A REVIEW ON DIFFERENT ON DIFFERENT TECHNIQUES FOR BALANCING OF THE IMBALANCED DATA.

[1]Er. Hemlata,[2]Gagandeep Kaur
Department of Computer Engineering
Guru Kashi University
Talwandi Sabo Bathinda,Punjab India.

Department of Computer Engineering
Guru Kashi University
Talwandi Sabo Bathinda,Punjab India.

Abstract: Data mining is performed on to the large repository of the data for generating the relevant data. In current time there are large number of applications where large repository can be used to store large generated data from itself. Like Facebook, Whatsapp. This much large repository of data will be very difficult to store and then later on process to extract useful data. Various data mining techniques are used for this purposes. But the data mining result generating ability will be reduced if the data is imbalanced. Because imbalanced data will be having imbalanced classes. One class can have substantial amount of data and other class can have very few data items. Various researchers are researching on this for having balancing of the data. So that result generation can be optimized. In current review paper is to study different techniques used for balancing of the imbalance classes.

Keywords: Imbalance, classification, Stream, Folding.

I. INTRODUCTION

With the internet age the data and information explosion have resulted in the huge amount of data. Fortunately to gather knowledge from such abundant data there exist data mining techniques. Data mining has been used in various areas like Health care, business intelligence, financial trade analysis, network intrusion detection etc.

General process of knowledge discovery from data involves data cleaning, data integration, data selection, data mining, pattern evaluation and knowledge presentation. Data cleaning, data integration constitute data preprocessing. Here data is processed so that it becomes appropriate for the data mining process. Data mining forms the core part of the knowledge discovery process. There exist various data mining techniques viz. Classification , Clustering, Association rule mining etc.

Classification is one of the important technique of data mining. It involves use of the model built by learning from the historical data to make prediction about the class label of the new data/observations. Formally, it is task of learning a target function f, that maps each attribute set x to a set of predefined class labels y. Classification model learned from historical data is nothing but the target function. It can serve as a tool to distinguish between the objects of different classes as well as to predict class label of unknown records. Fig 1 shows the classification
task which maps attribute set x to its class label y.

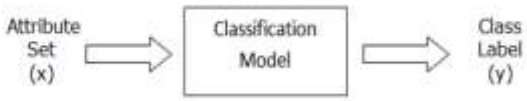

Fig. 1 Classification as a task of mapping input attribute set x into its class label y

1.1 An overview of data streams

Many real world applications, such as network traffic monitoring, credit card transactions, real time surveillance systems, electric power grids, remote sensors, web click streams etc, generate continuously arriving data known as data streams [4]. Unlike the traditional data sets, data streams arrive continuously at varying speeds. Data streams are fast changing, temporally ordered, potentially infinite and massive[8]. It may be impossible to store the entire data stream into memory or to go through it more than once due to its voluminous nature. Thus there is need of single scan, multidimensional, online stream analysis methods. In today's world with data explosion the data is increasing by terabytes and even petabytes, stream data has rightly captured our data mining needs of today. Even though complete set of data can be collected and stored its quite expensive to go through such huge data multiple times.

Data Stream Classification Since classification could help decision making by predicting class labels for given data based on past records, classification on stream data has been extensively studied in recent years with many interesting algorithms developed. Some of them are cited here: [4], Fig

2 depicts the classification model in data streams. As shown in fig. 2 data chunks C1;C2;C3::::Ci arrive one by one.

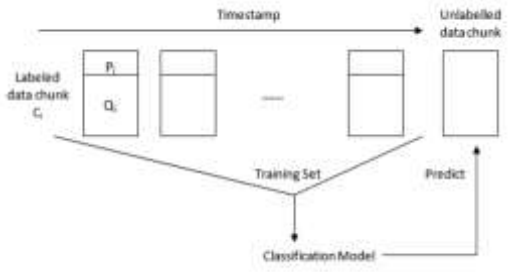

Fig. 2 Classification model in data streams

Each chunk contains positive instances Pi and negative instances Qi. Suppose C1;C2;C3::::Ci are labeled. At the time stamp m + 1, when an unlabelled chunk Cm+1 arrives, the classification model predicts the labels of instances in Cm+1 on basis of previously labeled data chunks. When experts give true class labels of the instances in Cm+1, the chunk can join the training set, resulting in more and more labeled data chunks. Because of storage constraints, it is critical to judiciously select labeled examples that can represent the current distribution well. Most studies on stream mining assume relatively balanced and stable data streams. However, many applications can involve concept-drifting data streams with skewed distributions. In data with skewed distributions each data chunk has many fewer positive instances.

