Expert Systems With Applications 46 (2016) 405–416

Contents lists available at ScienceDirect

Expert Systems With Applications

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

Adaptive semi-unsupervised weighted oversampling (A-SUWO) for imbalanced datasets

Iman Nekooeimehr, Susana K. Lai-Yuen*

Industrial and Management Systems Engineering, University of South Florida, 4202 East Fowler Avenue, ENB 118, Tampa, Florida 33620, USA

ARTICLE INFO

Keywords:
Imbalanced dataset
Classification
Clustering
Oversampling

ABSTRACT

In many applications, the dataset for classification may be highly imbalanced where most of the instances in the training set may belong to one of the classes (majority class), while only a few instances are from the other class (minority class). Conventional classifiers will strongly favor the majority class and ignore the minority instances. In this paper, we present a new oversampling method called Adaptive Semi-Unsupervised Weighted Oversampling (A-SUWO) for imbalanced binary dataset classification. The proposed method clusters the minority instances using a semi-unsupervised hierarchical clustering approach and adaptively determines the size to oversample each sub-cluster using its classification complexity and cross validation. Then, the minority instances are oversampled depending on their Euclidean distance to the majority class. A-SUWO aims to identify hard-to-learn instances by considering minority instances from each sub-cluster that are closer to the borderline. It also avoids generating synthetic minority instances that overlap with the majority class by considering the majority class in the clustering and oversampling stages. Results demonstrate that the proposed method achieves significantly better results in most datasets compared with other sampling methods.

1. Introduction

Many datasets in various applications are imbalanced where some classes contain many more instances than others. Some examples where imbalanced datasets need to be classified include detection of fraudulent bank account transactions or telephone calls (Akbani, Kwek, & Japkowicz, 2004; Wei, Li, Cao, Ou, & Chen, 2013), biomedical diagnosis (He & Garcia, 2009; Li, Chan, Fu, & Krishnan, 2014), text classification (Zheng, Wu, & Srihari, 2004), information retrieval and filtering (Piras & Giacinto, 2012) and college student retention (Thammasiri, Delen, Meesad, & Kasap, 2014). In two-class problems, the class that contains many instances is the majority class whereas the class that contains fewer instances is the minority class. When the dataset is imbalanced, conventional classifiers typically favor the majority class thus failing to classify the minority observations correctly and resulting in performance loss (Prati, Batista, & Silva, 2014). When the training data is highly imbalanced, the minority class may not even be detected. This kind of imbalance that exists between two different classes is called *between-class* imbalance. Another kind of imbalance that results in performance loss is *within-class* imbalance, which happens when the minority or majority instances have more than one concept (sub-cluster of data) and some of these concepts have less number of instances than others. In addition, the presence of high overlapping among the concepts is another factor that leads to classifiers' performance loss on minority instances (Alshomrani, Bawakid, Shim, Fernández, & Herrera, 2015). Current methods developed for imbalanced dataset problems do not address both *within-class* imbalance and *between-class* imbalance at the same time. Most of these methods also exacerbate the overlapping among the concepts after trying to address the imbalance problem.

Traditionally, the objective of supervised learning is to optimize the accuracy for the whole dataset, which may cause the classifier to ignore the performance on each individual class. In particular, in an imbalanced dataset, if a random classifier predicts all instances as the majority class, a very high accuracy can be achieved despite incorrectly classifying all minority instances. Therefore, it is strongly suggested to use measurements that are suitable for imbalanced dataset classification.

In this paper, a new oversampling method called Adaptive Semi-Unsupervised Weighted Oversampling (A-SUWO) is presented for imbalanced binary dataset classification. A-SUWO finds hard-to-learn instances by first clustering the minority instances and then assigning higher weights to those instances from each sub-cluster that are closer to the majority class. This approach enables the identification of all instances that are close to the decision boundary and

* Corresponding author. Tel.: +18139745547.
E-mail addresses: nekooeimehr@mail.usf.edu (I. Nekooeimehr), laiyuen@usf.edu (S.K. Lai-Yuen).

http://dx.doi.org/10.1016/j.eswa.2015.10.031
0957-4174/

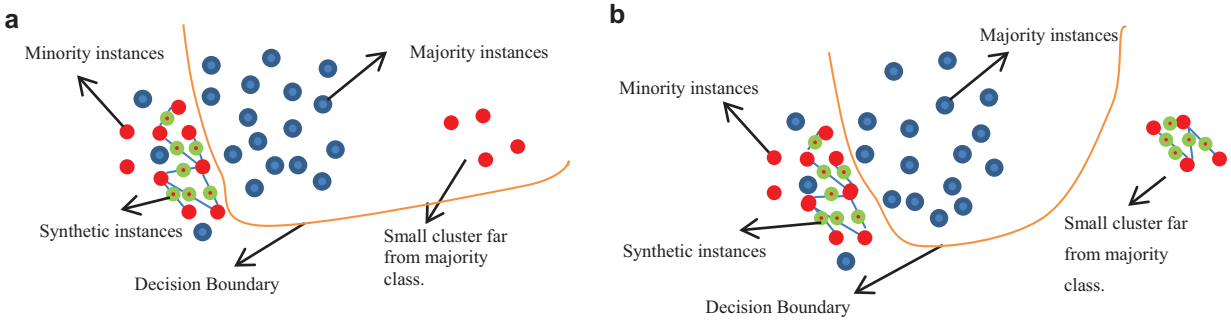

Fig. 1. (a) Minority cluster that is far from the majority class is ignored and not oversampled. (b) All minority clusters are considered for oversampling based on their misclassification complexity.

