Efficient Clustered Based Oversampling Approach to Solve Rare Class Problem

Mr.Chaitanya D. Sambare[1] Ms.Snehlata S. Dongre[2]
[1]M.Tech Student [2]Assistant Professor
[1,2]Department of Computer Science & Engineering
[1,2]G.H. Raisoni College of Engineering Nagpur, India

Abstract— In data mining, problems are arises when we are applying some data mining techniques to real-life data, which frequently shows rare class problem. Another name of rare class is class imbalance. Class imbalance problem encounters when one class having more samples than other classes. . The minority class data are those data which occur less frequently. Most of the researcher used two techniques to tackle the issue of rare class. Those techniques are the preprocessing techniques which are named as under sampling and oversampling. In under sampling, Important data belongs to mjority has been eliminated. Oversampling is applied to solve rae class issue by replicating the sample, data belongs to minority. Rare class issue leads to misclassification of minority class sample. To overcome this issue, we proposed an efficient clustered based oversampling technique i.e. ECMO which will minimize the problem of mirroring of data associated with the oversampling technique and analyse the performance of ECMO technique.

Key words: Rare Class Problem, Class Imbalance Problem, Under-Sampling, Oversampling

I. INTRODUCTION

Now a days, there are many real life problems occurred which belongs to data mining. Some of these are involved learning a classifier from rare class data., rare class data problem is also known by the name Skewed Data Problem. As the name suggest rare class problem is a problem where rare no sample belonging to one class which known as minority class and large number of samples belonging to other class known as majority class. Minority samples are those samples which are occurred less frequently. Rare class problem are encountered in various application such as banking, medical, oil etc. [2], [3]. In Banking, problems encountered such as sometimes invalid transaction are also consider as valid transaction due to rare class problem which will cause major problem. Rare Class constitutes a problem which will cause difficulty for most of the learning algorithms, which are focusing towards learning and prediction of the majority class sample. Due to such nature of classifier, performance of classifier becomes very poor. The classification process gives best result if numbers of samples in each class having nearly same quantity of data. To make quantity of samples equal in both the classes there are various techniques used. Sampling is one the method which is based on preprocessing of data [16]. There are two pre-processing techniques which are used to balance the dataset [4]. Over-sampling technique adds new data sample to existing dataset, While, under-sampling is to downsize the number of the majority class samples [5] [10]. But it may lose some samples, which will result in dropping the amount of information due to this performance of classifier becomes poor. There are other methods which emphasis on the algorithm based technique, which introduces certain medium to rare class data are balanced and improve the classification. Due to the simplicity, Oversampling and under sampling are used by many researcher to solve rare class issue. But the disadvantage associated under sampling is that it eliminates minority class sample which results in dropping of some crucial information [5]. Disadvantage of undersampling like dropping of crucial information, undersampling is avoided by most of the researchers [6]. Second technique is to apply oversampling, but the dis advantage associated with this technique is that it replicates many unnecessary minority class samples to counter the rare class issue. So, to eliminate this disadvantage we proposed Efficient clustered based over sampling technique.

II. RELATED WORK

Many researches on the rare class issue have been performed. Techniques for solution to rare class issue come under 3 different categories:

- Preprocessing Method
- Ensemble Method
- Combination of Sampling and Ensemble Method.

In Rare class issue, the aim of each and every technique which is used to solve the rare class issue is to balance i.e. to make the count of sample in each class near to equal. This is done because the classification task give more precise result when the number of examples belongs to each class is equal [8], [9]. Data preprocessing technique pre-process the data by performing under-sampling on the majority class instances and performing oversampling on minority class thereby creating a dataset where no of samples belongs to each class are equal, which in turn solve the Rare class Problem. Popular Preprocessing methods are is RUS and ROS. RUS i.e. Random under sampling is one of the simplest types of under-sampling technique which randomly select majority class examples and eliminate it to balance the distribution of data over the class [11] [13]. Fig 1 shows dataset after under sampling.

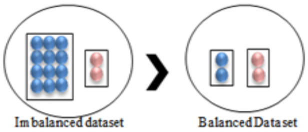


Fig. 1: Randomly removes sample

Over-sampling is another data pre-processing technique which is inverse of under sampling. It balances the distribution of data by duplicating or generating new samples which belongs to minority data. Fig 2 shows distribution of data after applying oversampling.

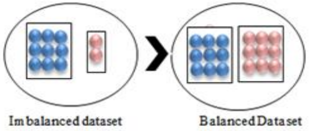


Fig. 2: Reproduce samples

However, both oversampling and under sampling are capable of solving the rare class problem and both of them having their own advantage and disadvantage Comparing oversampling and under resampling, 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 have problem of that it sometimes add unnecessary information.

