Imputação de dados faltantes em base de dados de classificação hierárquica multirrótulo usando regressão linear e polinomial

Alvaro Mateus Santana[a], Helyane Bronoski Borges[a]

[a]*Universidade Tecnológica Federal do Paraná, Universidade de Tecnologia, Rua Doutor Washington Subtil Chueire, 330, Jardim Carvalho, Ponta Grossa, PR, Brasil*
alvarosantana@alunos.utfpr.edu.br

Resumo

Algoritmos de aprendizagem de máquina normalmente necessitam de um conjunto sem dados faltantes, fato que nem sempre ocorre. Considerando isso, se torna necessário usar métodos para imputar dados em conjuntos de dados, tornando, em muitas vezes, o processo de aprendizagem de dados mais eficiente. Este trabalho propõe a imputação de dados faltantes no cenário de classificação hierárquica multirrótulo, em base de dados do projeto Gene Ontology, utilizando regressão linear e regressão polinomial. Por meio dos experimentos foi possível observar que houve melhora na métrica AUPRC em uma das bases de dados testadas.

Abstract

Machine learning algorithms usually require a set with no missing data, which is not always the case. Considering this, it becomes necessary to use methods to impute data into datasets, often making the data learning process more efficient. This work proposes the imputation of missing data in the multi-label hierarchical classification scenario, in a Gene Ontology project database, using linear regression and polynomial regression. Through the experiments it was possible to observe that there was an improvement in the AUPRC metric in one of the tested databases.

Palavras-chaves: dados faltantes, classificação hierárquica multirrótulo, regressão

1. Introdução

Segundo Cerri [1], um dos problemas enfrentados pelas tarefas de classificação é quando os dados apresentam atributos faltantes. 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.

Para Farhangfar et al. [2], como os dados podem ser gerados de formas variadas, muitas vezes são

ISSN:2526-1371 DOI:10.5281/zenodo.6090322

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.

Neste trabalho é proposto um método de imputação de dados faltantes em bases de dados com classificação hierárquica multirrótulo, utilizando regressão linear e polinomial.

2. Referencial teórico

Dados faltantes

Conjuntos de dados podem apresentar dificuldades relacionadas à qualidade, exemplos mais frequentes dessas dificuldades são dados ruidosos, inconsistentes, redundantes ou incompletos, a Figura 1 ilustra um conjunto de dados onde alguns atributos apresentam dados faltantes, sendo esses representados pelo caractere "?". No exemplo, caso fosse preciso aplicar determinado algoritmo de aprendizagem, como por exemplo Rede Neural Artificial, será necessário realizar alguma abordagem para tratar estes 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

Figura 1: Conjunto com dados faltantes. Tan et al. [3]

Classificação hierárquica multirrótulo

Para Cerri [1], existe um grande número de problemas em que uma ou mais classes podem ser divididas em subclasses ou agrupadas em superclasses, sendo um problema de classificação hierárquica. Segundo Borges [4], a hierarquia das classes pode ser representada por uma árvore ou DAG (*Direct Acyclic Graph*), sendo que a segunda abordagem se difere da primeira devido a ser possível uma classe possuir mais de um pai.

Carvalho et al. [5] diz que, normalmente na tarefa de classificação, cada instância é atribuída a somente um rótulo ou classe. Porém, em determinadas situações é necessário que dois ou mais rótulos sejam atribuídos para um exemplo, este tipo de classificação é denominado multirrótulo.

Por fim há determinados tipos de problemas que necessitam de uma representação utilizando ambas

ISSN:2526-1371 DOI:10.5281/zenodo.6090322

as abordagens, sendo então um problema de Classificação Hierárquica Multirrótulo.

Regressão

Segundo Tan et al. [3], 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. A regressão tenta incorporar o conhecimento de outras variáveis, buscando produzir predições mais inteligentes.

Este trabalho propõe solucionar o problema de imputação utilizando um modelo de regressão linear ou polinomial.

Além dos conceitos de regressão é importante destacar o conceito de correlação que será de grande importância dentro do estudo proposto. O coeficiente de correlação mostra a relação entre as variáveis dependente e independente, sendo atribuído valor entre 0 e 1, sendo que quanto mais próximo de 1, maior será a correlação.

Base de dados

Para realização dos experimentos foram utilizados duas bases de dados do projeto *Gene Ontology* [6], elas constam nos trabalhos de Vens et al [7] e Borges [4]. O projeto possui três categorias de funções denominadas de: processo biológico, função molecular e componente celular que são implementadas em três ontologias independentes. Nessas ontologias, genes e proteínas podem ser classificadas de forma hierárquica, podendo apresentar mais de uma função ou característica. Nesta área a principal motivação do uso da classificação hierárquica multirrótulo é a possibilidade de realizar diagnóstico de doenças e desenvolvimento de medicamentos.

