: média e não imputação de dados faltantes.

A regressão apresentou AUPRC superior em três das sete bases de dados, sendo que em duas destas o classificador apresentou AUPRC superior quando os dados não foram imputados. Além disso, a média proposta por Borges (2012) apresentou AUPRC superior nas duas bases de dados restantes.

Foram realizados testes estatísticos, onde o teste de Friedman mostrou que a soma dos postos médios da regressão polinomial foi superior as demais abordagens e o teste de Wilcoxon mostrou que o melhor resultado do IR-HMC

apresentou posições positivas com soma superior as das abordagens de não imputação e média.

Apesar das diferenças que os testes estatísticos mostraram e de o IR-HMC apresentar AUPRC superior em algumas bases de dados, os testes de Friedman e Wilcoxon mostraram que a hipótese nula não pode ser rejeitada, indicando que não há diferença estatística suficientemente significativa entre os resultados dos algoritmos.

6.1 Trabalhos futuros

Muitas extensões deste trabalho podem ser tratadas, como a aplicação do método IR-HMC em bases de dados de outros domínios diferentes da bioinformática, afim de comparar o desempenho do método. Também é possível utilizar outros classificadores hierárquicos multirrótulo, assim como outras medidas de avaliação afim de validar os resultados obtidos.

Um recurso que pode ser implementado no método é a gravação dos modelos de imputações criados, possibilitando seu uso em cenários semelhantes em situações futuras.

Outro possível trabalho futuro é a otimização de parâmetros de entrada, como por exemplo alterar o valor mínimo do coeficiente de correlação testando valores superiores ao limiar definido nos experimentos (0,3). Estes novos parâmetros podem ser testados em cada um dos métodos (regressão polinomial ou múltipla).

Outro estudo acerca dos parâmetros é sobre o grau do polinômio quando a regressão polinomial é utilizada, buscando estabelecer uma relação entre os dados e o grau do polinômio a ser utilizado. Ainda nesta linha, é possível estudar a relação entre a quantidade de vezes em que a regressão é utilizada e a disposição dos dados, buscando entender em que tipo de base de dados o uso da regressão é mais vantajosa.

Por fim, um trabalho futuro que também pode ser realizado é a análise profunda dos resultados de novos experimentos que busquem métricas em que seja possível estabelecer uma relação entre a quantidade de vezes em que a regressão é utilizada e o valor do AUPRC obtido, tentando entender em qual tipo de bases de dados a regressão é mais eficiente como método de imputação.

REFERÊNCIAS

BARRETO, C. A. S. **Uso de Técnicas de Aprendizado de Máquina para Identificação de Perfis de Uso de Automóveis Baseado em Dados Automotivos**. 2018. Dissertação (Mestrado) - Programa de Pós-graduação em Engenharia de Software, Universidade Federal do Rio Grande do Norte, Natal, 2018.

BARROS, R. C.; CERRI, R; FREITAS, A. A.; CARVALHO, A. C. P. Probabilistic Clustering for Hierarchical Multi-Label Classification of Protein Functions. *In*: EUROPEAN CONFERENCE ON MACHINE LEARNING AND PRINCIPLES AND PRACTICE OF KNOWLEDGE DISCOVERY IN DATABASES. 2013, Berlim. **Anais** [...]. Berlim: Springer, 2013. p. 385-400.

BENAHMED, L. E HOUICHI, L. The effect of simple imputations based on four variants of PCA methods on the quantiles of annual rainfall data. **Environmental Monitoring and Assessment.** Maine, v. 190, p. 569, set. 2018.

BLOCKEEL, H.; SCHIETGAT, L; STRUYF, J; DŽEROSKI, S. Decision trees for hierarchical multilabel classification: A case study in functional genomics. *In:* KNOWLEDGE DISCOVERY IN DATABASES**,** 2006, Berlim. **Anais** [...]. Berlim: Springer, 2006. p. 18-29.

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

BORGES, H. B.; NIEVOLA, J. C. Multi-Label Hierarchical Classification using a Competitive Neural Network for protein function prediction. *In:* The 2012 International Joint Conference on Neural Networks (IJCNN), Brisbane, 2012. **Anais** [...]. Brisbane: IEEE, 2006.

BUUREN, S. **Flexible Imputation of Missing Data**. 2. ed. Leiden: CRC Press, 2018.

CARVALHO, A. E FREITAS, A. A tutorial on multi-label classification techniques. **Foundations of Computation Inteligence**, v. 5, p. 177- 195, jul. 2009.

