IOSR Journal of Computer Engineering (IOSR-JCE)
e-ISSN: 2278-0661,p-ISSN: 2278-8727, Volume 21, Issue 3, Ser. VI (May - June 2019), PP 23-30
www.iosrjournals.org

Study on the Performance of Classification Algorithms for Data Mining

Md. Humayun Kabir
Department of Computer Science and Engineering, Jahangirnagar University, Bangladesh
Corresponding Author: Md. Humayun Kabir

Abstract: *This paper explores briefly some classification algorithms and their application in data mining to discover data mining models by analyzing a training data set. Various classifier models e.g., Naïve Bayes classifier, two functional models- Multilayer perceptron and SMO model, three decision tree models- ID3, J48 and Random Forests are generated from the training data set. A decision tree model represents the decision making knowledge embedded within a database as classification hierarchy to determine the class label of a particular data record. A car evaluation database is used in investigating the functioning of the various classifiers. The research work demonstrates how efficiently the classification algorithms can identify the similarities and differences among different car objects by analyzing their characteristics represented with the attribute values within the training data set in building the models using Weka data mining tool. The generated models can be applied on the test data set in predicting unknown class labels. A comparative analysis of the performance of the classifiers is also presented.*

Keywords: *Classifier, Training data set, Decision making knowledge, Data mining model, Class label*

Date of Submission: 16-06-2019 Date of acceptance: 02-07-2019

I. Introduction

In the modern technology world, a variety of modern vehicles are used for transportation in the roads and highways, air and water ways both in the private and public sectors. Millions of data records can be collected from the car sales and travelling records, which can be stored in databases for later analysis using intelligent software systems for decision making. Data mining [1-3] can be used to extract knowledge by analyzing transport data set using data mining algorithms. Various data mining methods and algorithms are used for data analysis. Classification [1-6] is a data mining method frequently used in Knowledge Discovery in Databases (KDD) processes. Classification is a supervised approach which is applied on a training data set to build a classification model. Among the various models, decision tree, neural network, and if-then rules are commonly used. Several classification algorithms have been investigated to study their performance and several improvements are proposed in [4-11]. The information and knowledge hidden in the training data set stored in a database can be extracted using classification algorithms e.g., ID3, C4.5, CART [9] and Naive Bayes to see insight into the data records to build classification models which can be applied on the test data set of unknown class labels to predict their class labels. In the decision tree based classification algorithms [1-6], the learning algorithm is applied to discover the optimized decision tree model by analyzing the attribute values by applying attribute test condition to select the best attributes of the training database to use as the nodes of the decision tree to be generated using each of the database partitions.

In this paper, a car evaluation database of UCI repository [12, 13] is analyzed by applying six classifiers using Weka [14] data mining tool to study their performance in data mining. In this study, classification efficiency, time requirement, precision etc. of the selected classifiers are obtained using Weka by analyzing the car evaluation database of 1728 instances, and a comparative analysis is presented.

The paper is organized as follows. Section II provides a brief overview of some data mining classification algorithms. In section III, a car evaluation database is analyzed using some selected classifiers to discover classification models using Weka data mining tool and presents a comparative analysis of various features of accuracy by class which are obtained for each classifier. Section IV presents results and performance analysis of the selected classifiers. Finally, section V concludes with limitation, potential application and future work.

II. Classification Algorithms

The decision tree based algorithm ID3 [2] was presented by J. R. Quinlan and later he presented C4.5 algorithm in 1993 [4]. The ID3 algorithm is applied on forensic data, and several improvements have been proposed in [5]. An improved version of C4.5 algorithm of higher efficiency is proposed in [6] for large data set.

Data mining technique employs many methods, e.g., Decision tree, Bayesian classification, Neural network etc. for classification of training dataset. A new improved synthesized data mining algorithm CA for high dimensional dataset using Principal Component Analysis (PCA), CURE and C4.5 is presented in [9]. Detail investigation and applications of C4.5 are presented in [4, 6, 10]. A few research works [15, 16] investigate the relationship and application of association rules and classification in data mining.

Decision tree method employs decision tree construction from large data set using decision tree induction algorithms. Several algorithms e.g., ID3, C4.5, CART build decision tree model for classification [9]. The ID3 algorithm [2] of J. R. Quinlan based on decision tree induction constructs a decision tree model by applying machine learning for data mining. The algorithm works by partitioning a training data set into subsets by choosing an attribute depending on attribute-based test [2]. The branches of the decision tree are formed with the outcomes of the test on the selected attribute by placing the partitions at different branches. The process terminates when a correct decision tree is formed for the records of the training data set [2].

