IJCRT.ORG ISSN : 2320-2882

INTERNATIONAL JOURNAL OF CREATIVE RESEARCH THOUGHTS (IJCRT)

An International Open Access, Peer-reviewed, Refereed Journal

Managing Class Imbalance Problems in Class Level and Data Level

[1]Phyo Thu Thu Khine, [2]Htwe Pa Pa Win

[1]Associate Professor, [2]Associate Professor

[1,2] University of Computer Studies, Hpa-an, Myanamar

Abstract: In most real-world today applications need to make large volume of transactions in a small period over networks and this raw big data struggle data analytics to handle many problems. Solving class imbalance problems in real-world datasets pose a great challenge in research fields as performance of the classifiers are suffered from this very large different ratio of class labels. Many mechanisms tried to handle these problems with reducing high rate in data level while degrading the accuracy performance of the classifiers and vice versa. Hence, a framework to solve imbalance distribution of class labels in data level with the best accompanied classifier in supervised mechanisms, is proposed in this paper. Effectiveness of proposed framework is assessed by using the real-world dataset and the empirical results are compared to other well-known systems and the former works. The proposed work achieved the acceptable results and outperform than the other works.

Index Terms **– Big Data, Class Distribution, Class Imbalance, Data Analytics, Data Level, Supervised Mechanism.**

I. INTRODUCTION

The rapid developments of online transactions systems lead to store high volume of data and these datasets need to be in safety state and have unequal distribution between classes. The harshness of class imbalance in dataset may cause from minor difficulties to severe problems to real-world applications. A dataset became imbalance when the classes as in example, fraud and normal, non-fraud labels are not equally contained in credit card financial database. The most labels create majority class instance while the limited label present minority class in dataset. These imbalance distributions of class labels can be viewed in most real-world dataset. If the ratio of class imbalance distribution is high, then the classifiers may yield performance of the prediction accuracy owing to the classification model is likely to anticipate the most majority class instances in training set. That prediction model is not applicable in real-world domain since minority class is most important and interested portion for the related experts [1, 2].

The properties of big data make difficulties to some traditional machine learning methodologies to make analysis and model for such data. The traditional methods suffer from touching high volume of data, facing with different data formats, the reaching of different speed from various sources when making filtering process and making required transformation of data. The increasing development of big data applications lead to develop more effective and efficient knowledge extraction mechanism from this data type [2].

In machine learning paradigm, classifiers are attempted to minimize incorrectly categorized instances or misclassification errors while to maximize the accuracy of predication. These classifiers can produce describe acceptable desired when the input dataset has roughly balanced in class labels. Therefore, performance assessment of classification algorithms is depending on the underlying dataset, especially need to care in imbalanced environments [3].

The class imbalance problems can create many difficulties to most of the classifiers during the learning procedure. Although, most of the systems tried to develop based on the traditional methods from direct learning of imbalanced data, they may not achieve acceptable results. Even they achieved the significant results, these results are unreliable because cardinality of minority class is extremely small. Therefore, maintaining the distribution of class is remarkably important issues and the role of minority class is awfully essential in preprocessing stage of the modeling of the classifiers and they need to handle with special methods [4].

Therefore, this paper proposed a mechanism to solve the imbalanced class problems in large datasets by using the most effective classification of Random Forest decision tree and oversampling technique. The rest of paper is described as the following section: Section 2 explores the relative works that manage the problems of imbalanced class. Section 3 establishes the background methodology of the proposed mechanisms. The proposed framework is described in section 4. The experimental evaluations are performed in section 5 and section 6 is the conclusion of proposed works.

II. RELATED WORKS

There are very many works that the previous system proposed to handle the distribution of class imbalance problems. These proposals of solving tactics can be classified into two main categories: the data level and the algorithm level. The data level methods sample the instances of training sets and these methods can be assembled into two main types: undersampling and oversampling. Improvements of these are described as the ensemble and application-oriented approaches.

Most studies that focus on the undersampling based works would be reviewed firstly.

The authors in [5] introduced a novel algorithm called WIMUS using undersampling methods to tackle the bottle necked of skewed distributed data. They separated the dataset into two pieces, minority and majority subsets. Then remove the noisy borderline instances of majority class that create the imbalance problems. Then they are reforming the strong dataset and classified with random forest classifier. They achieve the highest accuracy rate, 0.996% for UCI small imbalanced dataset.

