Research Article

Performance Improvement of Classification Model with Imbalanced Dataset

Vipin Khattri[a], Sandeep Kumar Nayak[b]

[a,b]Department of Computer Application, Integral University, Dasauli, Bas-ha Kursi Road, Lucknow – 226026, India.
vipinkhattri@gmail.com; nayak.kr.sandeep@gmail.com

Article History: Received: 10 January 2021; Revised: 12 February 2021; Accepted: 27 March 2021; Published online: 4 June 2021

Abstract: Classification models based on machine learning for the application of real life carry out classification tasks using real life dataset. Classification models have class imbalance problems when the dataset is imbalanced in nature. Classification models show biases for performing the classification towards the majority class due to class imbalance issues. The purpose of the study is to examine and control the class imbalance problem using the cluster centroid undersampling technique with the motive of improving the performance of machine learning classification. In order to accomplish the goal, this study performs experimental analysis for examining and controlling the class imbalance problem before and after applying the cluster centroid undersampling technique on imbalanced dataset. The experimental study is performed by using five different imbalanced datasets, cluster centroid undersampling technique and decision tree classification model. The results of this study are promising that supports this study and confirm that the class imbalance problem can be handled using undersampling techniques very effectively with performance improvement of a classifier from 11% to 67%. This study highlights the influences of class imbalance problems on machine learning classification models and experimental results with analysis provide an appropriate conclusion with an improvement in the performance of machine learning based classification models.

Keywords: Class Imbalance, Undersampling, Imbalanced Dataset, Binary Classification, Knowledge-based Classifier

1. Introduction

The accuracy of a predictive classifier is based on an algorithm of machine learning that depends on the dataset using which the classifier is trained [1]. The classifier shows 95% precision, which can be considered as a good classifier. Despite an excellent performance, the classifier gives an incorrect classification in many cases. The researchers try to find out the reason behind this misclassification and find out that the imbalanced dataset is the mainstay of this misclassification. When the dataset is analyzed, it is found that more than 95% of data samples belong to the one class, which is known as the majority class (negative instances). This means that the classifier shows its accuracy based on the majority class. In other words, the classifier does 95% correct classification if samples belong to the majority class, and 95% incorrect classification if samples belong to the minority class. In the imbalanced dataset (figure 1), samples of one class are significantly less than samples of the other class [1, 2].

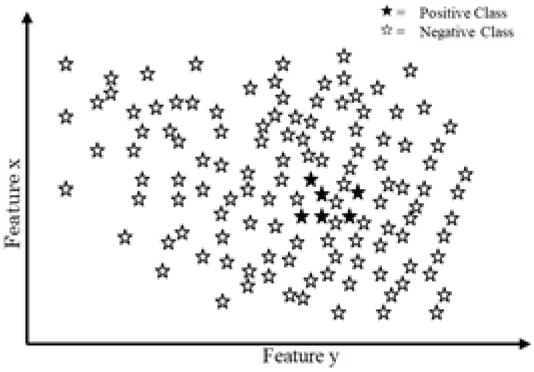

Figure 1. Imbalanced Dataset

Classifier is trained based on each class sample [3, 4]. Therefore, balanced samples of each class are required. Every class is significantly important in classification prediction. Building a classification model using imbalanced dataset may cause errors in classification [5]. Most classifiers demonstrate their accuracy in favour of the majority class because the classifier shows more interest in samples of the majority class [6]. Since algorithms of machine learning are designed to work on balance class distribution, which means the ratio between the majority class and minority class should be 1:1. However, in the imbalanced dataset, the ratio between the majority class and minority class is not 1:1, due to which the classifier tends to be more inclined towards the majority class [7]. Therefore, classification error arises with respect to the minority class samples.

This is a common problem in the application of the real world. The imbalanced dataset is very prevalent in all fields of human life [8]. In many areas like engineering [9], information technology [10], medical [9, 11] and finance [9, 12] have class imbalance problems. Whatever data are collected in the dataset by real-world applications is mostly imbalanced in nature.

The class imbalance issue creates a challenge for researchers in developing real-life classification applications with high accuracy for both majority and minority class [2, 4].

This study includes two main aspects

1) To examine the class imbalance issue with a decision tree classifier.

2) To control the class imbalance issue with decision tree classifier using the cluster centroid undersampling technique.