In the decision tree method, a partition containing the data records belonging to the same class is used to form a leaf of the decision tree, and the leaf is labelled with that class. Each path starting at the root that ends up at a leaf node following a branch, represents a class rule [2, 5] for determining a class satisfying some attribute-value relationships for some particular data records. It is expected that the number of unique classes that can be defined from the decision tree model generated from the training data set are equal to the number of branches labelled with unique conditions with which the decision tree terminates i.e., *n* branches, with the labelled leaf. Thus, the constructed decision tree model can be used to determine *n* different classes of the test data set. J48 algorithm is an extension of ID3 algorithm, and it is a Java implementation of C4.5 classification algorithm [11] in Weka.

Naive Bayes Classifier applies Bayes theorem for classification on training data set [8, 17]. Random Forest [18] are based on random vectors that are generated for the growth of trees. Its classification accuracy is resulted from an ensemble of growing trees leading to the most popular class. These are a combination of tree predictors where each tree depends on the values of a random vector [18]. A random forest [18] is a collection of tree structured classifiers. A tree is grown using the training set and the random vector. Various research works [19, 20, 21] using classification and decision tree have been presented to investigate the application of classification algorithms in different real life systems to analyze the relevant database to discover decision tree models, e.g., students behavior and failure analysis. Sequential Minimal Optimization (SMO) algorithm [22, 23] can be used to train support vector machines (SVMs), which requires the solution of a very large Quadratic Programming optimization problem. The Multilayer Perceptron [24, 25] works using back-propagation algorithm, which consists of hidden layers within the input and output layers.

III. Analysis of a Car Evaluation Database Using Data Mining Classifiers

In this paper, the data analysis task is performed using a car evaluation database [12, 13] consisting of 1728 instances of a car data set with 7 attributes by applying data mining classifiers using Weka data mining tool. Among the 7 attributes, first six attributes about the buying price, maintenance cost, number of doors, number of persons to carry, size of the luggage boot, and the safety level are used as input attributes [12, 13]. The seventh attribute *class* is used as the classification output to represent 4 car classes, i.e., unacc, acc, vgood and good based on the various car features suitable for a particular group of customers. TABLE I, II, and III show the classification performance obtained by executing Weka 3.4.3 and Weka 3.8.3 [14] on the car evaluation database by applying Naive Bayes, ID3, J48, Random Forests, SMO and Multilayer Perceptron classifiers. For experimental purpose, the .arff file was obtained by using Weka 3.4.3, and for data mining, all of the classifiers used are of Weka 3.8.3, and only ID3 classifier is used using Weka 3.4.3. The data analysis is performed using a computer system with Intel Core i5 of 2.4 GHz, RAM 4.0 GB with HDD 1 TB executing Weka [14] data mining tool under Windows environment by applying the classification algorithms as shown in TABLE I.

As an example, in percentage calculation, 88.83% of the 1728 instances are correctly classified, 4.17% are incorrectly classified, where 7% are left unclassified using ID3 algorithm for Fold = 4. Similarly, the percentage calculation can be obtained for Fold = 5. The Multilayer Perceptron algorithm has the highest number of correctly classified instances than that of all other algorithms for both Fold values 4 and 5, which is 99.02% approximately. Naive Bayes algorithm has the lowest percentage of correctly classified instances for Fold = 5, which is approximately 85.19%. For both of the Fold values, Multilayer Perceptron takes the highest computation time, e.g., 5.31 second for Fold = 4, whereas Naive Bayes and J48 algorithms require minimum computation time among the classifiers excluding ID3 as shown in TABLE I. Random Forest classifier has very small classification time 0.08 sec. which is nearly the half of the classification time required by SMO classifier. Though ID3 classifier has also nearly the smallest classification time 0.02 sec., it left some instances unclassified- 7% and 6.31% for Folds 4 and 5 respectively. TABLE I demonstrates that Fold numbers has no large impact on classification time required for each classification algorithm.

Table I. Classification Statistics for Different Classifiers Applying on the Car Evaluation Database [12, 13] Using Weka [14].

