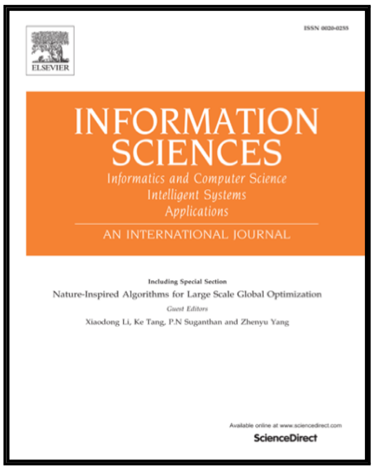

Improving Imbalanced Learning Through a Heuristic Oversampling Method Based on K-Means and SMOTE

Georgios Douzas, Fernando Bacao, Felix Last

PII: S0020-0255(18)30499-7
DOI: 10.1016/j.ins.2018.06.056
Reference: INS 13749

To appear in: *Information Sciences*

Received date: 15 January 2018
Revised date: 11 May 2018
Accepted date: 21 June 2018

Please cite this article as: Georgios Douzas, Fernando Bacao, Felix Last, Improving Imbalanced Learning Through a Heuristic Oversampling Method Based on K-Means and SMOTE, *Information Sciences* (2018), doi: 10.1016/j.ins.2018.06.056

Improving Imbalanced Learning Through a Heuristic Oversampling Method Based on K-Means and SMOTE

Georgios Douzas[a], Fernando Bacao[a,*], Felix Last[a]

[a] *Nova Information Management School, Campus de Campolide, 1070-312 Lisboa, Portugal*
Telephone: +351 21 382 8610

Abstract

Learning from class-imbalanced data continues to be a common and challenging problem in supervised learning as standard classification algorithms are designed to handle balanced class distributions. While different strategies exist to tackle this problem, methods which generate artificial data to achieve a balanced class distribution are more versatile than modifications to the classification algorithm. Such techniques, called oversamplers, modify the training data, allowing any classifier to be used with class-imbalanced datasets. Many algorithms have been proposed for this task, but most are complex and tend to generate unnecessary noise. This work presents a simple and effective oversampling method based on k-means clustering and SMOTE (synthetic minority oversampling technique), which avoids the generation of noise and effectively overcomes imbalances between and within classes. Empirical results of extensive experiments with 90 datasets show that training data oversampled with the proposed method improves classification results. Moreover, k-means SMOTE consistently outperforms other popular oversampling methods. An implementation[1] is made available in the Python programming language.

Keywords: Class-Imbalanced Learning, Oversampling, Classification, Clustering, Supervised Learning, Within-Class Imbalance

1. Introduction

The class imbalance problem in machine learning describes classification tasks in which classes of data are not equally represented. In many real-world applications, the nature of the problem implies a sometimes heavy skew in the class distribution of a binary or multi-class classification problem. Such applications include fraud detection in banking, rare medical diagnoses, and oil spill recognition in satellite images, all of which naturally exhibit a minority class [4, 18, 28, 29].

The predictive capability of classification algorithms is impaired by class imbalance. Many such algorithms aim at maximizing classification accuracy, a measure which is biased towards the majority class. A classifier can achieve high classification accuracy even when it does not predict a single minority class instance correctly. For example, a trivial classifier which scores all credit card transactions as legit will score a classification accuracy of 99.9% assuming that 0.1% of transactions are fraudulent; however in this case, all fraud cases remain undetected. In conclusion, by optimizing classification accuracy, most algorithms assume a balanced class distribution [29, 40].

*Corresponding author
Email addresses: gdouzas@icloud.com (Georgios Douzas), bacao@novaims.unl.pt (Fernando Bacao), mail@felixlast.de (Felix Last)

[1] The implementation of k-means SMOTE can be found at https://github.com/felix-last/kmeans_smote.