JETIR.ORG ISSN: 2349-5162 | ESTD Year : 2014 | Monthly Issue
JOURNAL OF EMERGING TECHNOLOGIES AND INNOVATIVE RESEARCH (JETIR)
An International Scholarly Open Access, Peer-reviewed, Refereed Journal

Classification of Imbalanced Datasets Using Hybrid Approach

[1]Hardikkumar Harishbhai Maheta
[1]Assistant Professor
[1]Information Technology Department
[1]Shantilal Shah Engineering College, Bhavnagar,India

Abstract : Classification is predicting the class labels of data samples in datasets. It is considered an essential and challenging problem in many real-life applications. In the class imbalance problem, some classes represent small numbers of data samples (minority class), and others classes represent large numbers of data samples (majority class). In the imbalanced dataset, data samples of the majority class dominate data samples of the minority class. Conventional classifiers are only interested in improving overall classification accuracy and ignore the importance of minority class. So, they are not sufficient enough for the classification of imbalanced datasets. In real life, the high classification accuracy of minority classes is essential. Research studies used various techniques such as sampling, feature selection, cost-sensitive learning, one-class learning, and ensemble methods to handle imbalanced environments in datasets. In this paper, we have employed a hybrid approach that combines the hybrid feature selection approach and use the optimal feature set to improve the classification accuracy of both majority and minority classes.

IndexTerms **- Imbalanced datasets, Hybrid approach, Classification, Feature Selection.**

I. INTRODUCTION

Classification is predicting the class labels of data samples in datasets. Predicting the class label in the balanced dataset is simple. On the contrary, it becomes difficult in the case of an imbalanced dataset. In the imbalanced dataset, the number of instances in one class is higher than in the other class in the same dataset. The class with a higher number of instances is called the majority class, and the class with a small number of instances is called the minority class. For example, consider a database having 90% of instances from the majority class and the rest 10% of instances from the minority class. These kinds of datasets are known as the imbalanced datasets (Spelmen, V. S., & Porkodi, R., 2018). Imbalanced dataset classification poses a challenge for predictive modeling because most machine learning algorithms exist for classification purposes and assume equal class distribution. These algorithms perform poorly in imbalanced datasets where the number of instances of each class is unequal. It is a severe problem as the minority class is more important, and the problem is more sensitive to classification errors for the minority class rather than the majority class. Researchers have developed many techniques to handle imbalanced datasets. It becomes difficult for the learning system to learn the instances related to the minority class of the imbalanced dataset. Therefore, most existing techniques do not perform well on the minority class instances when the dataset is imbalanced. Data processing practitioners encounter class imbalance problems in various fields like fraudulent telephone call detection, activity recognition, behavior analysis, diagnosis of rare diseases, etc. The imbalanced dataset contains unequal distribution between its classes. The main difficulty with the highly imbalanced dataset is extracting knowledge from it. Several issues carried out while using the data intrinsic characteristics within the classification problem are the imbalanced dataset characteristics such as small disjuncts, lack of density, dataset shift, noisy data, overlapping, and borderline examples. In many applications, such as medical diagnosis, minority class instances are more important than the majority class instances. Suppose the dataset of cancer patients contains 1,000 instances, where 900 instances are from the negative class (majority class) and 100 instances of the positive class (minority class). If the classification rule predicts all data as the majority class, the rule acquires 90% accuracy. In this case, the accuracy is not a proper representation of the classification performance because nothing from the minority class instances is classified accurately. The motivation behind this work is that the class imbalance problem reduces the classification accuracy for the minority class instances. Because their main focus is to gain high accuracy, sometimes the minority class instances are not classified accurately(Spelmen, V. S., & Porkodi, R., 2018). In this paper, we have used hybrid feature selection using genetic search and linear forward selection. We take the union of the optimal features from both approaches and apply various classification algorithms to those optimal features to improve the classification accuracy of both majority and minority classes.

