Performance Analysis of Ensemble Learning Methods with Base Classifiers for Data Mining

[1]Aman Chauhan, [2]A.J. Singh

[1]Research Scholar, [2]Professor
Department of Computer Science,
Himachal Pradesh University, Shimla, India

Abstract: In present scenario the rise of digital technology, the availability of raw data over the network has increased awfully and this data is continuously increasing day by day. The data which is generated by these technologies needs to be processed and analyzed in some efficient ways that the information thus extracted can help serving for better results. Data Mining is one of the hot topics in which researchers and academician putting an interest. Data mining offers various techniques and methods that may be used to predict the future trends and patterns in available data. There are various number of techniques which can be used to predict the class in which the particular data falls .Classification is used for analyzing the data and the comparison has been done over the set of parameters like correctly and incorrectly classified instances .The unbounded size and imbalance nature of data creates difficulty in Classification so there are ensemble methods to solve these problems. Problems of over-fitting and bias are also solved with Ensemble Learning Methods. This research focuses on analyzing the behavior of Ensemble Learning Methods (Bagging, Boosting) with the help of five base classifiers, i.e. Naïve Bayes, Decision Table, J48, OneR and Decision Stump over the varying size data. WEKA tool has been used for the experimental purpose and the parameters that helped in evaluating the performance are accuracy, error rate, precision and recall.

Keywords: DATA MINING, CLASSIFICATION, WEKA TOOL, ENSEMBLE LEARNING METHOD, BAGGING, BOOSTING.

I. INTRODUCTION

With the advancement and improvement in Digital technology, many companies and organizations usually gather huge amount of data from operational activities and after which raw data are left to waste in data repositories. This raw data is by itself of no relevance if it remains unprocessed. Data mining is all about the analysis of that large amount of data that are found in data repositories of many organizations. Data mining is the process of extracting unknown information from a huge amount of data through Knowledge discovery in databases (KDD) process. These KDD process in the data mining methods are used for extracting useful information or finding new patterns from the data. Numerous kinds of data mining tools such as WEKA, Knime and Rapid Miner etc. are available to analyze the data. These tools help us to analyze data with a collection of methods and techniques in their appropriate ways. WEKA tool has been used to analyze various classification techniques in this research. Also, the performance of Ensemble Learning Methods is analyzed with some base classifiers over the varying data sizes.

1.1 Classification:

Classification is the most common data mining technique which do primarily a data analysis task where the model is constructed in order to help predicting the class of data objects whose class labels are yet unspecified. It is the act of finding for a model that defines a class label in such a way that such a model can be used to predict an unknown class label. In simple words, classification is usually used to predict an unknown class label.

Classification techniques that have been focused in this paper:

1.1.1 Decision Table: Decision Table classifier is used to summarize the dataset by using a decision table containing same attributes as that of original. This method is used to numerically predict the data from decision tree. Decision table is a rule-based classifier which is an ordered set of IF-THEN rules that are much more compact and are much easier to comprehend than that of decision trees [2].

1.1.2 Naive Bayes: Naïve Bayes classifiers are the simplest classification algorithms and practically used when the dimensionality of input is high [3]. Naïve Bayes classifier is also used in complex situation that deals with large data size. It is purely based on Bayes Theorem and uses supervised learning technique**.**

1.1.3 Decision Stump: Decision Stump Classifier is the type of Decision Tree Classification Algorithm in which single input feature is enough to make the prediction of new data. [4].

1.1.4 J48: J48 classifier is a type of Decision Tree method which is used to create a model that predicts value of target variables in terms of multiple input variables. C4.5 algorithm which builds decision tree is implemented in WEKA tool as a classifier known as J48.

1.1.5 OneR: OneR Stands for One-Rule, is a simple but accurate classification algorithm that generates one rule for each predictor in data and selects the rule with the smallest total error as its "one rule". In simple words it uses the minimum error attribute for prediction.

1.2 Ensemble Learning Methods

Ensemble learning is also known as learning multiple classifiers system or committee-based learning. Multiple learning algorithms are used by Ensemble methods together for doing the same task to get better results than individual learning model. Ensemble learning is also known as committee-based learning or Meta learning techniques which gives better prediction than individual model. An ensemble includes no. of learners called as base learners. These base learners are produced by base learning algorithms such as neural network, decision tree or any other type of learning algorithms from training dataset. The advantages of using these methods leads to Better Accuracy (low error), Higher Consistency (avoid overfitting), Reduced Variance, Reduced Bias and Improved Prediction [5].