Nearly all real-life applications are facing the concept of imbalanced dataset. Hence, this study demonstrates an advantage in terms of outcomes by assessing the performance of a classifier with five different imbalanced datasets and a cluster centroid undersampling technique. The performance result fully demonstrates that the performance of a classifier is enhanced. It also supports the research work of this study. The result of this research shows that an improvement of approximately 11% to 67% has achieved on the accuracy of a classifier. The improvement in the performance of a classifier has not achieved only with a single dataset rather than it has achieved with all the imbalanced dataset. In view of the significance of the result and the study, it is an essential requirement that undersampling technique should be carried out with imbalanced dataset before construction the classification model.

The remaining paper has the following sections. Section 2 examines previous work to handle the class imbalance issue. Section 3 implements the methodology for handling the class imbalance issue. Section 4 analyses and discusses the result and section 5 gives the conclusion of the study.

2. Literature Review

This section reviews literature based on the class imbalance problem, handling class imbalance problem, imbalanced dataset and classifiers. Buda *et al.* [13] analyzed the experimental performance of convolutional neural networks (CNNs) in reference of the class imbalance problem. This paper discussed the different methods to handle the class imbalance issues. Three datasets of images were used in the experiment and finally found that oversampling can be used to handle class imbalance, undersampling also useful but performance depended on ratio of class imbalance.

The study of Santiso *et al.* [14] showed the performance of detection of adverse drug reaction using imbalanced dataset. The study also discussed various techniques like sampling, cost-sensitive learning and Ensemble learning to handle imbalanced datasets. The result of the study showed that the classifier produced better results using a sampling method to handle imbalanced dataset.

The study of imbalance class [15] analyzed and showed the degradation of performance of predictor due to the imbalance dataset. For handling the degradation due to imbalanced dataset, this study performed an experiment for classifier with 10-fold cross validation. This study also addressed the issue of handling imbalanced dataset in pre-training phase classifier for improving the performance.

Vuttipittayamongkol and Elyan [16] build a framework for handling imbalanced binary classification dataset due to overlapping classes. The basic work of the proposed method was to remove data from overlapping with the aim of the least loss of information. Neighbourhood searching methods were used in the proposed method to correctly find the overlapped samples. The result of the proposed method performed exceptionally well.

Fahrudin *et al.* [17] proposed new algorithm and enhance the working of SMOTE with reference to the balance the imbalanced dataset. The proposed algorithm used Attribute Weighted and KNN Hub with SMOTE. For practical implementation, this study utilized nine datasets that were acquired from the Keel repository. The outcome of this study showed better performance compared to other over-sampling algorithms.

3. Materials & Methodology with Experimental Setup

This study encompasses cluster centroid undersampling technique and decision tree to examine and control the class imbalance problem. This study also includes five distinct imbalanced datasets with different imbalance ratios.

3.1. Cluster Centroid Undersampling Technique

Cluster centroid is an undersampling technique of data mining. The task of this technique makes a balance between majority and minority class instances. In clustered centroid undersampling, insignificant instances are discarded among the majority class [18]. The differentiation of instances as significant and insignificant is achieved by using the concept of clustering. Cluster centroid utilizes the concept of the centroid cluster search. Clusters are created around the data points belonging to the majority class. The cluster centroid is identified by achieving the average feature vectors for all the features, on the data points pertaining to the majority class in the feature space. After locating the cluster centroid of the majority class, the instance pertaining to the cluster (majority class), which is farther away from the cluster centroid in feature space, is regarded to be the most insignificant instance. On the reverse, the instance pertaining to the majority class, that is closer to the cluster centroid in the feature space, is regarded as the most significant instance. Instances pertaining to the majority class are eliminated on the basis of their significance. In this process, the samples of majority class continue to be removed until the number of majority class samples equal to the number of minority class samples [19].

3.2. Decision Tree Classifier

The decision tree performs classification using a representation of the information depiction to classify the instances in two or more different class labels [20]. It works on supervised learning and can perform classification as well as regression tasks. The decision tree is used in different types of applications like medical, scientific, and business areas [8]. The decision tree can handle numerical and categorical values [9]. It uses decision rules, which are concluded from the training dataset [21].

