Pattern Recognition 57 (2016) 164–178

Contents lists available at ScienceDirect

Pattern Recognition

journal homepage: www.elsevier.com/locate/pr

# Analyzing the oversampling of different classes and types of examples in multi-class imbalanced datasets

José A. Sáez [a], Bartosz Krawczyk [b,*], Michał Woźniak [b]

[a] ENGINE Centre, Wrocław University of Technology, Wybrzeże Wyspiańskiego 27, 50-370 Wrocław, Poland
[b] Department of Systems and Computer Networks, Wrocław University of Technology, Wybrzeże Wyspiańskiego 27, 50-370 Wrocław, Poland

ARTICLE INFO

*Article history:*
Received 29 July 2015
Received in revised form
5 March 2016
Accepted 8 March 2016
Available online 16 March 2016

*Keywords:*
Machine learning
Imbalanced classification
Multi-class imbalance
Oversampling
Minority class types

ABSTRACT

Canonical machine learning algorithms assume that the number of objects in the considered classes are roughly similar. However, in many real-life situations the distribution of examples is skewed since the examples of some of the classes appear much more frequently. This poses a difficulty to learning algorithms, as they will be biased towards the majority classes. In recent years many solutions have been proposed to tackle imbalanced classification, yet they mainly concentrate on binary scenarios. Multi-class imbalanced problems are far more difficult as the relationships between the classes are no longer straightforward. Additionally, one should analyze not only the imbalance ratio but also the characteristics of the objects within each class. In this paper we present a study on oversampling for multi-class imbalanced datasets that focuses on the analysis of the class characteristics. We detect subsets of specific examples in each class and fix the oversampling for each of them independently. Thus, we are able to use information about the class structure and boost the more difficult and important objects. We carry an extensive experimental analysis, which is backed-up with statistical analysis, in order to check when the preprocessing of some types of examples within a class may improve the indiscriminate preprocessing of all the examples in all the classes. The results obtained show that oversampling concrete types of examples may lead to a significant improvement over standard multi-class preprocessing that do not consider the importance of example types.



## 1. Introduction

Typically classifier learning methods are designed to work with reasonably balanced datasets, but many real-world applications have to face imbalanced data, i.e., when several classes are under-represented (minority classes) in comparison to others (majority classes) [1,2]. Most of the classifier learning algorithms try to find the best decision boundaries. However, for imbalance datasets, when some of the classes are not well represented, setting meaningful boundaries is very difficult [3]. Furthermore, most of the classifier learning methods do not take into consideration a skewed data distribution and use as the training objective the overall accuracy, which guides them toward the prediction of classes that are over represented in the dataset [4]. As an example, let us consider the fraud detection problem, in which the percentage of the fraudulent transactions comparing to the legitimate ones is very low, as 0.01%. The classifier which will always make a decision that a given transaction is legitimate gains the 99.99% of accuracy, which seems to be the perfect result. Although the incidence of fraudulent transaction is limited to 0.01% of all the transactions, it may result in huge financial losses. Therefore, it is strongly recommended to use dedicated performance measures as the precision or recall to evaluate such methods.

Class imbalance makes the learning task more complex [5], but the disproportion between class examples is not the sole source of potential difficulties. For instance, the number of minority class examples may be insufficient to train a classifier, leading to overfitting [6]. In some imbalanced problems, a high classification error may be also affected by the validation scheme used to estimate the classifier performance [7]. Another problem is related to the case if the minority class examples form small distributed groups [8], which causes difficulties due to the lack of uniform structure in the minority class. All these problems related to imbalanced data have been intensively researched in the last decade [9–11] –for recent reviews in the topic, the reader may consult, e.g., [12,13]. However, most of the works are focusing on binary problems, which only involve two classes [14–16]. The main research lines in this scenario include the development of (i) *inbuilt mechanisms* [14], which change the classification strategies

* Corresponding author. Tel./fax: +48 071 320 21 06.
*E-mail addresses:* jose.saezmunoz@pwr.edu.pl (J.A. Sáez), bartosz.krawczyk@pwr.edu.pl (B. Krawczyk), michal.wozniak@pwr.edu.pl (M. Woźniak).

http://dx.doi.org/10.1016/j.patcog.2016.03.012
0031-3203/

to impose a bias toward the minority class, (ii) *data preprocessing methods* [15,9], which modify the data distribution to change the balance between classes, and (iii) *cost sensitive methods* [17] that assume a higher misclassification costs for examples of the minority class.

