Pattern Recognition 48 (2015) 1653–1672

Contents lists available at ScienceDirect

Pattern Recognition

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

Classifying imbalanced data sets using similarity based hierarchical decomposition

Cigdem Beyan*, Robert Fisher

School of Informatics, University of Edinburgh, G.12 Informatics Forum, 10 Crichton Street, Edinburgh EH8 9AB, UK

ARTICLE INFO

Article history:
Received 8 January 2014
Received in revised form
17 October 2014
Accepted 27 October 2014
Available online 26 November 2014

Keywords:
Class imbalance problem
Hierarchical decomposition
Clustering
Outlier detection
Minority–majority classes

ABSTRACT

Classification of data is difficult if the data is imbalanced and classes are overlapping. In recent years, more research has started to focus on classification of imbalanced data since real world data is often skewed. Traditional methods are more successful with classifying the class that has the most samples (majority class) compared to the other classes (minority classes). For the classification of imbalanced data sets, different methods are available, although each has some advantages and shortcomings. In this study, we propose a new hierarchical decomposition method for imbalanced data sets which is different from previously proposed solutions to the class imbalance problem. Additionally, it does not require any data pre-processing step as many other solutions need. The new method is based on clustering and outlier detection. The hierarchy is constructed using the similarity of labeled data subsets at each level of the hierarchy with different levels being built by different data and feature subsets. Clustering is used to partition the data while outlier detection is utilized to detect minority class samples. The comparison of the proposed method with state of art the methods using 20 public imbalanced data sets and 181 synthetic data sets showed that the proposed method's classification performance is better than the state of art methods. It is especially successful if the minority class is sparser than the majority class. It has accurate performance even when classes have sub-varieties and minority and majority classes are overlapping. Moreover, its performance is also good when the class imbalance ratio is low, i.e. classes are more imbalanced.

1. Introduction

In recent years, learning and classification with imbalanced data sets has become one of the key topics in pattern recognition due to its challenges especially for real-world applications where the data sets are dominated by normal examples in addition to a small amount of unusual examples [1–3]. Usually, the samples are grouped into binary classes. The well-represented class is called the majority class and the under-represented class is called the minority class. In such a case, a problem usually occurs because traditional classification algorithms tend to be biased towards to the majority class [4,5]. On the other hand, even though being imbalanced is not always a problem (such as where the classes are separable) imbalanced data sets usually contain overlapping regions where the prior probabilities of the two classes are almost equal [6]. Moreover, small disjuncts, and small sample size with high feature dimensionality [7] are frequently observed challenges in imbalanced data sets causing classification errors as well.

The appropriate evaluation criteria (such as the feature selection criterion to lead the training process and/or the criterion to evaluate the performance of classifiers) are also important issues when dealing with imbalanced data sets. For evaluation, many metrics exists in the literature. Accuracy is the most frequently used metric which is the sum of correctly predicted minority and majority samples over the total amount of samples. However, for imbalanced data sets, it is obvious that using accuracy might misguide the classifier and the importance of the minority class can be ignored since it is under-represented. This might be worse (total misclassification of the minority class) if the ratio between the classes is huge and the data is highly overlapping. Based on this, many alternative metrics has been proposed for evaluation of imbalanced classification. The geometric mean of sensitivity and specificity [8], adjusted geometric mean [9], Area Under Receiver Operating Characteristic curve (AUC) [10], and the *F*-measure which uses precision and recall (useful especially if we are highly interested in effective classification of a specific class) [5] are examples of effective metrics in this area.

Applications utilizing imbalanced data sets are diverse such as text categorization, medical diagnosis, fault detection, fraud detection, video surveillance, image annotations, anomaly detection [2,4,11]. Inherently, the diversity in applications has led to different

* Corresponding author. Tel.: +44 131 651 3441; fax: +44 131 650 6899.
E-mail addresses: C.Beyan@sms.ed.ac.uk (C. Beyan), rbf@inf.ed.ac.uk (R. Fisher).

http://dx.doi.org/10.1016/j.patcog.2014.10.032
0031-3203/

solutions over the years. Approaches are traditionally divided into four categories: (i) algorithmic level, (ii) data level, (iii) cost-sensitive methods and (iv) ensembles of classifiers.

