Preprocessing Imbalanced Dataset Using Oversampling Approach

Dasharath C. Magar1, S.M. Rokade2
1 Computer Engineering, S.V.I.T / Savitribai Phule University of Pune, India
2 Professor, Computer Engineering, S.V.I.T / Savitribai Phule University of Pune, India

Received 7 November 2015; Accepted 26 November 2015

ABSTRACT: The problem that contains an unequal distribution of data samples among different classes is known as imbalanced learning problem and it becomes a big challenge to any classifier as it becomes very hard to learn the minority class samples. It is because the classifier learned from the imbalanced data tends to favor the majority class samples, resulting in a large classification error over the minority class samples. This becomes very costly when identification of the minority class samples is crucial. Thus, the classifier learned from the imbalanced data needs to perform equally well both on the minority class and the majority class samples. Synthetic oversampling methods add samples by generating the synthetic minority class samples. The generated samples add essential information to the original data set that may help to improve classifiers’ performance. But most of the oversampling methods may generate the wrong synthetic minority samples in some scenarios and make learning tasks harder. So a new oversampling method is presented for efficiently handling imbalanced learning problems. This method first identifies the hard-to-learn informative minority class samples and assigns them weights according to their Euclidean distance from the nearest majority class samples. It then generates the synthetic samples from the weighted informative minority class samples using a clustering approach.

KEYWORDS – clustering, Imbalanced learning, oversampling, synthetic sample generation, undersampling.

1. Introduction

Imbalanced learning problems contain unequal distribution of data samples among different classes, where most of the samples belong to some classes and rest to the other classes. If such samples come only from two classes, the class having most of the samples is called the majority class and other the minority class. The primary goal of any classifier is to reduce its classification error, i.e., to maximize its overall accuracy[2]. However, imbalance learning problems pose a great challenge to the classifier as it becomes very hard to learn the minority class samples. It is because the classifier learned from the imbalanced data tends to favor the majority class samples, resulting in a large classification error over the minority class samples. This becomes very costly when identification of the minority class samples is crucial. Thus, the classifier learned from the imbalanced data needs to perform equally well both on the minority class and the majority class samples. Imbalance that exists between the samples of two classes is usually known as between-class imbalance. The actual cause for the bad performance of conventional classifiers on the minority class samples is not necessarily related to only on the between-class imbalance. Classifiers’ performance have been found to depreciate in the presence of within-class imbalance and small disjuncts problems. Besides, the complexity of data samples is another factor for the classifiers’ poor performance. If the samples of the majority and minority classes have more than one concepts in which some concepts are rarer than others and the regions between some concepts of different
classes overlap, then the imbalance problem becomes very severe.

Some of the most popular approaches to deal with imbalanced learning problems are based on the synthetic oversampling methods. In this paper, I am illustrating that in some scenarios many of these methods become inappropriate and fail to generate the useful synthetic minority class samples. It also shown scenarios for which the methods will generate wrong and unnecessary synthetic samples, thus will make learning tasks difficult. In this respect, proposed a novel synthetic oversampling method, whose goal is to alleviate the problems of imbalanced learning and generate the useful synthetic minority class samples. The essences of the proposed method are:

1) Selection of an appropriate subset of the original minority class samples
2) Assigning weights to the selected samples according to their importance in the data.
3) Using a clustering approach for generating the useful synthetic minority class samples.

2. Literature Survey
Yanping Yang, Guangzi M. Suggested an ensemble-based active learning algorithm to address the class imbalance problem. The artificial data are created according to the distribution of the training dataset to make the ensemble diverse, and the random sub-space resampling method is used to reduce the data dimension. In selecting member classifiers based on misclassification cost estimation, the minority class is assigned with higher weights for misclassification costs, while each testing sample has a variable penalty factor to induce the ensemble to correct current error. In our experiments with UCI disease datasets, instead of classification accuracy, F-value and G-means are used as the evaluation rule. Compared with other ensemble methods, their method shows best performance[3].