By contrast, this paper focuses on imbalanced classification for multi-class problems, because it is a more complex scenario and usually the reported solutions for binary cases are not applicable to it [18,19]. Here we deal with a more complicated situation, as the relations among the classes are no longer obvious. A class may be a majority one when it is compared to some other classes, but a minority or well-balanced for the rest of them. Therefore, developing efficient solutions to deal with this scenario is a needed research direction. There are some proposals within the data preprocessing field [20–22]. For example, Static-SMOTE [20] tries to increase the importance of minority classes within the dataset by resampling their instances. Another resampling approach based on clustering was introduced in [21]. Other techniques are particularly designed to work with specific classifiers [23,24]. Thus, a dynamic sampling method for multilayer perceptrons was proposed in [23], whereas other cost-sensitive neural networks based on resampling and moving thresholds were studied in [24]. Finally, some works address the multi-class imbalanced classification problem by using ensembles over the data [18,25,26]. An important research contribution within this field is that of Wang and Yao [18]. They studied the challenges posed by the multi-class imbalance problems and investigated the generalization ability of some ensemble solutions in the multi-majority and multi-minority cases. Other authors also proposed the decomposition of the original problem as binary ones [25].

The main contributions of this work are:

- a thorough study on oversampling approaches for handling multi-class imbalanced datasets. Note that this paper will not propose a new concrete data preprocessing method nor a classification algorithm for multi-class imbalanced data, but focuses on analyzing the aspects these approaches may consider in order to improve the results obtained.
- the proposition to analyze the structure of the classes in multi-class imbalanced problems in order to detect underlying structures and subsets of examples that may inform us about the characteristic of the considered task.
- a methodology of using oversampling techniques for the multi-class imbalanced classification that relies on the extracted knowledge about class and imbalance distribution types.
- an experimental evaluation of the proposed concept presenting the detailed results that offers an in-depth insight into the importance of selecting proper examples for the oversampling procedure. A dedicated website[1] presents detailed results.
- a set of conclusions that will allow to design efficient preprocessing methods and classifiers for multi-class imbalanced datasets; all these conclusions will be presented in Section 6.

The remaining part of this paper is organized as follows. Next section presents an overview of the imbalanced classification domain from the two-class and multi-class perspectives. Section 3 discusses the characteristic of different examples groups in multi-class imbalanced scenarios, whereas Section 4 describes in detail the proposed methodology. Section 5 presents the experimental study and Section 6 provides a summary including the conclusions that can be extracted from the analysis of the results. Finally, Section 7 presents the concluding remarks.

[1] http://www.kssk.pwr.edu.pl/krawczyk/multi-over

## 2. Imbalanced classification

This section is devoted to the main topic addressed in this paper, that is, imbalanced classification. First, Section 2.1 presents the possible scenarios for imbalanced datasets depending on their number of classes. Then, Section 2.2 discusses the case of binary imbalanced problems, whereas Section 2.3 focuses on multi-class imbalanced data.

### 2.1. Imbalanced datasets and cardinality of classes

The performance and quality of machine learning algorithms is conventionally evaluated using predictive accuracy. However, this is not appropriate when the data under consideration is strongly imbalanced, since the decision boundary may be strongly biased towards the majority class, leading to a poor recognition of the minority class. This fact causes most of the canonical classifiers to fail in such scenarios and requires developing dedicated algorithms for handling such difficulties. In the imbalance learning we can consider two scenarios: binary and multi-class. Their examples are given in Fig. 1.

In the binary case we have a well-defined relationship between classes: one is considered as the majority class and other as the minority one. This allows us to easily identify the potential bias and on which class we should focus when designing a pattern classification system. In the multi-class case this is no longer obvious, as the pairwise relationships between classes do not reflect the entire problem. A given class can at the same time be a majority group to a given subset of classes, minority group to others, or even of similar distribution to some of them.

Let us now present in the following sections an overview of these two possible class imbalance scenarios: binary and multi-class imbalanced datasets.

### 2.2. Binary imbalanced problems

Binary problems are well-researched in the imbalanced domain. Here the relationship between classes is obvious, the predominant one being the majority class and other being the minority class. Basically, techniques that address the problems associated with the binary imbalanced datasets may be divided into three groups [27]:

1. *Inbuilt mechanisms*, which try to adapt the existing classifiers to the problem of imbalanced datasets and bias them towards favoring the minority class.
2. *Data preprocessing methods* work to equalize the number of the examples of the classes.
3. *Hybrid methods*, which try to integrate the preprocessing techniques with the inbuilt mechanisms, especially with classifier ensembles [28].

Let us shortly describe the mentioned above approaches.

*Inbuilt mechanisms*: Implementing such mechanisms requires in-depth knowledge about the nature of the classifiers and underlying reasons of their failure in minority class recognition. One possibility is to perform one-class classification. Most of the works are trying to produce the one-class classifier, which can learn the minority class model and treating majority objects as outliers. Japkowicz et al. [29] propose to use autoencoder, while Kubat et al. [30] develop rule-based algorithm SHRINK. The ensemble of one-class classifiers applied to the imbalanced classification task is also described in [31]. Another approach, called cost-sensitive methods, is based on a loss function which informs about the misclassification cost. It has its root in probabilistic approach, where a learning algorithm tries to produce such a