Um método para seleção de atributos em bases de dados de classificação hierárquica multirrótulo

Raimundo Osvaldo Vieira[a,b], Helyane Bronoski Borges[a,c]

[a]*Programa de Pós-Graduação em Ciência da Computação, UTFPR Câmpus Ponta Grossa*

[b]*Departamento de Computação, IFMA Campus Monte Castelo*

Contato: [c]helyane@utfpr.edu.br

Introdução

Problemas de classificação hierárquica multirrótulo envolvem a manipulação de grandes volumes de dados e um elevado número de atributos e rótulos. Este contexto requisita a aplicação de técnicas para reduzir a dimensionalidade dos dados sem alterar seu signficado intrínseco, de modo a melhorar o desempenho do classificador. Este trabalho propõe um método para redução de dimensionalidade em bases de dados de classificação hierárquica multirrótulo, utilizando a técnica de seleção de atributos. Para isso, faz uso do Fisher Score e do cálculo da correlação entre atributos.

Classificação Hierárquica Multirrótulo

O problema de classificação é um dos mais importantes no campo da aprendizagem de máquina e consiste em atribuir um rótulo a um elemento do conjunto de dados alvo da tarefa de classificação. De modo particular, há contextos nos quais um objeto de um conjunto pode ser associado a mais de uma classe e, além disso, é possível, ainda, que as classes mantenham entre si relações taxonômicas. Neste caso, tem-se a classificação hierárquica multirrótulo, que pode ser aplicada para a predição de proteínas em dados de bioinformática, por exemplo.

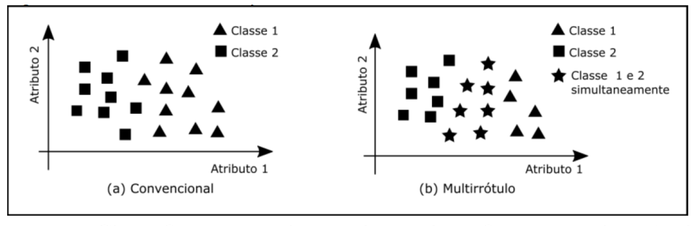

Figura 1: Classificação tradicional vs. classificação multirrótulo

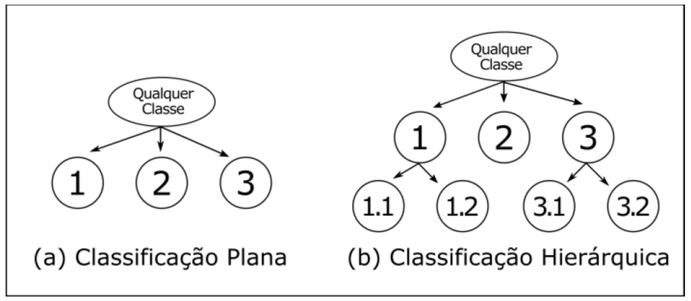

Figura 2: Classificação plana vs. classificação hierárquica

Seleção de Atributos

A seleção de atributos é uma das abordagens de redução de dimensionalidade em bases de dados, que compreende a escolha dos atributos mais relevantes a partir dos originais. As técnicas de seleção de atributos podem ser classificadas em três categorias: filtro, *wrapper* e embutida. A abordagem filtro corresponde à escolha de um subconjunto de atributos por meio da estimação da qualidade dos atributos com base apenas nos dados.

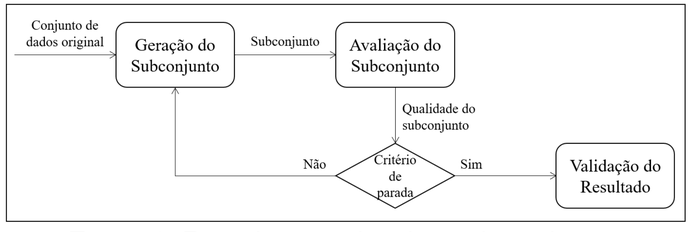

Figura 3: Procedimento de seleção de atributos

Método Proposto

O método consiste em ranquear os atributos a partir do cálculo do *Fisher Score*, adaptando-se esse cálculo para considerar a natureza hierárquica e multirrótulo da tarefa de classificação. A fórmula básica para o cálculo do *Fisher Score* do j-ésimo atributo é a seguinte:

$$F(x^j) = \frac{\sum_{k=1}^{c} n_k(\mu_j^k - \mu^j)^2}{(\sigma^j)^2} \quad (1)$$

onde, $(\sigma^j)^2 = \sum_{k=1}^{c} n_k(\sigma_k^j)^2$; μ_k^j e σ_k^j são a média e o desvio padrão do j-ésimo atributo da k-ésima classe; μ^j e σ^j são a média e o desvio padrão do j-ésimo atributo em relação a todo o conjunto de dados.

O método faz uso do coeficiente de correlação entre os atributos para lidar com redundância entre eles.

Por fim, conjunto de dados reduzido é avaliado por meio da aplicação de um classificador hierárquico multirrótulo e de sua medida de desempenho.

Para testar o método, serão realizados experimentos em bases de dados da Gene Ontology e os resultados serão avaliados com base na taxa de acerto dos classificadores e validados estatisticamente.

Referencias

[1] M. Dash y H. Liu, «Feature selection for classification,» en *Intelligent Data Analysis*, vol. 1, Elsevier, 1997.

[2] Q. Gu, Z. Li y J. Han, «Generalized Fisher Score for Feature Selection,» en *Proceedings of the 27th Conference on Uncertainty in Artificial Intelligence, UAI 2011*, 2012.

[3] R. F. Siqueira, *Redução de dimensionalidade em bases de dados de classificação hierárquica multirrótulo usando autoencoders*, Ponta Grossa, PR, 2019.

[4] A. Freitas y A. Carvalho, «A Tutorial on Hierarchical Classification with Applications in Bioinformatics,» en *Advances in Data Warehousing and Mining*, IGI Publishing, 2007.