DOI: http://dx.doi.org/10.26483/ijarcs.v8i7.4530

ISSN No. 0976-5697

Volume 8, No. 7, July – August 2017

International Journal of Advanced Research in Computer Science

RESEARCH PAPER

Available Online at www.ijarcs.info

RARE CLASS PROBLEM IN DATA MINING: REVIEW

Snehlata S. Dongre
PhD Scholar,
Computer Science and Engineering Department,
GHRCE Nagpur, India

Dr. Latesh G. Malik
Associate Professor,
Computer Science and Engineering Department,
Govt. Engineering College, Nagpur, India

Abstract: Class imbalance problem is getting so much attention of researchers now a days. In real life there are number of applications that generates imbalanced data sets. Imbalance nature of data makes classification task difficult. Dealing with these kinds of imbalanced dataset is the one of the biggest challenge in the data mining. Imbalanced dataset means the ratio of positive and negative classes is not balanced. The class that is having more number of samples is known as majority class and the class that is having less number of samples is known as minority class samples. Minority class samples are less but important. In the classification task, most of the times, we are ignoring minority class samples and more concentrating on majority class samples. This leads to good overall accuracy but poor minority class detection rate. Many algorithms have been proposed to deal with the imbalanced data problem but each has its prons and corns. Different techniques used for handling imbalance data are discussed here.

Keywords: class imbalanced problem, skewed data, rare class problem, data mining

I. INTRODUCTION

Imbalanced data is the one of the main issue in the data classification. Class imbalanced problem, Skewed data problem and rare class problems all are same terms and interchangeably used in this paper. Skewed nature of data makes the data classification so difficult. There are number of datasets that suffers the problem of imbalance data like Medical data [1] where the number of normal data is out number than the example of disease but minority samples are more important than the majority data. Misclassification of minority samples has very high risk in the field of Medical Science. Likewise the intrusion detection [2], fault detection [3], anomaly detection [4], detection of fraudulent telephone calls etc. are the other examples of imbalanced datasets. The classification method is having good overall accuracy as it is able to classify the majority class samples properly. But the analysis shows that the minority class samples are misclassified. Though the overall classification accuracy is good but the algorithm is failed to classify the minority class properly which is more important because the less number of examples are not sufficient for training the classification model. So we need techniques that can deal with the imbalance datasets. Also there is a need of the evaluation parameters or techniques that shows the clear classification status of minority samples. Most of the algorithms are not bothering about the distribution of classes in the dataset. So the same dataset will directly use for classification and results bad class detection rate for minority class. Fig No. 1 shows Class imbalanced problem. Where rectangle shows the minority class samples and circle shows the majority class samples. We have drawn a decision boundary to show you the separation of minority and majority classes but it is not easy to classify the minority class efficiently through classifier as it is trained with majority class samples better than the minority class samples.

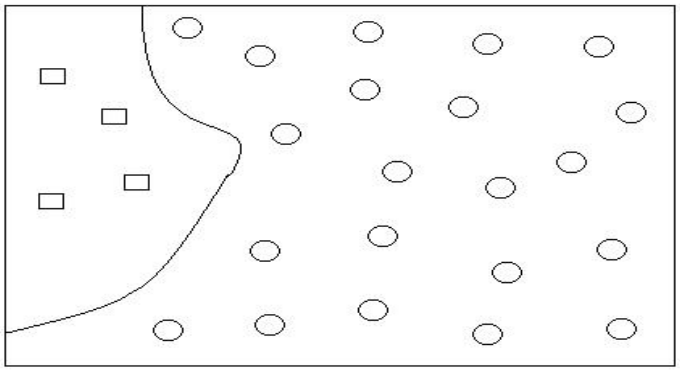

Fig. No.1: Imbalanced dataset

One of the approaches is the data level approach that tries to balance the dataset first and then applying classification. Second approach is algorithmic approach where the especially algorithm is designed to handle the rare class problem. Third approach is the hybrid approach in which the data level approach as well as algorithmic approach combinely used to handle rare class problem.

Rest of the paper is organized as follows, Section II explains various applications of Rare Class problem, section III discussed related work to rare class problem, and the last section concluded this paper.