III. Cluster-Based Efficient Minority Oversampling (ECMO)

Rare class issue is effectively handled or deal by applying Oversampling technique. However the only drawback associated with is that it replicates unnecessary information. To downsize this weakness we propose an approach which will avoid the drawback unnecessary replication of data. In this section, we propose a technique which will identify the samples whose chances of getting miss classification is more and after that generating new minority class samples from this minority class sample. In order to select the minority class samples whose chances of getting miss-classified is more we use KNN clustering. For our cluster-based efficient minority oversampling algorithm, we first divide all samples that is majority and minority in the data set into k=3 different clusters. Let X be the imbalance dataset and Xmin is minority and Xmaj is majority samples.

A. Algorithm:

- Step1: Check whether dataset contains rare class issue or not.
- IF dataset contains rare class data then GOTO Step2 Otherwise EXIT.
- Step2:.Divide the Dataset into 3 clusters.
- Step3: Select the minority whose chances of getting miss classified is from 3rd cluster.
- Step4: Generate new minority class samples from the Xmin selected in 3rd step.
- Step5: Add the samples obtained 4th step to get balanced dataset.

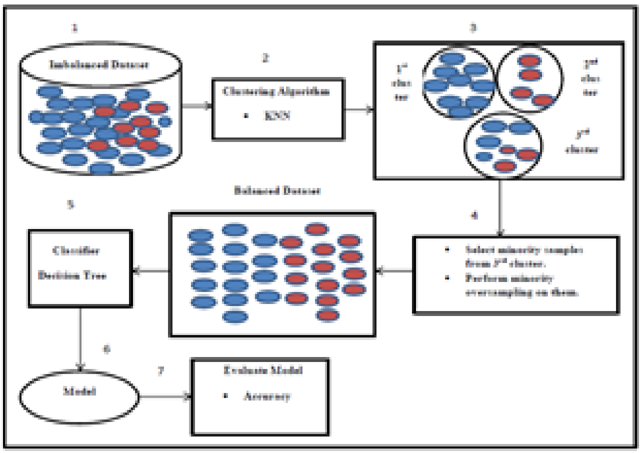


Fig. 3: Workflow of ECMO

Fig below shows the block diagram of Cluster-based efficient minority oversampling technique.

The first step is to check whether the dataset is balanced or not, if dataset is balanced then there is no need to do further process as our aim is to balance the dataset so that classifiers classify it correctly. Then if dataset is imbalanced then it is given as a input to perform clustering on it. For clustering purpose KNN clustering algorithm is used [11]. KNN clustering divides the dataset into 3 different clusters depending upon the distance of each data point or data sample. For clustering purpose value of k =3. Fig 3 shows the process of how we improve the task and how KNN separates the imbalanced data into 3 different clusters and the data after performing oversampling on minority sample of third clusters.

So, Each data sample will check their nearest 3 data sample, if all 3 neighbour are majority data samples then put that data sample into 1st cluster or if all 3 neighbouring samples are minority samples then put that data sample into 2nd cluster or if some neighbouring data sample are minority as well as minority then put that data sample into the 3rd cluster. After that select only the minority class samples form 3rd cluster and perform oversampling on them by generating new data samples. Adds the newly generated sample to original dataset and check whether the data set balanced or not. If dataset is balanced then give it as the input to the classifier. Classifier then create model based on balanced input dataset. After that, finally evaluate the model by test data and calculate performance factor such accuracy, precision etc.

IV. Experiment and Results

Performance of classifier that is the accuracy of classifier is to accurately judge the class of data where it should belong. However, while considering imbalance data problem other evaluation parameter of the classifier are also important beacause of the characterstics of rare class data. In this Paper, other calculating factors of the classifier are also considered such as recall, f-Measure, precision etc. All these are parrameters calculated by using confusion matrix which is shown in table below. Here we take positive class as Majority class and Negative class as minority class.

Class	Positive Class (Predicted)	Negative Class (Predicted)
Positive Class (Actual)	TP	FN
Negative Class (Actual)	FP	TN

Table 1: Fusion Matrix

$$\textbf{Sensitivity} = \textbf{TP} / (\textbf{TP} + \textbf{FN})$$

$$\textbf{Recall} = \textbf{Sensitivity}$$

$$\textbf{Specificity} = (\textbf{FP}) / (\textbf{FP} + \textbf{TN})$$

$$\textbf{Precision} = (\textbf{TP}) / (\textbf{TP} + \textbf{TN})$$

$$\textbf{F-measure} = \frac{2 * \textbf{Recall} * \textbf{Precision}}{\textbf{Recall} + \textbf{Precision}}$$

$$\textbf{G-mean} = \sqrt{\textbf{Sensitivity} * \textbf{Specificity}}$$

Above four calculated parameter used in our experiment to evaluate the improvement done by applying ECMO. Sensitivity is the proportion of positive class samples that are correctly classified among actual positive class samples which is nothing but Recall.