Ensemble Learning Methods that have been focused in this paper:

1.2.1 Bagging

Bagging is an Ensemble algorithm which improves the accuracy and stability of learning algorithms used in regression and statistical classification. The issue of overfitting and reducing variance are avoided by using bagging [5]. Bagging is also known as Bootstrap Aggregation.

1.2.2 Boosting

Boosting is a little variation of bagging. It emphasis on selecting points which gives wrong prediction in order to give accuracy. Boosting is a vigorous ensemble algorithm which is capable of reducing both bias & variance, and helps in converting weak learners (i.e., classifiers with weak correlations to the true classification) into strong learners (i.e., well-correlated classifiers).

II. RELATED WORK

In this section, we introduce some of the related works that have been done for performance of ensemble learning methods.

Eibe Frank et al. (2004) did a study on the WEKA tool and introduce the machine learning workbench which provides an environment for clustering, classification and regression feature selection. The study for data exploration and the experimental comparison of different machine learning techniques are contained in extensive collection of machine learning algorithms and data pre-processing methods complemented by (GUI) graphical user interfaces. Data given in the form of single relational table can be easily processed by WEKA. The objectives behind the study are to guide users to extract useful information from data by identifying suitable algorithm.[6]

S.B. Kotsiantis et al. (2006) proposed boosting technique for localized weak learners. Boosting methods were used for regression and classification problems that works locally. A comparison is executed with other established combining methods on standard classification and regression benchmark datasets in which decision stumps acts as base learner and the purposed technique give the efficient result [7].

Swati singhal et al. (2013) studied WEKA tool and introduce a brief introduction of the key principle of data-preprocessing, classification, clustering and introduction of WEKA tool and described various steps how to use WEKA tool for particular technologies.[8].

Priyanka kumara et al. (2018) studied "Analysis of credit card fraud detection using fusion classifiers". They analyzed some ensemble classifiers such as Random forest, Bagging, classification by regression, voting and compared them with some effective single classifiers like K-NN, naïve Bayes, SVM, RBF classifiers, MLP, Decision Tree. The evaluation of these algorithms was carried out through three different datasets and treated with SMOTE, to deal with the class imbalance problem. The comparison was based on metrics like accuracy, precision, true positive, true positive rate or recall, and false positive rate. They have concluded that there is no single classifier in data mining that can perform better than the ensemble classifiers. The classification via Regression ensemble classification technique performs well on both German Data with accuracy of 95.21% and Australian data with accuracy of 91.17% [9].

Kuldeep Randhawa et al. (2018) analyzed the AdaBoost ensemble techniques and Majority Voting techniques for credit card fraud detection. In this work, firstly a publicly available data is analyzed over some base classifiers and then the Meta techniques like AdaBoost and Majority Voting techniques are applied. Then, these techniques are implemented over the real-world data and the results are analyzed. It was indicated that the majority voting technique achieves good accuracy rate than adaboost in detecting the frauds [10].

III. WEKA TOOL

In the presented paper, the experimentation has been done by using WEKA tool (3.8.2 version). WEKA stands for Waikato Environment for Knowledge Analysis which is developed by The University of Waikato, Hamilton, New Zealand. WEKA is an open source and influential tool for data mining. ARFF (Attribute Relation File Format) file format is supported in WEKA. Some other formats like CSV, C 4.5 data files are also converted into ARFF format with the help of WEKA tool. WEKA has a vast

collection of algorithms for: Classification, Regression, Clustering, Association, Data preprocessing and Visualization. WEKA also provides interfaces like Explorer, Experimenter, Knowledge Flow, Simple CLI and Workbench [11].

IV. RESULTS AND DISCUSSION

WEKA tool has been used for analyzing the behavior of varying size data. Ensemble learning algorithms along with Different classification algorithms have been analyzed using the 10-fold cross validation method.

4.1 Datasets

The datasets used have been downloaded from the UCI repository [12] and Kaggle [13] websites. Three varying size datasets have been used.

Table 4.1: Description of Datasets

Dataset Name	Abbreviation Used	Data Types	Default Task	Instances	Attributes	Size
Hepatitis [14]	DS1	Multivariate	Classification	155	20	16.7KB
Chronic Kidney Disease [15]	DS2	Multivariate	Classification	400	25	43.6KB
Diabetes [16]	DS3	Multivariate	Classification	768	9	36.5KB

- In dataset 1, Hepatitis is to be classified and causes for it, is analyzed on the bases of different attributes.
- In dataset 2, prediction of the early onset of chronic kidney disease is done.
- In dataset 3, diabetes is classified and occurrence on age range and some other parameters.