II. APPLICATIONS

Rare class problems are widely available in the world. There are wide variety of applications in which rare class problem exists. We have enlisted few of them which are rarely known. Researchers are trying to handle this by using data mining techniques in different domains. Application of Rare class problem:-

- **Scene Classification:-** Rong Jin et. al.[7], have discussed ensemble Supper vector machine(SVM) for handling rare class problem in scene classification. They have shown that single SVM classifier is not able to deal with rare class problem whereas the SVM ensemble is performed well. Authors have discussed so many approaches but considered only different SVM approaches only for comparisons.
- **Remote sensing data classification:-** Mine classification[8] is one of the example for remote sensing classification. Mine classification is one of the best example of rare class problem as number of target (mines) are very few than other normal data. Technique named Infinitely Imbalanced logistic Regression (IILR) has been used to solve the rare class problem in mine classification.
- **Predicting Final Grades:-** Now a days educational mining is very much popular. Applying data mining techniques in educational domain raise the standard of education system. Raisul Islam rashu et al.[16] have discussed different resampling approaches like SMOTE, Random over-sampling and Random under-sampling for balancing the imbalanced dataset and improves the prediction rate for final grades prediction.
- **Customer Behavior Prediction:-** Most of the real world applications suffers by a problem known as imbalanced problem. Nengbao Liu et al. [18] have discussed different rare class balancing techniques for customer Behavior Prediction.

III. RELATED WORK

Many approaches are mentioned in the literature to address the rare class problem. Most of the authors have classified these approaches in mainly two categories [5]- 1) Sampling based solution and 2) Algorithmic based solution. Few authors also consider the 3) Feature selection as the solution for rare class problem. But we believe that there are three main categories of approaches to handle the rare class problem-1) Data level approach- here the main concentration is to balance the data set and after that classify the data. Normal classifier even gives the good results as the dataset is balanced. This mainly uses two techniques- Undersampling [5][6] majority samples and Oversampling [5][6] minority samples. In undersampling, majority class samples are removed to balance dataset whereas oversampling adds few samples to balance it. Each has its prons and cons. Best is to use the combination of oversampling & undersampling. 2) Algorithmic approaches, in which algorithms are designed such that they can handle rare class problem, this category includes single classifier and ensemble classifier and 3) Hybrid Approach, which is the combination of data level approach and algorithmic approach. Most of the strategies fall under this category as many strategies uses data sampling approach for balancing the data as well as classifier that have the capability to handle rare class problem. So this is combination of data sampling and classifier both.

A. Data level Approach

Synthetic minority oversampling technique (SMOTE) [21] is a very popular approach that synthetically generates the minority samples to balance dataset. It uses the oversampling technique. Modified synthetic minority oversampling technique (MSMOTE) [22] is the improvement in the SMOTE. It is quite similar as SMOTE the difference is in selection of Nearest Neighbor Synthetic minority oversampling technique (SMOTE) [21] is a very popular approach that synthetically generates the minority samples to balance dataset. It uses the oversampling technique. Modified synthetic minority oversampling technique (MSMOTE) [22] is the improvement in the SMOTE. It is quite similar as SMOTE the difference is in selection of Nearest Neighbor for the generation of synthetic minority samples. W. Jindaluang et. al. [9] have discussed under-sampling using cluster based approach to balance data. Majority class samples are clustered and prominent samples are selected as the majority samples. These prominent samples are merged with minority samples to create balanced dataset. Authors have concluded that their approach is good for imbalanced dataset. Though they have not justified that why they have used 5 fold method and 5 nearest neighbor approach? Why they consider value as 5. B. Das et al. [10] have suggested the oversampling approaches based on probably distribution called as RACOG and wRACOG. These approaches generate synthetic minority samples to balance dataset. RACOG runs for fixed number of iterations and generate the samples whereas wRACOG is going to add samples and observed the performance. It stops when performance has no more improvement. Authors have shown through results that RACOG and wRACOG is one of the big steps in handling class imbalanced. Scalable Instance Selection (OLIGOIS) method [20] is for class imbalanced datasets. In this approach dataset is divided in different parts and in each partition they have applied instance selection. All results are combined to create the final balanced dataset by using some voting technique. This approach has the property of scalability.