CLARE, A. E KING, R.D. Knowledge Discovery in Multi-label Phenotype Data. *In*: PRINCIPLES OF DATA MINING AND KNOWLEDGE DISCOVERY, 2001, Freiburg. **Anais** [...]. Freiburg: Springer, 2001. p. 42-53.

CERRI, R. **Técnicas de classificação hierárquica multirrótulo**. 2010. 241 f. Dissertação (Mestrado em Ciência da Computação) - Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo, São Carlos, 2010.

CERRI, R; BARROS, R. C. Hierarchical classification of Gene Ontology-based protein functions with neural networks. *In:* INTERNATIONAL JOINT CONFERENCE ON NEURAL NETWORKS (IJCNN). 2015, Killarney. **Anais** [...]. Killarney: IEEE, 2015.

CERRI, R.; BARROS, R. C.; CARVALHO, A. C.; JIN, Y. Reduction strategies for hierarchical multi-label classification in protein function prediction. **BMC Bioinformatics**, Nova Iorque, v. 17, set. 2016.

CELSTON, M.; MALPERTUY, A.; LELANDAIS, G.; BREVERN, A. G. Comparative analysis of missing value imputation methods to improve clustering and interpretation of microarray experiments. **BMC Genomics**, v. 11, p. 15, jan. 2010.

COHEN, J. **Statistical Power Analysis for the Behavioral Sciences**. 2. ed. New York: Laurence Erlbaum Associates, 1988.

DESMAR, J. Statistical comparisons of classifiers over multiple data sets. **Journal of Machine Learning Research**, v.7, p. 1-30, 2006

ESULI, A., FAGNI, T. TreeBoot.mh: A boosting algorithm for multi-label hierarchical text categorization. *In:* PROCEDING OF THE 13TH INTERNATIONAL SIMPOSIUM ON STRING PROCESSING AND INFORMATION RETRIEVAL**,** Glasgow, 2006. **Anais** [...]. Glasgow: Springer, 2006. p. 13-24.

FABRIS, F; FREITAS, A. A. Dependency network methods for Hierarchical Multi-label Classification of gene functions. *In:* IEEE Symposium on Computational Intelligence and Data Mining (CIDM). Orlando, 2014. **Anais** [...]. Orlando: IEEE, 2014.

FARHANGFAR A.; KURGAN, L. A.; PEDRYCZ, W. Experimental analysis of methods for imputation of missing values in databases. *In:* PROCEEDINGS OF

SPIE - THE INTERNATIONAL SOCIETY FOR OPTICAL ENGINEERING (PROCEEDINGS OF SPIE), Orlando, 2004. **Anais** [...]. Orlando: SPIE, 2004.

FARHANGFAR A.; KURGAN, L. A.; PEDRYCZ, W. A novel framework for imputation of missing values in databases. **IEEE Transactions on Systems, Man, and Cybernetics - Part A: Systems and Humans**, v. 37, no. 5, p. 692-709, ago. 2007.

FACELI, K.; LORENA, A. C.; GAMA, J.; CARVALHO, A.; **Inteligência Artificial**: Uma abordagem de Aprendizagem de Máquina. Rio de Janeiro: LTC, 2011.

FREITAS, A. A.; CARVALHO, A. C. P. F. A Tutorial on Hierarchical Classification with Applications in Bioinformatics. **Research and Trends in Data Mining Technologies and Applications**. v. 1, p. 175-208, 2007.

FREITAS, A. A.; CARVALHO, A. C. P. F. A Tutorial on Hierarchical Classification with Applications in Bioinformatics. **Research and Trends in Data Mining Technologies and Applications**. v. 1, p. 175-208, 2007.

FREUND, Y. E MASON, L. The The alternating decision tree learning algorithm. *In*: PROCEDINGS OF THE SIXTEENTH INTERNATIONAL CONFERENCE, San Francisco, 1999. **Anais** [...]. San Francisco: ACM DL, 1999. p. 124-133.

FREUND, Y. E SCHAPIRE, R. E. A short introduction to boosting. **Japanese Society for Artificial Intelligence**, v. 14, no. 5. p. 771-785, set. 1999.

FRIEDMAN, M. A comparison of alternative tests of significance for the problem of m rankings. *In:* THE ANNALS OF MATHEMATICAL STATISTICS, Nova Iorque, 1940. **Anais** [...]. Nova Iorque: Institute of Mathematical Statistics, 1940. n. 1, p. 86-92.

GALVÃO, L. E MERSCHMANN, L. HSIM: A supervised imputation method for hierarchical classification scenario. *In:* Discovery Science. 19., 2016, Bari. **Anais** [...]. Bari: Springer, 2016. p. 134-148.

GENE ONTOLOGY. **Gene Ontology overview.** 10 nov. 2021. Disponível em: http://geneontology.org/docs/ontology-documentation/. Acesso em: 05 nov. 2021.