The remainder of this paper is organized as follows. In Section II, imbalanced data distribution, its effect on classification performance, and various feature selection methods are discussed. In section III, existing techniques to handle imbalanced datasets for classification with their comparative analysis are discussed. Section IV describes the proposed methodology with experimental design. Section V compares the simulation results of the proposed approach with some conventional classifiers. Finally, Section VI concludes this paper.

II. BACKGROUND THEORY

2.1 Classification

Classification, a fundamental data mining technique, assigns items in a collection to target categories or classes(Jiawei, H. & Micheline, K., 2006). It performs classification on structured or unstructured data to accurately predict the target class for each data sample. Binary classification is the simplest type of classification problem, where the target attribute only has two possible values. For example, positive or negative. On the other hand, multiclass targets have more than two values. For instance, low, medium, high, or unknown credit rating. For a classification project, we use two data sets to divide the data. We use one set to build the model and another set for testing it. In the model build (training) process, a classification algorithm finds relationships between the values of the predictors and target values (class labels). The classification algorithms use different methods for finding relationships. These relationships are summarized in a model. This model is then applied to a different data set in which the class assignments are unknown. Classification models are tested by comparing the predicted values to known target values in a set of test data. The classification has many applications in customer segmentation, business modeling, marketing, credit analysis, and biomedical and drug response modeling.

2.2 Imbalanced Dataset

In an Imbalanced Dataset, one class consists of a larger number of instances than other classes in the same dataset (He, H. & Garcia, E. A.,2009). If there are two classes, then balanced data would mean 50% data instances of both classes. For most techniques, little imbalance is not a problem. So, if there are 60% instances for one class and 40% for the other, it should not cause any significant performance degradation. Only when the class imbalance is high, such as 90% instances for one class and 10% for the other, standard optimization criteria or performance measures may not be as effective and would need modification. A typical example of imbalanced data is an email classification problem where emails are classified as useful or spam. The spam emails are usually lower than the useful emails. So, using the original distribution of two classes leads to an imbalanced dataset. There are two types of imbalanced datasets, two-class imbalanced dataset and multi-class imbalanced dataset. The imbalanced dataset exists in many areas, such as medical diagnosis, detection of fraudulent calls, activity recognition, behavioral analysis, sentiment analysis, etc (He, H. & Garcia, E. A.,2009). Many factors influence the performance of a classifier model for imbalanced datasets, like the size of datasets, imbalance ratio, characteristics of imbalanced data like small disjunctions, the ambiguous boundary between classes, noisy data, and dimensionality of data. Conventional classifiers have difficulty accurately classifying instances of the minority class in an imbalanced dataset. If the classifier predicts all the instances as the majority class according to the classification rule, then the classifier gets high accuracy, but there are no instances of the minority class that are classified accurately because the main goal of classifiers is to gain high accuracy. In this case, conventional classifiers ignore the minority class instances. This is a major concern in many applications like medical diagnosis.

2.3 Feature Selection

Feature selection is the process of selecting features that can contribute most to your desired output in which you are interested (Pant, H. & Srivastava, R. , 2015). The dataset having some irrelevant features can decrease the accuracy of the models and make your model learn based on irrelevant features. Feature selection is important in data mining, especially for high-dimensional datasets. There are many advantages of feature selection techniques, such as reducing the complexity of a model and making it easier to interpret, reducing overfitting, and improving the accuracy of a model if the right subset is chosen. The feature selection technique can be divided into three categories; filter methods, wrapper methods, and embedded methods (Pant, H. & Srivastava, R. , 2015). Filter methods are generally used as a pre-processing step. In this, features are selected based on their scores in various statistical tests for their correlation with the outcome variable. There are many filter methods, such as the information gain method, chi-square method, correlation coefficient method, etc. The filter-based approach is more efficient because this technique does not depend on the induction algorithm and requires less computation. Filter methods do not use any learning algorithm for feature selection. Wrapper methods evaluate all possible combinations of the features and select the combination that produces the best result for the algorithm. The wrapper approach intention produces a good result, but it depends on the learning algorithms. These methods are called greedy algorithms because they find the best combination of features that can improve the performance of learning algorithms. Sometimes, the features selected are not suitable for other classification algorithms. The wrapper approach normally has a high computational cost and risk of model over-fitting. Some common examples of wrapper methods are forward feature selection, backward feature elimination, recursive feature elimination, etc. Embedded methods combine the qualities of filter methods and wrapper methods. It is implemented by algorithms with built-in feature selection methods. The feature selection algorithm is integrated as a part of the learning algorithm. Embedded methods did the feature selection process within the construction of the machine learning algorithm itself (Maldonado, S. et al, 2014).

