Performance Analysis of Classification Techniques for Car Data Set Analysis

Madhusmita Das and Rasmita Dash

***Abstract*—Data mining which is also known as data or knowledge discovery is an important technique to analyze data and discover important information from different data bases. Data mining consists of various classification and prediction algorithms. Classification is an important technique with a broad area of application. To analyze different algorithms of classification is a complex task. In this paper Waikato Environment for Knowledge Analysis (WEKA) tool is used for classification of an automobile data set. Here, 66 different classifiers are applied for this data set. The aim of this work is to compare and analyze different classification algorithms on WEKA tool and find out which classification algorithm is most suitable to work with automobile data set.**

***Index Terms*—Data mining; classification; WEKA tool; Naive Bayes.**

I. INTRODUCTION

DATA bases are loaded with concealed information and these are intended for intellectual decision making. For the prediction of future data and for the description of the data classes, different forms of data analysis, such as classification and prediction are used. Classification is a technique which predicts categorical class labels. This class labels can be of discrete or nominal. The classification technique classifies data based on the training set and the class labels [1]. As there is an increasing number of implementation being made for different algorithms of classification and prediction, there is a need to have a single platform that can compare the performance of all the classification algorithms and should give the information , which classifier is the best one[2]-[4]. In this paper authors have compare different techniques of data mining which are used for classification, namely decision tree induction ,neural networks, lazy learners, functions classifiers, meta learners, rules classifiers, and Bayesian classifiers. WEKA tool was used to find out correctly classified instances, confusion matrix, kappa statistics, True Positive (TP) rate, False Positive (FP) rate and ROC area and mean absolute error [3]-[5]. The comparison was made among 66 classifiers. The first step in this comparison study was running each of the 66 classification algorithms on car datasets using the full attribute set and 10-fold cross validation. 10-fold cross validation was used, because it provides a less biased result. The results were compared to see which algorithm performed better. To determine this, the following criteria were being taken into account such as, time taken to build the model, correctly classified instances, kappa statistics, TP rate, FP rate, ROC area.

The rest of the paper organize as follows, in section II, introduction to WEKA tool was presented. Various classifier used in this paper were discussed in section III. The data set used, in this paper was discussed in section IV. In section V, results were given and finally conclusion is discussed in section VI.

Madhusmita Das is with the Department of Computer Science and Information Technology, Institute of Technical Education and Research, Siksha 'O' Anusandhan (Deemed to be) University, Bhubaneswar, Odisha, India. (e-mail :madhusmitadas@soa.ac.in)

Rasmita Dash is with the Department of Computer Science and Engineering, Institute of Technical Education and Research, Siksha 'O' Anusandhan (Deemed to be) University, Bhubaneswar, Odisha, India. (e-mail: rasmitadash@soa.ac.in)

II. WEKA TOOL

WEKA stands for Waikato Environment for Knowledge analysis. It is created by researchers at the University of Waikato in New Zealand. WEKA is a user friendly tool. To work with WEKA, user does not require having a deep knowledge in different algorithms of data mining. The graphical user interface feature of WEKA, make it a popular data mining tool. In this paper, authors used WEKA as a tool to compare various classification techniques using automobile data set. Normally Attribute Relation File Format ARFF, is used for the data file in WEKA. ARFF is an ASCII text file [6]. ARFF is also developed at University of Waikato, to be used with WEKA.WEKA is an open source software. WEKA implemented more than 75 classification algorithms. It also include more than 45 preprocessing tool, clustering algorithms, attribute valuators, around 10 feature selection algorithms and some algorithm for association rule[7].

III. CLASSIFIER

This section describes the classification methods used in this paper. Total 66 algorithms are categorized into seven classifier such as Naïve Bayes, Functions, Lazy, Meta, Misc, Rules and Trees classifiers.

i) Naive Bayes Classifier: In data mining Bayes methods are used as one of the classification techniques [2]. In this research work authors used eight main Bayesian methods, namely AODE, AODEsr, Naive Bayes, Bayesian net, HNB, Naive Bayes simple, Naive Bayes updateable and WAODE.

