ISSN (Online) 2394-2320

International Journal of Engineering Research in Computer Science and Engineering (IJERCSE)
Vol 5, Issue 3, March 2018

A Systematic Performance of Various Multi Class Imbalance Data Classification in Data Mining

[1] Ganesh.T, [2] Nirmal Kumar.A, [3] Sankara Gomathi.S
[1] Assistant Professor, Department of IT, Vel Tech High Tech Dr.Rangarajan Dr.Sakunthala Engineering College, Chennai.
[2] Associate Professor, Department of IT, Vel Tech High Tech Dr.Rangarajan Dr.Sakunthala Engineering College, Chennai.
[3] Professor, Department of ECE, Adhi College of Engineering and Technology, Kancheepuram

Abstract: **- The usage of data is increasing day by day. There is a huge amount of data storage is required for handling the millions of twits, shares in social networks (twitter, facebook, WhatsApp, and youtube) per second. Databases are playing a vital role in data warehousing and mining. The process of storing the data in large repository place is known as Data Warehousing. Nowadays, Search Engines are struggling to follow the Search Engine Optimization techniques. So there is a pressure for the data analyst to fetch the data from the data warehouse efficiently. The task of classification with imbalanced datasets have attracted quite an interest from researchers in the recent years. Accordingly, various classification techniques are used to handle the newly arrived large amount of data. So many applications have been designed to address this problem from the different perspective such as data pre-processing, algorithm modification and sensitive learning. The problem of constructing fast and accurate classifiers in large data set is an important task in data mining and knowledge discovery. This paper illustrates the various classification techniques and also to improve the correctness of classifier for Classification Techniques in Data Mining.**

Keywords: **Data Mining, Data Warehousing, Classification Techniques, Classifiers, Imbalanced datasets.**

I. INTRODUCTION

If one class has significantly more samples than the others, then a data set is said to be imbalanced. In previous days, has highlighted the imbalanced problem significant interest in real-life applications in different domains such as fraud detection, medical diagnosis and text classifications. For imbalanced data the classification problem is interesting and challenging to researchers because most of the standard data mining methods claim their guess for balanced data and are not applicable for imbalanced one. Research scholars given two set of solutions to data classifications dealing with imbalanced problems: solving in data level by re-sampling, and solving in algorithm level by using design sophisticated classification approaches, where the past one is mainly preferred.

II. RELATED WORK

[1] In recent years multi-class classification field for data imbalancing is front scenario for researchers in the field of machine learning. Various authors propose a technique for the improvement of accuracy and prediction of class in multi-class classification. Some works are summarized here in the form of title and their contribution. Salvador Garcia, Jose Ramon Cano, Alberto Fernandez and Francisco Herrera entitled a method of Prototype Selection for Class Imbalance Problems as [1] classification algorithms is said to be unbalanced when one of the classes is represented by a very small number of cases compared to the other classes when a set of input is provided. In such cases, standard classifiers tend to be flooded by the large classes and ignore the small ones. A number of solutions have been proposed at the data and algorithmic levels. At the data level, we found forms of re-sampling such as over-sampling, where replication of examples or generation of new instances is performed; or under-sampling, where elimination of examples is performed. At the algorithmic level, an adjust of the operation of the algorithm is carried out to treat with unbalanced data.

[2] Zhi-Hua Zhou and Xu-Ying Liu entitled study of Training Cost-Sensitive Neural Networks with Methods Addressing the Class Imbalance Problem as [2] the effect of sampling and threshold-moving in training cost-sensitive neural networks. These techniques modify the distribution of the training data such that the costs of the examples are conveyed explicitly by the appearances of the examples. Threshold moving tries to move the output threshold toward inexpensive classes such that examples with higher costs become harder to be misclassified. In

classical machine learning or data mining settings, the classifiers usually try to minimize the number of errors they will make in dealing with new data.

[3] Piyasak Jeatrakul, Kok Wai Wong, and Chun Che Fung entitled an analysis of Data Cleaning for Classification Using Misclassification [3] as in most classification or function approximation problems; the establishing of an accurate prediction model has always been a challenging problem. When constructing a prediction model, it is always difficult to have an exact function or separation that describes the relationship between the input vector, X and target vector, Y. This paper presents the proposed misclassification technique to increase the confidence of cleaning noisy data used for training. In this paper, we focus our study for classification problem using ANN. The
CMTNN is applied to detect misclassification patterns. For our proposed technique, the training data is cleaned by eliminating the misclassification patterns discovered by both the Truth NN and Falsity NN. After misclassification patterns are removed from the training set, a neural network classifier is trained by using the cleaned data.