II. LITERATURE SURVEY

Data sampling has received much attention in data mining related to class imbalance problem. Data sampling tries to overcome imbalanced class distributions problem by adding samples to or removing sampling from the data set [2]. This method improves the classification accuracy of minority class but, because of infinite data streams and continuous concept drifting, this method cannot suitable for skewed data stream classification. Most existing imbalance learning techniques are only designed for two class problem. Multiclass imbalance problem mostly solve by using class decomposition. AdaBoost.NC [1-4] is an ensemble learning algorithm that combines the strength of negative correlation learning and boosting method. This algorithm mainly used in multiclass imbalance data set. The results suggest that AdaBoost.NC combined with random oversampling can improve the prediction accuracy on the minority class without losing the overall performance compared to other existing class imbalance learning methods. Wang et al. proposed the classification algorithm for skewed data stream in [2], which shows that clustering sampling outperforms the traditional undersampling, since clustering helps to reserve more useful information. However, the method cannot detect concept drifting. Chris [2] proposed that both sampling and ensemble technique are effective for improving the classification accuracy of skewed data streams. SVM based one-class skewed data streams learning method was proposed in [6], which cannot work with concept drifting. Liu et al. [16] proposed one class data streams algorithm, which follows the single classifier approach and can be used to classify text streams. One of the most common data sampling techniques is Random Under-sampling**.** RUS simply removes examples from the majority class at random until a desired class distribution is achieved. RUSBoost is a new hybrid sampling and boosting algorithm for learning from skewed training data. RUSBoost provides a simpler and faster alternative to SMOTEBoost which is another algorithm that combines boosting and data sampling [2]. RUS decreases the time required to construct a model, which is benefit when creating an ensemble of models that is use in boosting. RUSBoost presents a simpler, faster, and less complex than SMOTEBoost for learning from imbalanced data. SMOTEBoost combines a popular oversampling technique (SMOTE) with AdaBoost, resulting in a hybrid technique that increases the performance of its components. Infinitely imbalanced logistic regression [8] a recently developed classification technique that is named infinitely imbalanced logistic regression (IILR) acknowledges the problem of class imbalance in its formulation. Logistic regression (LR) is a commonly used approach for performing binary classification. It learns a set of parameters, {w0, and w1}, that maximizes the likelihood of the class labels for a given set of training data. When the number of data points belonging to one class far exceeds the number belonging to the other class, the standard LR approach can lead to poor classification performance. Cost-sensitive neural networks use sampling and threshold-moving method [8], this technique modify the distribution of training data such that cost of example calculated based on appearance of example. Threshold moving tries to move the output threshold toward low cost classes such that examples with higher costs become harder to be misclassified. Threshold-moving is a good choice which is effective on all the data sets and can perform cost-sensitive learning even with seriously imbalanced data sets. Boosting SVM [20] in this algorithm, the classifier is produced from the current weight observation. For given instance, class prediction function which is design in terms of kernel function K. Algorithm calculates the G-mean of classifier by applying different weight and generates new set of classifier. The weight is calculated in iteration of boosting algorithm. Finally, G-mean is used for prediction of good classifier from ensemble classifier. SVM boosting algorithm is still unable to handle the issue of imbalance distribution of data. For online classification of data streams with imbalanced class distribution, Lei [7] proposed an incremental LPSVM termed DCIL-IncLPSVM that has robust learning performance under class imbalance. Linear Proximal support vector machines [LPSVM], like decision trees, classic SVM, etc. are originally not design to handle drifting

data streams that exhibit high and varying degrees of class imbalance. Learning from class imbalance data stream, incremental learning algorithm is desirable to pose a capability for dynamic class imbalance learning (DCIL), i.e. learning from data to adjust itself adaptively to handle varied class imbalances. Lei [7] proposes a new incremental learning of wLPSVM for DCIL, where non-stationary imbalanced stream data mining problem is formalized as learning from data chunks of imbalanced class ratio, which are becoming available in an incremental manner. The proposed DCILIncLPSVM updates its weights and LPSVM simultaneously whenever a chunk of data is presented or removed.

III. COMPARATIVE ANALYSIS

S.No.	Algorithm	Advantages	Disadvantages
1	AdaBoost.NC[1]	Improve prediction accuracy of minority	have not considered overall performance of classifier
2	RUSBoost [2]	Simple, faster and less complex than SMOTE Boost algorithm	this technique has not been able to solve the multiclass imbalance problem in the overall context.
3	Infinitely imbalanced logistic regression [6]	This technique is most of the time is used for binary level classification.	the performance of the technique depends upon this issue that how much outlier stands in the data which is currently being delt with.
4	Linear Proximal support vector machines[7]	Handle dynamic class imbalance problem	this technique has not considered the redistribution of the data.
5	BoostingSVM[16]	this technique has improved the SVM based technique for prediction of minority sample in the overall sample context.	Ignore imbalance class distribution.

