International Journal of Trend in Research and Development, Volume 6(4), ISSN: 2394-9333
www.ijtrd.com

Analysis of Imbalance Classification Problem: An Assessment

[1]Nitesh Kumar and [2]Dr. Shailja Sharma,
[1]PG Scholar, [2]Associate Professor,
[1,2]Department of Computer Science and Engineering, Rabindranath Tagore University, Bhopal, India

Abstract: In last few years there are major changes and evolution has been done on classification of data. As the application area of technology is increases the size of data also increases. Classification of data becomes difficult because of unbounded size and imbalance nature 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. There are different methods available for classification of imbalance data set which is divided into three main categories, the algorithmic approach, data preprocessing approach and feature selection approach. Each of this technique has their own advantages and disadvantages. In this paper systematic study of each approach is define which gives the right direction for research in class imbalance problem.

Keywords: *Class imbalance problem, Skewed data, Imbalance data, rare class mining*

I. INTRODUCTION

In many real time applications large amount of data is generated with skewed distribution. A data set said to be highly skewed if sample from one class is in higher number than other [1] [16]. In imbalance data set the class having more number of instances is called as major class while the one having relatively less number of instances are called as minor class [16]. Applications such as medical diagnosis prediction of rare but important disease is very important than regular treatment. Similar situations are observed in other areas, such as detecting fraud in banking operations, detecting network intrusions [10], managing risk and predicting failures of technical equipment. In such situation most of the classifier are biased towards the major classes and hence show very poor classification rates on minor classes. It is also possible that classifier predicts everything as major class and ignores the minor class. various techniques have been proposed to solve the problems associated with class imbalance [9], which divided into three basic categories, the algorithmic approach, data-preprocessing and feature selection approach. In data-preprocessing technique sampling is applied on data in which either new samples are added or existing samples are removed. Process of adding new sample in existing is known as over-sampling and process of removing a sample known as under-sampling. Second method for solving class imbalance problem is creating or modifying algorithm. The algorithms include the cost sensitive method and recognition-based approaches, kernel-based learning, such as support vector machine (SVM) and radial basis function [16]. Applying an algorithm alone is not good idea because size of data and class imbalance ratio is high and hence a new technique i.e. the combination of sampling method with algorithm is used [12]. In classification, algorithm generally gives more important to correctly classify the majority class samples. In many applications miss-classifying a rare event can be result in more serious problem than common event [11]. For example in medical diagnosis in case of cancerous cell detection, misclassifying non-cancerous cells may leads to some additional clinical testing but misclassifying cancerous cells leads to very serious health risks. However in classification problems with imbalanced data, the minority class examples are more likely to be misclassified than the majority class examples, due to their design principles, most of the machine learning algorithms optimizes the overall classification accuracy which results in misclassification minority classes. The paper is organized as follows: section 2 contains current approaches which gives basic techniques that used to solve the problem of imbalance dataset. Section 3 gives the review of related work that handle class imbalance problem. Section 4 gives comparative study of some algorithm and finally end with concluding conclusion in Section 5.

II. EXISTING TECHNIQUES

The literature survey suggests many algorithm and techniques that solve the problem of imbalance distribution of sample. These approaches are mainly dividing into three methods such as sampling, algorithms, and feature selection.

1. Sampling

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 preprocessing method. 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: The 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. Figure 1 show the random under sampling method [4]. The problem with this method is loss of valuable information.

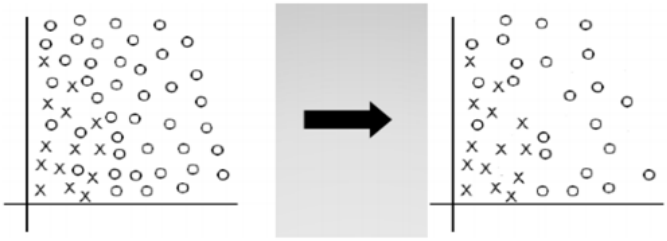

Figure 1: Randomly removes the majority sample

Over-sampling: Random Oversampling methods also help to achieve balance class distribution by replication minority class sample.