4.2 Parameters

Classification algorithms are evaluated by using confusion matrix. The columns signify class prediction while the rows signify actual class.
TN: TN (True negative) denotes number of correctly classified negatives patterns.
FP: FP (False positive) denotes number of misclassified negative patterns predicted as positives.
FN: FN (False negative) denotes number of positive patterns that are misclassified as negative.
TP: TP (True positive) denotes the number of correctly classified positive patterns.

4.2.1 Accuracy

Accuracy is overall performance of the classifier which is defined as percentage of number of correctly classified instances.

$$Accuracy = \frac{TP+TN}{TP+TN+FN+FP}$$

4.2.2 Error Rate

The frequency of occurring of an error refers to Error rates and also defined as “the ratio of the total number of data units in error to the total number of data unit transmitted.” As the error rate increases, the data transmission reliability decreases. [14]

$$Error\ rate = 1\text{-}accuracy$$

$$OR$$

$$FP+FN/\ TP+FP+TN+FN$$

.

4.2.3 Precision

The closeness of two or more measurements with each other is known as Precision. It is also known as positive predictive value.

$$Precision = \frac{TP}{TP + FP}$$

True Positives (TP): Number of instances predicted positive that are actually positive.

False Positive (FP): Number of instances predicted positive that are actually negative.

4.2.4 Recall

The percentage of total relevant results correctly classified refers to Recall. It is the true positive rate (also referred to sensitivity). It can be represented as:

$$Recall = \frac{TP}{TP + FN}$$

True Positive (TP): Number of instances predicted positive that are actually positive.

False Negative (FN): Number of instances predicted negative that are actually positive.

4.3 Comparing the Performance of Ensemble Algorithms on the Basis of the Varying Size of Datasets

- **Accuracy**

Table 4.2 shows the accuracy rates of various algorithms. The results show that:

- In case of Base classifiers: Naïve Bayes gives high accuracy for DS1, DS3. J48 and Decision Table gives high accuracy for DS2.
- In case of Boosting classifiers: Decision Table gives high accuracy for DS1 and Decision stump gives high accuracy for DS2, DS3.
- In case of Bagging: Decision Stump gives better accuracy for DS1 and Decision Table gives better accuracy for DS3.While DS2 gives similar accuracy for all classifications.

Table 4.2: Accuracy using Base Classifiers and Ensemble Algorithms

Algorithm	DS1	DS2	DS3
Naïve Bayes	84.52	95	76.3
Boosting (Naïve Bayes)	85.81	98	76.56
Begging (Naïve Bayes)	85.81	95.5	76.56
Decision Table	76.13	99	71.22
Boosting (Decision Table)	89.55	99.5	71.22
Begging (Decision Table)	80.65	98.75	75.78
J48	83.87	99	73.83
Boosting(J48)	85.81	99.5	72.4
Begging(J48)	83.87	98.75	74.61
OneR	80	92	71.48
Boosting (OneR)	80	97.5	70.18
Begging (OneR)	83.87	92.5	70.96
Decision Stump	77.42	92	71.88
Boosting (Decision Stump)	82.58	99	74.35
Begging (Decision Stump))	81.94	92.75	55.34