978-1-7281-4988-2/20/$31.00 ©2020 IEEE

IEEE Advancing Technology for Humanity

Authorized licensed use limited to: Carleton University. Downloaded on October 04,2020 at 09:49:35 UTC from IEEE Xplore. Restrictions apply.

ii) Function Classifier: The concept of neural network and regression are used in function classifier. In this work three different function classifiers, such as SMO, RBF Network and logistic are used.

iii) Rule Classifier: Under Rule classifier, nine classifiers are used, namely Decision table, DTNB, JRip, NNge, One Rule, PART, PRISM, Ridor and Zero Rules.

iv) Lazy Classifier: Lazy classifier is a computational expensive classifier. They require well-organized storage techniques and also require suitable parallel hardware for the implementation of the algorithms. The structure of the data is complicated in lazy classifier and it provides very less explanation towards the structure of data [8]-[10]. One of the major advantages of this classifier is the incremental learning nature of it. These classifiers are able to form multifarious assessment spaces having hyper multilateral shapes which may not be as effortlessly describable by other learning algorithms [10]. The methods of this algorithm, which are used in this research work are IBI, IBK, K- Star, LBK and LWL.

v) Meta Classifier: Meta classifier includes a large range of classifier. In this paper 24 different algorithms of Meta classifier are used.

vi) Decision Tree: In this paper 12 different algorithms of decision tree classifier are used for the experimental purpose.

vii) MISC: Five different MISC algorithms, such as Hyper pipes, min max extension, OLM, OSDL and VFI are used in this paper.

IV. DATA SET

The title of the data set which is used in this work is" Car Evaluation Database" detailed in Table I. The creator of this data set is Marko Bohanec. This data set is created in June 1997. Car Evaluation Database was derived from a simple hierarchical decision model [11]. The following concept structure or attributes are used by any model for the evolution of car.

TABLE I
CAR EVALUATION DATABASE

Attribute	Attribute meaning
CAR	It gives car acceptability
PRICE	The overall price of the car
buying	The buying price of the car
maint	The maintenance price of the car
TECH	Technical characteristics
COMFORT	The comfort level of the car
doors	Number of doors in the car
persons	The capacity in terms of person to carry in the car
lug boot	The size of luggage boot
safety	Estimated safety of the car

Input attributes are printed in lowercase. Besides the target concept CAR, the model includes three intermediate concepts: PRICE, TECH, and COMFORT. Total number of Instances is 1728 and total numbers of attributes are six. In this data set, there is no missing attribute value [12].The attribute values are given in Table II.

TABLE II
ATTRIBUTE VALUES

Attribute	Attribute Values
buying	v-high, high, med, low
maint	v-high, high, med, low
doors	2, 3, 4, 5 ,more
persons	2, 4, more
lug boot	Small, med, big
safety	low, med, high

The class distribution (number of instances per class) is given in Table III.

TABLE III
CLASS DISTRIBUTION

Class	Number of instance	Number of instance in %
unacc	1210	70.023
acc	384	22.222
good	69	3.993
v-good	65	3.762

V. RESULTS AND DISCUSSION

Applying eight Bayesian classifiers on the car dataset, the outputs are summarized in Table IV.

TABLE IV
RESULTS SHOWING OF APPLYING BAYESIAN CLASSIFIERS ON THE CAR DATA SET

Name of the classifier	Time taken to build the model	Correctly classified instances (in %)	TP rate	FP rate	ROC area
AODE	0.02	91.8981	0.919	0.054	0.996
AODESr	0.02	92.4769	0.925	0.05	0.993
Bayes Net	0.02	85.706	0.857	0.162	0.976
HNB	0	93.287	**0.933**	0.05	0.991
NAÏVE BAYES	0.02	85.5324	0.855	0.164	0.976
NAÏVE BAYES SIMPLE	0.03	85.5324	0.855	0.164	0.976
NAÏVE BAYES UPDATEABLE	0.03	85.5324	0.855	0.164	0.976
WAODE	0	91.1458	0.911	0.044	0.99

Authorized licensed use limited to: Carleton University. Downloaded on October 04,2020 at 09:49:35 UTC from IEEE Xplore. Restrictions apply.

