Knowledge-Based Systems 89 (2015) 385–397

Contents lists available at ScienceDirect

Knowledge-Based Systems

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

MLSMOTE: Approaching imbalanced multilabel learning through synthetic instance generation

Francisco Charte [a,*], Antonio J. Rivera [b], María J. del Jesus [b], Francisco Herrera [a,c]

[a] *Department of Computer Science and A.I., University of Granada, 18071 Granada, Spain*
[b] *Department of Computer Science, University of Jaén, 23071 Jaén, Spain*
[c] *Faculty of Computing and Information Technology – North Jeddah, King Abdulaziz University, 21589 Jeddah, Saudi Arabia*

ARTICLE INFO

Article history:
Received 14 April 2015
Received in revised form 16 June 2015
Accepted 19 July 2015
Available online 23 July 2015

Keywords:
Multilabel classification
Imbalanced learning
Oversampling
Synthetic instance generation

ABSTRACT

Learning from imbalanced data is a problem which arises in many real-world scenarios, so does the need to build classifiers able to predict more than one class label simultaneously (multilabel classification). Dealing with imbalance by means of resampling methods is an approach that has been deeply studied lately, primarily in the context of traditional (non-multilabel) classification.

In this paper the process of synthetic instance generation for multilabel datasets (MLDs) is studied and MLSMOTE (Multilabel Synthetic Minority Over-sampling Technique), a new algorithm aimed to produce synthetic instances for imbalanced MLDs, is proposed. An extensive review on how imbalance in the multilabel context has been tackled in the past is provided, along with a thorough experimental study aimed to verify the benefits of the proposed algorithm. Several multilabel classification algorithms and other multilabel oversampling methods are considered, as well as ensemble-based algorithms for imbalanced multilabel classification. The empirical analysis shows that MLSMOTE is able to improve the classification results produced by existent proposals.

1. Introduction

Classification is one of the main supervised learning applications, an important field in Machine Learning [1]. The goal is to train a model using a set of labeled data samples, obtaining a classifier able to label new, never seen before, unlabeled samples. The datasets used in traditional classification have only one class per instance. By contrast, in multilabel datasets (MLDs) [2] each instance has more than one class assigned, and the total number of different classes (labels) can be huge.

In many real world scenarios, such as text classification [3] and fraud detection [4], the number of instances associated to some classes is much smaller (greater) than the amount of instances assigned to others. This problem, known as imbalanced learning, has been widely studied over the last decade [5] in the context of classic classification. It is also present in multilabel classification (MLC), since labels are unevenly distributed in most MLDs. To deal with imbalance in MLC, methods based on algorithmic adaptations [6–8], the use of ensembles [9,10], and resampling techniques [11–13] have been proposed.

Among the existent resampling techniques, those based on the generation of new samples (oversampling) have shown [14] to work better than others. The new samples can be clones of existent ones, or be synthetically produced as in SMOTE (Synthetic Minority Over-sampling Technique) [15]. Multilabel oversampling algorithms based on the cloning approach have been proposed in [12,13], being demonstrated its capability to deliver an improvement in classification results. A synthetic approach to produce new samples in MLDs is still to be faced. SMOTE is the most popular algorithm for this task in non-multilabel datasets, so it would be a good starting point.

Imbalance in MLC has been faced mainly through algorithmic adaptations and the use of ensembles, while the resampling approach is the least examined path until now. Nevertheless it is an interesting way and deserves to be taken into account, as the results in [12] have shown. Since oversampling algorithms seem to produce better results, designing a more advanced method to produce new data samples could be worth the effort. This is the motivation behind MLSMOTE (Multilabel Synthetic Minority Over-sampling Technique), a novel multilabel oversampling algorithm designed to create synthetic instances associated to minority labels.

The popular SMOTE algorithm takes all the samples belonging to the minority class, picks a random instance among the nearest

* Corresponding author. Tel.: +34 953 212 892; fax: +34 953 212 472.
E-mail addresses: francisco@fcharte.com (F. Charte), arivera@ujaen.es (A.J. Rivera), mjjesus@ujaen.es (M.J. del Jesus), herrera@ugr.es (F. Herrera).

http://dx.doi.org/10.1016/j.knosys.2015.07.019
0950-7051/© 2015 Elsevier B.V. All rights reserved.

neighbors of each one, and produces a new sample with the same minority class. Both the number of nearest neighbors and the amount of synthetic instances used for each minority sample can be adjusted. In a multilabel context there will always be more than one minority label, thus a strategy for choosing the appropriate instances has to be established. Moreover, the synthetic instances need a set of labels (labelsets) instead of being associated to an individual class. Therefore, a method to generate these synthetic labelsets has also to be settled.