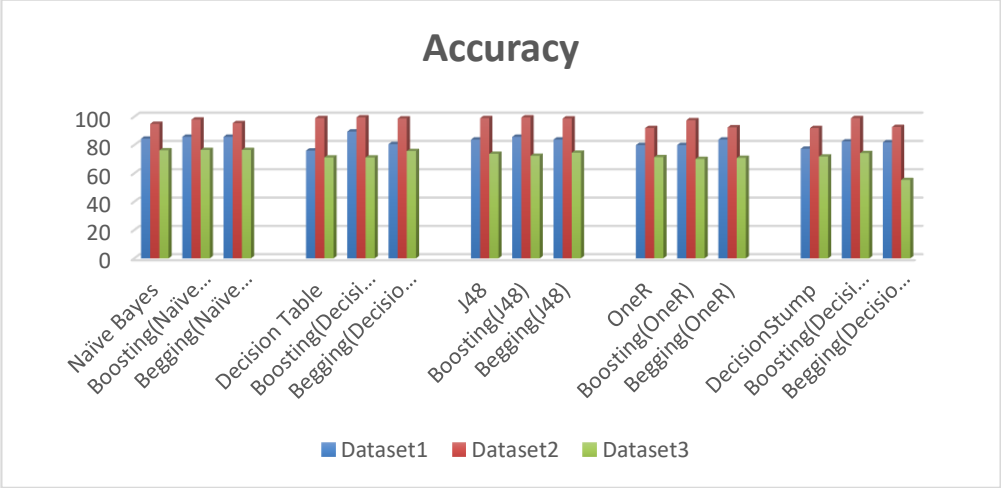

Fig. 4.1: Graphical Representation of Accuracy

- **Error Rate**

Table 4.3 shows the Error Rate of various algorithms. The results show that:

- In case of Base Classifiers: Naïve Bayes gives low error rate For DS1 and DS3. Decision Table gives low error For DS2.
- In case of Boosting: Decision Table gives low error rate for DS1, Decision Stump for DS2 and J48 for DS3 also gives low error rate.
- In case of Bagging: Decision Table and Decision Stump gives low error rate for DS1, Decision Stump for DS2 and Decision Table for DS3.

Table 4.3: Error Rate using Base Classifiers and Ensemble Algorithms

Algorithms	**DS 1**	**DS 2**	**DS3**
Naïve Bayes	15.48	5.0	23.70
Boosting (Naïve Bayes)	14.19	2.0	23.44
Bagging (Naïve Bayes)	14.19	4.50	23.44
Decision Table	23.87	1.0	28.78
Boosting (Decision Table)	10.45	.50	28.78
Bagging (Decision Table)	19.35	1.25	24.22
J48	16.13	1	26.17
Boosting (J48)	14.19	.5	27.60
Bagging (J48)	16.13	1.25	25.39
OneR	20.00	8.00	28.52
Boosting (OneR)	20.00	2.50	29.82
Bagging (OneR)	16.13	7.50	30.04
Decision Stump	22.58	8.00	28.13
Boosting (Decision Stump)	17.42	1.00	25.65
Bagging (Decision Stump)	18.06	7.25	44.66

Fig. 4.2: Graphical Representation of Error Rate

- **Precision**

Table 4.4 shows the precision values of various algorithms. The results show that:

- In case of Base Classifiers: Naïve Bayes resulted in improving the precision for DS1 AND DS3, Decision Table and J48 Improves for DS2.
- In case of Boosting: Decision Table improves precision for DS1 while Decision Stump improves precision for DS2 and DS3.
- In case of Bagging: Naïve Bayes improves precision for DS1 and Decision Stump improves precision for DS2, DS3.

Table 4.4: Precision using Base Classifiers and Ensemble Algorithms

Algorithms	**DS 1**	**DS 2**	**DS 3**
Naïve Bayes	91.60	88.24	67.77
Boosting (Naïve Bayes)	89.15	94.94	68.64
Bagging (Naïve Bayes)	93.16	89.29	68.64
Decision Table	84.13	99.32	59.92
Boosting (Decision Table)	98.23	100	60
Bagging (Decision Table)	82.07	92.39	69.90
J48	86.57	99.32	63.24
Boosting (J48)	88.55	99.33	60.37
Bagging (J48)	86.57	96.77	64.54
OneR	83.33	86.88	63.39
Boosting (OneR)	83.82	96.05	58.99
Bagging (OneR)	84.51	87.50	64.90
Decision Stump	79.73	84.71	60.16

Boosting (Decision Stump)	88.10	97.40	65.78
Bagging (Decision Stump)	81.88	86.23	64.97

