IJARCCE

ISSN (Online) 2278-1021
ISSN (Print) 2319 5940

International Journal of Advanced Research in Computer and Communication Engineering

ISO 3297:2007 Certified

Vol. 5, Issue 8, August 2016

# A Study of Imbalanced Classification Problem

**P. Rajeshwari**[1], **D. Maheshwari**[2]

Research Scholar, Department of Computer Science, Dr.N.G.P. Arts and Science College, Coimbatore[1]

Department of Computer Technology, Dr.N.G.P. Arts and Science College, Coimbatore[2]

**Abstract:** Recently there are major changes and evolution has been done on classification of data. Class imbalance problem become greatest issue in data mining. Imbalance problem occur where one of the two classes having more sample than other classes. The most of algorithm are more focusing on classification of major sample while ignoring or misclassifying minority sample. The minority samples are those that rarely occur but very important.. Class imbalance is one of the most challenges of machine learning and data mining fields. In this paper we focused on more techniques involve solving rare class or imbalanced problem.

**Keywords:** Class imbalance problem, data pre-processing, machine learning, and algorithm.

## 1. INTRODUCTION

### 1.1 Data mining

Data mining is defined as the process of extracting previously unknown, valid, and actionable information from large databases and then using the information to make crucial business decisions. Finding information hidden in data is as theoretically difficult as it is practically important.

It has lot of technique such as artificial intelligent, neural networks, genetic algorithms, K-nearest neighbour method, decision trees, and data reduction. These are all basic techniques in data mining.

### 1.2. Imbalance classification

In class imbalanced problem consider where one class is under-represented relative to another, remains among the leading challenges in the development of prediction models. In majority class will have high accuracy prediction as well as minority class will have low accuracy prediction. Classification of data becomes more difficult because of unlimited size and imbalance nature of data.

In the Imbalance problem occur where one of the two classes having more sample than other classes. Classification of imbalance data set which is divided into three main categories, the algorithmic approach, data pre-processing approach and feature selection approach.

Each of this technique has their own advantages and disadvantages. Data pre-processing technique sampling is applied on data in which either new samples are added or removed existing samples. Process of adding new sample is known as over-sampling and process of removing a sample known as under-sampling. Another method for solve problem of class imbalance is boosting. Boosting is powerful learning algorithm that Improve the performance of weak classifier. Feature selection is a key step for many machine learning algorithms, especially when the data is high-dimensional.

## 2. LITERATURE REVIEW

Classification is a problem frequently encountered when a categorical dependent variable analyzed and its relation to a set of independent variables is explored. The ever-increasing growth in data quantity and computation complexity has largely deteriorated the performance and accuracy of classification models. In order to deal with such situations [1]

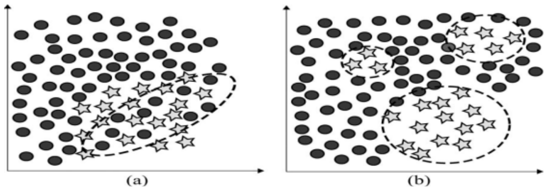
Hung-Yi Lin proposed multivariate statistical analyses. Multivariate statistical analyses have two advantages. First, they can explore the relationships between variables and find the most characterizing features of the observed data. Second, they can solve problems which are stalled by high dimensionality.

Our learning model advances three new distinguishing characteristics including evaluation method, feature selection, and feature extraction. Hence, the main contributions of this paper are threefold. First, the enhanced relevance analysis is proposed for feature evaluation process.

Second, the synergistic classification effect is enhanced by our heuristic feature selection algorithm. Finally, the generation of coarse-grained classifier composed of lowly correlated relevant features is successfully realized. Bagging method consists in training different classifiers with bootstrapped replicas of the original training data-set. [2]

Breiman et al proposed concept of bootstrap aggregating to construct ensembles. That is, a new data-set is formed to train each classifier by randomly drawing instances from the original data-set. Hence, diversity is obtained with the re-sampling procedure by the usage of different data subsets.
Finally, when an unknown instance is presented to each individual classifier, a majority or weighted vote is used to infer the class.

