Neurocomputing 163 (2015) 3–16

Contents lists available at ScienceDirect

Neurocomputing

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

Addressing imbalance in multilabel classification: Measures and random resampling algorithms

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 29 October 2013
Received in revised form
28 February 2014
Accepted 11 August 2014
Available online 16 April 2015

Keywords:
Multilabel classification
Imbalanced classification
Resampling algorithms
Undersampling
Oversampling

ABSTRACT

The purpose of this paper is to analyze the imbalanced learning task in the multilabel scenario, aiming to accomplish two different goals. The first one is to present specialized measures directed to assess the imbalance level in multilabel datasets (MLDs). Using these measures we will be able to conclude which MLDs are imbalanced, and therefore would need an appropriate treatment. The second objective is to propose several algorithms designed to reduce the imbalance in MLDs in a classifier-independent way, by means of resampling techniques. Two different approaches to divide the instances in minority and majority groups are studied. One of them considers each label combination as class identifier, whereas the other one performs an individual evaluation of each label imbalance level. A random undersampling and a random oversampling algorithm are proposed for each approach, giving as result four different algorithms. All of them are experimentally tested and their effectiveness is statistically evaluated. From the results obtained, a set of guidelines directed to show when these methods should be applied is also provided.

1. Introduction

Multilabel classification (MLC) [1] is receiving significant attention lately, and it is being applied in fields such as text categorization [2] and music labeling [3]. In these scenarios, each data sample is associated with several concepts (class labels) simultaneously. Therefore, MLC algorithms have to be able to give several outputs as result, instead of only one as in traditional classification.

The data used for learning a classifier is often imbalanced. Thus, the class labels assigned to each instance are not equally represented. This is a profoundly examined problem in binary datasets [4] and to a lesser extent to multiclass datasets [5]. A measure called *imbalance ratio* (IR) [4] is used to know the datasets' imbalance level. Traditionally, imbalanced classification has been faced through techniques [6] such as resampling, cost-sensitive learning, and algorithmic-specific adaptations.

That most MLDs suffer from a high level of imbalance is a commonly accepted fact in the literature [7]. However, there are not specific measures to assess the imbalance level in MLDs. Thus, the imbalanced nature of MLDs is more an assumption than an established fact. To date, there are some proposals to deal with imbalanced MLDs focused in algorithmic adaptations of MLC algorithms [7–9], so they are classifier-dependent solutions. An alternative classifier-independent way to address the imbalance in MLDs would be by means of preprocessing techniques, with resampling algorithms in particular. This approach would allow the use of any state-of-the-art MLC algorithm.

In this paper, we tackle the mentioned imbalanced problem for MLDs from a double perspective, the analysis of the imbalance level and proposals for reducing the imbalance in MLDs.[1]

There is a need for specific measures that can be used to obtain information about the imbalance level in MLDs. Three measures directed to assess the MLDs imbalance level are introduced and discussed.

Four resampling methods aimed at reducing the imbalance in MLDs are proposed. The measures will offer a convenient guide to know if an MLD suffers from imbalance or not, and therefore when it could benefit from the preprocessing. Regarding the resampling methods, undersampling and oversampling were the reasonable techniques to follow, although the difficulty on how to deal with

* 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).

[1] This paper is an expanded version of our previous work [33] from HAIS'13, including new preprocessing proposals and a vastly extended experimental study.

http://dx.doi.org/10.1016/j.neucom.2014.08.091
0925-2312/

multiple labels has to be solved. We examine two different approaches:

- One of them is based on the Label Powerset (LP) transformation, evaluating the frequency of full labelsets. Two algorithms founded on this approach were introduced in [33], one performs random undersampling (LP-RUS) and the other one random oversampling (LP-ROS).
- The second approach evaluates the frequency of individual labels, instead of full labelsets, isolating the instances in which one or more minority labels appear. Based on this approach another two algorithms are proposed, one for random undersampling (ML-RUS) and the other one for random oversampling (ML-ROS).

The usefulness of the measures and effectiveness of the methods are proven experimentally, using different MLDs and MLC algorithms, and the results are thoroughly analyzed using statistical tests. The conducted experimentation is used as an exploratory test on how known resampling algorithms could be adapted to the multilabel scenario.

The rest of this paper is structured as follows: Section 2 briefly describes the MLC task and the learning from imbalanced data problem. Section 3 introduces the imbalance problem in MLC, and describes the proposed measures to assess the imbalance level in MLDs. The resampling methods proposal is presented in Section 4. In Section 5, the experimental framework is described, and the results obtained are analyzed. Finally, the conclusions are given in Section 6.