The group in [6] found that Convolutional Neural Network, CNN face the problems in big data classification of imbalanced set of images and undersampling method and cost-sensitive methods can be used with that CNN. They showed that the work with CNNs achieved 0.98% maximum in image datasets.

Secondly, the research on oversampling based works has been studied. Typical oversampling method is the random sampling process to enhance the minority class distribution. Base on this mechanism many other methods are advanced.

The most popular oversampling technique, Synthetic Minority Over-sampling Technique called SMOTE, is developed in [7] to increase the minority class. They also mentioned that the ensemble uses of undersampling and SMOTE may be better for some environments. There are many advancements of the SMOTE technologies exist. In SMOTEBoost, the oversampling method is used with the combination of ADoBoost classifier is proposed in [8]. The Borderline-SMOTE based on k-nearest neighbor to generate synthetic samples of the minority label is improved in [9]. The combination of undersampling and SMOTE called RUSBoost, the better and faster sampling method to skew the class distribution is developed in [10].

The researchers in [11] invented a novel oversampling algorithm called multiclass radial-based oversampling, MC-RBO based on the extensive experiments. The main concept of their method is to produce the artificial instances as the extensive ways of existing oversampling method. Their assembling ways of both near-Bayesian support vector machines, NB-SVM and Decision Tree with oversampling method outperform the previous popular methods.

The scientists in [12] created an oversampling method applying conditional generative adversarial network, cGAN based on SMOTE. They used a discriminator for training reasons in network, as an outlier examiner to check the distributions of class instances between minority and majority natures. Therefore, the bias of classification boundary from the outliers can be prevented. They generate artificial minority class to the initial dataset. They named their algorithm as OD-GAN, to detect outliner using oversampling on adversarial network and their method outperform than other methods that used outliner detection.

Later, the other methods that used the ensemble techniques of both, would be discussed.

The authors in [13] proposed ensemble techniques of both oversampling and undersampling and use the heterogeneous ensembles based on multiple learning algorithms for having a higher potential to generate diverse members not to be homogeneous ones. They evaluated the performance with AUC and F1 measures from their experiment and better results produce than the other homogenous ones and achieve with 0.93 AUC as though the highest percentage.

The rest methods are the application oriented previous works for the fraud detecting technique that used the same dataset as this paper.

The authors in [14] proposed the mechanism to resolve class imbalance problem of fraud detection with focus on class distribution, the sample size, the separation of class and within the class concept. They use clustering tree for class separation and construct a tree. They compared their results with other five popular algorithms and prove the effectiveness of their system.

The group in [15] proposed a fraud detection technique using SMOTE technique with different machine learning methods. They show all the results and achieve 99.96% as the highest accuracy with the combination of Naïve Bayes and SMOTE.

III. BACKGROUND METHODOLOGY

3.1 Maintenance Mechanism for Class-Imbalance Learning at Data Level

To handle the distribution of class imbalance problems in learning stages, there are many mechanisms characterized into two core groups: resampling and making cost sensitive. The data level sampling methods modifies distribution of the class among the dataset to get the appropriate weights between classes. These data level method can be assembled into oversampling, undersampling and balancing methods. The brief explanation of the methods is as the following [16].

3.1.1 Oversampling

The oversampling methods form the additional minority instances of training set to enlarge the total size of minority class in the dataset.

Normal Oversampling: The normal resampling maintains the distribution in the subsamples of training set to get a uniform one. This mechanism over sample the minority class instead of being undersampled the majority class in order to obtain the same class of the training set.

Randomize Oversampling: It uses randomization to select the subsamples of minority class in the original dataset with replacement.

There may be different improvement mechanisms for oversampling methods as described in related works. Among them, SMOTE technique is the most popular well-known method for imbalance problems.

3.1.2 Undersampling

The undersampling methods make the additional majority instances of training set to enlarge the total size of majority class in the dataset.

Normal Undersampling: It calculates the total number of instances in class that consists in the dataset and percentage of minority of the class. The purpose is to undersample class of majority so that both classes in the training set have the same number of instances without replacement.

Randomize Undersampling: It uses the randomization to produce the subsamples of original dataset. It can spread from the rarest to the most common class with resampled to be different in the class frequencies.

3.1.3 Balancing

The balancing methods balance the data instance weight in training set to have equal weight. Some of the methods of these types are as follows.

Reweighting: This algorithm is a supervised method for data cleaning process in order to battle the class imbalanced problems in large dataset. It calculates the weight of instance data and sums in training dataset to the same weight in each class. As the changes of the weight of data occur in the first batch, it can be continued with the classifiers. If the input class is numeric, equal-with discretization method is used to discretize the label for weighting as the pseudo class.