DOI 10.17148/IJARCCE.2016.5812

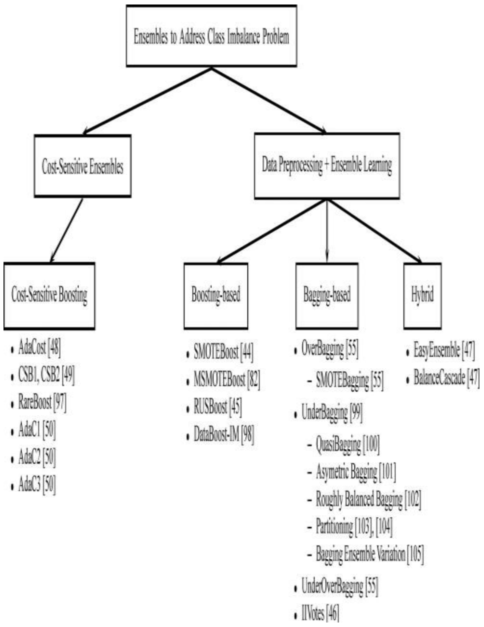


Example of difficulties in imbalanced data-sets. (a) Class overlapping (b) Small disjuncts

Boosting (also known as ARCing, adaptive resampling and combining) was introduced by [3]Schapire et al. Schapire proved that a weak learner can be turned into a strong learner in the sense of probably approximately correct (PAC) learning framework. AdaBoost is the most representative algorithm in this family, it was the first applicable approach of Boosting, and it has been appointed as one of the top ten data mining algorithms. AdaBoost is known to reduce bias (besides from variance), and similarly to support vector machines (SVMs) boosts the margins. AdaBoost uses the whole data-set to train each classifier serially, but after each round, it gives more focus to difficult instances, with the goal of correctly classifying examples in the next iteration that were incorrectly classified during the current iteration.

Igloo-Plot tool addresses some of the key limitations of in existing multivariate visualization and analysis tools. Visual of data mining tool to cover the secure information from the data set, it convey the identify trend to end user. [4]Bhusan K. Kuntal, Tarini Shankar Ghosh et al. Proposed multivariate Here present a new GUI based tools analyze for multivariate datasets. It can capture three main things for any multidimensional data. Firstly, identify cluster based on similarity within the analyze data point. Secondly, user to visualize the large amount of data and find different feature for every data point using based on color code string. Finally, it involves quickly identification of cluster of data group. Feature advantage to collect even extract information from any multi dimensional data destroy the information. It focused to solve the complicated of the feature characterized this database.

Some problem such as infinite population of data and feature selection process of classifier is critical problem found in review. [5]Mahendra Sahare, et al. proposed lot of technique to solving multivariate problem. Implementation of binary classifier in the form of liner classifier generate such a problem, the first approach relied on extending binary classification problems to handle the multiclass case directly. This included neural networks, decision trees, support vector machines, naive bayes, and k-nearest neighbours. The second approach decomposes the problem into several binary classification tasks. Several methods are used for this decomposition: one versus- all, all-versus-all, error-correcting output coding, and generalized coding. The third one relied on arranging the classes in a tree, usually a binary tree, and utilizing a number of binary classifiers at the nodes of the tree till a leaf node is reached. In future it minimized the problem of feature reduction problem and error correcting code for binary classifier.



Taxonomy for ensembles to address the class imbalance problem

Unbalanced classification technique lots of incremental training algorithm, which is suitable for problems of sequentially arriving data and fast constraint parameter variation. For multiclass SVM, there are mainly two types of approaches for training and classification, one of which is to combine several binary classifiers and another is to consider all the classes in one big optimization problem. We have to adopt some techniques to deal with the worst situation. When a node encounters the worst situation, we record the trained pair and its classifier to the trained list, delete the two sub nodes, and select another pair of the node to train. [6]Bee Wah Yap, Khatijahhusna Abd Rani et al. Proposed lot of methods. This paper applied four methods: Oversampling, under sampling, Bagging and Boosting in handling imbalanced datasets. Document classification, loan default prediction, fraud detection or medical Classification which involve a binary response variable, the dataset are often highly imbalanced. For a binary response variable with two classes, when the event of interest is underrepresented, it is referred to the positive or minority class. Thus, the number of cases for the negative or majority class is very much higher than the minority cases. A simulation study should be carried out whereby data are generated and then the different approaches are compared so as to obtain a conclusive decision on the best strategy to handle imbalanced data.