2.4 Genetic Algorithm

The genetic algorithm is based on the principles of Genetics and Natural selection (Lambora, A et al, 2019). The genetic algorithm is a heuristic search-based optimization technique and iterative technique. Genetic algorithms are a subset of a branch of computation known as Evolutionary Computation. Genetic algorithms are randomized in nature. Genetic algorithms are used to solve both constrained and unconstrained optimization problems, which are based on natural selection. Genetic algorithms are faster and more efficient compared to other traditional methods and have good parallel capabilities. Genetic algorithms provide a list of the best solutions for one problem. Genetic algorithms differ from classical algorithms in two main ways. Classical algorithms generate a single point at each iteration, whereas genetic algorithms generate a population of points in every iteration. Classical algorithms select the next point in the sequence, whereas genetic algorithms use a random generator to decide the next point. Genetic algorithms simulate natural selection, the procedure that drives biological evolution. The genetic algorithm constantly modifies a population of individual solutions (Lambora, A et al, 2019). The genetic algorithms randomly select individuals from the current population to be parents and use that parent to produce the children for the next generation in every step.

2.5 Forward Selection

Forward selection is a popular attribute selection method for classification (Jović, A. et al, 2015). It is a wrapper method and an iterative process. It is an attractive method because it is tractable and gives an optimal sequence of models. It starts with having no attribute in the model, and in each iteration, it keeps adding the feature that improves the learning model. It continues the addition

of the new feature until the model performance does not improve. Only the feature with the highest best performance is added to the selection. It is an extension of the Best First method. The method takes a fixed number of k attributes. The search uses the initial ordering to select the top k attributes or performs a ranking. In this technique, the search direction is selected in the forward selection or floating forward selection. The linear forward selection is a method to reduce the feature expansion in each forward selection iteration. The main advantage of linear forward selection is the reduction in computational efforts (Jović, A. et al, 2015).

III. LITERATURE SURVEY

3.1 Methods to Handle Imbalanced Dataset

Many standard algorithms fail to classify imbalanced data because the classification error in the majority class dominates the classification error in the minority class. The techniques to handle imbalanced datasets can be categorized as a data-level approach, algorithmic-level approach, and hybrid approach (Yusof, R. et al , 2017) (Ramyachitra, D., & Manikandan, P. , 2014).

3.1.1 Data-level Approach

In the data-level approach, the pre-processing step has been employed to balance the dataset before applying any classification algorithm. Data level methods are also known as external methods. The data level methods are classified into two types: sampling methods and Feature selection methods (Yusof, R. et al , 2017). Sampling methods try to balance the dataset by removing or eliminating instances of the majority class or adding instances into the minority class called under sampling and over-sampling respectively. Both methods provide an equal class distribution in an imbalanced dataset. The most encountered problem is the loss of important instances in the under-sampling. The main problem with over-sampling is that sometimes new instances may lead to over fitting of the classifier. From these basic sampling techniques, many techniques have been extended such as random sampling. Random sampling is a combination of both the under-sampling and over-sampling methods (Yusof, R. et al , 2017). Feature selection aims to reduce the unwanted or less related attributes or features from the dataset based on some criteria (Yuan, Z., & Zhao, P., 2019). Feature selection is a key step for many algorithms when the dataset is high dimensional. Because the class imbalance problem suffers from the issue of high dimensionality, hence feature selection is an essential technique. Feature selection methods are categorized into three types: filter methods, wrapper methods, and embedded methods. The filter methods use independent criteria to evaluate the subset of features. Wrapper methods try to evaluate all possible combinations of the features and select the optimal combination that produces the best result for the algorithm. Embedded methods are a combination of both filter and wrapper methods and, use the advantages of both methods. It's implemented by algorithms that have built-in feature selection methods.