Distributed Reweighting: This is nearly the same with the Reweighting process and used for distribution of imbalanced class in a dataset. The only difference is the changes of the weight of data occur distributed in every batch of the training set. This will probably yield the training set where both imbalanced classes are nearly balanced.

3.2 Performance Measures

The performance of classification system is measured as follows [2]:

Metric	Formula
Accuracy	$\frac{TP + TN}{TP + FP + FN + TN}$
Sensitivity	$\frac{TP}{TP + FN}$
Specificity	$\frac{TN}{TN + FP}$
Precision	$\frac{TP}{TP + FP}$
Recall	$\frac{TP}{TP + FN}$
F-measure	$\frac{2 * Precision * Recall}{Precision + Recall}$
G-MEAN	$\sqrt{sensitivity * specificity}$
AUC	$\frac{1 + TPR - FPR}{2}$

IV. PROPOSED FRAMEWORK

The proposed framework for the distribution of the class imbalance problems emphasize on both of two categories: at data level and at classifier level instead of handling only at one level in other systems. In the data level, the normal oversampling method is proposed to reduce the imbalance in class distribution of the fraud dataset. At the classifier level, the Random Forest classifier is proposed as the best cooperating partner of the data level's solving method to improve the performance of the overall system. The best cooperative work of the proposed framework for the fraud dataset can be seen in the experiment section.

V. EXPERIMENTAL EVOLUTIONS

The dataset for this experiment is used from real world public Kaggle dataset of credit card transaction of European cardholders between two days duration in September 2013. The dataset includes 31 features including the class label. It contains total of 284,807 transaction instances, but just only a small transaction of 492 are frauds, only 0.193% of the total instances. Therefore, the distribution of classes creates a problem for the classifier, and it is extremely important for that real-time transaction data of financial process.

Firstly, the experiments are carried out for the classifier oriented in accuracy and begin with the classifiers that the previous systems recommend that these classifiers achieved the higher performance in many fields. The accuracy results for the credit card information are computed and demonstrated in Figure 1, in propose to compare to clearly see the best classifier among them. All the experiment performances are measured with the standard 10-fold cross-validation.

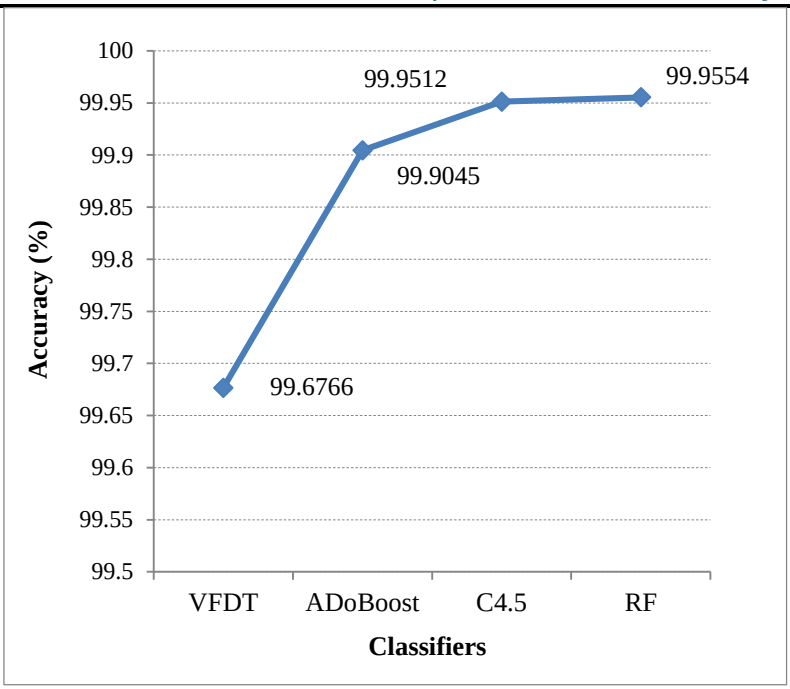

Figure 1. Comparison of accuracy results for the classifiers

As could be seen from the Figure 1, the popular very fast decision tree is the worst accuracy producer between the described recommended classifiers. Although, ADoBoost is the better classifier for the other system, it is worse than the other two for the imbalance set of fraud data. Among the best two, C4.5 and Random Forest, the Random Forest is the best with 99.9554% and higher 0.0032% than the other. The experiments are continued to increase the classification accuracy of the classifier and to handle the classifier's troubles of class imbalanced problems at the data level. The performance results for the supervised sampling methods for the best classifier for the fraud data are shown detail in Table 1. The increment of the proposed combination of the oversampling method over the baseline Random Forest Classifier are calculated and described in Table 2.