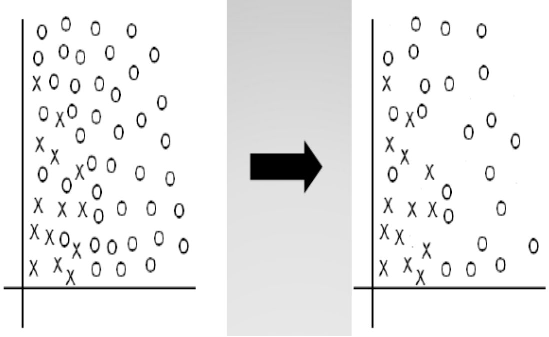
The main problem in imbalanced data is that the majority classes that are represented by large numbers of patterns rule the classifier decision boundaries at the expense of the minority classes that are represented by small numbers of patterns. It is a non heuristic method that aims to balance class distribution through the random elimination of majority class examples. Its major drawback is that it can discard potentially useful data, which could be important for the induction process. In the same way as random oversampling, it tries to balance class distribution, but in

DOI 10.17148/IJARCCE.2016.5812

this case, randomly replicating minority class instances. Several authors agree that this method can increase the likelihood of occurring over fitting, since it makes exact copies of existing instances. In this hybrid method both under sampling and oversampling will be applied for the datasets so as to make it a balance dataset.

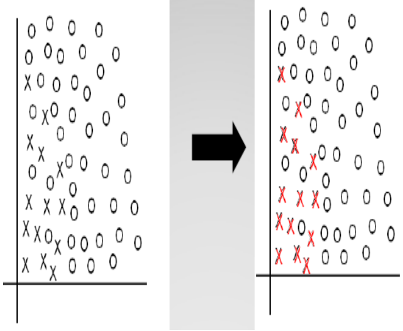
A Support Vector Machine (SVM) performs classification by finding the hyper plane that maximizes the margin between the two classes. The Vectors (Cases) define the hyper plane are the support vectors. Support Vector Machines are based on the concept of decision planes it define decision boundaries. Aditya Tayal [8] et al proposed RankSVM to modify to take advantages of the rare class situation to linear combination of rare class kernel function.These approaches are mainly dividing into three methods such as sampling, algorithms, and feature selection. Sampling techniques used to solve the problems with the distribution of a dataset, sampling techniques involve artificially re-sampling the data set, it also known as data pre-processing method.

Sampling can be achieved by two ways, [13] Under-sampling the majority class, oversampling the minority class, or by combining over and under sampling techniques. Under-sampling techniques used to most important method in under sampling is random under-sampling method which trying to balance the distribution of class by randomly removing majority class sample.



Randomly removes the majority sample.

Random Oversampling methods also help to achieve balance class distribution by replication minority class sample.



Replicate the minority class samples

A several new algorithms have been created for solving the class imbalance problem. The goal of this approach is to optimize the performance of learning algorithm on unseen data. One-class learning methods recognized the sample belongs to that class and reject others. Under certain condition such as multi-dimensional data set one class learning gives better performance than others.

| Techniques | Financial | Banking | business | ATM | Prevent ad fraud | Credit card |
|---|---|---|---|---|---|---|
| Support vector machine(SVM) | 80% | 85% | 95% | 87% | 99% | 100 % |
| Random Forest (RF) | 55% | 40% | 70% | 55% | 20% | 35% |
| Logistic Regression [LR] | 60% | 45% | 10% | 35% | 67% | 50% |
| Pre-Processing | 70% | 58% | 63% | 57% | 87% | 63% |
| Peer-To-Peer (P2p) Lending. | 100 % | 73% | 90% | 95% | 69% | 25% |

To comparison of imbalanced classification for various techniques

## III. CONCLUSION

There are lot of method proposed to solve imbalanced classification problem. Such that data pre-processing method, Bagging, Boosting, CART algorithm, multiclass classification and high dimensional data set. Support vector machine and Synthetic Minority Over-sampling Technique (SMOTE) these techniques also using solving the imbalanced problem.