2. Preliminaries

MLC usually demands more complex models than traditional classification to be faced. As traditional datasets, class distribution in MLDs frequently involves some imbalance level. The imbalance level in MLDs tends to be higher indeed. This characteristic makes this task even more challenging. In this section, MLC and classification with imbalanced data problems are introduced.

2.1. Multilabel classification

In many application domains [2,3,10] each data sample is associated with a set of labels, instead of only one class label as in traditional classification. Therefore, with Y being the total set of labels in an MLD D and x_i a sample in D, a multilabel classifier h must produce as output a set $Z_i \subseteq Y$ with the predicted labels for the i-th sample. As each distinct label in Y could appear in Z_i, the total number of potential different combinations would be $2^{|Y|}$. Each one of these combinations is called a *labelset*. The same labelset can appear in several instances of D.

There are two main approaches [1] to accomplish an MLC task: data transformation and algorithm adaptation. The former aims to produce from an MLD a dataset or group of datasets that can be processed with traditional classifiers, while the objective of the latter is to adapt existent classification algorithms in order to work with MLDs. Among the transformation methods, the most popular are those based on the binarization of the MLD, such as *Binary Relevance* (BR) [11] and *Ranking by Pairwise Comparison* [12], and the LP [13] transformation, which produces a multiclass dataset from an MLD considering each labelset as class. In the algorithm adaptation approach there are proposals of multilabel C4.5 trees [14], algorithms based on nearest neighbors such as ML-kNN [15], multilabel neural networks [2,16], and multilabel SVMs [17].

In the literature there are some specific measures to characterize MLDs, such as label cardinality *Card*, defined as shown in Eq. (1), and label density *Dens*, Eq. (2). The former is the average number of active labels per sample in an MLD, while the latter is designed to obtain a dimensionless measure:

$$Card(D) = \sum_{i=1}^{|D|} \frac{|Y_i|}{|D|}. \tag{1}$$

$$Dens(D) = \frac{Card(D)}{|Y|}. \tag{2}$$

A recent review on multilabel learning algorithms can be found in [18].

2.2. Classification with imbalanced data

The learning from imbalanced data problem is founded on the different distributions of class labels in the data [19], and it has been thoroughly studied in traditional classification. In this context, the measurement of the imbalance level in a dataset is obtained as the ratio of the number of samples of the majority class and the number associated with the minority class, being known as IR [4]. The higher the IR, the larger the imbalance level. The difficulty in the learning process with this kind of data is due to the design of most classifiers, as their main goal is to reduce some global error rate [4]. This approach tends to penalize the classification of minority classes.

Generally, the imbalance problem has been faced with three different approaches [6]: data resampling, algorithmic adaptations [5], and cost sensitive classification [20]. The former is based on the rebalancing of class distributions through resampling algorithms, either deleting instances of the most frequent class (undersampling) or adding new instances of the least frequent one (oversampling). Random undersampling (RUS) [21], random oversampling (ROS) and SMOTE [22] are among the most used resampling methods to equilibrate imbalanced datasets. The advantage of this approach is in that it can be applied as a general method to solve the imbalance problem, independent of the classification algorithms used once the datasets have been preprocessed. A general overview on imbalanced learning can be found in [23].

2.3. Learning from imbalanced MLDs

Conventional resampling methods are designed to work with one output class only. Each sample in an MLD is associated with more than one class, and this is a fact to be taken into account. Furthermore, those methods usually assume that there are only one minority label and one majority label, whereas in MLDs with hundreds of labels many of them can be considered as minority/majority cases. Thus, an approach to resample MLDs, which have a set of labels as output and several of them could be considered minority/majority labels, is needed.

Most of the published algorithms aim to deal with the imbalance problem by means of algorithmic adaptations of MLC classifiers, or the use of ensembles of classifiers. Furthermore all of them are classifier-dependent, instead of general application methods able to work with another MLC learning algorithms. Some of the existent proposals are the following:

- *Ensemble Multilabel Learning* [7] is a method based on the use of heterogeneous algorithms to build an ensemble of MLC classifiers. The authors aim to face two problems simultaneously, learning from imbalanced data and capturing correlation information among labels. The ensemble is composed of five well-known MLC algorithms, being able to improve classification results in some configurations.