A Tabela 1 lista as duas bases de dados utilizadas, apresentando também características de cada.

Tabela 1 - 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
3	Eisen	2418	79	3573	3686	1823

O estado da arte

Foi realizada revisão da literatura acerca de métodos de imputação, porém não foi possível verificar abordagens desenvolvidas exclusivamente para o cenário de classificação hierárquica multirrótulo. Como alternativa, em um segundo eixo de revisão foi inspecionado como foi realizada a etapa de

ISSN:2526-1371 DOI:10.5281/zenodo.6090322

preparação dos dados nos trabalhos relacionados a classificadores hierárquicos multirrótulo, sendo que neste eixo foram encontradas menções ao uso da média como método de imputação.

3. Método proposto e resultados dos experimentos realizados

A figura 2 ilustra as três etapas principais do método de imputação proposto. Inicialmente é criado um subconjunto de dados com outros exemplos que compartilham classes com o exemplo com dado faltante. Nesta fase, caso não seja encontrado outro exemplo que compartilhe a classe com dado faltante, também é verificado classes ascendentes dentro da hierarquia.

Na segunda etapa é realizada a busca do atributo com melhor coeficiente de correlação com o atributo com dado faltante. Caso o coeficiente de correlação seja pequeno, a regressão de dados será substituída pela média. Será utilizada a definição de Cohen [8] para baixas correlações, onde serão aceitos somente valores acima de 0,3.

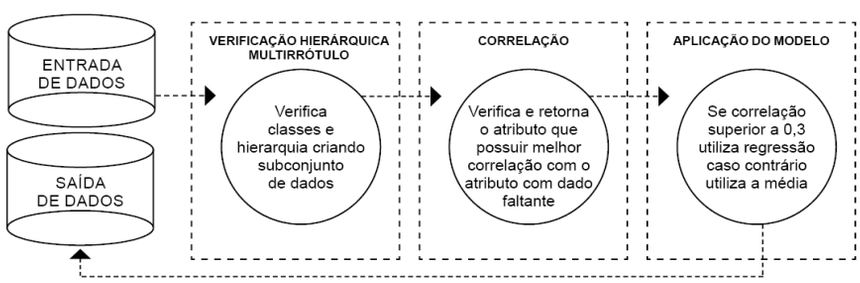

Figura 2: Fases do método de imputação

4. Experimentos e resultados

Os experimentos foram realizados em três etapas distintas e são ilustrados na Figura 3, inicialmente é utilizado a abordagem para tratar dados faltantes seguido do uso do classificador Clus-HMC proposto por Vens et al [7], na base de dados após a imputação de valores faltantes e por fim verificar a acurácia da classificação por meio da medida baseada na curva de precisão e revocação (AUPRC).

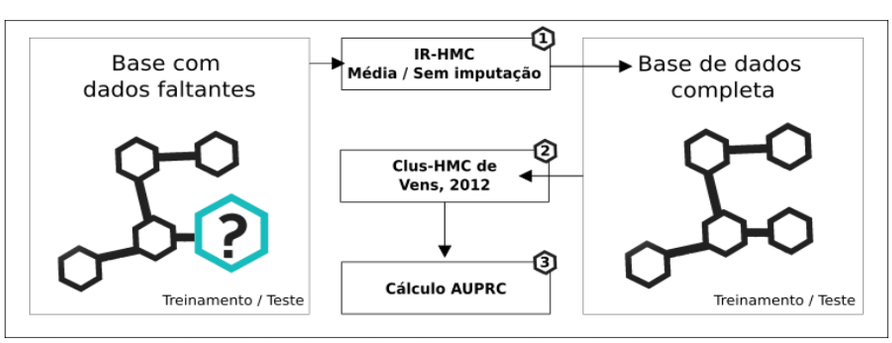

Figura 3: Metodologia utilizada na realização dos experimentos

ISSN:2526-1371 DOI:10.5281/zenodo.6090322

Na execução do classificador Clus-HMC, as bases de dados foram normalizadas utilizando a técnica MinMax e divididas em 2/3 para o treinamento e 1/3 para o teste.

Os resultados foram comparados entre as duas abordagens do método (regressão linear e polinomial), com a média ponderada proposta por Borges [4] e também com a execução do algoritmo Clus-HMC sem imputação de dados faltantes. A Figura 4 ilustra os resultados obtidos.