Table 1: Performance Results for Different Sampling Techniques

Classifier	Time(s)	Accuracy	Precision	Recall	F1_score	MCC
BaseLine	465.08	99.9554	1.000	1.000	1.000	0.862
Reweighting	212.84	90.4417	0.920	0.904	0.904	0.824
Oversampling and Reweighting	166.06	95.7287	0.961	0.957	0.957	0.918
Distributed Reweighting	0.03	96.6667	0.969	0.967	0.967	0.935
SMOTE	473.43	99.9369	0.999	0.999	0.999	0.904
Normal Undersampling	445.37	99.9575	1.000	1.000	1.000	0.869
Randomized Undersampling	484.4	99.9586	1.000	1.000	1.000	0.872
Normal Oversampling	329.14	99.9803	0.999	0.999	0.999	0.942

Table 2: Increment Performance of the Resampling over Normal Classifier

Classifier	Time(s)	Accuracy	Precision	Recall	F1_score	MCC
BaseLine RF	465.08	99.9554	0.999	0.999	0.999	0.862
Normal Oversample+RF	329.14	99.9803	0.999	0.999	0.999	0.942
Increasement	-135.94	+0.0249	Nil	Nil	Nil	+0.08

Then the tests are stepping to compare the outcomes of other systems with the same format of the training and the testing set. There are total of 170,884 instances for training set with 170583 for normal samples and 301 instances for abnormal, and the balance rate is 566.7:1. The remainder of the 113731 instances of normal samples and 191 instances of abnormal are for testing set, with 595.5:1 balance rate. Performance comparison of different methodologies of the previous system and the proposed method for this data set are shown in Table 3.

The results from the Table 3 gives a proof that the proposed methodology outperformed the previous proposed methods. The results for these popular SMOTEBoost and RUSBoost are calculated with the same environments and test set as other methods in the table.

Table 3: Performance Comparisons with Other Previous System

Model	F1	Precision	Recall
Logistic Regression [15]	0.726	0.893	0.612
SVM [15]	0.739	0.921	0.617
Decision Tree [15]	0.753	0.711	0.801
AdaBoosting [15]	0.779	0.847	0.723
Random Forest [15]	0.842	0.914	0.781
Clustering Tree [15]	0.852	0.916	0.796
SMOTEBoost [8]	0.998	0.998	0.998
RUSBoost [10]	0.998	0.998	0.998
Proposed Method	0.999	0.999	0.999

Again, the experiments are continued with the other system in different criteria of test set. In this process the training set contains 284807 instances and 56861 instances for test set. The results for this set with different mechanisms of previous system are described in Table 4. For these testing criteria, the proposed system provides the better accuracy result than the other mechanisms.

Table 4: Performance Comparison with the Previous System

Model	Accuracy (%)	Precision	Recall
LR+SMOTE [16]	97.46	58.82	91.84
NB+SMOTE [16]	99.23	16.17	82.65
RF+SMOTE [16]	99.96	96.38	81.63
MLP+SMOTE [16]	99.93	79.21	81.63
Proposed Method	99.98	98.2	93.9

VI. Conclusion

This paper proposed a mechanism to solve imbalanced class problems in fraud detection by the combination of oversampling technique and Random Forest classification. Even the oversampling method can produce better results in some classifier while the accuracy is degraded in other classifier. Likewise, the most popular advancement of SMOTE methods can provide better results in other system with the combination different classifiers, but it reduces with the proposed Random Forest classifier for the fraud dataset. That reasons assure that the effective sampling method to support the best companion classifier for the target domain is very important. Therefore, the proposed mechanism of the best combination of sampling method and classification is the most effective approach for the real-world application of fraud dataset. The advancement of sampling mechanism and the improvement of multi class imbalanced environments may be the future works.