Figure 2: Replicate the minority class samples

There is no need to add extra information, it reuse the data [12]. This problem can be solving by generating new synthetic data of minority sample. SMOTE generates synthetic minority examples to over-sample the minority class. In this method learning process consume more time because original data set contain very small number of minority samples.

2. Algorithms

A several new algorithms have been created for solving the class imbalance problem. The goal of this approach is to optimize the performance of learning algorithm on unseen data. One-class learning methods recognized the sample belongs to that class and reject others. Under certain condition such as multi-dimensional data set one class learning gives better performance than others [5]. Instated of changing class distribution applying cost in decision making is another away to improve the performance of classifier. Cost-sensitive learning methods try to maximize a loss function associated with a data set. These learning methods are motivated by the finding that most real-world applications do not have uniform costs for misclassifications. The actual costs associated with each kind of error are unknown typically, so these methods need to determine the cost matrix based on the data and apply that to the learning stage. A closely related idea to cost-sensitive learners is shifting the bias of a machine to favor the minority class [8]. The goal of cost sensitive classification is to minimize the cost of misclassification, which can be realized by choosing the class with the minimum conditional risk. Table 1 gives the cost matrix which contains two classes I & j. λ_{ij} cost of misclassification. Diagonal element are Zero indicate that cost of correct classification has no cost. Another algorithmic approach for skewed distribution of data is modifying the classifier [8]. Kernel based approach borrows the idea of support vector machines to map the imbalanced dataset into a higher dimension space. Then by combining with oversampling technique or ensemble method, the classifier is supposedly to perform much better than learning from the original dataset.

Table 1: Cost matrix

		Prediction	
		Class i	Class j
True	Class i	0	λ_{ij}
	Class j	λ_{ji}	0

In terms of SVMs, several changes have been made to improve their class prediction accuracy and result suggests that SVM have ability to solve the problem of skewed vector without introducing noise [9]. Boosting methods can be combined with SVMs very effectively in the presence of imbalanced data [16].

3. Feature Selection

The goal of feature selection, in general, is to select a subset of j features that allows a classifier to reach optimal performance, where j is a user-defined parameter. For high-dimensional data sets, it uses filters that score each feature independently based on a rule. Feature selection is a key step for many machine learning algorithms, especially when the data is high-dimensional. Because the class imbalance problem is commonly accompanied by the issue of high dimensionality of the data set, hence applying feature selection techniques is essential. Sampling techniques and algorithmic methods may not be enough to solve high dimensional class imbalance problems [5]. Feature selection as a general part of machine learning and data mining algorithms has been thoroughly researched, but its importance to resolving the class imbalance problem is a recent development with most research appearing in the previous several years [18]. In this time period, a number of researchers have conducted research on using feature selection to combat the class imbalance problem. Ertekin [17] studied the performance of feature selection metrics in classifying text data drawn from the Yahoo Web hierarchy. They applied nine different metrics to the data set and measured the power of the best features using the naive Bayes classifier.

III. LITERATURE SURVEY

Miho Ohsaki et. al, 2017[1], There have been many attempts to classify imbalanced data, since this classification is critical in a wide variety of applications related to the detection of anomalies, failures, and risks. Many conventional methods, which can be categorized into sampling, cost-sensitive, or ensemble, include heuristic and task dependent processes. In order to achieve a better classification performance by formulation without heuristics and task dependence, we propose confusion-matrix-based kernel logistic regression (CM-KLOGR). Its objective function is the harmonic mean of various evaluation criteria derived from a confusion matrix, such criteria as sensitivity, positive predictive value, and others for negatives.

[2] Alberto Fernández et. al, 2017[2], BigData applications are emerging during the last years, and researchers from many disciplines are aware of the high advantages related to the knowledge extraction from this type of problem. However, traditional learning approaches cannot be directly applied due to scalability issues. To overcome this issue, the MapReduce framework has arisen as a “de facto” solution. Basically, it carries out a “divide-and conquer” distributed procedure in a fault-tolerant way to adapt for commodity hardware. Being still a recent discipline, few researches has been conducted on imbalanced classification for Big Data.