The decision tree comprises various parts such as the root node, decision node, leaf or terminal node, and sub-tree. Root node contains all complete nodes, and it divides into more sub-nodes that are similar. The decision node divides into sub-nodes, and the leaf or terminal node does not have sub-nodes. A tree has branches that are known as a subtree. Parent node divides into sub-node and sub-node is called a child node. Two steps are used in the decision tree classifier to solve classification problems. The first step is executed for creating a tree, and the second step is acted for tree pruning. The decision tree uses a statistical approach to place the attribute on the decision tree as a root node and decision node. The performance of the decision tree depends on the strategy of how the tree splits. During the construction of the decision tree, efficient attribute selection is required for a root node or the decision node. The decision tree uses different criteria for efficient attribute selection like entropy or information gain or Gini index.

3.3 Dataset

The complete study is moving around the imbalanced dataset. This study used five different imbalanced datasets such as CM1, Credit Card Fraud, Glass-3, JM1, Loan Data (Kaggle dataset). Before conduction of the experiment, all datasets were pre-processed with respect to the categorical data, missing values, normalization and binary class. Each dataset exhibits following characteristics (Table 1).

Table 1. Characteristics of Imbalanced Dataset

S.No.	Dataset	Number of Instances	Instances Class		(%) Minority Class	Imbalance Ratio
			Majority	Minority		
1	CM1	498	449	49	9.84	9.16
2	Credit Card Fraud	284807	284315	492	0.17	577.88
3	Glass-3	214	197	17	7.94	11.59
4	JM1	10885	8779	2106	19.35	4.17
5	LoanData	9578	8045	1533	16.01	5.25

3.4. Experimental Setup

This study has a clear aim to show the improvement in the performance of the classification task by applying cluster centroid undersampling technique on the imbalanced dataset. The aim of the study is accomplished by completing the four following objectives.

- Find the performance of a classifier using different imbalanced dataset.
- Prepare a balanced dataset by applying a cluster centroid undersampling technique on each imbalanced dataset.
- Find the performance of a classifier using different balanced dataset.
- Compare the performance of a classifier before and after applying the cluster centroid undersampling technique using imbalanced and balanced dataset.

A framework (figure 2) is designed to achieve the objectives mentioned above. This framework has four parts and each part maps each objective of the study separately.

[1] **Part-I.** This part is designed to achieve the objective I and has three steps. The first step trains a classifier using each imbalanced training dataset. The second step tests the classifier using each imbalanced test dataset. The third step evaluates the performance of a classifier using standard performance metrics against each imbalanced dataset.

[2] **Part-II.** This part is designed to achieve the objective II, and prepares the balanced dataset from all imbalanced datasets using a cluster centroid undersampling technique separately.

[3] **Part-III.** This part is designed to complete objective III, and has three steps. The first step trains a classifier using each balanced training dataset. The second step tests the classifier using each balanced test dataset. The third step evaluates the performance of a classifier using standard performance metrics against each balanced dataset.

[4] **Part-IV.** This part is designed to map the objective IV. This part compares the performance results that are created during the execution of part-I and part-III of a classifier based on parameters of standard performance metrics separately.