In rare class classification Recall is also one of the important performance measure [8]. For rare class, more the recall lowers the precision rate. F-measure and G-mean are also the important measure used in rare class problem.

UCI Dataset is used by us to apply ECMO technique [17]. The Pie chart below shows the imbalance ratio of the dataset taken for ECMO experiment, the three datasets are used Shuttle, Haberman, Ecoli. Haberman datasets has 3 attribute with imbalance ratio of 23%. Shuttle dataset has 9 attribute with imbalance ratio of 9% and Echoli dataset has 9 attribute and imbalance ratio of 3%.

Dataset Name	Sensitivity	Precision	Specificity	F-measure	G-Mean
Haberman	0.49	1.0	1.0	0.65	0.69
Echoli	0.76	1.0	0	0.86	0
Shuttle	0.87	1.0	1.0	0.93	0.95

Table 4.1: Performance on the dataset before applying ECMO

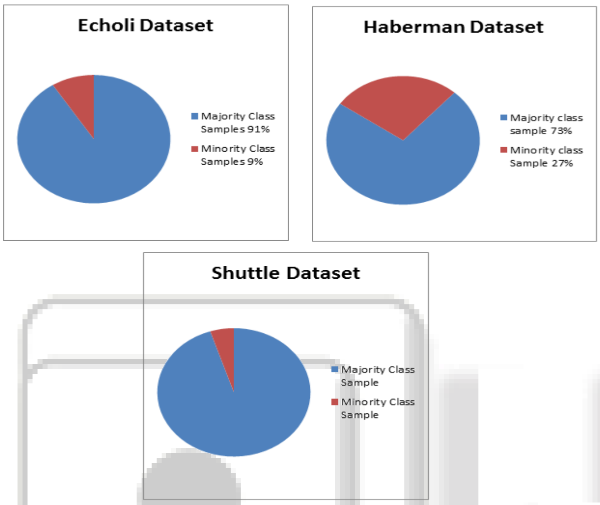


Fig. 4: Imbalance Ratio

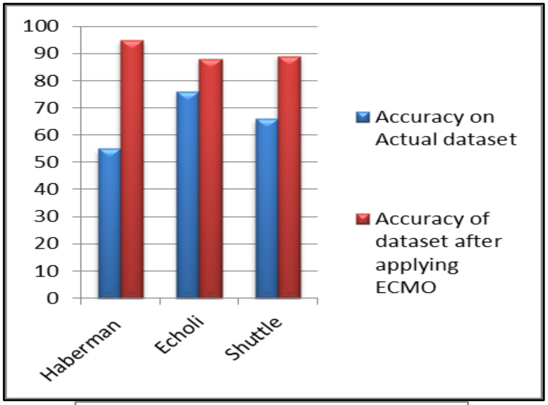
A. Result Analysis

We evaluate our technique on 3 different dataset and for the classification task in our technique we used C4.5 decision tree classifier, as we discussed above the performance meter of classifiers are Sensitivity, Precision, Recall, G-mean, and Specificity [18]. Analytical measures of the performance of classifier are Sensitivity and specificity. True positive rate (TPR) is also called as Sensitivity. Specificity measures the proportion of negatives which are correctly identified. If value of G-mean is close or near to one, it means that TP samples and TN samples are well balances. From this we can conclude that having a higher G-mean by a classifier increases its chances to correctly predict the information that from which class it belongs. Following table2 give the values before applying ECMO technique. Table 2 shows the performance means of a classifier before applying the ECMO and Table 3 shows the performance means of classifier after applying ECMO.

Dataset Name	Sensitivity	Precision	Specificity	F-measure	G-Mean
Haberman	0.86	1.0	1.0	0.92	0.93
Echoli	1.0	0.93	0.8	0.96	0.89
Shuttle	0.83	1.0	1.0	0.91	0.90

Table 4.2: Performance on the dataset after applying ECMO

Fig 4 shows Accuracy graph of the classifier improve after applying Effective Cluster based Minority oversampling technique as compared to performance of the classifier on the original dataset.



