IJARIIT

INTERNATIONAL JOURNAL OF ADVANCE RESEARCH, IDEAS AND INNOVATIONS IN TECHNOLOGY

ISSN: 2454-132X

Impact factor: 4.295

(Volume2, Issue6)

Available online at: www.ijariit.com

Class Imbalance Problem in Data Mining using Probabilistic Approach

Disha Gupta[*]
Dept. of CSE, RTMNU, Nagpur
disha.g14@gmail.com

Reetu Gupta
Dept. of IT, Pune University
reetugupta.rs@gmail.com

Prashant Khobragade
Dept. Of CSE, RTMNU, Nagpur
prashukhobragade@gmail.com

Abstract— Class imbalance problem are raised when one class having maximum number of examples than other classes. The classical classifiers of balance datasets cannot deal with the class imbalance problem because they pay more attention to the majority class. The main drawback associated with it majority class is loss of important information. The Class imbalance problem is a difficult due to the amount and nature of data. This paper focuses different methods of class imbalance problem. It is been consider the majority class to achieve the class imbalanced problem. This paper mainly focuses the minority class sample to achieve the problem and proposed method for class imbalance problem using minority sample data. The oversampling and under sampling both concept were used to identify the correct class label of the sample using probabilistic approach, the main objective of this paper, to proposed method to minimize the misclassification rate of minority class sample, balance and classify the data more accurately thereby improving the performance of classifier.

Keywords— Class Imbalance, Data Mining, Oversampling, Classification, KNN Clustering.

I. INTRODUCTION

The classification techniques typically consider a balanced class distribution between two or more than classes in data mining. The task of construct such classifier is to predict and classify the data based on their class label for an unobserved input objects based on a certain number of observations. Usually, a classifier performs better, when the classification technique is applied to a dataset evenly or equally distributed to different classes. The problem of imbalanced class division occurs when one class is represented by a large number of data while the remaining other is represented by only a few[1]. The majority of work has been done with majority sample class and simply avoid of minority class sample. But some time minority sample data also be used for the class definition and may improve the classification result. The class imbalance problem can appear either from between classes (inter class) or within a single class (intra class) [4]. Inter-class imbalance refers to the case when one class has larger number of example than another class. The degree of imbalance can be represented by the ratio of size of the minority class to size of the majority class. However, for classification of imbalanced data, other performance factor should be calculated such as Precision, Recall, f-measure and G-mean etc. most of algorithms are more focusing on classification of majority class sample while not considering or misclassifying minority sample. Misclassification of minority class affects in many real time cases as fraudulent credit card transactions, medical diagnosis and e-mail filtering.

II. LITERATURE REVIEW

An easy Bernan Das et al. [1] introduced two probabilistic approaches, namely RACOG and wRACOG to synthetically generating and strategically selecting new minority class samples. The proposed approaches use the joint probability distribution of data attributes and Gibbs sampling to generate new minority class samples, While RACOG selects samples produced by the Gibbs sampler based on a predefined lag, wRACOG selects those samples that have the highest probability of being misclassified by the existing learning model.

Xiaowan Zhang and Bao-Gang Hu paper [2] define cost-free learning (CFL) formally in comparison with cost-sensitive learning (CSL). The main difference between them is that a CFL approach seeks optimal classification results without requiring any cost information, even in the class imbalance problem. Using the strategy can handle binary/multi-class classifications with/without abstaining. Significant features are observed from the new strategy. While the degree of class imbalance is changing, the proposed strategy is able to balance the errors and rejects accordingly and automatically.

Chris Seiffert et al. [3] present a new hybrid sampling/boosting algorithm, called RUSBoost, for learning from skewed training data. This algorithm provides a simpler and faster alternative to SMOTE Boost, which is another algorithm that combines boosting and data sampling. This paper evaluates the performances of RUSBoost and SMOTE Boost, as well as their individual components (random under-sampling, synthetic minority oversampling technique, and AdaBoost). RUSBoost and SMOTE Boost both outperform the other procedures, and RUSBoost performs comparably to and often better than SMOTE Boost while being a simpler and faster technique.