	AUPRC - Média ponderada (Borges 2012)	AUPRC Sem imputação	AUPRC Regressão Linear	AUPRC Regressão Polinomial
Cellcycle	0,4375	0,4377	0,4394	0,4397
Eisen	0,4541	0,4570	0,4536	0,4537

Figura 4: Resultados dos experimentos

Observa-se que na base de dados Cellcycle a regressão linear e polinomial apresentou melhores resultados da métrica AUPRC quando comparados a média ponderada e quando não houve imputação, já na base de dados Eisen não foi possível observar esta melhora. Um possível motivo é o fato da base de dados Eisen possuir uma quantidade menor de amostras, o que pode ter afetado a acurácia do cálculo da correlação. Outro fato importante a ser analisado nos resultados é que a regressão foi utilizada 2.748 vezes nas amostras da base de dados Eisen que corresponde a 79% dos dados faltantes, enquanto na base Cellcycle o uso da regressão foi acionado 6.302 vezes que corresponde a 12% dos dados faltantes.

A avaliação dos resultados obtidos foi feita utilizando o teste de hipótese de Friedman [9], sendo que este teste é indicado para comparar o desempenho de vários algoritmos em diferentes bases de dados. O uso deste teste é indicado pois ele é não paramétrico, já que há dificuldade em se conhecer a distribuição dos dados. De acordo com os autores Iman e Davenport [10] o teste de Friedman é considerado muito conservador, e por isso, sugere o uso da estatística FF distribuída de acordo com a distribuição F de Snedecor.

Analisando os resultados apresentados na Figura 4, têm-se $F_x^2 = 1{,}2$. Usando a estatística F_F para correção de F_x^2 têm-se $F_F = 0{,}25$.

Para os graus de liberdade $k-1$ e $(k-1)*(n-1)$ em que $k=2$ e $n=4$, encontra-se o seguinte valor tabelado na distribuição de F de Snedecor $F(1{,}3) = 10{,}13$. Como $F_F < F(1{,}3)$, logo a hipótese nula não pode ser rejeitada.

ISSN:2526-1371 DOI:10.5281/zenodo.6090322

5. Conclusão

Com base nestes resultados, o método apresentou melhora na métrica AUPRC no conjunto de dados que possui quantidade maior de exemplos, além disso é importante observar o fato de a quantidade maior de exemplos também reduzir a quantidade de vezes que a regressão foi utilizada.

Porém, apesar de haver melhora na métrica AUPRC, o teste de Friedman mostrou que a hipótese nula não pode ser rejeitada, indicando que não há diferença estatística entre os resultados dos algoritmos.

Como trabalhos futuros podem ser realizados experimentos nas demais bases de dados do projeto Gene Ontology, além de serem testadas outras abordagens de regressão, como por exemplo regressão múltipla. Nestes experimentos a regressão polinomial foi testada com um polinômio quadrático, porém, podem ser realizados experimentos com polinômios cúbicos e superiores.

Referências

[1] R. Cerri. 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.

[2] A. Farhangfar, L. A. Kurgan, W. Pedrycz. Experimental analysis of methods for imputation of missing values in databases. Proceedings of SPIE, Orlando, v. 5421, p. 172-182, 2004.

[3] P. Tan, M. Stenbach, V. Kumar, Introdução ao Data Mining Mineração de Dados. Rio de Janeiro: Editora Ciência Moderna, 2009.

[4] H. B. Borges, Classificador hierárquico multirrótulo usando uma rede neural competitiva. 2012. 188 f. Tese (Doutorado) - Universidade Católica do Paraná, Curitiba.

[5] A. C. P. L. F. Carvalho, J. Gama, A. C. Lorena, K. Faceli. Inteligência Artificial: uma abordagem de aprendizado de máquina. Rio de Janeiro: LTC Editora, 2011.

[6] Gene Ontology, Gene Ontology Overview, disponível em http://geneontology.org/docs/ontology-documentation/. Acessado em 22/08/2021.

[7] C. Vens, J. Sruyf, L. Schietgat, S. Džeroski, Decision trees for hierarchical multi-label classification. Machine learning, Springer, v.73, n.2, p.185.

[8] J, Cohen, Statistical Power Analysis for the Behavioral Sciences. 2. ed. New York: Laurence Erlbaum Associates, 1988.

[9] J. H. Friedman, J. W. Tukey. A Projection Pursuit Algorithm for Exploratory Data Analysis. IEEE Transactions on Computers, IEEE, v. 100, n. 9, p. 881-890, 1974.

[10] R. L. Iman, J. M. Davenport. Approximations of the critical region of the Friedman statistic. Communications in Statistics. p. 571-595, 1980.

ISSN:2526-1371 DOI:10.5281/zenodo.6090322