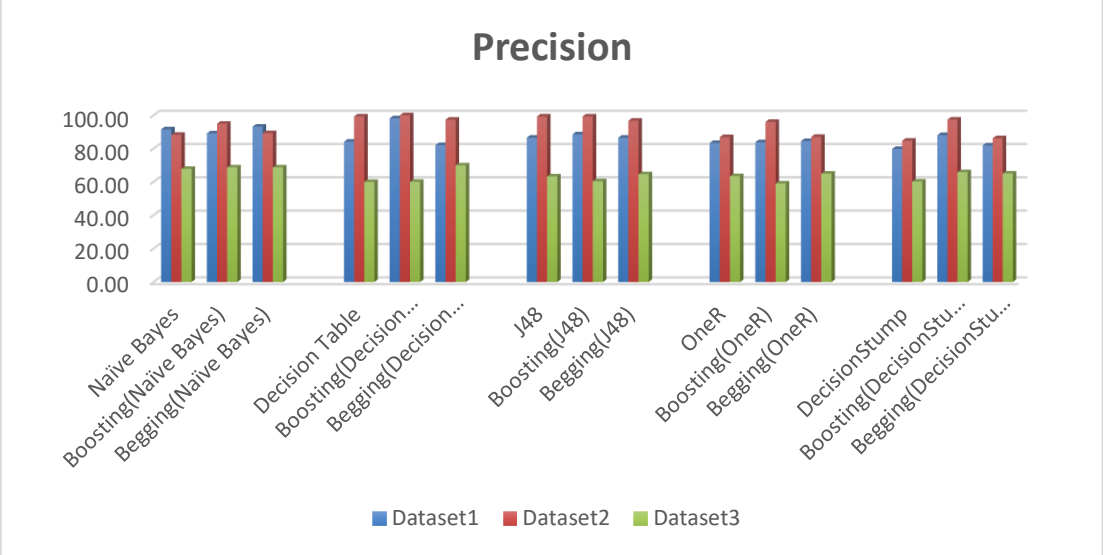

Fig. 4.3: Graphical Representation of Precision

- **Recall**

Table 4.5 shows the recall value of various algorithms. The results show that:

- In case of Base Classifier: Decision stump and J48 improves recall for DS1 and Naïve Bayes Improves recall for DS2 and DS3.
- In case of Boosting: Naïve Bayes improves the recall for DS1 and OneR improves recall for DS2 and DS3.
- In case of Bagging: Decision Table improves recall for DS1 and DS3. While J48 improves recall for DS2.

Table 4.5: Recall using Base Classifiers and Ensemble learning Algorithms

Algorithms	**DS 1**	**DS 2**	**DS 3**
Naïve Bayes	88.62	100	61.19
Boosting (Naïve Bayes)	93.5	100	60.45
Bagging (Naïve Bayes)	88.62	100	60.45
Decision Table	86.18	98	52.99
Boosting (Decision Table)	90.24	98.67	52.61
Bagging (Decision Table)	96.75	99.33	53.73
J48	94.31	98	59.7
Boosting (J48)	94.31	99.33	60.82
Bagging (J48)	94.31	100	60.45
OneR	93.5	92.67	43.28
Boosting (OneR)	92.68	97.33	47.76

Begging (OneR)	97.56	93.33	36.57
Decision Stump	95.93	96	57.46
Boosting (Decision Stump)	90.24	100	55.22
Bagging (Decision Stump)	99.19	96	47.76