3.1.2 Algorithmic Approach

Algorithmic methods employ the new design of classification algorithms or enhance the existing algorithms by applying some changes. The algorithmic approach aims to optimize learning performance. There are three algorithmic methods: one-class learning methods, cost-sensitive methods, and ensemble methods (Yusof, R. et al , 2017) (Ramyachitra, D., & Manikandan, P. , 2014). One-class classification always deals with a binary classification problem. It is a recognition-based approach, and here each class is learned separately. In this approach, training is done only on samples of one class (target class). The one-class learning is specifically used on the dataset that is extremely imbalanced and contains noisy features. One-class classification method can be used for binary imbalanced datasets where the negative case is considered normal, and the positive case is an outlier. The objective of this approach is to create a decision surface that covers all the available instances of the dataset. All the represented data outside the target class are labeled as outliers. This approach focuses on the data instances of a minority class as a target class while handling imbalanced datasets. Cost-sensitive learning is a part of machine learning that takes the costs of prediction errors when training a model. The cost is a penalty associated with an incorrect prediction. Cost-sensitive learning works based on the constructed cost matrix. Cost-sensitive learning is a technique to define different misclassification errors such as false positive and false negative patterns. Cost-sensitive learning aims to minimize the misclassification cost, test cost, and other type of cost. Cost-sensitive techniques can be divided into data re-sampling, algorithm modification, and ensemble methods. An ensemble technique is a technique that combines or aggregates more than one technique, such as two learning algorithms, a classifier with a sampling technique or feature selection, and others. Ensemble methods are categorized into two types, such as Bagging and Boosting methods. The bagging technique is also called Bootstrap Aggregating. Bagging uses bootstrap sampling to get the data subsets for training the base learners, and for aggregating the outputs of base learners it uses voting for classification. Boosting method is an iterative process. If instances from the dataset are misclassified in one round, then more cost is given to them in the next round.

3.1.3 Hybrid Approach

Hybrid methods are a combination of data-level methods and algorithmic-level methods (Ramyachitra, D., & Manikandan, P. , 2014). The hybrid approach is required to overcome the problems of the data-level methods and algorithmic-level methods and achieve better accuracy(Ramyachitra, D., & Manikandan, P. , 2014). We have designed a new hybrid algorithm that takes advantage of two or more existing algorithms.

In Table 1, I have compared all three above discussed approaches.

Table 1 Comparative analysis of existing methods

Method	Strength	Limitation
Under Sampling	It can be easily implemented	It suffers from the problem of data loss
Over Sampling	It can be easily implemented	It suffers from the problem of over fitting
Feature Selection	It can work with high-dimensional dataset. Reduce the irrelevant features from the dataset	More time required for implementation
Cost-sensitive Learning	This technique is good when the dataset is extremely imbalanced.	The misclassification costs often are unknown.
Ensemble Methods	Gives better classification performance than individual classifiers.	Requires more computation time

.

