

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 Campus Ponta Grossa, Rua Doutor Washington Subtil Chueire, 330, Ponta Grossa, PR, Brasil*

^b*Depto. de Computação, IFMA Campus Monte Castelo, Av. Getúlio Vargas, 04, São Luís, MA, Brasil*

^c*Helyane Bronoski Borges: helyane@utfpr.edu.br*

Palavras-chaves: classificação hierárquica multirrótulo, redução de dimensionalidade, seleção de atributos

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. Em geral, este tipo de aplicação possui um grande volume de dados e um número elevado de atributos e rótulos, o que interfere de forma negativa no desempenho do classificador. Para lidar com este problema é necessária a aplicação de técnicas para reduzir a dimensionalidade dos dados sem alterar seu significado intrínseco. 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. Esta pesquisa tem como objetivo propor um método para seleção de atributos em bases de dados de classificação hierárquica multirrótulo utilizando a abordagem filtro. 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. Além disso, o método faz uso do coeficiente de correlação entre os atributos para lidar com redundância entre eles. Em seguida, o 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.