INTERNATIONAL JOURNAL OF CURRENT MULTIDISCIPLINARY STUDIES
Available Online at http://www.journalijcms.com
VOL. 3, ISSUE,03, PP.631-632, MARCH, 2017
IJCMS

RESEARCH ARTICLE

SURVEY ON CLASS IMBALANCE PROBLEM USING DATA MINING TECHNIQUES

Mundle R.V[1] and Chaudhari M.S[2]

[1,2]Department of CSE, PBCOE, Nagpur

ARTICLE INFO

Received 10th December, 2016
Received in revised form 8th January, 2017
Accepted 5th February, 2017
Published online 28th March, 2017

Keywords:

Term—Imbalance class distribution, class imbalance problem, undersampling, oversampling.

ABSTRACT

As technology is increasing,the data storage regarding each technology also increases.Many problems occur when large data is supposed to be handled.One such problem in data mining and machine learning techniques is class imbalance.The overall emphasis on how the class imbalance problem is solved using various techniques. Several techniques such as data processing, algorithmic approach and others are used for solving the skewed data problem of class imbalance. Data processing can be further classified into oversampling and undersampling which deals with the majority and minority class samples.These techniques are very useful for solving the class imbalance problem.

INTRODUCTION

Large number of real world applications give rise to data sets with imbalance in between classes.Examples are medical diagnosis, fraud detection, text classification etc.The classification technique usually consider a balanced class for distribution.An unbalanced calss cannot take part in class distribution. Imbalanced class distribution in datasets occur when one class, often the one that is of more interest, that is the positive or minority class, is insufficiently represented.The corresponding literature survey focuses on how class imbalance problem is solved and what are the techniques used for solving those problems.

LITERATURE REVIEW

All paper discus about class imbalance problem which is very common in many domains. The survery leads to how the class imbalance problem can be solved using various algorithms and techniques. The major techniquies such as oversampling and undersampling are broadly discussed.

Rushi Longadge, Snehlata S. Dongre,Dr. Latesh Malik [1] examines that the dataset containing unmatched proportion of samples from one class to the other is termed as skewed dataset. The author states that imbalanced dataset occurs when one class is having more samples than the other classs. Minority and majority class are those which play a major role in class imbalance problem. There are many algorithms and techniques that solve the problem of imbalance distribution of sample. These approaches are mainly divided into three methods such as sampling, algorithms, and feature selection.

Chris Seiffert, Taghi, M. Khoshgoftaar[2] proposed that several techniques have been used to solve the problem of class imbalance, including data sampling and boosting. The author had presented a new hybrid sampling/boosting algorithm, called RUSBoost, for learning from skewed training data. This algorithm provides a simpler and faster alternative to SMOTEBoost, which is another algorithm that combines boosting and data sampling. They had evaluated the performances of RUSBoost and SMOTEBoost, as well as their individual components. RUSBoost as an attractive alternative for improving the classification performance of learners built using imbalanced datasets.

P. Rajeshwari, D. Maheshwari [3] states that 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 they focused on more techniques involve solving rare class or imbalanced problem

Hung-Yi Lin[4] 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.

*✉ ***Corresponding author:*** **Mundle R.V**
Department of CSE, PBCOE, Nagpur

Breiman[5] presented a 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.

Aditya Tayal[6] proposed Rank SVM 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.

Ali Mirza Mahmood [7] Sampling can be achieved by two ways, 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.

Sotiris Kotsiantis, et.al[8] A number of solutions to the class-imbalance problem were previously proposed both at the data and algorithmic levels. At the data level, these solutions include many different forms of re-sampling such as random oversampling with replacement, random undersampling, directed oversampling (in which no new examples are created, but the choice of samples to replace is informed rather than random), directed undersampling, oversampling with informed generation of new samples, and combinations of the above techniques.

N.V.Chawla[9] Oversampling can introduce an additional computational task if the data set is already fairly large but imbalanced. SMOTE generates synthetic minority examples to over-sample the minority class.Its main idea is to form new minority class examples by interpolating between several minority class examples that lie together. For every minority example, its k (which is set to 5 in SMOTE) nearest neighbors of the same class are calculated, then some examples are randomly selected from them according to the over-sampling rate. After that, new synthetic examples are generated along the line between the minority example and its selected nearest neighbors. Thus, the overfitting problem is avoided and causes the decision boundaries for the minority class to spread further into the majority class space.

Aida Ali, Siti Mariyam Shamsuddin, and Anca L. Ralescu [10] states that in class imbalanced classification, the training set for one class (majority) far surpassed the training set of the other class (minority), in which, the minority class is often the more interesting class. In this paper, we review the issues that come with learning from imbalanced class data sets and various problems in class imbalance classification. A survey on existing approaches for handling classification with imbalanced datasets is also presented. They discuss current trends and advancements which potentially could shape the future direction in class imbalance learning and classification.They also found out that the advancement of machine learning techniques would mostly benefit the big data computing in addressing the class imbalance problem which is inevitably presented in many real world applications especially in medicine and social media.

CONCLUSION

In literature survey, research in data mining, techniques for finding the result of imbalance problem is reviewed. Different researchers will overcome problem in Class imbalance, in order to improve the rules applied on large size database. Several techniques of mining such as Algorithms, RUSboost, SMOTE have some advantages and disadvantages hence there is scope for improving the efficiency of mining.

From the study, it is found out that the user can get desired result without wasting the time, money or any other factor.

References

1. Mr.Rushi Longadge, Ms. Snehlata S. Dongre, Dr. Latesh Malik "Class Imbalance Problem in Data Mining: Review", *International Journal of Computer Science and Network* (IJCSN) Volume 2, Issue 1, February 2013.
2. Chris Seiffert, Taghi M. Khoshgoftaar, *M,* Jason Van Hulse, "RUSBoost: A Hybrid Approach toAlleviating Class Imbalance", ieee transactions on systems, man, and cybernetics—part a: systems and humans, vol. 40, no. 1, january 2010.
3. P. Rajeshwari, D. Maheshwari "A Study of Imbalanced Classification Problem" *International Journal of Advanced Research in Computer and Communication Engineering* ISO 3297:2007 Certified Vol. 5, Issue 8, August 2016 .
4. Hung-Yi Lin "Efficient classifiers for multi-class classification problems" Decision Support Systems, Volume: 53, Page no: 41073–481, Year: 2012..
5. L. Breiman, "Bagging predictors" Machine Learning, volume: 4, page no: 123–140, Year: 1996.
6. 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.
7. 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.
8. Sotiris Kotsiantis, Dimitris Kanellopoulos, Panayiotis Pintelas "Handling imbalanced datasets: *A review" GESTS International Transactions on Computer Science and Engineering*, Vol.30, 2006.
9. N.V. Chawla, L. O. Hall, K. W. Bowyer, and W. P. Kegelmeyer. SMOTE: Synthetic Minority Oversampling TEchnique. *Journal of Artificial Intelligence Research*, 16:321-357, 2002.
10. Aida Ali, Siti Mariyam Shamsuddin, and Anca L. Ralescu " Classification with class imbalance problem:A Review" *Int. J. Advance Soft Compu. Appl*, Vol. 7, No. 3, November 2015.