Vaibhav et. al, 2014[3], classification of the data collected from students of polytechnic institute has been discussed. This data is pre-processed to remove unwanted and less meaningful attributes. These students are then classified into different categories like brilliant, average, weak using decision tree and naïve Bayesian algorithms. The processing is done using WEKA data mining tool. This paper also compares results of classification with respect to different performance parameters.

Kaile Su et. al, 2014[4], Rough set theory provides a useful mathematical concept to draw useful decisions from real life data involving vagueness, uncertainty and impreciseness and is therefore applied successfully in the field of pattern recognition, machine learning and knowledge discovery. This paper presents an overview of basic concepts of rough set theory. The paper also surveys applications of rough sets in feature selection and classification.

Senzhang Wang et. al, 2012[5], Re-sampling method is a popular and effective technique to imbalanced learning. However, most re-sampling methods ignore data density information and may lead to over fitting. A novel adaptive over-sampling technique based on data density (ASMOBD) is proposed in this paper. Compared with existing re-sampling algorithms, ASMOBD can adaptively synthesize different number of new samples around each minority sample according to its level of learning difficulty. Therefore, this method makes the decision region more specific and can eliminate noise. What's more, to avoid over generalization, two smoothing methods are proposed.

IV. COMPARATIVE STUDY

Analysis drawn from comparative study of each of the algorithm is shown in following table.

Table 2: Comparative Study

Authors	Title	Publication	Methodology	Outcomes
Miho ohsakil	Confusion matrix based kernel logistic regression for imbalanced data classification	IEEE transactions on knowledge and data engineering2017	Confusion matrix based kernel logistic regression	Inconsistency exist
Alberto Fernandez	An insight into imbalanced big data classification outcome and challenges	Springer journal of big data	Map reduce	Low precision
Vaibhav	Classification and performance evaluation using data mining algorithms	International journal of innovative research in science engineering and technology	Decision tree and naïve Bayesian	High complexity

Many areas are affected by class imbalance problems. The solution provided by many techniques in data mining is helpful but not enough. The consideration of which technique is best for handling a problem of data distribution is highly depends upon the nature of data used for experiment.

CONCLUSIONS

Practically, it is reported that data preprocessing provide better solution than other methods because it allow adding new information or deleting the redundant information, which helps to balance the data. Another method that helpful to solve the problem of class imbalance is boosting. Boosting is powerful ensemble learning algorithm that improved the performance of weak classifier. The algorithm such as RUSBoost, SMOTEBoost is an example of boosting algorithm. Feature selection method can also used for classification of imbalance data. The performance of a feature selection algorithm depends on the nature of the problem. Finally, this paper suggests that applying two or more technique i.e. hybrid approach gives better solution for class imbalance problem.

References

[1] Miho Ohsaki, Peng Wang, Kenji Matsuda, Shigeru Katagiri, Hideyuki Watanabe, and Anca Ralescu, "Confusion-matrix-based Kernel Logistic Regression for Imbalanced Data Classification", IEEE Transactions on Knowledge and Data Engineering, 2017.

[2] Alberto Fernández, Sara del Río, Nitesh V. Chawla, Francisco Herrera, "An insight into imbalanced Big Data classification: outcomes and challenges", Springer journal of bigdata, 2017.

[3] Vaibhav P. Vasani1, Rajendra D. Gawali, "Classification and performance evaluation using data mining algorithms", International Journal of Innovative Research in Science, Engineering and Technology, 2014.

[4] Kaile Su, Huijing Huang, Xindong Wu, Shichao Zhang, "Rough Sets for Feature Selection and Classification: An Overview with Applications", International Journal of Recent Technology and Engineering (IJRTE) ISSN: 2277-3878, Volume-3, Issue-5, November 2014.

[5] Senzhang Wang, Zhoujun Li, Wenhan Chao and Qinghua Cao, "Applying Adaptive Over-sampling Technique Based on Data Density and Cost-Sensitive SVM to Imbalanced Learning", IEEE World Congress on Computational Intelligence June, 2012.

