Information Sciences 409–410 (2017) 17–26

Contents lists available at ScienceDirect

Information Sciences

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

Clustering-based undersampling in class-imbalanced data

Wei-Chao Lin[a], Chih-Fong Tsai[b,*], Ya-Han Hu[c], Jing-Shang Jhang[b]

[a] Department of Computer Science and Information Engineering, Asia University, Taiwan
[b] Department of Information Management, National Central University, Taiwan
[c] Department of Information Management, National Chung Cheng University, Taiwan

ARTICLE INFO

Article history:
Received 15 July 2016
Revised 5 May 2017
Accepted 6 May 2017
Available online 8 May 2017

Keywords:
Class imbalance
Imbalanced data
Machine learning
Clustering
Classifier ensembles

ABSTRACT

Class imbalance is often a problem in various real-world data sets, where one class (i.e. the minority class) contains a small number of data points and the other (i.e. the majority class) contains a large number of data points. It is notably difficult to develop an effective model using current data mining and machine learning algorithms without considering data preprocessing to balance the imbalanced data sets. Random undersampling and oversampling have been used in numerous studies to ensure that the different classes contain the same number of data points. A classifier ensemble (i.e. a structure containing several classifiers) can be trained on several different balanced data sets for later classification purposes. In this paper, we introduce two undersampling strategies in which a clustering technique is used during the data preprocessing step. Specifically, the number of clusters in the majority class is set to be equal to the number of data points in the minority class. The first strategy uses the cluster centers to represent the majority class, whereas the second strategy uses the nearest neighbors of the cluster centers. A further study was conducted to examine the effect on performance of the addition or deletion of 5 to 10 cluster centers in the majority class. The experimental results obtained using 44 small-scale and 2 large-scale data sets revealed that the clustering-based undersampling approach with the second strategy outperformed five state-of-the-art approaches. Specifically, this approach combined with a single multilayer perceptron classifier and C4.5 decision tree classifier ensembles delivered optimal performance over both small- and large-scale data sets.

1. Introduction

In data mining and machine learning, it is difficult to train an effective learning model if the class distribution in a given training data set is imbalanced. This is known as the class imbalance problem. One class might be represented by a large number of examples, whereas the other might be represented by only a few. In addition, for most data mining algorithms, rare objects are much more difficult to identify than common objects [3,33,37]. This is a problem encountered in numerous real-world applications such as medical diagnosis, financial crisis prediction, and e-mail filtering [12]. Furthermore, the primary class of interest in the data mining task is usually the minority (or rare) class.

Without consideration of the class imbalance problem, learning algorithms or constructed models can be overwhelmed by the majority class and can ignore the minority class. For example, consider a two-class data set with an imbalance ratio of 99%, where the majority class constitutes 99% of the data set and the minority class contains only 1%. To minimize the

* Corresponding author.
E-mail address: cftsai@mgt.ncu.edu.tw (C.-F. Tsai).

http://dx.doi.org/10.1016/j.ins.2017.05.008
0020-0255/

error rate, the learning algorithm classifies all of the examples into the majority class, which yields an error rate of 1%. In this case, all of the examples belonging to the minority class are paramount and must be identified as incorrectly classified [24].

Bankruptcy prediction is a practical class imbalance problem [35,39]. In particular, the numbers of bankruptcy cases (i.e. the minority class) are usually much smaller than those of nonbankruptcy cases (i.e. the majority class). The type I error rate, which means that a prediction model incorrectly classifies the bankruptcy case into the nonbankruptcy class, is more critical than the average rate of classification accuracy. This is because higher type I error rates are likely to increase bad debts for financial institutions.

A variety of methods have been proposed to solve this problem. Such methods can be divided into four types: algorithmic-level methods, data-level methods, cost-sensitive methods, and ensembles of classifiers [4,12] (cf. Sections 2.2 and 2.3). In particular, the data-level methods, which focus on preprocessing the imbalanced data sets before constructing the classifiers, are widely considered in the literature. This is because the data preprocessing and classifier training tasks can be performed independently. In addition, according to Galar et al. [12], who conducted a comparative study of numerous well-known approaches, combinations of data preprocessing methods with classifier ensembles perform better than other methods.

Data preprocessing methods are based on resampling the imbalanced training data set before the model training stage. To create balance, the original imbalanced data set can be resampled by oversampling the minority class [8,15] and/or undersampling the majority class [17,23]. Some representative approaches combine oversampling and undersampling data preprocessing with classifier ensembles through boosting [31] or bagging [6] techniques; for example SMOTEBoost [9], RUSBoost [32], OverBagging [36], and UnderBagging [2].

Most of these approaches perform several rounds of random resampling for the majority (i.e. undersampling) or minority (i.e. oversampling) class. In the next group of methods, different balanced training sets are used to train a number of specific classifiers for later combination as classifier ensembles. Of these two resampling strategies, undersampling has been shown to be a better choice than oversampling [5,12]. This is because the oversampling strategy may increase the likelihood of overfitting in the model construction process. However, with the undersampling strategy, some useful data present in the majority class might be eliminated [34].

To overcome the limitations of undersampling, we propose replacing the random undersampling strategy with a clustering technique. The aim of clustering analysis is to group similar objects (i.e. data samples) into the same clusters; the objects in different clusters are different in terms of their feature representations [16]. Therefore, using clustering analysis to undersample the majority class generates a number of clusters, with each cluster containing similar data. Specifically, each cluster centroid (or center), which is based on the mean of similar data in the same group calculated by the k-means algorithm [14], can be used to represent the data in the whole group. In other words, the original data in the same groups are replaced by the cluster centers, thereby reducing the size of the majority class.

In this paper, we demonstrate that this type of clustering-based undersampling strategy can reduce the risk of removing useful data from the majority class, enabling the constructed classifiers (including both single classifiers and classifier ensembles) to outperform classifiers developed using a random undersampling strategy.

The contribution of this paper is twofold. First, we present two strategies of using the *k*-means clustering technique for undersampling in the class imbalance domain problem, which has never been done before. Second, several combinations of the clustering-based undersampling approach with different classification techniques, including five single classifiers and five classifier ensembles, are compared over a large number data sets to identify the optimal solution.

The remainder of this paper is organized as follows. Section 2 overviews the class imbalance problem and some widely used representative approaches that have been compared in the literature. Section 3 describes the research methodology including the clustering-based undersampling method and model construction. Section 4 presents the experimental results, and Section 5 concludes the paper.

2. Literature review

2.1. The class imbalance problem

Class imbalance (or imbalanced classification) is a problem in data sets with skewed distributions of data points. This has the following characteristics [7,12].

- Class overlapping [3]: When the data samples belonging to different classes overlap (as shown in Fig. 1), classifiers have difficulty effectively distinguishing between different classes. In most cases, instances belonging to the minority class are classified into the majority class.
- Small sample size: In practice, collecting sufficient data for class imbalanced data sets is challenging. One solution is to balance the imbalance ratios of the data sets to reduce the misclassification error.
- Small disjuncts: The data samples in the minority classes are distributed in numerous feature spaces, as shown in Fig. 2. This causes a high degree of complication during the classification stage.

Due to a significant difference between the sample sizes of two different classes (i.e. high imbalance ratios), classifiers may treat some of the data points in the minority class as outliers, which produces a very high misclassification error rate