Raisul Islam Rashu et al. [4] gives data mining approaches that have been used in business purposes since its inception however, at present it is used successfully in new and emerging areas like education systems. In this paper, it uses data mining approaches to predict students' final outcome, i.e., final grade in a particular course by overcoming the problem of imbalanced dataset. Several re-sampling techniques are given to balance the dataset so that to get better performance. Re-sampling techniques include SMOTE, RUS, ROS.

Jueun Kwak et al. [5], when the class sizes are highly imbalanced, the standard algorithm tend to strongly favor the majority class and provide notably low detection of the minority class as a result. The method proposes an online fault detection algorithm based on incremental clustering. The algorithm accurately finds wafer faults even in severe class distribution skews and efficiently processes massive sensor data in terms of reductions in the required storage.

Wattana Jindaluang et al. gives the class imbalance problem using under-sampling [6] has a drawback that it throws away important information in a majority class. To overcome this problem, this paper proposed a cluster based under-sampling method. This used a clustering algorithm that is performance guaranteed, named k-centers algorithm, which clusters the data in the majority class and selects a number of representative data in many proportions, and then combines them with all the data in the minority class as a training set.

Shuo Wang et al. [7], they improve the resampling strategy inside OOB (Oversampling based online bagging) and UOB (Under sampling based online bagging), and look into their performance in both static and dynamic data streams. They give the first comprehensive analysis of class imbalance in data streams, in terms of data distributions, imbalance rates and changes in class imbalance status.

They find that UOB is better at recognizing minority-class examples in static data streams, and OOB is more robust against dynamic changes in class imbalance status. Then they propose two new ensemble methods that maintain both OOB and UOB with adaptive weights for final predictions, called WEOB1 and WEOB2. They are shown to possess the strength of OOB and UOB with good accuracy and robustness.

Shuo Wang et al. [8] Studies the challenges posed by the multiclass imbalance problems and investigate the generalization ability of some ensemble solutions, including their recently proposed algorithm AdaBoost.NC, with the aim of handling multiclass and imbalance effectively and directly.

III. EXISTING METHODOLOGY

All The imbalanced data problem in classification can appear in two different types of data sets: binary problems, where one class having more number of samples than the other and multi-class problems, where the applications have more than two classes and unbalanced class distribution hinder the classification performance.

One of the common approaches to tackle class imbalance problem is sampling. Sampling methods modify the distributions of the majority and minority class in the training data set to obtain a more balanced number of instances in each class. To minimize class imbalance, there are two basic methods, under sampling and over sampling.

A. Under-sampling

It removes data from the original data set by randomly select a set of majority class example and then remove this sample [2]. Hence, an under sample approach is aim to decrease the skewed distribution of majority class by lowering the size of majority class [9]. Under-sampling is suitable for large application where the number of majority samples is very large and lessening the training instances reduces the training time and storage. The drawback of this technique is that there is no particular technique to remove patterns of the majority class, thus it can discard data potentially important for the classification process which degrade classifier performance.

B. Oversampling

It is a method to adding a set of sampled from minority class by randomly select minority class examples and then replicating the selected examples and adding them to data set [9]. The advantage is that no information is lost, all instances are employed. However, the major problem of this technique is leads to a higher computational cost. The drawback of this technique is if some of the small class samples contain labeling error, adding them will actually deteriorate the classification performance on the small class [12].

However, both oversampling and under sampling are capable of solving the imbalance class problem and both of them having their own advantage and disadvantage. Comparing oversampling and under re-sampling, observation simply favoring oversampling is that under-sampling removes some data from the original data, that data may be important so it result in loss of information while oversampling does not suffer from this problem.

C. Hybrid Method

In this method, both oversampling and under sampling method is used to balance the dataset. In hybrid method, oversampling is usually done on minority class samples and under sampling is usually done on majority class samples. Hybrid method can also utilize using bagging and boosting [11].

D. Random Oversampling (ROS)

Random oversampling balances a data set by duplicating examples of the minority class until a desired class balance ratio is achieved. The benefit of using under sampling to balance the class distribution in the data set is that the time required to train the classifier will be relatively short, but a smaller training data set also has their drawback that is loss of valuable information [10].

