Imbalanced Data Classification Using a Relevant Information-Based Sampling Approach

Keider Hoyos[1]
Jorge Fernández[1]
Beatriz Martinez[1]
Óscar Henao[1]
Álvaro Orozco[1]
Genaro Daza[2]

[1]**Automatic Researh Group, Universidad Tecnológica de Pereira, Pereira, Colombia. jkhoyos@utp.edu.co**
[2]**nstituto de Epilepsia y Parkinson del Eje Cafetero, Pereira, Colombia.**

Abstract

The imbalanced data refer to datasets where the number of samples in one class (majority class) is much higher than the other (minority class) causing biased classifiers in favor of the majority class. Currently, it is difficult to develop an effective model using machine learning algorithms without considering data pre-processing to balance the imbalanced data sets. In this paper, we propose a Relevant Information based under-sampling (RIS) approach to improve the classification performance for the minority class by selecting the most relevant samples from the majority class as training data. Our RIS approach is based on a self-organizing principle of relevant information, which allows extracting the underlying structure of the majority class preserving different levels of detail of the original data with a smaller number of samples. Additionally, the RIS captures the data structure beyond second order statistics by estimating information theoretic measures which quantify the statistical structure of the majority class accurately, decreasing the consequences of the imbalanced classes distribution problem. We test our methodology on synthetic and real-world imbalanced datasets. Finally, we use a cross-validation scheme to quantify the classifier performance by evaluating the geometric mean. Results show that our proposal outperforms the state of the art methods for imbalanced class distributions regarding classification geometric mean, especially in highly imbalanced datasets.

Disponible en https://www.springer.com/gp/book/9783030011314