B. Algorithmic Approach

Xiaowan Zhang et al. [11] have proposed Cost Free Learning strategy for handling the class imbalanced problem. They said that there are mainly two categories Cost Free Learning (CFL) and Cost Sensitive Learning (CSL). Mutual information is used here. But this strategy adds additional computational cost to different approaches. He He et. al [19] uses the SVM for handling class imbalanced problem. Authors have suggested two modifications in SVM so that it can deal with data imbalanced problem. Yubin Park et al [23] have proposed ensemble of decision trees for handling class imbalance problem. Properties α-divergence is used for this purpose. Peng Wang et al [24] have discussed the concept of granularity for classifier and by using this concept they have proposed a low granularity classifier that can deal with concept drift as well as class imbalance problem. Cost-sensitive XCS Classifier System [25] is another approach that handles the class imbalanced problem. This uses the rewards for correctly classified samples. The correctly classified samples of minority class outweigh the majority correctly classified samples. They proposed that by proper reward setting one can cope up with the data imbalance problem. XCS classifier is used here.

C. Hybrid Approach

BalancedBoost Technique is a Hybrid approach that uses data level approach along with the algorithmic approach proposed by H. Wei et al. [12]. It uses feature selection method named weighted symmetrical uncertainty with ensemble algorithm BalancedBoost to deal with the rare class problem. This approach gives good results for network traffic

data. S. Wang et al. [13][14] have proposed two approaches Oversamplingbased Online Bagging(OOB) and Undersamplingbased Online Bagging(UOB). Both are based on bagging ensemble classifier approach. Performed good for imbalanced dataset but these have not bothered about the imbalanced ratio even the resampling rate is not at all correlated with imbalanced rate. OOB and UOB have been analyzed by the authors and improved to overcome its disadvantages. Weighted Ensembles of OOB & UOB is denoted as WEOB1 [14] and WEOB2 [14]. These approaches give better performance as compared to the OOB and UOB. But it cannot work for multiclass problem. C. Seiffert et al. [15] have proposed hybrid approach named as RUSBoost of data sampling with boosting which is a combination of data level approach and algorithmic approach. RUSBoost uses the Random undersampling Technique with Boosting techniqe for handling data imbalanced problem. Random undersampling deletes the instances randomly from the dataset until it balance the dataset. RUSBoost is reducing training time also increasing the performance. But it is using under sampling technique, so it deletes few important data from the dataset that may play vital role in the data classification. Hybrid sampling SVM [17] uses oversampling & undersampling both techniques with SVM to balance the dataset first and classify. This technique uses SVM hyperplane boundary to identify less important majority class samples that needs to be undersampled and most important minority class samples that needs to be oversampled to balance the dataset. But this approach has considered two class problems only.

IV. CONCLUSSION

Rare class problem is very common in the real world applications. The minority class samples are very less in number, but it very important to classify them correctly. In this paper few applications have been discussed that have rare class problem and different approaches that are used to handle it. Different approaches are reviewed here. We have broadly classified different approaches in three main categories. Data level approach, Algorithmic approach and Hybrid approach which is the combination of above two approaches. Data level approach is important as it includes all resampling approaches like undersampling, oversampling to balance dataset. Algorithmic approach also works well for rare class problem. Hybrid approach, most of the techniques fall under this category as it involves sampling techniques as well as algorithms that can handle rare class problem. Hybrid approach is the best solution for rare class problem.

V. REFERENCES

[1] M. Mazurowski, P. Habas, J. Zurada, J Lo, J. Baker and G. Tourassiet, "Training neural network classifiers for medical decision making: The effects of imbalanced datasets on classification performance," Neural networks Vol. 21, No. 2, pp 427-436, 2008.

[2] S. Dongre, K. Wankhade, "Intrusion Detection System Using New Ensemble Boosting Approach," International Journal of Modeling and Optimization, Vol. 2, No. 4, pp 488-492. August 2012

[3] Z. yang, W. Tang, A. shintemirov, and Q. wu, "Association rule miningbased dissolved gas analysis for fault diagnosis of power transformers," IEEE Transaction of Sytem, Man, Cybernatics. C, Appl. Rev., vol. 39, no. 6, pp. 597-610, 2009.

[4] W. Khreich, E. Granger, A. Miri, R. Sabourin, "Iterative Boolean combination of classifiers in the ROC space: An application to anomaly detection with HMMs," Pattern Recognition Vol. 43, No. 8, pp 2732-2752, 2010.

[5] R. Longadge, S. Dongre, L. Malik, "Class Imbalance Problem in Data Mining: Review," International Journal of Computer Science and Network (IJCSN) Volume 2, No 1, February 2013.

