Information Sciences 454–455 (2018) 161–177

Contents lists available at ScienceDirect

Information Sciences

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

A hybrid evolutionary preprocessing method for imbalanced datasets

Ginny Y. Wong[a,*], Frank H.F. Leung[a], Sai-Ho Ling[b]

[a] Centre for Signal Processing, Department of Electronic and Information Engineering, Hong Kong Polytechnic University, Hung Hom, Hong Kong

[b] Centre for Health Technologies, Faculty of Engineering and Information Technology, University of Technology Sydney, NSW, Australia

ARTICLE INFO

Article history:
Received 24 April 2015
Revised 17 April 2018
Accepted 21 April 2018
Available online 1 May 2018

ABSTRACT

Imbalanced datasets are commonly encountered in real-world classification problems. Many machine learning algorithms are originally designed for well-balanced datasets, therefore re-sampling has become an important step to pre-process imbalanced data. This aims to balance the datasets by increasing the samples of the smaller class or decreasing the samples of the larger class, which are known as over-sampling and under-sampling, respectively. In this paper, a sampling strategy that is based on both over-sampling and under-sampling is proposed, in which the new samples of the smaller class are created based on fuzzy logic. Improvement of the datasets is done by the evolutionary computational method of Cross-generational elitist selection, Heterogeneous recombination and Cataclysmic mutation (CHC) that under-samples both the minority and majority samples. Consequently, a hybrid preprocessing method is proposed to re-sample imbalanced datasets. The evaluation is done by applying the Support Vector Machine (SVM), C4.5 decision tree and nearest neighbor rule to train a classification model from the re-sampled training sets. From the experimental results, it can be seen that our proposed method improves both the $F-measure$ and AUC. The over-sampling rate and complexity of the classification model are also compared. Our proposed method is found to be superior to all other methods under comparison and it is more robust in different classifiers.

1. Introduction

The classification of imbalanced datasets has recently been a popular topic [22,27]. Most machine learning tools, such as neural networks and support vector machines (SVMs), were originally designed for well-balanced datasets. Therefore, if the dataset is imbalanced, the performance of the classifier can be poor. The reason for this is apparent. For example, considering a dataset with 99% of data from class A and only 1% of data from class B, then the accuracy is 99% if the classifier ignores the data from class B and labels the whole dataset as class A. It is already very hard to achieve an accuracy above 99% by using most of the learning algorithms. However, the minority class of datasets is usually more important and meaningful. For example, in a medical problem, there are much fewer samples of people with a particular disease than those of healthy people. If a classifier is needed to label whether some people are infected or not, then the minority class (i.e. people with a particular disease) is the class of interest.

* Corresponding author.
E-mail addresses: ginnyyk.wong@connect.polyu.hk (G.Y. Wong), frank-h-f.leung@polyu.edu.hk (F.H.F. Leung), steve.ling@uts.edu.au (S.-H. Ling).

https://doi.org/10.1016/j.ins.2018.04.068
0020-0255/

Problems with imbalanced datasets can easily be found in the real world, such as intrusion detection [9], speech recognition [26], identification of power distribution fault causes [41], and bioinformatics problems [16]. There are two main approaches to solve problems caused by imbalanced datasets: the first is the data level approach and the second is the algorithm level approach. The data level approaches [3,8,18,28] include balancing the class distribution by over-sampling the minority class or under-sampling the majority class. The algorithm level approaches improve the existing machine learning methods by adjusting the probabilistic estimate [38], modifying the cost per class [32], adding some penalty constants [25], or learning from one class instead of two classes [30,35].

Many experiments show that re-sampling is a good data level approach to handle imbalanced data; see, for example, [12,15,42]. Moreover, it is more flexible because it does not depend on the chosen classifier. Therefore, we will focus on re-sampling in this paper. There are three main types of strategies for re-sampling data. The first is over-sampling, which can be done randomly or by the Synthetic Minority Over-sampling Technique (SMOTE) [8]. The second is under-sampling, which includes Tomek links [37] and the Neighborhood Cleaning Rule (NCL) [24]. The last is the hybrid method, which combines the two previous methods (over-sampling and under-sampling methods).