Classification Algorithm	Fold	Correctly classified		Incorrectly classified		Unclassified		Classification Time (sec.)
		Instances	% (approx.)	Instances	% (approx.)	Instances	% (approx.)	
Naive Bayes	4	1487	86.05	241	13.95	0	0	0.02
Naive Bayes	5	1472	85.19	256	14.81	0	0	0.01
ID3	4	1535	88.83	72	4.17	121	7.00	0.02
ID3	5	1536	88.89	83	4.80	109	6.31	0.02
J48	4	1574	91.09	154	8.91	0	0	0.02
J48	5	1582	91.55	146	8.45	0	0	0.01
Random Forest	4	1621	93.81	107	6.19	0	0	0.08
Random Forest	5	1617	93.58	111	6.42	0	0	0.08
SMO	4	1608	93.06	120	6.94	0	0	0.14
SMO	5	1610	93.17	118	6.83	0	0	0.14
Multilayer Perceptron	4	1711	99.02	17	0.98	0	0	5.31
Multilayer Perceptron	5	1711	99.02	17	0.98	0	0	5.22

TABLE II shows that the structural properties of the classification tree generated using tree type classifier J48 do not change for a particular algorithm with the changes in fold numbers.

Table II. Classification Tree Properties for Tree Type Classifier J48 Using Weka [14] Applying on the Car Evaluation Database [12, 13].

Classification Algorithm	Fold	Class Labels	Size of the Tree	Number of Leaves
J48	4	4	182	131
J48	5	4	182	131

TABLE II shows that the number of class labels, tree size and the number of leaves in the generated decision tree do not change with Fold numbers as demonstrated by J48 decision tree classification algorithm. TABLE III shows the values for some computed features e.g. TP Rate, FP Rate, Precision, Recall and F-Measure of accuracy by Class where the features have their usual meaning [14, 26] for various data mining classifiers with Fold = 4 obtained by analyzing the car evaluation database [12, 13] using Weka [14] for 4 car Classes. For Fold = 5, there is a small change in the computed values of the features for each algorithm, and the detail is skipped here for simplicity.

IV. Results and Performance Analysis

In this paper, a car evaluation database of 1728 instances consisting of 7 attributes of a car is analyzed using the classification algorithms Naive Bayes, ID3, J48, Random Forests, SMO and Multilayer Perceptron for comparative performance analysis of the algorithms. Features representing the classification performance of six classifiers are determined as percentage of total number of instances with time taken to build the model in second(s) for Fold = 4 and 5 using Weka [14] for each of the 4 car classes: *acc* for accepted, *unacc* for unaccepted, *good* for good and *vgood* for very good classes [12, 13], which are summarized in TABLE I. TABLE II shows the structural properties of the classification decision tree model generated using the tree type classifier J48 using Weka. Several classification features of accuracy for each of the 4 car classes for Fold = 4 are summarized in TABLE III. Classification performance, classification accuracy by class, and precision of various classifiers are shown using bar charts for each of the 4 car classes for comparative analysis, which are graphically shown and explained below.

Table III. Some Computed Features of Accuracy by Class for Fold = 4 for Various Classifiers Using Weka [14] Applying on the Car Evaluation Database [12, 13].

Algorithm	Class	TP Rate	FP Rate	Precision	Recall	F-Measure
Naive Bayes	Unacc	0.960	0.185	0.924	0.960	0.941
	Acc	0.719	0.096	0.681	0.719	0.700
	vgood	0.492	0.001	0.941	0.492	0.646
	Good	0.261	0.008	0.563	0.261	0.356
ID3	unacc	0.972	0.029	0.990	0.972	0.981
	Acc	0.922	0.026	0.897	0.922	0.909
	vgood	0.907	0.006	0.796	0.907	0.848
	Good	0.800	0.010	0.692	0.800	0.742
J48	unacc	0.956	0.066	0.971	0.956	0.964
	Acc	0.867	0.060	0.806	0.867	0.836
	vgood	0.800	0.014	0.684	0.800	0.738

	Good	0.464	0.010	0.667	0.464	0.547
Random Forests	unacc	0.973	0.050	0.978	0.973	0.976
	Acc	0.901	0.039	0.867	0.901	0.884
	vgood	0.846	0.008	0.797	0.846	0.821
	Good	0.623	0.008	0.754	0.623	0.683
SMO	unacc	0.959	0.058	0.975	0.959	0.967
	Acc	0.883	0.050	0.835	0.883	0.858
	vgood	1.000	0.006	0.867	1.000	0.929
	Good	0.638	0.008	0.772	0.638	0.698
Multilayer Perceptron	unacc	0.999	0.006	0.998	0.999	0.998
	Acc	0.979	0.004	0.984	0.979	0.982
	vgood	0.985	0.002	0.955	0.985	0.970
	Good	0.899	0.003	0.925	0.899	0.912

Fig. 1 and Fig. 2 graphically show the classification performance of various classifiers for Fold = 4 and Fold = 5 respectively based on TABLE I.

