Neurocomputing 000 (2017) 1–12

Contents lists available at ScienceDirect

Neurocomputing

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

Using sub-sampling and ensemble clustering techniques to improve performance of imbalanced classification

Samad Nejatian[a,b], Hamid Parvin[c,d,*], Eshagh Faraji[d,e]

[a] Department of Electrical Engineering, Yasooj Branch, Islamic Azad University, Yasooj, Iran
[b] Young Researchers and Elite Club, Yasooj Branch, Islamic Azad University, Yasooj, Iran
[c] Department of Computer Engineering, Nourabad Mamasani Branch, Islamic Azad University, Nourabad Mamasani, Iran
[d] Young Researchers and Elite Club, Nourabad Mamasani Branch, Islamic Azad University, Nourabad Mamasani, Iran
[e] Department of Electrical Engineering, Nourabad Mamasani Branch, Islamic Azad University, Nourabad Mamasani, Iran

ARTICLE INFO

Article history:
Received 26 January 2016
Revised 13 April 2017
Accepted 10 June 2017
Available online xxx

Keywords:
Imbalanced learning
Neural networks
Decision tree
Cancer diagnosis

ABSTRACT

Abundant data of the patients is recorded within the health care system. During data mining process, we can achieve useful knowledge and hidden patterns within the data and consequently we will discover the meaningful knowledge. The discovered knowledge can be used by physicians and managers of health care to improve the quality of their services and to reduce the number of their medical errors. Since by the usage of a single data mining algorithm, it is difficult to diagnose or predict diseases, therefore in this research, we take a combination of the advantages of some algorithms in order to achieve better results in terms of efficiency. Most of standard learning algorithms have been designed for balanced data (the data with the same frequency of samples in each class), where the cost of wrong classification is the same within all classes. These algorithms cannot properly represent data distribution characteristics when datasets are imbalanced. In some cases, the cost of wrong classification can be very high in a sample of a special class, such as wrongly misclassifying cancerous individuals or patients as healthy ones. In this article, it is tried to present a fast and efficient way to learn from imbalanced data. This method is more suitable for learning from the imbalanced data having very little data in class of minority. Experiments show that the proposed method has more efficiency compared to traditional simple algorithms of machine learning, as well as several special-to-imbalanced-data learning algorithms. In addition, this method has lower computational complexity and faster implementation time.

1. Introduction

Different methods of data mining can help predict diseases automatically with high accuracy rate. Moreover, additional costs of irrelevant clinical trials will be reduced through this process. It also reduces the wrong predictions due to human tiredness, and consequently improves the quality of services. Some of the data mining methods that have been successfully applied to medical data include: neural networks, decision trees (DT), association rule mining, Bayesian networks, support vector machines (SVM), clustering and etc. Depending on the type of their application, one of these methods will be more useful. However, it is very hard to choose only a data mining algorithm that is suitable to diagnose or predict all diseases. Some algorithms are better than the others for certain purposes. However, when we bring advantages of several algorithms together, it will result in a better performance. Performance criteria will be discussed later in this study. By the way, it is almost impossible to choose the best data mining method to predict diseases for a specific criterion like accuracy, sensitivity and characteristic.

Data analysis and the confusion among them is a problem preventing to achieve remarkable diagnostic results, because the knowledge within the data should be used properly. In fact, data mining is a response to the need of health care organizations. The more data and the complexity of their relations are, the more difficult is to access the hidden information among data. It is often assumed that distribution of classes is balanced or nearly balanced. In general, the cost of wrong classification for all classes is assumed to be the same as well. So when the dataset is imbalanced, these algorithms cannot properly display data distribution features. In a sense, these algorithms tend to put an unknown data into the

* Corresponding author at: Department of Computer Engineering, Nourabad Mamasani Branch, Islamic Azad University, Nourabad Mamasani, Iran.
E-mail addresses: parvin@alumni.iust.ac.ir, parvinhamid@gmail.com (H. Parvin).

http://dx.doi.org/10.1016/j.neucom.2017.06.082
0925-2312/

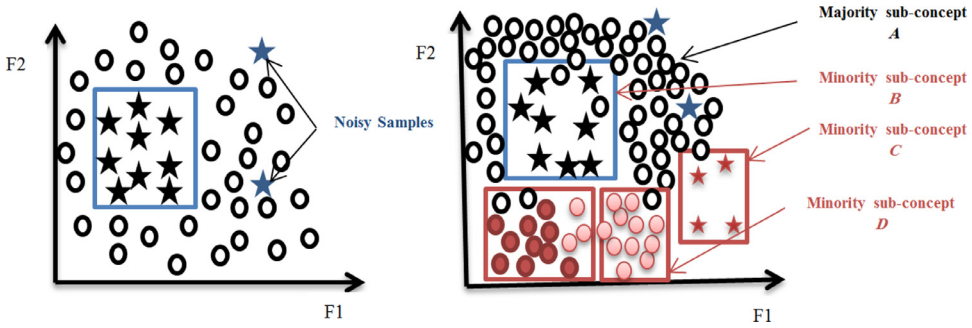

Fig. 1. (A) An imbalanced interclass dataset (left). (B) Dataset with high complexity, intra-class and interclass imbalance, multi-concept, overlapping of classes, noise (right).

classes with more frequency, and as a result, it provides unacceptable accuracy among data classes.