IV. PROPOSED MODEL

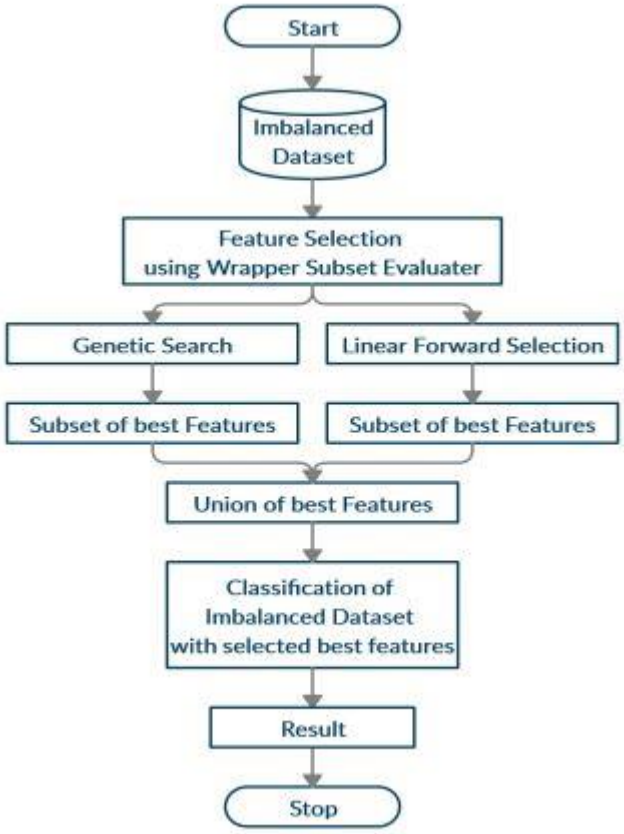

Figure 1: Proposed Model

The steps of Proposed Model are as follows:
Step1: Load the Imbalanced Dataset.
Step2: Select Optimal Feature using Wrapper Subset Evaluator.
Step3: Apply two Subset selection methods to identify optimal feature set. (1) Genetic Search (2) Linear Forward Selection
Step4: Identify optimal feature set from (1) Genetic Search and (2) Linear Forward Selection methods.
Step5: Take Union of the optimal features derived from (1) Genetic Search and (2) Linear Forward Selection methods.
Step6: Apply various classification algorithms on selected feature set.

V. EXPERIMENTAL DESIGN AND RESULTS

In this section, various experiments are conducted to test the performance of proposed model. The proposed model is applied on several binary classification problems. The proposed approach is compared with other traditional classification techniques in terms of overall classification accuracy and cross accuracy discussed below.

As discussed earlier, in classification of imbalanced data sets, overall classification accuracy is not a correct measure to compare different classifiers. In imbalance data sets, high classification accuracy of each class is desirable instead of high overall classification accuracy. In this research work, we have used term 'cross accuracy' to compare different classifiers shown in Eq. 1

Cross Accuracy = (Accuracy of Class 1) * (Accuracy of Class 2) * …… * (Accuracy of Class N) . (1)

To compute cross accuracy, we multiply accuracy of each class. So, if classifier shows very low accuracy for one of the class then the value of cross accuracy would be very low. To achieve high value of cross accuracy we need to achieve good classification accuracy for all the classes. For example if three classes are there in data set with following numbers of data instances. (Class 1: 100 Class 2: 50 Class3: 10) Here, Class 1 has majority training instances compare to Class 2 and 3. We will compare two classification results shown below.

Table 2 Representing Importance of Cross Accuracy over Overall Accuracy

	Pred class1	Pred class2	Pred class3		Pred class1	Pred class2	Pred class3
Actual class1	95	2	3	Actual class1	80	10	10
Actual class2	15	25	10	Actual class2	10	37	3
Actual class3	2	3	5	Actual class3	1	1	8
	Result1				Result2		

In Result1, overall accuracy = (95+25+5) / (160) = 0.78 and cross accuracy = (0.95*0.5*0.5) = 0.23 whereas in Result2, overall accuracy = (80+37+8) / (160) = 0.78 and cross accuracy = (0.95 * 0.74 * 0.80) = 0.47. In both results despite of having equal overall classification accuracy, result2 has better cross accuracy than Result1 as it gives better accuracy among all the classes of data set. So we have compared all classifiers using two measures (i) overall accuracy and (ii) cross accuracy. Solution having high overall accuracy along with high cross accuracy is considered better than solution having high overall accuracy only.