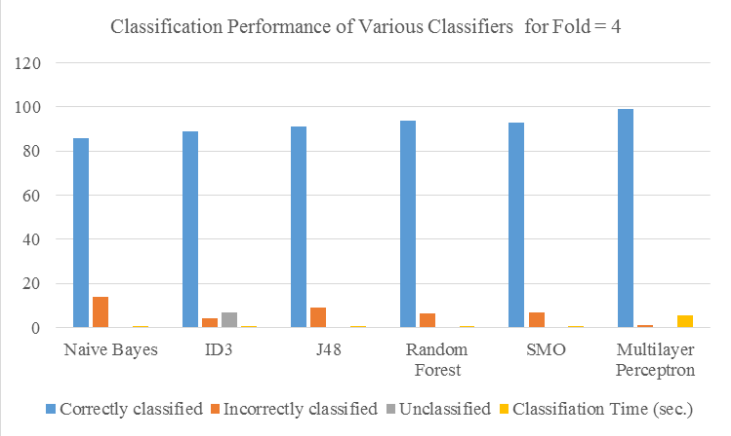

Fig. 1: Classification Performance of Various Classifiers Computed in Percentage of Total Number of Instances with Time Taken to Build the Model for Fold = 4 based on TABLE I Obtained Using Weka [14].

The charts shown in Fig. 3 to Fig. 8 demonstrates the accuracy by Class for the computed accuracy values shown in TABLE III obtained by applying different classifiers as labelled in the corresponding figure for Fold = 4 using Weka [14] on the car evaluation database [12, 13].

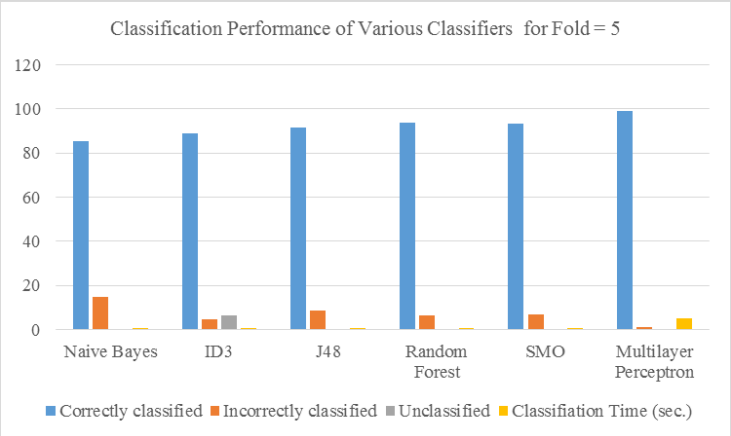

Fig. 2: Classification Performance of Various Classifiers Computed in Percentage of Total Number of Instances with Time Taken to Build the Model for Fold = 5 based on TABLE I Obtained Using Weka [14].

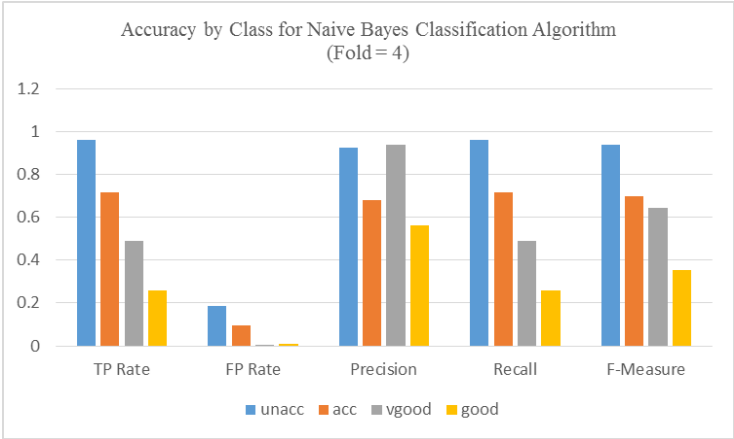

Fig. 3: Accuracy by Class for Naive Bayes Classifier (Fold = 4) Using the Computed Accuracy Values of TABLE III Obtained by Applying the Classifier on the Car Evaluation Database [12, 13] Using Weka [14].