FENG, S; FU, P; ZHENG, W. A hierarchical multi-label classification method based on neural networks for gene function prediction. **Biotechnology & Biotechnological Equipment**, v. 32, p. 1613-1621, nov. 2018.

FENG, S; FU, P; ZHENG, W. A postprocessing method in the HMC framework for predicting gene function based on biological instrumental data. **Review of Scientific Instruments**, v. 89, mar. 2018.

GOOGLE SCHOLAR. Disponível em: http://scholar.google.com.br. Acesso em: 05 nov. 2021

J. W. GRAHAM. Missing data analysis: making it work in the real world. **The Annual Review Of Psychology**, Pennsylvania, p. 549-576, jul. 2009.

LITTLE R. J. A. E RUBIN, D. B. **Statistical Analysis with Missing Data**. 2 ed. New Jersey: Wiley & Sons, 2002.

MA, Q. *et al*. MIDIA: exploring denoising autoencoders for missing data imputation. **Data Mining and Knowledge Discovery,** v. 34, p. 1859–1897, jul. 2019

MELO, LEONARDO L. **Deep learning para identificação de déficit hídrico em plantas com base em imagens térmicas**. 2021. 86 f. Tese (Doutorado) - Universidade de São Paulo, São Paulo, 2021.

MONTIEL, J., READ, J., BIFET, A., E ABDESSALEM, T. Scalable Model-Based Cascaded Imputation of Missing Data, *In*: ADVANCES IN KNOWLEDGE DISCOVERY AND DATA MINING**,** Melbourne, 2018. **Anais** [...]. Melbourne: Springer, 2018. p. 64-76.

MUKAKA, M. M. Um guia para o uso adequado do coeficiente de correlação em pesquisas médicas. **Malawi Medical Journal**, Malawi, v. 24, n. 3, p. 69-71, set. 2012.

NETO, P. L. O. C. **Estatística**. 2 ed. São Paulo: Edgard Blücher, 2002.

PAGANI, R. N.; KOVALESKI, J. L.; RESENDE, L. M. Methodi Ordinatio: a proposed methodology to select and rank relevant scientific papers encompassing the impact

factor, number of citation, and year of publication. **Scientometrics**, v. 105, n. 3, p. 2109- 2135, dez. 2015.

PEARSON, K. e HERON, D. On theories of association. **Biometrika**, v. 9, no. 1/2, p. 159-315, mar. 1913.

QUINLAN, J. R. C4.5: programs for machine learning. **Machine Learning,** v 16, p. 235-240, set. 1993.

REZENDE, S, O. **Sistemas Inteligentes: Fundamentos e Aplicações.** Barueri, SP: Manole, 2005.

RUBIN D. B. Inference and Missing Data. **Biometrika**, v. 63, p. 581-592, dez. 1976.

RUBIN D. B. **Multiple imputation for nonresponse in surveys**. New York: Wiley, 1987.

SARKAR, D; BALI, R; SHARMA, T. **Practical Machine Learning with Python:** A Problem-Solver's Guide to Building Real-World Intelligent Systems. Berkeley: Apress, 2018.

SILVA, M. J. C. **Imputação múltipla: comparação e eficiência em experimentos multiambientais**. 2012. 122 f. Dissertação (Mestrado) – Escola Superior de Agricultura Luiz de Queiroz, Piracicaba, 2012.

SCHAPIRE, R. E. E SINGER, Y. Improved boosting algorithms using confidence-rated predictions. **Machine Learning**, v. 37, 1999, p. 297-336.

TSOUMAKAS, G. e KATAKIS, I. Multi-label classification: An overview. **International Journal of Data Warehousing and Mining**, IGI Publishing, v. 3, n. 3, p. 2109- 2135, 2007.

TAN P. *et al*. **Introdução ao Data Mining Mineração de Dados**. Rio de Janeiro: Editora Ciência Moderna, 2009.

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

VERONEZE, R. **Tratamento de Dados Faltantes Empregando Biclusterização com Imputação Múltipla.** 2011. 238 f. Dissertação (Mestrado) - Faculdade de Engenharia Elétrica e de Computação, Universidade Estadual de Campinas. Campinas, 2011.

WEHRMANN, J; CERRI, R; BARROS, R. Hierarchical Multi-Label Classification Networks. *In*: INTERNATIONAL CONFERENCE ON MACHINE LEARNING. 35., 2018, Stockholm. **Anais** [...]. Stockholm, 2018.

WILCOXON, F. Individual comparisons by ranking methods. **Biometrics**, v. 1, p. 80-83, 1945.