IV. CONCLUSION

From the study of various research paper based on balancing of the data classes large amount of work is taking place. This is for analysis purpose. So that system of analysis does not fall in short for lack of data in one or more classes. There are various real life applications where this type of data analysis is required like traffic monitoring, credit card transactions, real time surveillance systems, electric power grids, remote sensors web click stream etc. generate continuously arriving data known as data streams. These data stream classification are helpful in decision making purposes. By predicting class labels for giving data based on past records. One of the researcher has worked on the balancing data stream by over sampling technique. Further large amount of work can be under taken to improve upon the work.

V. FUTURE WORK

Various research technique has been studied based in balancing the data stream. So that predicting the classes for the data stream can be possible. On researcher has worked on the technique named as over sampling technique. But further this work can be extended by using over sampling and folding technique.

REFERENCES

[1] Shuo Wang, Member, and Xin Yao, "Multiclass Imbalance Problems: Analysis and Potential Solutions", IEEE
Transactions On Systems, Man, And Cybernetics—Part B: Cybernetics, Vol. 42, No. 4, August 2012.
[2] Chris Seiffert, Taghi M. Khoshgoftaar, Jason Van Hulse, 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.
[3] Björn Waske, Sebastian van der Linden, Jón Atli Benediktsson, Andreas Rabe, and Patrick Hostert
"Sensitivity of Support Vector Machines to Random Feature Selection in Classification of Hyper-spectral Data", IEEE
Transactions On Geosciences And Remote Sensing, Vol. 48, No. 7, July 2010
[4] Xinjian Guo, Yilong Yin1, Cailing Dong, Gongping Yang, Guangtong Zhou,"On the Class Imbalance Problem" Fourth
International Conference on Natural Computation, 2008.

[5] Mike Wasikowski, Member and Xue-wen Chen, "Combating the Small Sample Class Imbalance Problem Using Feature
Selection", IEEE Transactions on Knowledge and Data Engineering, Vol. 22, No. 10, October 2010.
6] Rukshan Batuwita and Vasile Palade,"Fuzzy Support Vector Machines for Class imbalance Learning" IEEE Transactions
On Fuzzy Systems, Vol. 18, No. 3, June 2010.
[7] Lei Zhu, Shaoning Pang, Gang Chen, and Abdolhossein Sarrafzadeh, "Class Imbalance Robust Incremental LPSVM for Data Streams Learning" WCCI 2012 IEEE World Congress on Computational Intelligence June, 10-15,2012 - Australia.
[8] David P. Williams, Member, Vincent Myers, and Miranda Schatten Silvious, "Mine Classification With Imbalanced
Data", IEEE Geosciences And Remote Sensing Letters, Vol. 6, No. 3, July 2009.
[9] Chris Seiffert, Taghi M. Khoshgoftaar, Jason Van Hulse, Amri Napolitano "A Comparative Study of Data Sampling and Cost Sensitive Learning" , IEEE International Conference on Data Mining Workshops. 15-19 Dec. 2008.
[10] Mikel Galar,Fransico, "A review on Ensembles for the class Imbalance Problem: Bagging,Boosting and Hybrid-Based Approaches" IEEE Transactions On Systems, Man, And Cybernetics—Part C: Application And Reviews, Vol.42,No.4 July 2012

earning-2009.

[11] Yuchun Tang, Yan-Qing Zhang, Nitesh V. Chawla, , and Sven Krasser "Correspondence SVMs Modeling for Highly
Imbalanced Classification" IEEE Transactions On Systems, Man, And Cybernetics—Part B: Cybernetics, Vol. 39, No. 1,
February 2009
[12] Peng Liu, Lijun Cai, Yong Wang, Longbo Zhang "Classifying Skewed Data Streams Based on Reusing Data" International Conference on Computer Application and System Modeling (ICCASM 2010).
[13] Zhi-Hua Zhou, Senior Member, and Xu-Ying Liu "Training Cost-Sensitive Neural Networks with Methods Addressing
the Class imbalance Problem" IEEE Transactions On Knowledge And Data Engineering, Vol. 18, No. 1, January 2006
[14] Qun Song Jun Zhang Qian Chi " Assistant Detection of Skewed Data Streams Classification in Cloud Security", IEEE Transaction 2010.
[15] Nadeem Qazi,Kamran Raza, "Effect Of Feature Selection, Synthetic Minority Over-sampling (SMOTE) And Undersampling On Class imbalance Classification", 14th International Conference on Modeling and Simulation-2012.
[16] Benjamin X. Wang and Nathalie Japkowicz "Boosting Support Vector Machines for Imbalanced Data Sets" Proceedings of the 20th International Conference on Machine L