The aim of this paper is to present MLSMOTE. As mentioned above, its goal is to produce synthetic instances associated to minority labels. In order to know which labels are minority, MLSMOTE leans on the multilabel imbalance measures proposed in [16]. The features of the synthetic instances are obtained by interpolation of values belonging to nearest neighbors, as in SMOTE. The labelsets of these new instances are also gathered from nearest neighbors, taking advantage of label correlation information in the neighborhood. For this task three different methods were studied, the intersection of the labels which appear in the neighbors, the union of those, and a third method based on a ranking of label occurrences. An extensive experimentation, structured in two different phases that will be detailed later, has been conducted. From the analysis of this experimentation it can be concluded that MLSMOTE, our multilabel synthetic minority oversampling technique, accomplishes a general improvement in classification results when compared with previous proposals with the same purpose.

The rest of this paper is organized as follows. In Section 2 the MLC and imbalanced learning problems are introduced. Section 3 provides a comprehensive review on the published approaches to work with multilabel imbalanced datasets. Section 4 provides all the details about the MLSMOTE algorithm, its parameters and implementation. In Section 5 the experimental framework used is defined, and the results obtained from experimentation are analyzed. Section 6 provides a final discussion and conclusions.

2. Preliminaries

The algorithm proposed in this paper has ties with two different topics, multilabel classification and imbalanced learning. In this section a brief introduction to both is provided, along with some specific details regarding imbalanced learning in the multilabel context.

2.1. Multilabel classification

In traditional classification the datasets are composed of a set of input features and a unique value in the output attribute, the class or label. In MLC [2] each sample may contain more than one value (class) in the output feature. Thus, the output of the classifier is not an individual label but a set of them. As stated in [2], a multilabel classifier will usually generate its prediction using two different methods. One is giving as output a bipartition, composed as a set of *true/false* values for each label. Another is returning a label ranking. In any case, most MLC solutions are built around one of two different approaches:

- Data transformation methods aim to convert the original dataset in order to use traditional classification algorithms to process it. A complete taxonomy of transformation methods for MLDs can be found in [17]. The two most popular ones are called Binary Relevance (BR) [18] and Label Powerset (LP) [19]. The former generates multiple binary datasets, one for each label, while the latter produces only one multiclass dataset, using as class the set of active labels in each sample.
- The goal of the method adaptation approach is to modify known classification algorithms to make them able to work with MLDs. There are many proposals in this field, from MLC trees like ML-TREE [20] or a multilabel kNN called ML-kNN [21] to multilabel neural networks [22,23] and SVMs [24]. There are also several methods based on ensembles of classifiers, such as RAkEL [25], CLR [26], HOMER [27], CC [28], and ECC [29], as well as other approaches to the problem, such as the use of Error-Correcting Codes [30].

In addition to new algorithms, MLC also demanded specific measures to evaluate classification results, as well as measures aimed to assess some MLDs peculiarities. In [2] the definition of most of them can be found. A recent review on the state-of-the-art multilabel learning algorithms, as well as evaluation measures, can be found in [31]. The measures used in this study will be defined later, in SubSection 2.3 (characterization measures) and Section 5 (evaluation measures).

2.2. Imbalance in traditional classification

In general, most classifiers underperform when used with imbalanced datasets. As stated in [32] the reason lies in their design, aimed to reduce the global error rate. This is a design which tends to benefit the most represented class in the dataset (majority class), labeling new instances with this class at the expense of the minority class. Moreover, imbalanced distribution of classes can complicate other common problems, such as noisy labels [33].

Three main approaches [34] have been proposed to face the imbalance problem. Data resampling follows the preprocessing approach, rebalancing the class distribution by deletion [35] or creation [15] of instances. Resampling techniques are classifier-independent solutions to the imbalanced learning problem, albeit some proposals for specific classifiers exist [36], and have shown their general effectiveness [14]. The other two approaches, algorithmic adaptations [37] and cost sensitive classification [38], are classifier dependent. The goal of the former is to modify existent classifiers taking into account the imbalanced nature of the datasets. The latter combines the design of adapted classifiers with some data preprocessing techniques. The present study is focused in the first approach. A general introduction and additional details about these approaches can be found in [39]. In some cases, resampling techniques are used along with ensembles of classifiers to tackle the imbalance problem. A general overview on ensemble methods is provided in [40]. The use of ensembles in imbalanced classification was recently reviewed in [41], and some specific algorithms are proposed in [42].

Most resampling algorithms consider one majority (minority) class only. Thus, undersampling techniques remove instances from the most frequent class only, whereas oversampling methods create instances from the least frequent one only. SMOTE works this way, generating new samples associated to the least frequent class. Firstly, the set of instances belonging to the minority class is obtained. For each instance in this set, SMOTE gathers a small batch of nearest neighbors. Typically the size of this group is 5. For each synthetic instance to produce, one of these neighbors is randomly picked. The features of the new sample are interpolated along the line which connects the reference and the neighbor instances. The class of the synthetic sample is always the minority class.

2.3. Imbalance in multilabel classification

The total number of distinct labels tends to be quite large in nearly all MLDs. The most usual cases are in the range from several dozens to a few hundreds of labels. There also are some extreme