Sukarna Barua, Md. Monirul Islam, Xin Yao, Fellow, IEEE, and Kazuyuki Murase presented a new approach for Preprocessing Imbalanced Dataset. Imbalanced learning problems contain an unequal distribution of data samples among different classes and pose a challenge to any classifier as it becomes hard to learn the minority class samples. Synthetic oversampling methods address this problem by generating the synthetic minority class samples to balance the distribution between the samples of the majority and minority classes. This paper identifies that most of the existing oversampling methods may generate the wrong synthetic minority samples in some scenarios and make learning tasks harder. To this end, a new method, called Majority Weighted Minority Oversampling Technique (MWMOTE), is presented for efficiently handling imbalanced learning problems. MWMOTE first identifies the hard-to-learn informative minority class samples and assigns them weights according to their euclidean distance from the nearest majority class samples. It then generates the synthetic samples from the weighted informative minority class samples using a clustering approach. This is done in such a way that all the generated samples lie inside some minority class cluster. MWMOTE has been evaluated extensively on four artificial and 20 real-world data sets. The simulation results show that our method is better than or comparable with some other existing methods in terms of various assessment metrics, such as geometric mean (G-mean) and area under the receiver operating curve (ROC), usually known as area under curve (AUC)[2].

Cieslak and Chawla proposed a cluster-based algorithm, called local sampling, in which the Hellinger distance measure is used first for partitioning the original data set. A
sampling method is then applied to each partition and finally data of all partitions are merged to create the new data set[6].

Naheed Azeem, Shazia Usmani proposed different data mining techniques for identifying fault prone modules as well as compare the data mining algorithms to find out the best algorithm for defect prediction[7].

Renqing Li and Shihai Wang proposed tools like C4.5, SVM, KNN, Logistic, NaiveBayes, AdaBoost and SMOTEBoost based on software metrics for balancing data set[8].

In [11] Proposed sampling techniques that have been shown to be very successful in recent years. To address imbalanced learning issue oversampling of minority class is done. There are various Oversampling techniques which can be used to reestablish the class balance. Oversampling method is a data level method. The main advantage of data level methods is that they are self-sufficient. The methods at data level modify the distribution of the imbalanced datasets, and then these modified i.e. balanced datasets are provided to the algorithm to improve the Imbalanced learning.

In [10] proposed a method that finds minority samples which are difficult to learn and computes Euclidean distance between nearest majority class samples. Using clustering approach and weighted minority class samples it generates synthetic samples for oversampling purpose. Proposed work will evaluate this approach on real & artificial datasets.

Kehan Gao, Taghi M, Khoshgoftaar and Randall Wald employ two sampling techniques, random undersampling (rus) and synthetic minority oversampling technique (smote), and two ensemble boosting approaches, rusboost and smoteboost (in which rus and smote, respectively, are integrated into a boosting technique), as well as six feature ranking techniques. They apply the proposed techniques to several groups of datasets from two real-world software systems and use two learners[12].

Talayeh Razzaghi introduced a cost-sensitive learning method (CSL) to deal with the classification of imperfect data. Typically, most traditional approaches for classification demonstrate poor performance in an environment with imperfect data. They propose the use of CSL with Support Vector Machine, which is a well-known data mining algorithm. The results reveal that the proposed algorithm produces more accurate classifiers and is more robust with respect to imperfect data. Furthermore, they explore the best performance measures to tackle imperfect data along with addressing real problems in quality control and business analytics[13].

Hong Cao and Xiao-Li Li proposes a novel Integrated Oversampling (INOS) method that can handle highly imbalanced time series classification. They introduce an enhanced structure preserving oversampling (ESPO) technique and synergistically combine it with interpolation-based oversampling. ESPO is used to generate a large percentage of the synthetic minority samples based on multivariate Gaussian distribution, by estimating the covariance structure of the minority-class samples and by regularizing the unreliable eigen spectrum. To protect the key original minority samples, They used an interpolation-based technique to oversample a small percentage of synthetic population [14].
3 System Architecture

![System Architecture Diagram]