The above table shows that HNB classifier is giving the best result. From the ROC area, it is also clear that HNB is giving the best result.

Applying three function classifiers on the car dataset, the outputs are summarized in Table V.

TABLE V
RESULTS SHOWING THE OUTCOME OF FUNCTIONS CLASSIFIER ON CAR DATASET

Name of the classifier	Time taken to build the model	Correctly classified instances (in %)	TP rate	FP rate	ROC area
SMO	1.95	**93.75**	0.938	0.052	0.99
RBFNetwork	0.94	88.2523	0.883	0.123	0.974
logistic	1.66	93.1134	0.931	0.061	0.99

From the above table, it is clear that SMO classifier is giving best result among function classifier. According to the ROC area SMO and logistic both are giving the best result. Applying five lazy classifiers on the car data set, the outputs are summarized in Table VI.

TABLE VI
RESULTS SHOWING THE OUTCOME OF APPLYING LAZY LEARNERS ON THE CAR DATASET

Name of the classifiers	Time taken to build the model	Correctly classified instances (in %)	TP rate	FP rate	ROC area
IB1	0	77.2569	0.773	0.223	0.775
IBK	0	93.5185	0.935	0.059	0.997
K-STAR	0	87.5579	0.876	0.179	0.996
LBR	0.2	**94.1551**	0.942	0.047	0.99
LWL	0	70.0231	0.7	0.49	0.97

From the above table, it is found that LBR classifier is giving the best result. The ROC area is also showing that LBR is having the best result.

Applying 24 different Meta classifiers to the car data set, the results are summarized in Table VII.

TABLE VII
RESULTS SHOWING THE OUTCOME OF APPLYING META CLASSIFIERS ON THE CAR DATA SET

Name of the classifiers	Time taken to build the model	Correctly classified instances (in %)	TP rate	FP rate	ROC area
ADABOOSTM1	0.11	70.0231	0.7	0.7	0.874
ATTRIBUTE SELECTED CLASSIFIER	0.19	92.3611	0.924	0.056	0.976
BAGGING	0.19	92.1875	0.922	0.038	0.986
CLASSIFICATION VIA CLUSTERIG	0.06	54.8611	0.549	0.549	0.576
CLASSIFICATION VIA REGRESSION	2.19	96.7593	0.968	0.015	0.998
CV PARAMETER SELECTION	0.02	70.0231	0.7	0.7	0.497
DAGGING	7.41	89.8727	0.899	0.064	0.985
DECORATE	1.09	93.3449	0.933	0.042	0.986
END	0.16	92.3611	0.927	0.053	0.986
ENSEMBLE SELECTION	11.09	89.8148	0.898	0.056	0.981
FILTERED CLASSIFIER	0.02	92.3611	0.924	0.056	0.976
GRADING	0.02	70.0231	0.7	0.7	0.5
LOGITBOOST	0.17	86.6898	0.867	0.104	0.973
MULTIBOOST AB	0.09	70.0231	0.7	0.7	0.878
MULTICLASS CLASSIFIER	1.56	89.9306	0.899	0.082	0.98
MULTISCHEME	0	70.0231	0.7	0.7	0.497
ORDINAL CLASS CLASSIFIER	0.03	92.1875	0.922	0.052	0.978
RACED INCREMENTAL LOGITBOOST	0.03	70.0231	0.7	0.7	0.756
RANDOM COMMITTEE	0.2	91.6667	0.917	0.113	0.989
RANDOM SUBSPACE	0.08	70.0231	0.7	0.7	0.923
ROTATION FOREST	2.11	**98.6111**	0.986	0.01	**1**
STACKING	0.02	70.0231	0.7	0.7	0.497
STACKINGC	0.6	70.0231	0.7	0.7	0.497
VOTE	0	70.0231	0.7	0.7	0.497