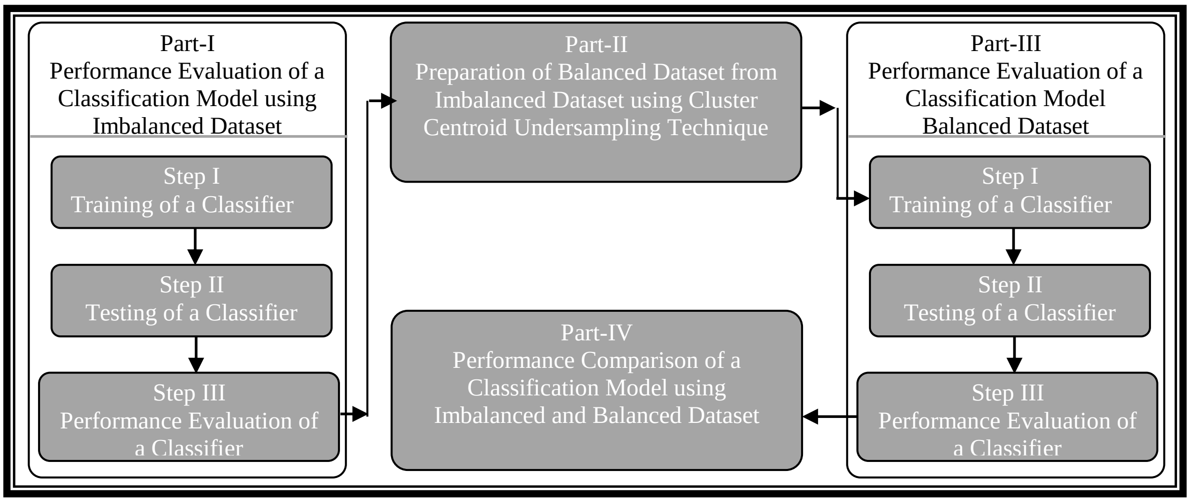
Part-I
Performance Evaluation of a Classification Model using Imbalanced Dataset
Step I
Training of a Classifier
Step II
Testing of a Classifier
Step III
Performance Evaluation of a Classifier
Part-II
Preparation of Balanced Dataset from Imbalanced Dataset using Cluster Centroid Undersampling Technique
Part-IV
Performance Comparison of a Classification Model using Imbalanced and Balanced Dataset
Part-III
Performance Evaluation of a Classification Model Balanced Dataset
Step I
Training of a Classifier
Step II
Testing of a Classifier
Step III
Performance Evaluation of a Classifier

Figure 2. Framework for Performance Evaluation of a Classification Model using Imbalanced and Balanced Dataset

3.5. Performance Evaluation Measure

The significance of the study is to find the performance of a classifier and compare it with previous results. The comparison and analysis are made on the basis of standard performance metrics. The performance metrics work on the basis of the confusion matrix [16, 22] and are used in order to evaluate the performance of a classifier using imbalanced and balanced dataset. In this study, accuracy, precision, recall, F-Score, G-mean and AUC ROC Score (find the most accurate performance measure of a classifier) are used as performance metrics.

4. Results and Discussion

The main goal of this study is to control the class imbalance problem of imbalanced dataset using undersampling technique. In order to accomplish the goal of this study, an effective experimental procedure was followed. The impact of undersampling technique on imbalanced dataset was evaluated. The decision tree executed with five imbalanced datasets before and after applying cluster centroid undersampling technique. This section shows the results of comparative performance of a decision tree classifier with an imbalanced and balanced dataset.

Table 1 shows the number of samples of a majority and a minority class of imbalanced data. After applying the cluster centroid undersampling technique on imbalanced dataset, the number of majority samples was reduced to the number of minority samples in each dataset and using decision tree found the performance with an imbalanced and balanced dataset.

The comparative results (Table 2) are showing the impact using six performance measures with reference to the before and after applying the undersampling technique using each dataset. When the value of each performance measure reaches to 1.0 then classifier should be considered as good classifier. This impact can also be seen in figure 3 using AUC-ROC Score most accurate performance measure of the classification model. Figure

4 is clearly seen that the performance of a classifier is improved from 11% to 67%. The analysis result of table 2 is shown that all the performance measures were improved after apply the cluster centroid undersampling technique but accuracy decreased. The drawback of a accuracy is that it can give the result of accuracy in terms of majority class. Therefore, for analysis this study has taken six performance measures.

Table 2. Performance Comparisons of a Decision Tree Classifier without and with using the Cluster Centroid Undersampling Technique on each Imbalanced Dataset.

Evaluation Measures	Dataset				
	CM1	Credit Card Fraud	Glass3	JM1	Loan
Without using the Cluster Centroid Undersampling Technique					
Accuracy	0.86000	0.99920	0.87692	0.75077	0.72651
Precision	0.28571	0.78417	0.33333	0.34268	0.19540
Recall	0.26667	0.74150	0.14286	0.35948	0.21795
F1 Score	0.27586	0.76224	0.20000	0.35088	0.20606
Geometric mean	0.49690	0.86095	0.37139	0.54983	0.42415
AUC ROC Score	0.59630	0.87057	0.55419	0.60024	0.52169
With using the Cluster Centroid Undersampling Technique					
Accuracy	0.70000	0.98649	0.90909	0.67801	0.82283
Precision	0.66667	0.99320	1.00000	0.65706	0.78728
Recall	0.80000	0.97987	0.85714	0.70521	0.84471
F1 Score	0.72727	0.98649	0.92308	0.68028	0.81498
Geometric mean	0.69282	0.98651	0.92582	0.67824	0.82412
AUC ROC Score	0.70000	0.98653	0.92857	0.67876	0.82437

