Filtragem horizontal de dados na predição de coautorias em redes sociais acadêmicas

Vitor Rodrigues da Cunha Estima, Luciano Antonio Digiampietri

Escola de Artes Ciências e Humanidades - Universidade de São Paulo

vrcestima@usp.br, digiampietri@usp.br

Objetivos

A predição de coautorias, com base nas informações sobre a produção científica do Brasil que podem ser obtidas da Plataforma Lattes, pode favorecer a comunicação entre pesquisadores e otimizar o processo de produção científica, identificando possíveis colaboradores.

Na predição de coautorias, tipicamente, a base de dados utilizada sofre de desbalanceamento de dados, ou seja, em uma rede social típica haverá muitos pares de pessoas que não se relacionarão e apenas uma pequena quantidade que irá se relacionar.

O objetivo deste trabalho é estudar e desenvolver técnicas de filtragem horizontal e classificação binária em inteligência artificial a fim de lidar com o desbalanceamento de dados e otimizar a predição.

Métodos e Procedimentos

O desenvolvimento deste trabalho foi iniciado por uma revisão bibliográfica de forma a identificar quais são as principais estratégias para tratar o desbalanceamento em grandes bases de dados. A base de dados utilizada foi obtida pelo Grupo de Análise de Redes Sociais e Cientometria (GARSC) da USP.

Os algoritmos de filtragem e classificação foram desenvolvidos com a linguagem Python, utilizando as bibliotecas Pandas e Scikit-learn.

Resultados

Como resultado, foi desenvolvida uma nova técnica de filtragem chamada GUTE (Genetic Undersampling TEchnique).

Baseada no paradigma de *undersampling*, a técnica utiliza de um algoritmo genético para realizar de forma inteligente a filtragem horizontal dos dados.

Sem a utilização de nenhuma técnica de pré-processamento, a classificação com *Random Forest* na base de currículos obteve um F1-Score de 0,62, após a filtragem com o GUTE, obtivemos um F1-Score de 0,75.

Conclusões

Foi obtido um ganho notável no problema de Predição de Coautorias utilizando o algoritmo de filtragem desenvolvida.

A técnica proposta neste trabalho tem potencial para ser aplicada em problemas reais de desbalanceamento de dados.

Referências Bibliográficas

DIGIAMPIETRI, L. A. et al. Um sistema de predição de relacionamentos em redes sociais. Anais do XI Simpósio Brasileiro de Sistemas de Informação (SBSI 2015). p.139-146, 2015.

MARUYAMA, W. T. Predição de coautorias em redes sociais acadêmicas. 2016. Dissertação - Escola de Artes, Ciências e Humanidades, Universidade de São Paulo, São Paulo, 2016.