For experiment purposes, we used six binary (two-class) imbalanced datasets. All the datasets are taken from the KEEL (Knowledge Extraction based on Evolutionary Learning) toolkit (Mittal, P. , 2012). The description of the datasets used for experiments is shown in the following table 3.

TABLE 3 Imbalanced Datasets used for Experiments

Dataset Name	**Imbalance Ratio**	**Number of Features**	**Number of data Instance**
glass4	15.47	9	Positive: 13, Negative: 201,Total: 214
page-blocks-1-3_vs_4	15.86	10	Positive: 28, Negative: 444,Total: 472
vehicle2	2.88	18	Positive: 218, Negative: 628,Total: 846
glass-0-1-2-3_vs_4-5-6	3.2	9	Positive: 51, Negative: 163,Total: 214
page-blocks0	8.79	10	Positive: 559, Negative: 4913,Total: 5472
yeast-2_vs_8	23.1	8	Positive: 20, Negative: 462,Total: 482

TABLE 4 Experiments results comparing accuracy for various imbalanced datasets

	Dataset: glass4 Imbalance Ratio : [15.47]				
CLASSIFICATION ALGORTITHMS		Training Data		Testing Data	
		Cross accuracy	Overall accuracy	Cross accuracy	Overall accuracy
	Naive Bayes	0.145	89.71%	0.228	94.39%
	ANN	0.749	96.26%	0.753	96.72%
	Decision Tree	0.516	93.45%	0.667	94.85%
	Random Forest	0.678	96.26%	0.833	97.66%
	Logistic Regression	0.225	92.99%	0.231	95.32%
	Dataset: page-blocks-1- 3_vs_4 Imbalance Ratio : [15.86]				
SIFICATIO N ALGO RITH		Training Data		Testing Data	
		Cross	Overall	Cross	Overall

		accuracy	accuracy	accuracy	accuracy
	Naive Bayes	0.550	94.06%	0.964	96.61%
	ANN	0.962	99.57%	0.998	99.78%
	Decision Tree	0.813	98.09%	0.849	98.30%
	Random Forest	0.924	99.15%	0.998	99.78%
	Logistic Regression	0.666	96.39%	0.639	97.45%
	Dataset: Vehicle2		**Imbalance Ratio : [2.88]**		
CLASSIFICATION ALGORITHMS		Training Data		Testing Data	
		Cross accuracy	Overall accuracy	Cross accuracy	Overall accuracy
	Naive Bayes	0.483	79.43%	0.540	84.63%
	ANN	0.945	97.87%	0.955	98.22%
	Decision Tree	0.889	95.39%	0.894	95.50%
	Random Forest	0.969	98.81%	0.967	98.69%
	Logistic Regression	0.914	96.45%	0.921	96.69%
	Dataset: glass-0-1-2- 3_vs_4-5-6		**Imbalance Ratio : [3.2]**		
CLASSIFICATION ALGORITHMS		Training Data		Testing Data	
		Cross accuracy	Overall accuracy	Cross accuracy	Overall accuracy
	Naive Bayes	0.731	89.71%	0.791	91.58%
	ANN	0.811	93.45%	0.911	96.26%
	Decision Tree	0.793	90.65%	0.841	92.52%
	Random Forest	0.838	93.45%	0.887	95.32%
	Logistic Regression	0.774	92.52%	0.836	94.39%
	Dataset: page-blocks0		**Imbalance Ratio : [8.79]**		
CLASSIFICATION ALGORITHMS		Training Data		Testing Data	
		Cross accuracy	Overall accuracy	Cross accuracy	Overall accuracy
	Naive Bayes	0.428	90.20%	0.631	94.31%
	ANN	0.708	95.85%	0.742	96.03%
	Decision Tree	0.823	97.09%	0.843	97.35%
	Random Forest	0.858	97.38%	0.885	97.71%
	Logistic Regression	0.630	94.97%	0.630	94.97%
	Dataset: yeast-2_vs_8		**Imbalance Ratio : [15.47]**		
CLASSIFICATION ALGORITHMS		Training Data		Testing Data	
		Cross accuracy	Overall accuracy	Cross accuracy	Overall accuracy
	Naive Bayes	0.548	97.92%	0.598	98.13%
	ANN	0.548	97.92%	0.598	98.13%
	Decision Tree	0.499	97.71%	0.499	97.71%
	Random Forest	0.499	97.71%	0.548	97.92%
	Logistic Regression	0.548	97.92%	0.548	97.92%