Pre-processing method produced in the 25 % accurate result. The simulation study could investigate the effect of different methods of handling imbalanced problem. Logistic regression method provides 43 % results in accurately. It is also important to note that the classifiers performance depend on data quality.

Support vector machine (SVM) performs well in predictive data analysis most familiar to solve unbalanced classification. Support vector machine (SVM) will be providing 60 % result. It can also easily understandable. In the future research is to control the unbalanced data or classification using the Support Vector Machine (SVM) technique.

## REFERENCES

[1] Hung-Yi Lin "Efficient classifiers for multi-class classification problems" Decision Support Systems, Volume: 53, Page no: 41073–481, Year: 2012.

[2] L. Breiman, "Bagging predictors" Machine Learning, volume: 4, page no: 123–140, Year: 1996.

[3] Mikel Galar, Alberto Fernandez, Edurne Barrenechea, Humberto Bustince, "A Review on Ensembles for the Class Imbalance Problem: Bagging-, Boosting-, and Hybrid-Based Approaches", IEEE transactions on systems, Page no:1-22.

[4] Bhusan K. Kuntal, Tarini Shankar Ghosh, Sharmila S. Mande" Igloo-Plot: A tool for of multidimensional datasets" genomics, Volume: 103, Page no: 11 – 20, Year: 2014.

[5] Mahendra Sahare and Hitesh Gupta," A Review of Multi-Class Classification for Imbalanced Data", International Journal of Advanced Computer Research, Volume:2,Issue:3 page no:163-168,year:2012.

[6] Clifton Phua, Vincent Lee1, Kate Smith1 and Ross Gayler" A Comprehensive Survey of Data Mining-based Fraud Detection Research"

[7] Mr.Rushi Longadge, Ms. Snehlata S. Dongre, Dr. Latesh Malik "Class Imbalance Problem in Data Mining: Review "International Journal of Computer Science and Network, Volume: 2, Issue: 1, Year: 2013.

[8] Aditya Tayal, Thomas F. Coleman, and Yuying Li" RankRC: Large-ScaleNonlinear Rare Class Ranking" IEEE transactions on knowledge and data engineering, Volume: 27,Page no: 3347 – 3359.Year: 2015.

[9] M.A.H. Farqu ad, Indranil Bose" Pre-processing unbalanced data using support vector machine" Decision Support Systems, Volume: 53, Page no: 226 -233, Year: 2012.

[10] Wei-Jiun Lin and James J.Chen "Class imbalanced classifiers for High-dimensional data" briefings in bioinformatics, Volume: 14, Issue: 1, Page no: 13 – 26, Year: 2012.

[11] Shaza M. Abd Elrahman and Ajith Abraham" A Review of Class Imbalance Problem" Journal of Network and Innovative Computing, ISSN: 2160-2174, Volume:1, Page no: 332-340, Year: 2013.

[12] Ali Mirza Mahmood" Class Imbalance Learning in Data Mining – A Survey" International Journal of Communication Technology for Social Networking Services, Volume: 3, Issue: 2, Page no: 17-36, Year: 2015.

[13] Mr. Rushi Longadge, Ms. Snehlata S. Dongre, and Dr. Latesh Malik" Class Imbalance Problem in Data Mining: Review" International Journal of Computer Science and Network (IJCSN), Volume: 2, Issue: 1, Year: 2013.

[14] Ben Fei and Jinbai Liu" Binary Tree of SVM: A New Fast Multiclass Training and Classification Algorithm" IEEE transactions on neural networks, volume: 2, page no: 696-704, Year:2006.

DOI 10.17148/IJARCCE.2016.5812