also considers all sub-clusters, even small ones, for oversampling as shown in Fig. 1(b). A-SUWO avoids over-generalization using two strategies. First, it clusters the minority instances by considering the majority class to reduce overlapping between the generated minority instances and majority instances. A semi-unsupervised hierarchical clustering approach is proposed that iteratively forms minority sub-clusters while avoiding majority sub-clusters in between. Second, it oversamples minority instances based on their average Euclidean distance to majority instances to further decrease the chance of generating overlapping instances between classes. In addition, A-SUWO determines sub-cluster sizes adaptively based on their misclassification error. In our method, misclassification error is an indication of sub-cluster complexity and is determined using a new measurement based on the standardized average error rate and cross validation. Sub-clusters with higher misclassification error will be assigned a larger size while the ones with lower misclassification error will be assigned a smaller size.

In order to validate A-SUWO, an extensive experimental design is performed. The proposed A-SUWO method is evaluated on 16 publicly available datasets, classified using 4 classifiers and compared with eight other oversampling techniques. F-measure, G-mean and AUC are used as the performance measures. The performance measures are determined using 4-fold stratified cross validation and repeated three times.

The remainder of this paper is organized as follows. In the next section, a review of related previous works is presented. In Section 3, the A-SUWO methodology is described. Section 4 presents the experimental design and results, while conclusions are provided in Section 5.

2. Previous work

There is an increasing interest in addressing imbalanced dataset classification. These works can be categorized into four main types of techniques: data preprocessing, algorithmic modification, cost-sensitive learning, and ensemble of classifier sampling methods (Díez-Pastor, Rodríguez, García-Osorio, & Kuncheva, 2015; Galar, Fernandez, Barrenechea, Bustince, & Herrera, 2012). Although there is no one single method that works well for all imbalanced dataset problems, sampling methods have shown great potential as they attempt to improve the dataset itself rather than the classifier (Barua, Islam, Yao, & Murase, 2014), (Chawla, Bowyer, Hall, & Kegelmeyer, 2011), (Han, Wang, & Mao, 2005), (Yen & Lee, 2009). Sampling methods change the distribution of each class observation by either oversampling the minority samples or undersampling the majority samples. In the case of oversampling, sampling methods generate new minority instances to balance the dataset and in the case of undersampling, they remove some majority instances from the dataset. Undersampling methods have shown to be less efficient than oversampling methods because the removal of majority instances may eliminate important information from the dataset, especially in cases where the dataset is small (He, Bai, Garcia, & Li, 2008; Japkowicz & Stephen, 2002; Zhou, 2013).

The simplest oversampling method is random sampling. It randomly selects a minority instance and duplicates it until the minority class reaches a desired size. Random oversampling generates new instances that are very similar to the original instances resulting in over-fitting. To overcome this problem, Chawla et al. developed Synthetic Minority Oversampling Technique (SMOTE) (Chawla et al., 2011) where new synthetic instances are generated between randomly selected minority instances and their *NN*-nearest neighbors, where *NN* is a user-defined variable. However, this may cause over-generalization as the new instances are generated without considering the majority instances thus increasing the overlap between minority and majority classes (He & Garcia, 2009), (Yen & Lee, 2009), (López, Fernández, García, Palade, & Herrera, 2013). Over-generalization can be exacerbated when the dataset has higher imbalance ratio as the instances of the minority class are very sparse and can become contained within the majority class after oversampling. This can further deteriorate subsequent classification performance (Kotsiantis, Kanellopoulos, & Pintelas, 2006).

Various approaches have been proposed to address over-generalization. Safe-level SMOTE (Bunkhumpornpat, Sinapiromsaran, & Lursinsap, 2009) presents a method that calculates a "safe-level" value for each minority instance, then generates synthetic instances closer to the largest safe level. The safe-level value is defined as the number of other minority instances among its *NN*-nearest neighbors. Safe-level SMOTE can cause overfitting because synthetic instances are forced to be generated farther from the decision boundary. Borderline-SMOTE (Han et al., 2005) presents a method to identify the borderline between the two classes, and oversamples only the minority samples on the borderline. ADASYN (He et al., 2008) assigns weights to minority instances so that those that have more majority instances in their neighborhood have higher chance to be oversampled. However, Borderline-SMOTE and ADASYN do not find all the minority instances close to the decision boundary (Barua et al., 2014). MWMOTE (Barua et al., 2014) approaches this problem by presenting a two-step procedure to find candidate majority border instances and then candidate minority border instances. Then, weights are assigned to candidate minority instances based on their Euclidean distances to the candidate majority border instances so that those with higher weights have a higher chance to be oversampled. However, small concepts of minority instances that are far from the majority class are not detected even if they may contain important information as shown in Fig. 1(a). In general, it is necessary to find hard-to-learn instances to be used for oversampling because they contain important information for the classifier. These instances are usually near the decision boundary or belong to small concepts (He & Garcia, 2009; Japkowicz & Stephen, 2002). The presence of small concepts in the dataset is referred to as *within-class* imbalance.