From the above Table VII, it is clear that rotation forest classifier is giving the best result for Meta learners. From the

Authorized licensed use limited to: Carleton University. Downloaded on October 04,2020 at 09:49:35 UTC from IEEE Xplore. Restrictions apply.

above table, it is also clear that rotation forest is having the highest ROC area.

Applying five type of MISC classifiers, to the car data set, the outputs are summarized in Table VIII.

TABLE VIII
RESULTS SHOWING THE OUTCOME OF APPLYING MISC CLASSIFIERS ON THE CAR DATASET

Name of the classifiers	Time taken to build the model	Correctly classified instances (in %)	TP rate	FP rate	ROC area
HYPERPIPES	0	70.0231	0.7	0.7	0.881
MINMAX EXTENSION	0	**97.2801**	0.973	0.025	0.974
OLM	0.31	92.7083	0.927	0.126	0.901
OSDL	0.02	96.1806	0.962	0.013	0.974
VFI	0.02	81.5972	0.816	0.061	0.956

From the above table, it is clear that min-max extension is giving the best result but from the ROC area it is found that min-max extension and OSDL classifiers are giving the best result.

Applying 12 verities of tree classifiers on the car dataset, the outputs are summarized in Table IX.

TABLE IX
RESULTS SHOWING THE OUTCOME OF APPLYING TREE, CLASSIFIERS ON THE CAR DATASET

Name of the classifies	Time taken to build the model	Correctly classified instances (in %)	TP rate	FP rate	ROC area
BF TREE	1.67	97.0486	0.97	0.016	0.994
DECISION STUMP	0.02	70.0231	0.7	0.7	0.709
FT	2.34	95.081	0.951	0.053	0.962
ID3	0.05	89.3519	0.962	0.018	0.946
J48	0.02	92.3611	0.924	0.056	0.976
J48 GRAFT	0.11	92.3611	0.924	0.056	0.976
LAD TREE	0.75	90.6829	0.907	0.027	0.981
NB TREE	2.94	94.213	0.942	0.041	0.989
RANDOM FOREST	0.2	92.7083	0.927	0.056	0.988
RANDOM TREE	0.03	74.3634	0.744	0.314	0.857
REP TREE	0.06	87.3843	0.874	0.079	0.968
SIMPLE CART	1.39	**97.1065**	0.971	0.011	**0.993**

From the above table, it is clear that SIMPLE CART tree is having the highest accuracy and the ROC area also showing SIMPLE CART as the best classifier.

Applying nine different RULE classifiers on the car dataset, the outputs are summarized in Table X.

TABLE X
RESULTS SHOWING THE OUTCOME OF APPLYING RULE CLASSIFIERS ON THE CAR DATASET

Name of the classifiers	Time taken to build the model (in seconds)	Correctly classified instances (in %)	TP rate	FP rate	ROC area
DECISION TABLE	0.27	91.0301	0.91	0.11	0.973
DTNB	0.88	95.2546	0.953	0.03	0.993
JRip	I3.31	86.4583	0.865	0.064	0.947
NNge	0.39	94.5023	0.945	0.061	0.942
One R	0	70.0231	0.7	0.7	0.5
PART	0.8	95.7755	0.958	0.016	0.49
PRISM	0.34	89.294	0.962	0.02	0.947
RIDOR	0.17	**96.2963**	0.963	0.022	**0.99**
ZERO R	0	70.0231	0.7	0.7	0.497

From the above table, it is clear that RIDOR classifier is giving the highest accuracy. The ROC area is also highest for RIDOR classifier.