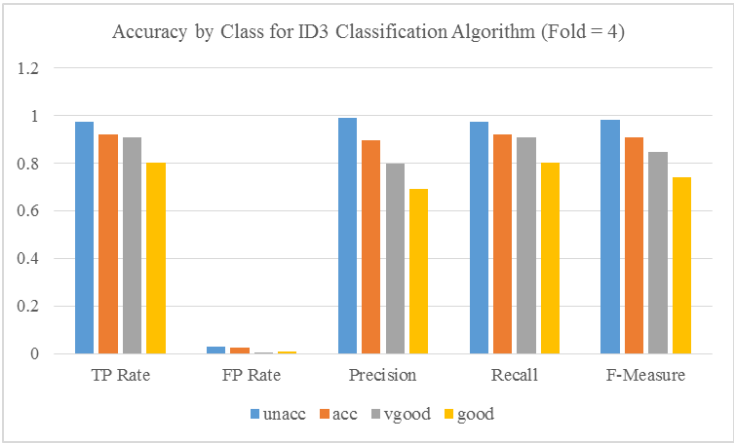

Fig. 4: Accuracy by Class for ID3 Classifier (Fold = 4) Using the Computed Accuracy Values of TABLE III Obtained by Applying the Classifier on the Car Evaluation Database [12, 13] Using Weka [14].

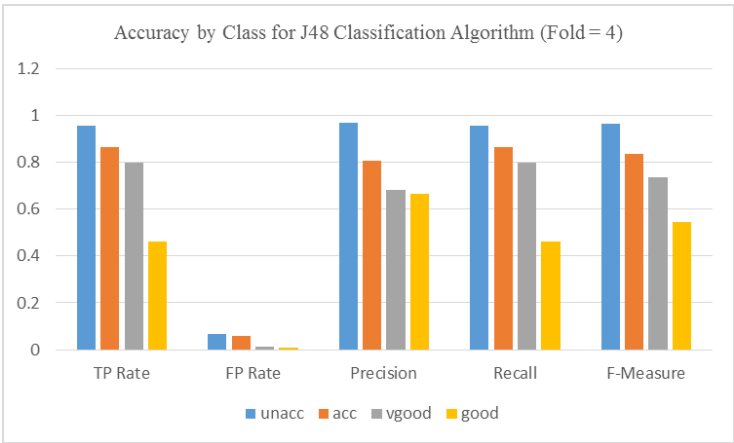

Fig. 5: Accuracy by Class for J48 Classifier (Fold = 4) Using the Computed Accuracy Values of TABLE III Obtained by Applying the Classifier on the Car Evaluation Database [12, 13] Using Weka [14].

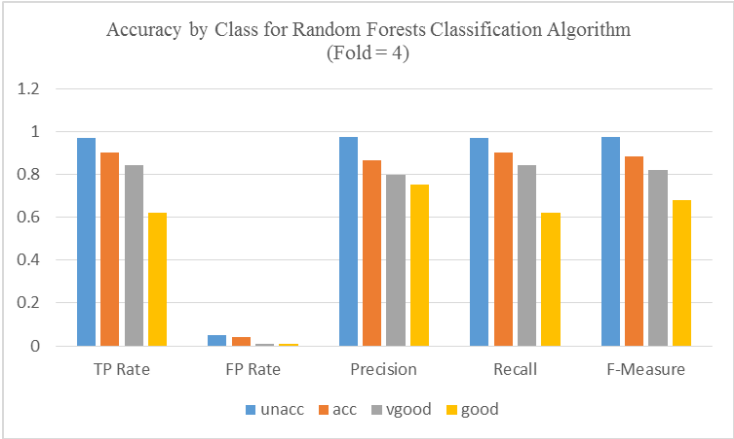

Fig. 6: Accuracy by Class for Random Forests Classifier (Fold = 4) Using the Computed Accuracy Values of TABLE III Obtained by Applying the Classifier on the Car Evaluation Database [12, 13] Using Weka [14].