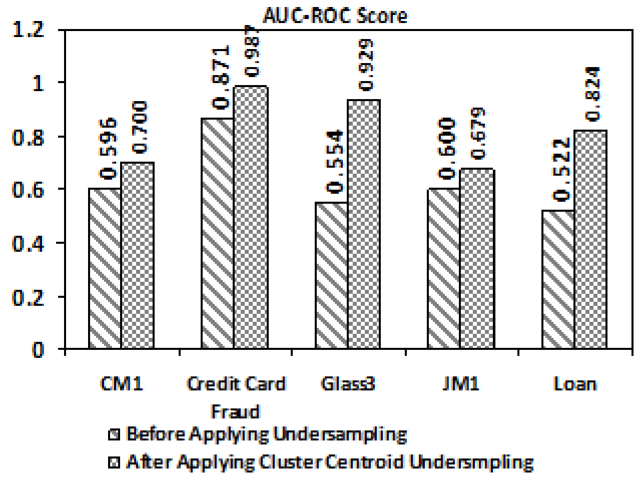

Figure 3. Performance Comparision of a Decision Tree Classifier using AUC-ROC Score before applying Undersampling Technique and after applying Cluster Centeroid on Imbalanced Dataset.

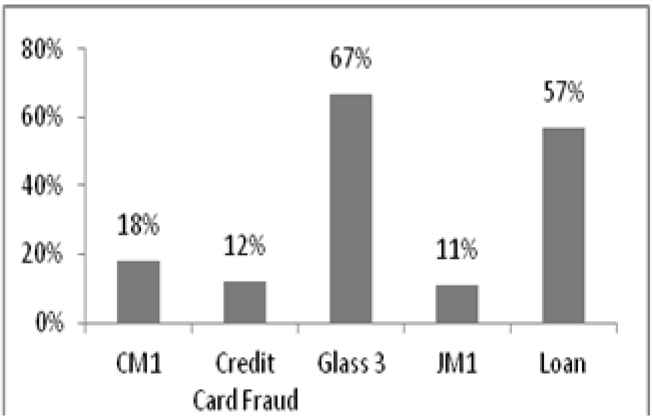

Figure 4. Performance Improvement of a Decision Tree Classifier using AUC-ROC Score after applying the Cluster Centeroid on Imbalanced Dataset.

After analyzing the results and graphs, it is evident that the performance of the classifier was enhanced after applying the cluster centroid technique on imbalanced dataset. This means that class imbalance problem can be controlled using undersampling technique.

5. Conclusion

The aim of this study was to examine and control the class imbalance problem by applying the cluster centroid undersampling technique on an imbalanced dataset and improve the performance of a classifier. In the direction of achieving the goal, this study reviewed various kinds of literature with different dimensions to find the reason behind the degradation of the performance of the prediction model (classifier) on the imbalanced dataset. A framework of an experimental setup including four parts was established for this study. The experimental setup used cluster centroid undersampling technique, which applied on five datasets with different imbalance ratios. In the experiment, a decision tree classifier used for evaluating the performance. In view of evaluating the performance of a classifier, this study also used six different performance evaluation measures.

The comparative results of this study were encouraging research for controlling class imbalance problems and proven that after applying cluster centroid undersampling technique on the imbalanced dataset, the performance of a classification model was improved. The result of this study gives the direction that the imbalanced dataset of a real-life application must be implemented with undersampling technique for building an unbiased and balanced classification system due to which system could find the accurate and balanced result of the prediction model (classifier).

In view of the significance of undersampling technique for this research study, the authors will implement this concept of credit card fraud detection system of real life application in future. It will help enhance the accuracy of the system.

Acknowledgment

The authors give our thanks and gratitude towards the Integral University for supporting research work and providing Manuscript Communication Number-IU/R&D/2020-MCN000952.

References

[1] M. A. H. Farquad and I. Bose, Preprocessing unbalanced data using support vector machine, Decision Support Systems, vol.53, no.1, pp.226-233, 2012.

[2] A. Fernández, S. García, M. Galar, R. C. Prati, B. Krawczyk and F. Herrera, Learning from imbalanced data sets, Springer, Berlin, 2018.