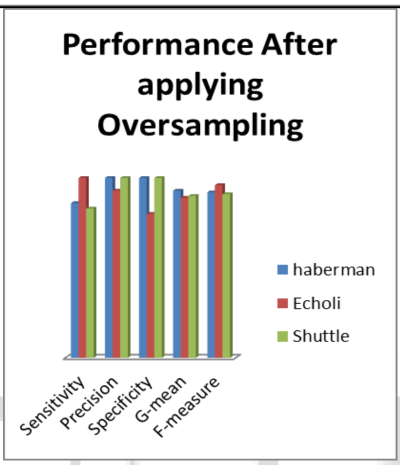


Fig. 5: Performance factor of a classifier after applying ECMO technique

From results it display that capacity of the classifier is boosted after applying Effective Cluster based Minority Oversampling technique (ECMO).

V. CONCLUSION

The results show that ECMO effectively improve the classification rate of rare class dataset. While considering rare class dataset G-mean is important measure of performance. From the result in 4 we see that in all dataset g-mean improved after applying ECMO. F-measure before applying ECMO algorithm on all dataset and after applying ECMO the comparison shows that Performance of C4.5 classifier improve the classification rate of classifier. Hence the performances of classifier improve if data is pre-process by applying ECMO technique then give it to classifier. In future, ECMO technique is applied on the multimedia data and multiclass data.

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] 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.

[3] Gang Wu and Edward Y. Chang, "KBA: Kernel Boundary Alignment Considering Imbalanced Data Distribution", IEEE TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING, VOL. 17, NO. 6, JUNE 2005.

[4] Xinjian Guo, Yilong Yin1, Cailing Dong, Gongping Yang, Guangtong Zhou,"On the Class Imbalance Problem", Fourth International Conference on Natural Computation.

[5] Show-Jane. Yen, and Yue-Shi. Lee,"Under-Sampling Approaches for improving Prediction of the Minority Class in an Imbalanced Dataset", Conference on the Intelligent Control and Automation, Lecture Notes in Control and Information Sciences (LNCIS), Vol.344, August 2006, pp. 731-740.

[6] 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.

[7] QiangWang, "Research Article on A Hybrid Sampling SVM Approach to Imbalanced Data Classification", Hindawi Publishing Corporation Abstract and Applied AnalysisVolume 2014, Article ID 972786, 7 pages,http://dx.doi.org/10.1155/2014/972786

[8] Jia Li, Hui Li *, Jun-Ling Yu, "Application of Random-SMOTE on Imbalanced Data Mining", 2011 Fourth International Conference on Business Intelligence and Financial Engineering

[9] Rushi Longadge, Snehlata S. Dongre, Latesh Malik "Multi-Cluster Based Approach for Skewed Data in Data Mining" E-ISSN: 2278-0661, P- ISSN: 2278-8727volume 12, Issue 6 (Jul. - Aug. 2013), Pp 66-73.

[10] Show-Jane. Yen, and Yue-Shi. Lee, ―Under-Sampling Approaches for improving Prediction of the Minority Class in an ImbalancedDataset,‖ In Proceedings of the Intelligent Control and Automation, Lecture Notes in Control and Information Sciences (LNCIS), Vol.344, August 2006, pp. 731-740.

[11] 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.

[12] N.V. Chawla, A. Lazarevic, L.O. Hall, and K.W. Bowyer, ―SMOTEBoost: Improving Prediction of the Minority Class in Boosting,‖ Proc. Seventh European Conf. Principles and Practice of Knowledge Discovery in Databases, pp. 107-119, 2003 .

[13] 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

[14] N. Japkowicz and S. Stephen, "The Class Imbalance Problem: A Systematic Study", Intelligent Data Analysis, vol. 6, no. 5, pp. 429-449, 2002.

[15] Li Zhang, WenXian Wang, "A Re-sampling Method for Class Imbalance Learning with Credit Data", 2011 International Conference of Information Technology, Computer Engineering and Management Sciences.

[16] Haibo He and Edwardo A,"Garcia ―Learning from Imbalanced Data", IEEE TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING, VOL. 21, NO. 9, SEPTEMBER 2009.

[17] P.M. Murphy and D.W. Aha, "UCI Repository of Machine Learning Databases," Dept. of Information and Computer Science, Univ. of California, Irvine, CA, 1994.

[18] S. R. Safavin and D. Landgrebe, "A survey of decision tree classifier methodology," IEEE Trans. Syst., Man, Cybern., vol. 21, no. 3, pp. 660–674, Jul. 1991.

[19] S. K. Murthy, "Automatic construction of decision trees from data: a multidisciplinary survey," Data Mining Knowl. Disc., vol. 2, no. 4, pp.345–389, 1998.

[20] R. Kohavi and J. R. Quinlan, "Decision-tree discovery," in Handbook of Data Mining and Knowledge Discovery, W. Klosgen and J. M. Zytkow, Eds.London, U.K.: Oxford Univ. Press, 2002, ch. 16.1.3, pp. 267–276.