[6] Mikel Galar, Alberto Fernandez, Edurne Barrenechea, Humberto Bustince and Francisco Herrera, "A Review on Ensembles for the Class Imbalance Problem: Bagging, Boosting, and Hybrid-Based Approaches", IEEE Transactions on Systems, Man and Cybernetics—Part C: Applications and Reviews, Vol. 42, No. 4, July 2012.

[7] Nada M. A. Al Salami, "Mining High Speed Data Streams". UbiCC Journal, 2011.

[8] Dian Palupi Rini, Siti Mariyam Shamsuddin and Siti Sophiyati, "Particle Swarm Optimization: Technique, System and Challenges", International Journal of Computer Applications (0975 – 8887) Volume 14– No.1, January 2011.

[9] Amit Saxena, Leeladhar Kumar Gavel, Madan Madhaw Shrivas, "Online Streaming Feature Selection", 27th International Conference on Machine Learning, 2010.

[10] Yuchun Tang, Member, Yan-Qing Zhang, Nitesh V. Chawla and Sven Krasser, "SVMs Modeling for Highly Imbalanced Classification", IEEE Transaction on Systems, Man and Cybernetics, Vol. 39, NO. 1, Feb 2009.

[11] Haibo He and Edwardo A. Garcia, "Learning from Imbalanced Data", IEEE Transactions on Knowledge and Data Engineering, September 2009.

[12] Thair Nu Phyu, "Survey of Classification Techniques in Data Mining", International Multi Conference of Engineers and Computer Scientists, IMECS 2009, March, 2009.

[13] Haibo He, Yang Bai, Edwardo A. Garcia and Shutao Li, "ADASYN: Adaptive Synthetic Sampling Approach for Imbalanced Learning", IEEE Transaction of Data Mining, 2009.

[14] Swagatam Das, Ajith Abraham and Amit Konar, "Particle Swarm Optimization and Differential Evolution Algorithms: Technical Analysis, Applications and Hybridization Perspectives", Springer journal on knowledge engineering, 2008.

[15] "A logical framework for identifying quality knowledge from different data sources", International Conference on Decision Support Systems, 2006.

[16] "Database classification for multi-database mining", International Conference on Decision Support Systems, 2005.

[17] Volker Roth, "Probabilistic Discriminative Kernel Classifiers for Multi-class Problems", Springer-Verlag journal, 2001.

[18] R. Chen, K. Sivakumar and H. Kargupta "Collective Mining of Bayesian Networks from Distributed Heterogeneous Data", Kluwer Academic Publishers, 2001.

[19] Shigeru Katagiri, Biing-Hwang Juang and Chin-Hui Lee, "Pattern Recognition Using a Family of Design Algorithms Based Upon the Generalized Probabilistic Descent Method", IEEE Journal of Data Minig, 1998.

[20] I. Katakis, G. Tsoumakas, and I. Vlahavas. Tracking recurring contexts using ensemble classifiers: an application to email filtering. Knowledge and Information Systems, Pp 371–391, 2010.

[21] J. Kolter and M. Maloof. Using additive expert ensembles to cope with concept drift. In Proc. ICML, Pp 449–456, 2005.

[22] D. D. Lewis, Y. Yang, T. Rose, and F. Li. Rcv1: A new benchmark collection for text categorization research. Journal of Machine Learning Research, Pp 361–397, 2004.

[23] X. Li, P. S. Yu, B. Liu, and S.-K. Ng. Positive unlabeled learning for data stream classification. In Proc. SDM, Pp 257–268, 2009.

[24] M. M. Masud, Q. Chen, J. Gao, L. Khan, J. Han, and B. M.Thuraisingham. Classification and novel class detection of data streams in a dynamic feature space. In Proc. ECML PKDD, volume II, Pp 337–352, 2010.

[25] P. Zhang, X. Zhu, J. Tan, and L. Guo, "Classifier and Cluster Ensembles for Mining Concept Drifting Data Streams," Proc. 10th Int'l Conf. Data Mining, 2010.