The importance of designing sampling strategies has been discussed in [31], which may affect the successful learning of different classes. Hybrid re-sampling methods are reported to have the advantage of treating datasets with a high imbalanced ratio, see [3,6]. Although some hybrid methods have been proposed to reduce the over-generalization problem from over-sampling methods, most of these methods are based on SMOTE and the results may be limited by the synthetic samples of SMOTE, see [3,34,40]. Therefore, a hybrid re-sampling method is proposed in this paper. Fuzzy logic, which is a useful tool to treat imbalanced datasets [12], is used to over-sample the minority class samples instead of SMOTE. A fuzzy rule base is formed based on the samples of the minority class. A rule is then selected randomly with reference to the effectiveness of each rule. The selected rule is used as the criteria to generate a new sample of the minority class. These steps will repeat until the majority class and minority class are the same size.

A large over-sampled training dataset will increase the complexity of the classification model and decrease the efficiency of the learning algorithm. It will also easily cause over-generalization, especially for some noisy datasets. This happens because the decision boundary could become narrow or the overlapping area between the majority class and minority class could become large after over-sampling. Therefore, an evolutionary algorithm (EA) is applied to both the synthetic samples and majority samples to under-sample the dataset. The chosen EA is the Cross-generational elitist selection, Heterogeneous recombination and Cataclysmic mutation (CHC) algorithm [11], which is able to select the most representative instances among the many algorithms studied in [5].

We will carry out experiments to compare our proposed method with three SMOTE-extended over-sampling methods, four hybrid re-sampling methods, and one under-sampling method, which are: SMOTE, Safe-Level-SMOTE [4], Adaptive Synthetic Sampling [21], SMOTE+Tomek Links [3], SMOTE+Rough Set [34], SMOTE+CHC (sCHC) [40], agglomerative hierarchical clustering [10], and EUSCHC [14]. A total of 44 imbalanced datasets from the UCI Repository [2] are used in the experiments. The SVM [7], C4.5 decision tree [33], and nearest neighbor rule (1NN) are used as tools to reach a classification model for each re-sampled dataset and evaluate each re-sampling method. The evaluation measures are based on the $F-measure$ and the area under the receiver operating characteristic curve (AUC). Although there are many hybrid pre-processing methods, only some of them are similar to our method, and consider and focus on the data size. In this paper, CHC is used to reduce the data size and achieve a good performance. Additionally, the proposed method enhances the performance in the over-sampling stage by taking advantage of the fuzzy rule base.

The rest of this paper is organized as follows. In Section 2, some preprocessing methods and CHC are reviewed. Section 3 presents the details of the proposed re-sampling strategy and the evaluation method. To show the effectiveness of our proposed approach, the comparisons with other methods and the results are discussed in Section 4. We will draw a conclusion in Section 5.

2. Previous work

This section describes some previous works that have used re-sampling methods, which will then be compared with our proposed method in the experiments. The concepts of the CHC will also be discussed.

2.1. Re-sampling methods

As discussed in the previous section, there are three main strategies for re-sampling data, which will be described in more detail in the following subsections.

2.1.1. Over-sampling methods

Some instances are produced for the minority class to balance the class distribution. The simplest is a non-heuristic method (random over-sampling) that replicates samples of the original minority class to generate the new instances. This method easily causes over-fitting because the new instances copy exactly from the original minority class. SMOTE [8] is a well-known method that creates the new instances by interpolating several minority samples that join together. This method makes use of each minority class sample and inserts synthetic samples along the line segments, joining any/all of the k minority class nearest neighbors to over-sample the minority class. An example is shown in Fig. 1. Five nearest