[3] G. Kovács, An empirical comparison and evaluation of minority oversampling techniques on a large number of imbalanced data sets, Applied Soft Computing, vol. 83, p. 105662, 2019.

[4] S. Perry, D. Delen and T. Liu, A synthetic informative minority over-sampling (SIMO) algorithm leveraging support vector machine to enhance learning from imbalanced data sets, Decision Support Systems, vol. 106, pp. 15-29, 2018.

[5] O. Gram and I. Rekik, Multi-view learning-based data proliferate for boosting classification using highly imbalanced classes, Journal of neuroscience methods, vol. 327, p. 108344, 2019.

[6] W. Lu, Z., Li and J. Chu, Adaptive ensemble undersampling-boost: a novel learning framework for imbalanced data, Journal of systems and software, Vol. 132, pp. 272-282, 2017.

[7] B. Krawczyk, Learning from imbalanced data: open challenges and future directions, Progress in Artificial Intelligence, vol. 5, no. 4, pp. 221-232, 2016.

[8] V. López, A. Fernández, S. García, V. Palade and F. Herrera, An insight into the classification with imbalanced data: Empirical results and current trends on using data intrinsic characteristics, Information sciences, vol. 250, pp. 113-141, 2013.

[9] Y. Sun, A. K. Wong and M. S. Kamel, Classification of imbalanced data: A review, International journal of pattern recognition and artificial intelligence, vol. 23, no. 4, pp. 687-719, 2009.

[10] I. H. Laradji, M. Alshayeb and L. Ghouti, Software defect prediction using ensemble learning on selected features, Information and Software Technology, vol. 58, pp. 388-402, 2015.

[11] S. Fotouhi, S. Asadi and M. W. Kattan, A comprehensive data level analysis for cancer diagnosis on imbalanced data, Journal of biomedical informatics, vol. 90, p.103089, 2019.

[12] D. Veganzones and E. Séverin, An investigation of bankruptcy prediction in imbalanced datasets, Decision Support Systems, vol. 112, pp. 111-124, 2018.

[13] M. Buda, A. Make and M. A. Mazurowski, A systematic study of the class imbalance problem in convolutional neural networks, Neural Networks, vol. 106, pp. 249-259, 2018.

[14] S. Santiso, A. Casillas and A. Pérez, The class imbalance problem detecting adverse drug reactions in electronic health records, Health informatics journal, vol. 25, no. 4, pp. 1768-1778, 2019.

[15] F. Thabtah, S. Hammoud, F. Kamalov and A. Gonsalves, Data imbalance in classification: Experimental evaluation, Information Sciences, vol. 513, pp. 429-441, 2020.

[16] P. Vuttipittayamongkol and E. Elyan, Neighbourhood-based undersampling approach for handling imbalanced and overlapped data, Information Sciences, Vol. 509, pp. 47-70, 2020.

[17] T. Fahrudin, J. L. Buliali and C. Fatichah, Enhancing the performance of smote algorithm by using attribute weighting scheme and new selective sampling method for imbalanced data set, Int J Innov Comput Inf Control, vol.15, pp.423-444, 2019.

[18] S. J. Yen. and Y. S. Lee, Cluster-based under-sampling approaches for imbalanced data distributions, Expert Systems with Applications, vol.36, no.3, pp.5718-5727, 2009.

[19] M. Pawlicki, M. Choraś, R. Kozik and W. Hołubowicz, On the Impact of Network Data Balancing in Cybersecurity Applications. In International Conference on Computational Science, Springer, pp. 196-210, 2020.

[20] C.C. Aggarwal, Data Classification: Algorithms and Applications, Data Mining and Knowledge Discovery Series. Chapman & Hall/CRC, New York, 2014.

[21] H. He and E. A. Garcia, Learning from imbalanced data, IEEE Transactions on knowledge and data engineering, Vol. 21, no. 9, pp. 1263-1284, 2009.

[22] H. Zhang, H. Zhang, S. Pirbhulal, W. Wu and V.H.C.D. Albuquerque, Active Balancing Mechanism for Imbalanced Medical Data in Deep Learning–Based Classification Models, ACM Transactions on Multimedia Computing, Communications, and Applications, vol.16, no.1s, pp.1-15, 2020.