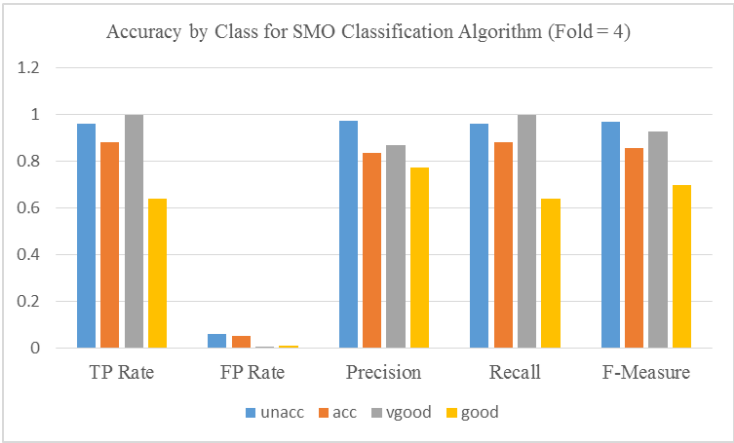

Fig. 7: Accuracy by Class for SMO Classifier (Fold = 4) Using the Computed Accuracy Values of TABLE III Obtained by Applying the Classifier on the Car Evaluation Database [12, 13] Using Weka [14].

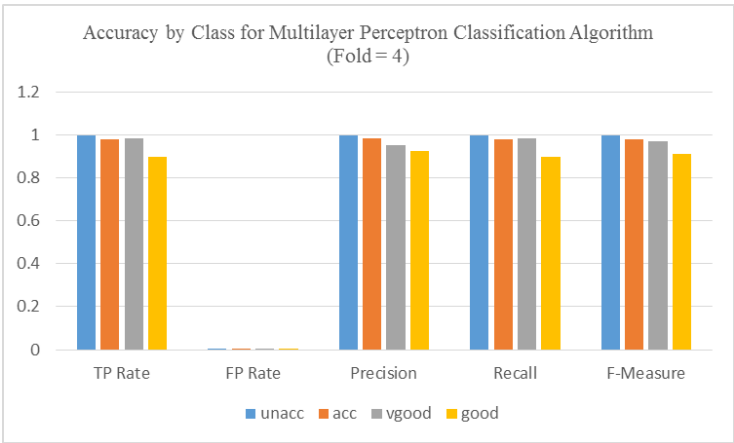

Fig. 8: Accuracy by Class for Multilayer Perceptron Classifier (Fold = 4) Using the Computed Accuracy Values of TABLE III Obtained by Applying the Classifier on the Car Evaluation Database [12, 13] Using Weka [14].

Fig. 9 depicts the comparative classification precision of various classifiers for Fold = 4 for 4 car classes unacc, acc, vgood and good. Among the classifiers, the Multilayer Perceptron classifier has the highest classification precision in classifying each of the 4 car classes.

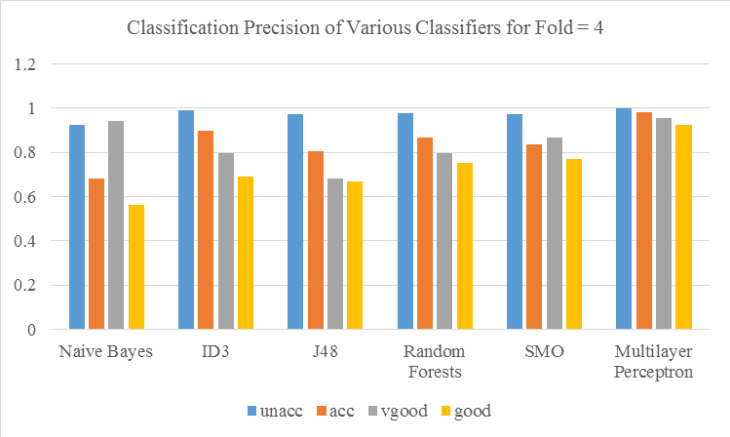

Fig.9. Classification Precision of Various Classifiers Obtained by Applying the Classifiers on the Car Evaluation Database [12, 13] Using Weka [14] with Rearranging Data of TABLE III (Fold = 4) for Precision Only.