IV. PROPOSED APPROACH

The proposed approach is to solve the class imbalance problem using oversampling technique. Oversampling can be done by replicating or by generating samples synthetically. Replicating samples causes unnecessary addition of samples, so proposed approach clusters the samples in such a way that the misclassified samples are organized in a cluster separating classified samples. Using minority misclassified samples; new samples are generated and oversampled to build a classifier model more accurate. Thus improves the accuracy as performance measure of classifier and enables the dataset become appropriately balanced. Figure 1.2 is Block diagram of proposed approach.

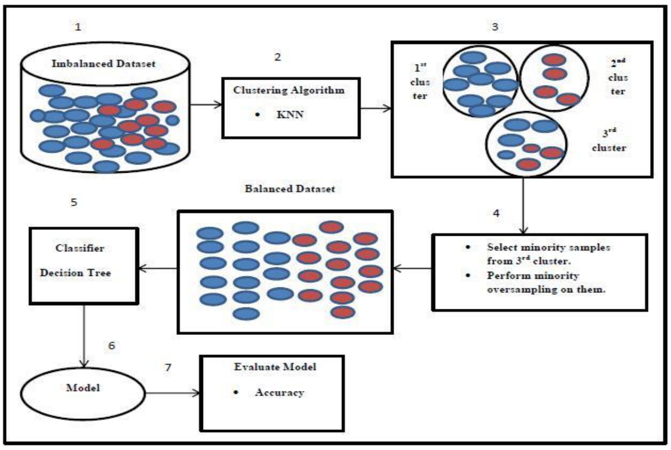

Figure 1.2 Block diagram of proposed approach

Main objective in this research is to try to increase the classification accuracy of minority class by avoiding the drawbacks of the existing approaches. For that, an efficient approach combines the clustering approach and oversampling approach to deal with the problem of class imbalance class.

To implement and evaluate this approach use the following methodology steps as presented in Figure 1.3.

Step 1. Collection data: Collect various numeric datasets from UCI machine learning repository. Here, Haberman, Shuttle and Ecoli dataset are used.

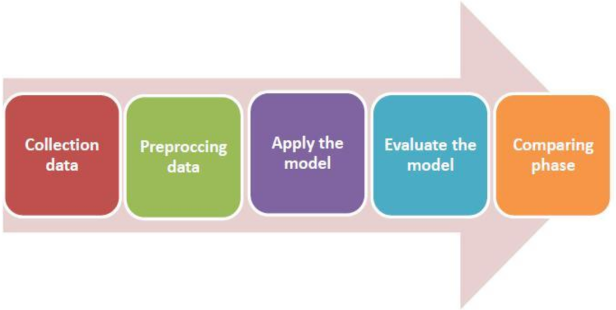

Figure 1.3 Methodology Steps

Step 2. Preprocessing data: Apply proposed approach which is KNN based selection of minority class, whose chances of getting miss classification is more [18]. By applying proposed approach dataset is divided into three different clusters then select the data sample belong to minority class and whose chances of getting misclassified is more. Generate the new minority class sample by applying SMOTE on selected minority class sample and there by balance the dataset in turn classify the data more accurately.

Step 3. Apply the model: Implement model by using one of the classification algorithms. Use C4.5 decision tree classifier for the classification purpose.

Step 4. Evaluate the model: Evaluate the classification performance of model. Use accuracy and F-measure, precision, recall, G-mean as performance metrices.

Step 5. Comparing phase: Compare performance before using proposed approach and after using it. In comparison phase, comparison will be done on performance parameter of classifier such as overall accuracy, precision, recall, G-Mean before applying proposed approach and after applying proposed approach. Figure 1.3 is of Methodology Steps which defines steps of project work.