VI. Conclusion

After analyzing the experimental results, we conclude that the class imbalance problem occurs when one class contains more data instances than another in the same dataset. The major challenge with the class imbalance problem is the prediction of minority class instances. The main goal of classifiers is to gain high accuracy, and because of this tendency sometimes they do not predict instances of minority class accurately in imbalanced datasets. We proposed a combination of two feature selection methods, namely genetic algorithm and linear forward selection, which gives the best set of features that helps to improve classifiers' performance on imbalanced datasets. We applied the proposed methodology to six imbalanced datasets. The experimental results suggest a hybrid combination of feature selection methods works well on different classifiers.

References

[1] He, H., & Garcia, E. A. (2009). Learning from imbalanced data. IEEE Transactions on knowledge and data engineering, 21(9), 1263-1284.

[2] Jiawei, H., & Micheline, K. (2006). Data mining: concepts and techniques. Morgan kaufmann.

[3] Jović, A., Brkić, K., & Bogunović, N. (2015, May). A review of feature selection methods with applications. In 2015 38th international convention on information and communication technology, electronics and microelectronics (MIPRO) (pp. 1200-1205). Ieee.

[4] Lambora, A., Gupta, K., & Chopra, K. (2019, February). Genetic algorithm-A literature review. In 2019 international conference on machine learning, big data, cloud and parallel computing (COMITCon) (pp. 380-384). IEEE.

[5] Maldonado, S., Weber, R., & Famili, F. (2014). Feature selection for high-dimensional class-imbalanced data sets using support vector machines. Information sciences, 286, 228-246.

[6] Mittal, P. (2012). Knowledge extraction based on evolutionary learning (KEEL): analysis of development method, genetic fuzzy system. Int. J. Comput. Appl. Inf. Technol, 1, 22-25.

[7] Pant, H., & Srivastava, R. (2015). A survey on feature selection methods for imbalanced datasets. International Journal of Computer Engineering and Applications, 9(2), 197-204.

[8] Ramyachitra, D., & Manikandan, P. (2014). Imbalanced dataset classification and solutions: a review. International Journal of Computing and Business Research (IJCBR), 5(4), 1-29.

[9] Spelmen, V. S., & Porkodi, R. (2018, March). A review on handling imbalanced data. In 2018 international conference on current trends towards converging technologies (ICCTCT) (pp. 1-11). IEEE.

[10] Yuan, Z., & Zhao, P. (2019, May). An improved ensemble learning for imbalanced data classification. In 2019 IEEE 8th joint international information technology and artificial intelligence conference (ITAIC) (pp. 408-411). IEEE.

[11] Yusof, R., Kasmiran, K. A., Mustapha, A., Mustapha, N., & MOHD ZIN, N. A. (2017). TECHNIQUES FOR HANDLING IMBALANCED DATASETS WHEN PRODUCING CLASSIFIER MODELS. Journal of Theoretical & Applied Information Technology, 95(7).