i) **The algorithmic level approaches** force the classifier to converge to a decision threshold biased to an accurate classification of the minority class such as by adjusting the weights for each class. For instance, in [3] a weighted Euclidean distance function was used to classify the samples using k-nearest neighbors (k-NN). Similarly, a Support Vector Machine (SVM) with a kernel function biased to the minority class is proposed in [12] to improve the minority class prediction.
ii) **The cost-sensitive approaches** assign different costs to training examples of the majority and the minority classes [13,14]. However, it is difficult to set the cost properly (can be done in many ways) and may depend on the characteristics of the data sets. The standard public classification data sets do not contain the costs [2] and over-training is highly possible when searching to find the most appropriate cost.
iii) Re-sampling the data in order to handle the problems caused by the imbalanced nature of data is another approach. This **data level approach** does not modify the existing classifiers and is applied as a pre-processing technique prior to the training of a classifier. The data set can be re-sampled by oversampling the minority class [3], and/or under sampling the majority class [8,15,16]. Even though being independent of the classifier seems like an advantage, it is usually hard to determine the optimal re-sampling ratio automatically. Additionally, it might be problematic to oversample minority classes yet keep the distribution the same, especially in real-world applications where overlaps between minority and majority classes are highly likely. Similarly, while under-sampling the majority class, it is usually difficult to keep the new distribution of the majority class as similar as the distribution that it is sub-sampled from.
iv) **Ensembles of classifiers** have been popular in the last decade [17]. There are two main approaches; bagging and boosting. Bagging contains different classifiers which are applied to subsets of the data [18]. Alternatively, in boosting, the whole set is used to train classifiers in each iteration while more attention is given to the classification of the samples that are misclassified in the previous iteration. This is done by adjusting the weights toward their correct classification. The most well known boosting method is AdaBoost [19]. Even though ensembles are frequently used for classification of imbalanced data sets, they are not able to handle the imbalanced data sets by themselves. And they require one or a combination of the approaches that are mentioned above such as re-sampling data (SMOTEBoost [20], EUSBoost [2] etc.).

In this study, we propose a new approach which is not completely defined by any of these categories presented above. The proposed method is a hierarchical decomposition which is based on clustering and uses outlier detection as the classifier. Following the standard approach in the literature, we consider only two-class problems with imbalanced data sets. The hierarchy is built using the similarity of data subsets while using the selected best feature subset (in terms of a chosen feature selection criterion) at each level of the hierarchy. Clustering of data based on the selected feature subset (without initially using known class labels) is the way to partition the data into separable subsets. Using outlier detection as the classifier is due to the assumption that the samples of the minority class are expected to be outliers and should be differentiated by the chosen outlier definition. For instance, outliers can be the samples that are far away from cluster center given a cluster composed of samples of the majority and the minority classes.

The basic steps of the proposed hierarchical decomposition are clustering (Section 3.1), outlier detection (Section 3.2), and feature selection (Section 3.3). The hierarchy is automatically generated using the similarities of data samples and does not use any class and/or feature taxonomy as hierarchical classifiers do. Many hierarchical classifiers are motivated by a taxonomy such as [21,22]. In contrast to approaches which use the same feature space for all classifications we use different feature subsets at different levels of the hierarchy. This allows us to use more specific features once the data has become more focused onto specific subclasses (which might occur in the lower levels of the hierarchy).

The proposed method is evaluated using data sets from different fields and is compared with popular supervised learning methods in combination with algorithmic level and data level approaches. Additionally, synthetic data sets are used to test the performance of the proposed method in detail and different conditions (Section 4).

The contributions of this paper are as follows:

- We present a novel method that uses outlier detection in combination with clustering to classify imbalanced data sets.
- We present a new hierarchical decomposition method which does not use any fixed hierarchy based on features and/or classes. By being based on clustering, it is different from common hierarchical methods which use supervised learning.
- We show that different feature spaces can be used to build the hierarchy.
- Results show that the proposed method is successful especially when the distribution of the minority class is sparser than the majority class. It performs well when the class imbalanced ratio (the number of minority class samples over majority class samples) is low. It is successful if the majority and minority samples are highly overlapping and even when both classes contain varieties (such as having a mixture of distributions or having subclasses).

The paper is structured as follows: Section 2 briefly discusses the previous research on the class imbalance problem, hierarchical classifiers and hierarchical decomposition methods. Section 3 introduces the proposed method including the hierarchy construction (training step), new sample classification using the hierarchy (testing step) and basic methods (clustering, outlier detection, and feature selection). The test data sets, experimental set up and the corresponding results are given in Section 4. Finally, Section 5 discusses the proposed method with its advantages and shortcomings.

2. Related research

The research related to the proposed method can be divided into two subsections: research considering the class imbalance problem and research about hierarchical classifiers and decomposition.

2.1. Class imbalance problem

One of the most popular *re-sampling approaches* is SMOTE [3] which creates synthetic minority class instances by pre-processing before classification. In this approach, for each minority class sample a new sample is created on the line joining it to the nearest minority class neighbor. Previously, it has been combined with many supervised methods such as SVM [5], Naive Bayes [3], C4.5 [2,3], Random Forest [23,24]. Even though it is popular and works better than only under-sampling the majority class, it does