References

[1] W. Qian, S. Li, "A novel class imbalance-robust network for bearing fault diagnosis utilizing raw vibration signals", Measurement, vol. 156, 2020. doi: https://doi.org/10.1016/j.measurement.2020.107567

[2] Leevy et al., "A survey on addressing high-class imbalance in big data", Journal of Big Data, 5:42, 2018. doi: https://doi.org/10.1186/s40537-018-0151-6

[3] F. Thabtah, S. Hammoud, F. Kamalov, and A. Gonsalvesv, "Data Imbalance in Classification: Experimental Evaluation", Information Sciences, vol. 513, March 2020, pp. 429-441. doi: https://doi.org/10.1016/j.ins.2019.11.004

[4] J. Hamidzadeh, N. Kashefi and M. Moradi, "Combined weighted multi-objective optimizer for instance reduction in two-class imbalanced data problem", Engineering Applications of Artificial Intelligence, vol. 90, April 2020, 103500. doi: https://doi.org/10.1016/j.engappai.2020.103500

[5] D. D. Prasad, D. V. Prasad, and K. N. Rao, "Imbalanced Data Using with-in Class Majority Under Sampling Approach", In 2019 IEEE International Conference on Electrical, Computer and Communication Technologies (ICECCT), Coimbatore, India, 20-22 February 2019, pp. 1-5. doi: 10.1109/ICECCT.2019.8869339

[6] J. M. Johnson, and T.M. Khoshgoftaar, "Survey on deep learning with class imbalance", Journal of Big Data, 6, article number 27, March 2019. doi: https://doi.org/10.1186/s40537-019-0192-5

[7] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, "SMOTE: Synthetic minority over-sampling technique", Journal of Artificial Intelligence Research, vol. 16, No. 1, June 2002, pp. 321–357. doi: https://dl.acm.org/doi/10.5555/1622407.1622416

[8] N.V. Chawla, A. Lazarevic, L.O. Hall, and K.W. Bowyer, "SMOTEBoost: Improving Prediction of the Minority Class in Boosting", In: N. Lavrač, D. Gamberger, L. Todorovski, H. Blockeel (eds), Knowledge Discovery in Databases: PKDD 2003, Lecture Notes in Computer Science, vol. 2838. Springer, Berlin, Heidelberg. doi: https://doi.org/10.1007/978-3-540-39804-2_12

[9] H. Han, W. Y. Wang, and B.H. Mao, "Borderline-SMOTE: A New Over-Sampling Method in Imbalanced Data Sets Learning," In: Huang DS., Zhang XP., Huang GB. (eds), Advances in Intelligent Computing. ICIC 2005. Lecture Notes in Computer Science, vol. 3644. Springer, Berlin, Heidelberg. doi: https://doi.org/10.1007/11538059_91

[10] C. Seiffert, T. M. Khoshgoftaar, J. V. Hulse, and A. Napolitano, "RUSBoost: A Hybrid Approach to Alleviating Class Imbalance", IEEE Transactions on Systems, Man, and Cybernetics - Part A: Systems and Humans, vol. 40, Issue. 1, 2010, pp. 185-197. doi: 10.1109/TSMCA.2009.2029559

[11] B. Krawczyk, M. Koziarski, M. Woźniak, "Radial-Based Oversampling for Multiclass Imbalanced Data Classification", IEEE Transactions on Neural Networks and Learning Systems (Early Access), June 2019, pp. 1-14. doi: 10.1109/TNNLS.2019.2913673

[12] J-H. Oh, J.Y. Hong, and J-G. Baek, "Oversampling method using outlier detectable generative adversarial network", Expert Systems with Applications, vol. 133, 1, November 2019, pp. 1-8. doi: https://doi.org/10.1016/j.eswa.2019.05.006

[13] H. G. Zefrehi, and H. Altınçay, "Imbalance learning using heterogeneous ensembles", Expert Systems with Applications, vol. 142, 15, March 2020, 113005. doi: https://doi.org/10.1016/j.eswa.2019.05.006

[14] Y. Zhang, G. Liu, L. Zheng, and C. Yan, "A Hierarchical Clustering Strategy of Processing Class Imbalance and Its Application in Fraud Detection", IEEE 5th International Conference on Data Science and Systems (HPCC/SmartCity/DSS), August. 2019. doi: 10.1109/HPCC/SmartCity/DSS.2019.00249

[15] D. Varmedja, M. Karanovic, S. Sladojevic, M. Arsenovic, and A. Anderla, "Credit Card Fraud Detection - Machine Learning methods", 18th International Symposium INFOTEH-JAHORINA (INFOTEH), 20-22 March 2019. doi: 10.1109/INFOTEH.2019.8717766

[16] I. Witten, E. Frank, M. Hall, and C. Pal, Data Mining: Practical Machine Learning Tools and Techniques, Fourth Edition, Morgan Kaufmann, 2016, ISBN: 978-0-12-804291-5.