In Fig. 9, Multilayer Perceptron classifier has the highest precision, approximately 99% for unacc car class. For vgood car class, Multilayer Perceptron classifier also has the highest precision whereas Naive Bayes has the second highest precision. For good car class, Multilayer Perceptron classifier also has the highest precision whereas Naive Bayes has the lowest precision than all other classifiers. Overall, Multilayer Perceptron classifier has the highest precision for all of the 4 car classes.

The extracted knowledge obtained by analyzing the training data set using data mining technique can be used in efficient decision making [27, 28] for car sales promotion. To recommend a particular group of customers to purchase a particular class of car, the knowledge represented by the generated model based on the training data set can be used.

V. Conclusion

This paper presents the application of three types of classifiers- Bayesian, decision tree and functional on a car evaluation database using Weka to discover data mining models. Various features of classification performance- some important ones of them, e.g., classification time required in seconds to generate each model, percentage of correctly classified instances and incorrectly classified instances, precision etc. are compared for six classifiers. A comparative analysis on these features are also presented for 4 car classes. Based on the analysis presented in this paper, the best classifier can be chosen for data mining classification of a particular relational database. A car evaluation data set of 1728 instances of UCI repository has been analyzed using classification algorithms- Naive Bayes, ID3, J48, Random Forests, SMO and Multilayer Perceptron for data mining using two Folds 4 and 5. The classification performance of each algorithm with the obtained values of the features for 4 car classes for all of the six algorithms have been presented. Among the decision tree classifiers used, though ID3 has comparatively very small classification time to generate the model, it lefts some instances unclassified, and hence a possible efficient solution may be attempted to find for this algorithm. Some computed features of accuracy by class for Fold = 4 for various classifiers are also shown. Finally, the classification precision of the selected six classifiers are compared, which are obtained using Weka for Fold = 4, and shown graphically for comparative analysis to choose the appropriate classifier for using in data mining classification task. The research work demonstrates the variety of classifiers and also their comparative performance. An intelligent software system may be attempted to develop for automation of car sales promotion and advisory intelligent information system based on using the extracted knowledge of the past sales database through generating data mining models by incorporating data mining systems.

Acknowledgements

The author thanks to the colleagues and experts for the valuable discussions occasionally made with them. The data set that is used for analysis for applying the classifiers using Weka data mining tool is obtained from UCI machine Learning Repository.

References

[1]. L. Zhang, Y. Chen, Y. Liang and N. Li, Application of Data Mining Classification Algorithms in Customer Membership Card Classification Model, *International Conference on Information Management, Innovation Management and Industrial Engineering 2008*. IEEE Computer Society.

[2]. J. R. Quinlan, Induction of Decision Trees. *Machine Learning. 1(1)*, March 1986, 81-106.

[3]. J. R. Quinlan and R. L. Rivest, Inferring Decision Trees Using the Minimum Description Length Principle, *Information and Computation, 80(3)*, 1989, 227-248.