The proposed architecture creates a classifier based on the input dataset. Run classifier and classify the dataset in binary class. In classification, decision tree classifier builds a model on training data and checks it with test data. The classifier model is built with C4.5 decision tree classifier which deploys the classification task and gives correctly and incorrectly classified data. Then dataset is given to Clustering. In KNN selection based clustering; dataset is clustered in such a way that it clusters misclassified data. Select only those clusters which have minority samples that are misclassified. Then selected minority samples are given to oversampling technique where it generates new samples. Dataset which is obtained after generating samples by applying oversampling gives back to check whether the data set is balanced or not. The probabilistic oversampling technique handles the imbalanced datasets to get balanced datasets.

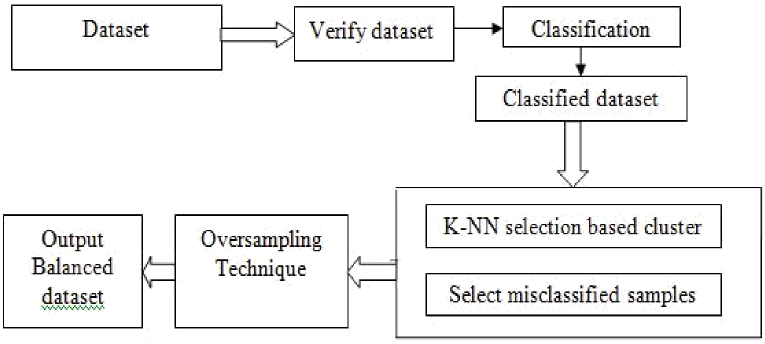

Figure 1.4 Proposed Architecture

Figure 1.4 is of proposed architecture where input dataset is imbalanced dataset then is preprocessed to verify whether it is balanced or not. Then next K-NN selection based clustering is done to select minority misclassified samples which are oversampled to obtain balanced dataset.

CONCLUSION

This paper reported different technique for class imbalance problem the data level process provides better approach to balance class using minority sample value by using the oversampling algorithm on the given dataset. In data preprocessing, oversampling has many advantages over under sampling, but oversampling also has some disadvantages that it replicates unnecessary information. The proposed a new approaches where it oversamples only those minority samples which gets misclassified. The approach also improves the accuracy of the classifier and balances the data more appropriately using KNN classification and clustering mechanism. In future work, researchers need to analyze approach on multiclass imbalance problem and also need to consider the problem of imbalance data with noisy dataset especially if the noise in class attribute.

REFERENCES

[1] Barnan Das, Narayanan C. Krishnan, Diane J. Cook, “RACOG and wRACOG Two Probabilistic Oversampling Techniques” IEEE TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING, VOL. 27, NO. 1, JANUARY 2015, PP 222-232.

[2] Xiaowan Zhang, Bao-Gang Hu, “A New Strategy of Cost-Free Learning in the Class Imbalance Problem” IEEE TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING, VOL. 26, NO. 12, DECEMBER 2014, PP 2872-2881.

[3] Chris Seiffert, Taghi M. Khoshgoftaar, Member, IEEE, Jason Van Hulse, Member, IEEE, and Amri Napolitano “RUSBoost: A Hybrid Approach to Alleviating Class Imbalance” IEEE TRANSACTIONS ON SYSTEMS, MAN, AND CYBERNETICS—PART A: SYSTEMS AND HUMANS, VOL. 40, NO. 1, JANUARY 2010, PP 185-194.

[4] Raisul Islam Rashu, Naheena Haq, Rashedur M Rahman “Data Mining Approaches to Predict Final Grade by Overcoming Class Imbalance Problem” 2014 17th International Conference on Computer and Information Technology (ICCIT), PP 215-222.

[5] Jueun Kwak, Taehyung Lee, Chang Ouk Kim “An Incremental Clustering-Based Fault Detection Algorithmfor Class Imbalanced Process Data” IEEE TRANSACTIONS ON SEMICONDUCTOR MANUFACTURING, VOL. 28, NO. 3, PP 212-220.

[6] Wattana Jindaluang, Varin Chouvatut,Sanpawat Kantabutra “Under-sampling by Algorithm with Performance Guaranteed for Class-imbalance Problem” 2014 International Computer Science and Engineering Conference (ICSEC), PP 812-823.