Figure 3.1: Architecture of system

3.1 Modules:
The system architecture is divided into three modules.

Module 1: Construction of Minority Class Samples (set $S_{\text{imin}}$):
In the first module, my oversampling method identifies the Most important and hard-to-learn minority class samples from the original minority set, $S_{\text{min}}$ and construct a set, $S_{\text{imin}}$, by the identified samples as shown in Figure 3.1. The method first filters the original minority set, $S_{\text{min}}$, to find a filtered minority set, $S_{\text{minf}}$. To do this, I compute $NN(x_i)$ for each $x_i \in S_{\text{min}}$. I then remove each $x_i$ if its $NN(x_i)$ contains only the majority class samples as shown in Figure 3.1. The removed minority class sample is likely to be noisy because it is surrounded only by the majority class samples. This removal prohibits the noisy minority class sample to take part in the synthetic sample generation process. Thus, the method will be able to remove existing noisy samples from the given data and at the same time will not add any new noise, i.e., noisy synthetic sample to the data.

For each $x_i \in S_{\text{minf}}$, the method constructs $N_{\text{maj}}(x_i)$. The samples in $N_{\text{maj}}(x_i)$ will be the borderline majorities and expected to be located near the decision boundary when $k_2$ is small. I combine all the $N_{\text{maj}}(x_i)$s to form the borderline majority set, $S_{\text{bmaj}}$.

For each $y_i \in S_{\text{bmaj}}$, the method constructs $N_{\text{maj}}(y_i)$ and combines all such $N_{\text{maj}}(y_i)$s to form $S_{\text{min}}$. The parameter $k_3$ used in $N_{\text{maj}}(y_i)$ needs to be large enough for including all the hard-to-learn minority class samples required to generate the synthetic samples. A large $k_3$ ensures the participation of many difficult samples in the sample generation process. The generated samples will likely add sufficient and essential information to learning Figure 3.1.

Module 2: Finding the Weights for the Informative samples of $S_{\text{imin}}$
In the second module, each member of $S_{\text{imin}}$ is given a selection weight, $S_w$, according to its importance in the data. Samples close to the decision boundary contain more information than those of further so assign more weights to such samples. The minority class samples in a sparse cluster are more important than those in a dense cluster so assign more weights to such samples. The minority class samples near a dense majority class cluster are more important than those near a sparse majority class cluster, so assign more weights to such samples Figure 3.1.

Module 3: Synthetic Sample Generation
In the third module, the method generates the synthetic samples from $S_{\text{imin}}$ using $S_w$s and produces the output set, $S_{\text{omin}}$, by adding the synthetic samples to $S_{\text{min}}$. My method finds the clusters of the minority data set, $S_{\text{min}}$, using a modified hierarchical clustering algorithm Figure 3.1.

Clustering $S_{\text{min}}$
1. Assign each sample to a separate cluster, i.e., initially, there will be $D$ clusters, each of size one.
2. Find the two closest clusters say, $L_i$ and $L_j$.
3. Merge the clusters $L_i$ and $L_j$ into a single cluster, $L_m$. This will reduce the number of clusters by one.
4. Update the distance measures between the newly computed cluster and all the previous cluster(s).
5. Repeat Steps 2-4 until all data
samples are merged into a single cluster of size $D$.

4 Conclusion

Many oversampling methods exist in the literature for imbalanced learning problems. These methods generate the synthetic minority class samples from the hard-to-learn minority class samples with an aim to balance the distribution between the samples of the majority and minority classes. However, in many conditions, existing methods are not able to select the hard-to-learn minority class samples effectively, assign relative weights to the selected samples appropriately, and generate synthetic samples correctly. Based on these observations, I propose a new method, for imbalance learning problems. The method not only selects the hard-to-learn minority class samples effectively but also assigns them weights appropriately. Furthermore, it is able to generate correct synthetic samples. The method takes an equal samples in majority and minority class so that dataset will be more accurate.

References


[13] Cost-sensitive Learning –Based Methods for Imbalanced Classification Problems with Applications, Talayeh Razzaghi
[14] Integrated Oversampling for Imbalanced Time Series Classification, Hong Cao and Xiao-Li Li.