[4]. J. R. Quinlan, Improved Use of Continuous Attributes in C4.5, *Journal of Artificial Intelligence Research, 4(1)*, 1996, 77-90.
[5]. L. Qin and W. Yongquan, Improved ID3 Algorithm Using Ontology in Computer Forensics, 2010 *International Conference on Computer Application and System Modeling (ICCASM 2010)*, 2010, 494-497.
[6]. R. Cao and L. Xu, Improved C4.5 Algorithm for the Analysis of Sales, in *2009 Sixth Web Information Systems and Applications Conference*, IEEE Computer Society, 2009, 173-176.
[7]. S. Ponmani, R. Samuel and P. VidhuPriya, Classification Algorithms in Data Mining - A Survey. *International Journal of Advanced Research in Computer Engineering & Technology. 6(1)*, January 2017, 1-6.
[8]. K. Deeba and B. Amutha, Classification Algorithms of Data Mining, *Indian Journal of Science and Technology. 9(39)*, October 2016, DOI: 10.17485/ijst/2016/v9i39/102065.
[9]. J. Dan, Q. Jianlin, G. Ziang, C. Li and H. Peng, A Synthesized Data Mining Algorithm Based on Clustering and Decision Tree, *Proc. 2010 10th IEEE International Conference on Computer and Information Technology (CIT 2010)*, IEEE Computer Society.
[10]. S. L. Salsberg, Book Review: C4.5: Programs for Machine Learning by J. Ross Quinlan. Morgan Kauffman Publishers Inc., *Machine Learning, 16(3)*, September 1994, 235-240.
[11]. G. Kaur and A. Chhabra, Improved J48 Classification Algorithm for the Prediction of Diabetes, *Internatioal Journal of Computer Applications. 98(22)*, July 2014, 13-17.
[12]. D. Dua and C. Graff (2019), UCI Machine Learning Repository [http://archive.ics.uci.edu/ml], Irvine, CA: University of California, School of Information and Computer Science.
[13]. M. Bohanec and B. Zupan, Car Evaluation Database, UCI Machine Learning Repository [http://archive.ics.uci.edu/ml/ datasets/Car+Evaluation], Irvine, CA: University of California, School of Information and Computer Science.
[14]. WEKA, Waikato Environment for Knowledge Analysis. The University of Waikato, Hamilton, New Zealand, URL: www.cs.waikato.ac.nz/ml/weka/downloading.html.
[15]. R. H. Rajon and J. P. Malar Dhas, A Method for Classification Based on Association Rules Using Ontology in Web Data, *International Journal of Computer Applications (IJCA), 49(8),* July 2012, 13-17.
[16]. B. Liu, W. Hsu and Y. Ma, Integrating Classification and Association Rule Mining, *Proc. Fourth International Conference on Knowledge Discovery and Data Mining*-1998, AAAI 1998.
[17]. S. Kharya and S. Soni, Weighted Naive Bayes Classifier: A Predictive Model for Breast Cancer Detection, *International Journal of Computer Applications (IJCA), 133(9)*, January 2016, 32-37.
[18]. L. Breiman, Random Forest, *Machine Learning, 45(1)*, 2001, 5-32.
[19]. A. Cufoglu, M. Lohi and K. Madani. A Comparative Study of Selected Classifiers with Classification Accuracy in User Profiling. 2009 World Congress on Computer Science and Information Engineering,
[20]. V. P. Bresferean, Analysis and Predictions on students behavior using decision trees in WEKA environment, *IEEE Proc. 29th International Conference on Information Technology Interfaces*, ITI 2007.
[21]. P. Lumbini Khobragade and P. Mahadik. Predicting Students' Academic Failure Using Data Mining Techniques, *International Journal of Advance Research in Computer Science and Management Studies. 3(5)*, May 2015.
[22]. J. C. Platt, Fast Training of Support Vector Machines Using Sequential Minimal Optimization, Advances in Kernel Methods – Support Vector Learning, (MIT Press, January 1998), Microsoft Research, url:https://www.microsoft.com/en-us/research/publication/fast-training-of-support-vector-machines-using-sequential-minimal-optimization.
[23]. S. Singaravelan, D. Murugan and R. Mayakrishnan. Analysis of Classification Algorithms J48 and SMO on Different Datasets, *World Engineering and Applied Sciences Journal, 6(2)*, 2015, 119-123.
[24]. S. K. Pal and S. Mitra, Multilayer Perceptron, Fuzzy Sets, and Classification, IEEE Transactions on Neural Networks, 3(5), IEEE Publisher, 683-697, 1992.
[25]. R. Collobart and S. Bengio, Links between Perceptron, MLPS and SVMs, IDIAP Publisher, 2004.
[26]. R. R. Bouckaert, E. Frank, M. Hall, R. Kirkby, P. Reutemann, A. Seewald and D. Scuse, WEKA Manual for Version 3-8-3, September 2018, The University of Waikato, Hamilton, New Zealand, URL: www.cs.waikato.ac.nz/ml/weka/downloading.html.
[27]. A. Kusiak, Data Mining and Decision Making, *Proceedings of the SPIE Conference on Data Mining and Knowledge Discovery 2002*.
[28]. N. A. Haris, Munaisyah Abdullah, Abu Talib Othman, and Fauziah Abdul Rahman, Optimization and Data Mining for Decision Making, *WCCAIS'2014 Congress*. IEEE Publication.

IOSR Journal of Computer Engineering (IOSR-JCE) is UGC approved Journal with Sl. No. 5019, Journal no. 49102.

Md. Humayun Kabir. "Study on the Performance of Classification Algorithms for Data Mining" IOSR Journal of Computer Engineering (IOSR-JCE) 21.3 (2019): 23-30.