An imbalanced dataset is any dataset representing an imbalanced distribution among its classes, in such a way that the imbalanced distribution is too much. This type of imbalance is called inter-classes imbalance (such as a one-to-one thousand distribution (1:1000) where in this case, one class completely eliminates the other one). The imbalanced distribution wasn't between two classes necessarily and there may be among several ones, though. In scientific communities, over 65% rate of a class may be even considered to be imbalanced data [14,19,23,24].

The distributions among many actual datasets are mainly imbalanced, so it is necessary to modify the learning algorithms in order to extract knowledge out of them. As one example of these imbalanced dataset, we can exemplify the data related with the patients with breast cancer. These data are often shown with positive (cancer) and negative (health) classes. As expected, the number of healthy people is much higher than cancerous patients. Therefore, a kind of classification is required which exploits appropriate and balanced prediction accuracy for both minority and majority classes.

As we know that medical diagnosis of a cancerous patient as a healthy individual is unacceptable (and similarly a diagnosis of a healthy person as a patient), so in order to generate decision support systems, modified classifications are required. Applied classifiers must be able to provide high validity for minority class, but also does not affect the validity of majority one. For example, in this case, a healthy sample may be diagnosed 100% correctly, while the correct classification accuracy of the patient is 10%. So, it is very possible that the patient's sample is diagnosed wrongly. In this regard, it is obvious that the single evaluation criteria such as overall accuracy and error rate do not provide enough information about the quality of imbalanced learning. This kind of imbalance is called inherently imbalanced. This means that the imbalance is a direct result of the nature of data space. It is worth mentioning that imbalanced data are not just inherent; and imbalance can be sometimes relative as well, that is, the number of minority samples is naturally large but their number is very low compared to the majority class.

The data complexity is an important issue which includes data overlapping, missing data and etc. This concept is shown in Fig. 1. In Fig. 1, the stars and circles represent the minority and majority classes, respectively. As it is clear, two distributions shown in parts (A) and (B) are imbalanced, but in part (B), there are sample overlapping and multi-concept, too. According to part (B) the sub-concept C may be not learned because of lack of data.

Another form of imbalance is intra-class which corresponds to the distribution of representation data for sub-concepts in a class. In Fig. 1(B), class B and C represent the dominant minority and majority sub-concept, respectively. In addition, A and D are dominant concept and dominant sub-concept for majority class, respectively.

For each class, the number of samples existing in the dominant cluster of that class eliminates the sub-concept. As it is clear, this data space represents inter-classes and intra-class imbalance.

In this paper, we present a new method to classify imbalanced training data, and we compare this method with standard methods such as the nearest neighbor, decision tree and multi-layer perceptron neural network (MLP).

In the following, we review the literature and introduce some works done in this area. Then, we examine the evaluation criteria of these methods and the manner of classification tests. Finally, we will discuss the results of the tests and conclude the paper. In general, contributions presented in this article include:

- A new method for learning from imbalanced data.
- An efficient method to be used in the decision support system for breast cancer diagnosis.
- The results of the proposed method on real dataset of breast cancer.
- A method for the diagnosis of cardiovascular patients.

2. Related works

In this section, we review the literature of topic and the previous works. In this paper, training set and the number of its samples is presented by S and m. $S = \{(x_i, y_j) |\ i = 1, \ldots, m\}$ where $x_i \in X$ is a sample in the n-dimensional characteristic space of $X = \{(f_1, f_2, \ldots, f_n) | f_i \in \mathbb{R}\}$and $y_i \in Y = \{1, \ldots, c\}$ is the label of the class associated with the sample x_i. For example, $c = 2$ indicates a classification with two classes. S_{min} and S_{max} are sample sets of the minority and majority classes that the union of them is the training set, and intersection of them is null. Also, we consider E as the acquired set of sampling from S.

As discussed earlier, when a standard learning algorithm is applied to an imbalanced data, the minority class is not often learned well, because the deductive rules describing the minority concept are often much weaker than the ones describing the majority concept. In order to show the effect of imbalanced learning problem on standard learning algorithms, consider the general decision tree algorithm.

Decision tree is built based on a recursive top-down greedy search method which uses a feature selection method for selecting the best feature as the separation criterion in each node. Next, nodes are created based on possible values of the separator feature. As a result, at each stage, the training set is divided into smaller subsets which can result in separate rules of the class concept. Finally, these rules are combined, and make the hypothesis which results in the lowest error rate in the classes.

The problem occurring by using this process in the presence of imbalanced data is in two directions. First, frequent partitioning of data space leads to smaller observations of minority samples which brings about a reduction in the number of leaves describing the concept of minority class, and its result contains less con-

Please cite this article as: S. Nejatian et al., Using sub-sampling and ensemble clustering techniques to improve performance of imbalanced classification, Neurocomputing (2017), http://dx.doi.org/10.1016/j.neucom.2017.06.082