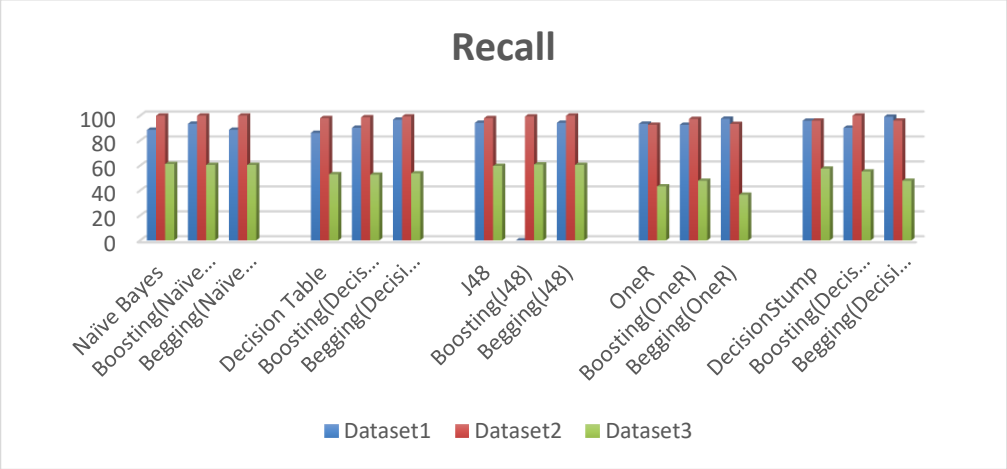

Fig. 4.4: Graphical Representation of Recall

CONCLUSION AND FUTURE SCOPE

In this work, the performance of five base classifiers i.e. Naïve Bayes, Decision Table, OneR, Decision Stump and J48 along with two Ensemble Learning techniques i.e. Bagging, Boosting have been analyzed with respect to the three datasets of varying sizes. All these algorithms have been compared on the basis of parameters like accuracy, error rate, precision and recall. Ensemble learning methods are used for the same task in order to have better prediction than that of the individual base classifier. The advantages of using Ensemble Learning methods leads to improved prediction, reduced variance and reduced bias. It has been observed form the results that J48 and Decision Table algorithm gave the highest accuracy rate among the base classifiers and the Ensemble methods helped improving the performance of J48 and Decision Table algorithm. From the results this has been observed that in majority of the cases Ensemble Learning algorithms have the highest accuracy rates, least error rate values except in Decision stump classifier which gives low accuracy rate, high error rate in DS3 after Bagging and high rates of Precision shows little decline in OneR classifier for DS3 after Bagging .In Recall parameter Decision Stump classifier gives low recall value after Bagging for DS3 when compared to various base classification algorithms .In some cases, Bagging and Boosting had the same values as that of the base algorithms but they never decreased the performance of base algorithms by a such a significant amount, though there were some negligible decreases at times. With respect to the varying size of the data, it has been observed that the accuracy rate of the algorithms does not increases that much with the increase in the data size but decreases a bit when the data grows at a larger scale; same behavior is experienced when the precision and recall rates are taken into consideration. However, few exceptional behaviors were also noticed. From our observation it is concluded that the behavior of classifier varies with the type of dataset rather than the increasing size of data sets. In case of errors, error rates decrease with the increasing size of data.

For the future scope, some new algorithms can be implemented on different tools like Rapid Minor, Orange instead of WEKA and their performance can be analyzed and improved with some other multiple learning techniques. More parameters also be taken for future work to check the variations in the results. The impact of change in the number of folds of cross-validation can also be observed.

REFERENCES

[1] Sagar S. Nikam, "A Comparative Study of Classification Techniques in Data Mining Algorithms", Oriental Journal of Computer Science and Technology, Vol 8(1), pp 13-19,2015.

[2] V. Krishnaiah, Dr. G. Narsimha and Dr. N. Subhash Chandra, "Survey of Classification Techniques in Data Mining", International Journal of Computer Sciences and Engineering, Vol 2(9), pp 65-74, 2014.

[3] Sagar S.Nikam, " A Comparitive Study of Classification Techniques in Data mining Algorithms," Oriental Journal of computer Science and technology, vol .8(1), pp 13-19,2015.

[4] V. Krishnaiah et al., "Survey of Classification Techniques in Data Mining," International Journal of Computer Sciences and Engineering, vol. 2(9), pp 65-74, 2014

[5] Data mining in bioinformatics using Weka", Eibe Frank Mark Hall Len Trigg Geoffrey Holmes Ian H. Witten,Bioinformatics, Volume 20, Issue 15, 12 October 2004, Pages 2479–2481
[6] Data mining in bioinformatics using Weka", Eibe Frank Mark Hall Len Trigg Geoffrey Holmes Ian H. Witten,Bioinformatics, Volume 20, Issue 15, 12 October 2004, Pages 2479–2481
[7] S.B. Kotsiantis et al., "Local Boosting of Decision Stump for Regression and Classification Problem," Journal of Computer, 2006
[8] Swati Singhal and Monika Jena, A Study on WEKA Tool for Data Preprocessing ,Classification and Clustering, International Journal of Innovative Technology and Exploring Engineering(IJITEE),Vol 2(6),2013
[9] Kumari, Priyanka, and Smita Prava Mishra. "Analysis of Credit Card Fraud Detection Using Fusion Classifiers." Computational Intelligence in Data Mining. Springer, Singapore, 2019. 111-122.
[10] Kuldeep Randhawa et al., "Credit Card Fraud Detection using AdaBoost and Majority Voting", IEEE, Vol 6, 2018
[11] WEKA, the University of Waikato, Available at: http://www.cs.waikato.ac.nz/ml/weka/
[12] UCI Repository, available at: https://archive.ics.uci.edu
[13] Kaggle Datasets, available at: https://www.kaggle.com
[14] https://archive.ics.uci.edu/ml/datasets/hepatitis
[15] https://archive.ics.uci.edu/ml/datasets/chronic_kidney_disease
[16] https://storm.cis.fordham.edu/~gweiss/data-mining/weka-data/diabetes.arff