[4] Amal S. Ghanem and Svetha Venkatesh, Geoff West entitled problem in Multi-Class Pattern Classification in Imbalanced Data [4] as the majority of multi-class pattern classification techniques are proposed for learning from balanced datasets. However, in several real-world domains, the datasets have imbalanced data distribution, where some classes of data may have few training examples compared for other classes. Despite the success of these techniques reported in different domains for various types of applications, such as text document classification, and speech recognition, most of these techniques are mainly proposed for learning from relatively balanced training data. In this paper we focused on two main challenges in pattern recognition: the imbalanced class problem and multi-class classification. We reviewed the different strategies proposed to solve these two challenges. Based on this research, we outlined a framework that can handle these challenges simultaneously. Our approach (Multi-IM) is based on a relational technique designed for the binary imbalanced problem (PRMs-IM). Multi-IM extends PRMs-IM to a generalized framework for multi-class classification.

III. BASIC CONCEPTS

A. Supervised Machine Learning

Supervised learning is where having input and output variables are X and Y an algorithm is used to learn the mapping function from the input to the output.
Main aim is to estimate the mapping function so well that when you have new input data (x) that you can predict the output variables (Y) for that data. It is said to be supervised learning because the algorithm process learning from the training dataset can be thought of as a teacher supervising the learning process.

B. Learning with class imbalance problem

To learning with class imbalance distribution causes major issues that is most standard algorithms are accuracy driven. We want to minimize the overall error that is tried to maximize the classification accuracy we use and operate many classification algorithms. We want to choose accuracy for performing criterion in class imbalance classification it gives inaccurate and misleading information about a classifier performance. Another concern with imbalanced class learning is that most standard classifiers assumed that the domain application datasets are equal in sight. Many classification algorithms do not take into account the underlying distribution of the datasets thus generate inaccurate model representation in class-learning task.

C. Challenges with class imbalance classification

Class imbalance ensures when there are significantly lesser training examples in one class compared to other class. The nature of class imbalance distribution could occur in two situations.
1) Intrinsic problem related to class imbalance happens naturally. A naturally imbalanced class distribution happens in the case of credit card fraud or in rare disease detection.
2) when the data is not naturally imbalanced, instead it is too expensive to acquire such data for minority class learning due to cost, confidentiality and tremendous effort to find a well-represented data set, like a very rare occurrence of the failure of a space-shuttle. Class imbalance involves a number of difficulties in learning, including imbalanced class distribution, training sample size, class overlapping and small disjuncts.

D. Imbalanced class distribution

In certain domain problems, the imbalance ratio could be as extreme as 1:10000. The study of investigated the correlation between ratio imbalances in training set with the classification results using decision tree classifier, and found out that a relatively balanced distribution between classes in datasets generally gives better results. However, as pointed out by the degree of imbalance class distribution that will start to hinder the classification performance is still not explicitly known. An experiment from the study discovered that a balance distribution among classes is not a guarantee to improve a classifier

performance since a 50:50 population ratio does not always the best distribution to learn from. This suggests that class imbalance distribution is not the only reason that deteriorates a classifier performance, other factors such as training sample size and class complexity also give influence

IV METHODOLOGY

A. SVM AND CLASS IMBALANCE

It gives solutions for balanced data sets; they are sensitive to the imbalance in the datasets and produce sub-optimal models. We proposed several possible reasons as to why SVMs can be sensitive to class imbalance.

1. Weakness of the soft margin optimization problem

SVM's separating hyper plane model developed with an imbalanced dataset can be skewed towards the minority class [8], and this skewness can degrade the performance of that model with respect to the minority class. Remember the objective function of the SVM soft-margin optimization problem. Maximizing the margin mainly focusing on first part of objective function, Minimizing the penalty term associated with the misclassifications of second part, where the regularization parameter C can also be considered as the assigned misclassification cost. Example the same misclassification cost for all the training examples in order to reduce the penalty term, the total number of misclassifications should be reduced. During the density of majority class would be higher than the density of minority class examples even around the class boundary region, that time data set is imbalanced, where the ideal hyper plane would pass through. This is also pointed out in [9], that the low presences of positive examples make them appear further from the ideal class boundary than the negative examples.