[6] T. Lakshi, Ch. Prasad, "A study on classifying Imbalanced Datasets," Proc. International Conference on Networks & Soft Computing. 2014

[7] R. Jin, A. Hauptmann, "On predicting rare classes with SVM Ensemble in scene classification," Proc. IEEE International conference on Acoustics, speech and signal processing (ICASSP-2003). PP: 21-24. 2003

[8] D. Williams, "Mine classification with Imbalanced Data," IEEE Geoscience & remote sensing letters. Vol. 6, No 3. July 2009.

[9] W. Jindaluang and V. Chauvatut, "Under-sampling by algorithm with performance Guaranteed for Class-imbalanced problem," Proc. IEEE International Computer Science and Engineering Conference (ICSEC 2014). PP 215-221

[10] B. Das, N. Krishnan and D. Cook, "RACOG and wRACOG: Two Probabilistic Oversampling Techniques," IEEE Transactions on Knowledge and Data Engineering. Vol. 27, No. 1, January 2015. PP 222-234

[11] X. Zhang and B. Hu, "A New Strategy of Cost-Free learning in the Class Imbalance Problem'" IEEE Transactions on Knowledge and Data Engineering. Vol. 25, No. 12, pp 2872-2885, December 2014

[12] H. Wei, B. Sun and M. Jing, "BlancedBoost: A Hybrid Approach for Real-time Network Traffic Classification," Proc. IEEE International Conference on Computer Communication and Networks (ICCCN), 2014

[13] S. Wang, L. Minku and X. Yao, "A Learning Framework for online class imbalance Learning," Proc. IEEE symposium Computational Intelligence and Ensemble Learning. pp 36-45, 2013.

[14] S. Wang, L. Minku and X. Yao, "Resampling-Based Ensemble Methods for Online Class Imbalanced Learning," IEEE Transactions on Knowledge and Data Engineering. Vol 27, No 5, May 2015. PP 1356-1368

[15] C. Seiffert, T. Khoshgoftaar, J. Hulse and A. Napolitano. "RUSBoost: A Hybrid Approach to Alleviating Class imbalance," IEEE Transactions on Systems, Man, and Cybernatics-Part A: Systems and Humans, Vol 40, No 1, Jan 2010.

[16] R. Rashu, N. Haq and R. Rahman, "Data Mining Approches to predict Final Grade by Overcoming Class imbalance Problem," Proc. International Conference on Computer and Information Technology (ICCIT) IEEE 2014

[17] Q. Wang, "A Hybrid Sampling SVM Approach to Imbalanced Data Classification," Research Article Hindawi Publishing Corporation, Abstract and Applied Analysis, Vol 2014.

[18] N. Liu, W. Woon and Z. Afshari, "Handling Class imbalance in Customer Behaviour Prediction," Proc. International Conference on Collaboration Technologies and Systems IEEE 2014. pp 100-103

[19] H. He and A. Ghodsi, "Rare class Classification by Support Vector Machine," Proc. International Conference on Pattern Recognition IEEE, pp 548-551, 2010

[20] N. Pedrajas, J. Rodríguez, and A. García, "OligoIS: Scalable Instance Selection for Class-Imbalanced Data Sets," IEEE Transactions On Cybernetics, Vol. 43, No. 1, February 2013

[21] N. Chawla, K. Bowyer, L. Hall, and W. Kegelmeyer, "SMOTE: synthetic minority over-sampling technique," Journal of Artificial Intelligence Research, vol. 16, pp. 321–357, 2002.

[22] S. Hu, Y. Liang, L. Ma, and Y. He, "MSMOTE: Improving classification performance when training data is imbalanced,"

Proc. International Workshop Computer Science and Engineering, vol. 2, pp. 13–17, 2009

[23] Y. Park and J. Ghosh, "Ensembles of α-Trees for Imbalanced Classification Problems" IEEE Transactions on Knowledge and Data Engineering, Vol. 26, No. 1, pp 131-143, 2014

[24] P. Wang, H. Wang, X. Wu, W. Wang, and B. Shi, "A Low-Granularity Classifier for Data Streams with Concept Drifts and Biased Class Distribution," IEEE Transactions on Knowledge and Data Engineering, Vol. 19, No. 9, pp 1202-1213, 2007

[25] N. Thach, P. Rojanavasu and O. Pinngern, "Cost-sensitive XCS Classifier System Addressing Imbalance Problems," Proc. International Conference on Fuzzy Systems and Knowledge Discovery IEEE. pp 132-136, 2008