Método de Extração de Atributos para Classificação Hierárquica Multirrótulo usando Autoencoders

Rafael Fernandes Siqueira[1], Helyane Bronoski Borges[1]

[1]Departamento Acadêmico de Informática
Universidade Tecnológica Federal do Paraná (UTFPR)
Av Monteiro Lobato, s/n - Km 04, CEP 84016-210 - Ponta Grossa – PR – Brasil

rafaelsiqueira@alunos.utfpr.edu.br, helyane@utfpr.edu.br

Palavras-chave: Classificação Hierárquica Multirrótulo; Redução de Dimensionalidade; Extração de Atributos; Autoencoders.

Resumo. A predição de proteínas em dados de bioinformática é um exemplo de problema de Classificação Hierárquica Multirrótulo no qual cada instância pode estar associada a múltiplas classes, e estas por sua vez, estão organizadas em uma hierarquia. A alta dimensionalidade dos atributos e das classes influência no desempenho dos classificadores, tanto no custo computacional quanto na performance, pois prejudica a busca por padrões e extração de conhecimento útil. A extração de atributos é uma das técnicas de redução de dimensionalidade em base de dados, que busca eliminar atributos irrelevantes e/ou redundantes que tendem a confundir o algoritmo de aprendizagem. Nessa técnica, por meio de combinações e/ou transformações dos atributos originais, geram-se novos atributos em um espaço de menor dimensão, que são mais significativos e que melhor representam a base de dados original. Desse modo, propõe-se um método de extração de atributos para classificação hierárquica multirrótulo, FEAE-HMC (*Feature Extraction based on Autoencoders for the Hierarchical Multi-label Classification*), baseado em conceitos de Deep Learning e adaptações em uma rede *Autoencoder* tradicional. Após o treinamento da rede, com o ajuste dos pesos, é possível reduzir a dimensionalidade das amostras da base de dados. O método é dividido em duas etapas principais: a extração de atributos e a avaliação do conjunto de dados reduzido, por meio de um classificador hierárquico multirrótulo e sua medida de desempenho. Nos experimentos são utilizados dados biológicos de 10 bases de dados da Ontologia Gênica, sendo a hierarquia das classes estruturada como um Grafo Acíclico Dirigido (DAG). As bases extraídas pelo FEAE-HMC, quando submetidas em um Classificador Hierárquico Multirrótulo, geram modelos nos quais se obtêm o desempenho preditivo, em relação a medida AUPRC, equivalente e até mesmo superior ao desempenho da base original.