B. EXTERNAL IMBALANCE LEARNING METHODS FOR SVMS: DATA PREPROCESSING METHODS

1. Re-sampling methods

This mainly focussing on random and focused oversampling methods and synthetic data generation methods. Resampling methods have been successfully applied to train SVMs with imbalanced datasets in different domains.

This concept focussing first the separating hyper plane found by training an SVM model on the innovative imbalanced dataset is used to select the most helpful examples for a given classification problem, which are the data points lying around the class boundary region. This concept impartial by oversampling as conflicting to blindly oversampling the entire dataset. SVM training time significantly can be reduced while obtaining the comparable classification results to the original oversampling method.

2. Ensemble learning methods

This concept separates majority class dataset into multiple sub datasets such that each of these sub-datasets has a related number of examples as the minority class dataset. This can be done by with the help of random sampling with or without replacement, or clustering methods. SVM classifiers can be developed for each one is to train with the same positive dataset and a different negative sub-dataset. Finally, the decisions made by the classifier ensemble are combined by using a method such as majority voting.

C. INTERNAL IMBALANCE LEARNING METHODS FOR SVMS

1. Different Error Costs (DEC)

The SVM algorithm to be very sensitive to class imbalance would be that the soft margin objective function assigns the same cost C for both positive and negative misclassifications in the penalty term. This would cause the separating hyper plane to be skewed towards the minority class, which would finally yield a suboptimal model. The DEC method is a cost-sensitive learning solution is proposed to overcome this problem in SVMs.

2. One class learning

We have trained an SVM model only with the minority class examples uses first method. Another method, extend the DEC method and to assign a $C- = 0$ misclassification cost for the majority class examples and $C+ = 1/N+$ misclassification cost for minority class examples, where $N+$ is the number of minority class examples. From the experimental results obtained on several heavily imbalanced synthetic and real-world datasets, these methods have been observed to be more effective than general data rebalancing methods.

V. CONCLUSION

This paper presents an overview on class imbalance classification and the inevitable challenges that come with it. It describes the main issues that hinder the classifier performance in managing highly imbalanced datasets and the many factors that contribute to the class imbalance problems. Research gap in previous works are discussed along with justifications to this research attempt. This paper also suggests several potential developments in the domain such as the machine learning for data mining and the boom of sentiment analysis from social media that could fuel the future research direction.

VI. REFERENCES

[1] Kotsiantis, S., D. Kanellopoulos, and P. Pintelas, Handling imbalanced datasets: a review. GESTS International Transactions on Computer Science and Engineering, 2006. Vol 30(No 1): p.25-36.

[2] Yang, Z., et al., Association rule mining-based dissolved gas analysis for fault diagnosis of power transformers. Systems, Man, and Cybernetics, Part C: Applications and Reviews, IEEE Transactions on, 2009. 39(6): p. 597-610.

[3] Zhu, Z.-B. and Z.-H. Song, Fault diagnosis based on imbalance modified kernel Fisher discriminant analysis. Chemical Engineering Research and Design, 2010. 88(8): p. 936-951.

[4] Tavallaee, M., N. Stakhanova, and A.A. Ghorbani, Toward credible evaluation of anomalybased intrusion-detection methods. Systems, Man, and Cybernetics, Part C: Applications and Reviews, IEEE Transactions on, 2010. 40(5): p. 516-524.

[5] T. Hastie, R. Tibshirani. Classification by pairwise coupling. The annals of statistics, 1998, vol. 26, no. 2, pp. 451-471.

[6] R. Rifkin, A. Klautau. In defense of one-vs-all classification. The Journal of Machine Learning Research, 2004, vol. 5, pp. 101-141.

[7] N. Garcia-Pedrajas, D. Ortiz-Boyer. Improving Multiclass Pattern Recognition by the Combination of Two Strategies. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2006, vol. 28, no. 6, pp. 1001-1006.

[8] Gustavo E. A. P. A. Batista, Ronaldo C. Prati and Maria Carolina Monard - A Study of the Behavior of Several Methods for Balancing Machine Learning Training Data‖ in Sigkdd Explorations.