On the car data set, 66 classifiers has been applied. Comparing the outcome of the 66 classifiers is a tedious task. So, it has been divided into subparts. The 66 classifiers come under seven categories, namely Bayesian classifiers, lazy classifiers, tree classifiers, functions classifiers, rules classifiers, Meta classifiers and MISC classifiers. According to the seven category of all the 66 classifier, the best classifiers for each category obtained from the above tables are summarized in Table XI.

TABLE XI
BEST CLASSIFIER OF EACH CATEGORY

NAME OF THE CLASSIFICATION METHOD	BEST CLASSIFIER
FUNCTIONS CLASSIFIERS	SMO
BAYES CLASSIFIER	NAÏVE BAYES
LAZY CLASSIFIERS	LBR
META CLASSIFIER	ROTATION FOREST
MISC CLASSIFIER	MIN-MAX EXTENSION
TREE CLASSIFIER	SIMPLE CART
RULES CLASSIFIER	RIDOR

Authorized licensed use limited to: Carleton University. Downloaded on October 04,2020 at 09:49:35 UTC from IEEE Xplore. Restrictions apply.

Finally for each category of classifier, the best classifier along with its time taken to build the model, correctly classified instances, TP rate, FP rate and ROC area are given in table IX. From the Table XII, it shows that rotation forest is the best classifier, when the car dataset is used. The Rotation forest classifier correctly classified 98.611 percentages of instances.

TABLE XII
COMPARISON OF BEST CLASSIFIERS USING THE CAR DATASET

Name of the classifier	Time taken to build the model (in seconds)	Correctly classified instances (in %)	TP rate	FP rate	ROC area
HNB	0	93.287	0.933	0.05	0.991
SMO	1.95	93.75	0.938	0.052	0.953
LBR	0.02	94.1551	0.942	0.047	0.992
ROTATION FOREST	2.11	**98.6111**	**0.986**	**0.01**	**1**
MIN-MAX EXTENSION	0	97.2801	0.973	0.025	0.974
SIMPLE CART	1.39	97.1065	0.971	0.011	0.993
RIDOR	0.17	96.2463	0.963	0.022	0.97

The TP rate is also high with FP rate low. The ROC area is also highest for Rotation forest classifier which conform that Rotation forest is the best classification algorithm when car data set is used.

VI. CONCLUSION

The performance of 66 selected classification algorithms, which are categorized into seven different classifier, namely, Bayes, Function, Lazy, Meta, Rules, Misc and Tree are analyzed using car data set in WEKA tool. The 75% of the data set is used for training and the remaining is used for testing reason. The comparison is done over the accuracy of correctly classifying instances. Time taken to build the model, TP rate, FP rate and ROC area are also consider for the performance analysis of different classification algorithms on the car data set.From the analysis, it is clearly visible that Rotation Forest classification algorithm, which is coming under Meta classifier category, is giving the better correctly classified instances (98.61%) as compared to other algorithms, when car data set is used in WEKA tool. However from the analysis, it is also visible that rotation forest classification algorithm require more time as compared to other classifier to build the model.

REFERENCES

[1] Mahendra Tiwari, Manu Bhai Jha, and OmPrakash Yadav. "Performance analysis of Data Mining algorithms in Weka.." IOSR Journal of Computer Engineering (IOSRJCE), ISSN : 2278-0661, Vol.6, Iss.3,2012.

[2] Saichanma, Sarawut, Sucha Chulsomlee, Nonthaya Thangrua, Pornsuri Pongsuchart, and Duangmanee Sanmun. "The Observation Report of Red Blood Cell Morphology in Thailand Teenager by Using Data Mining Technique." Advances in hematology 2014.

[3] Pankaj saxena & sushma lehri. "Analysis of various clustering algorithms of data mining on health informatics." International Journal of Computer & Communication Technology,ISSN (PRINT): Vol. 4, Issue. 2,2013.

[4] Arka Haldar, G.Prudhvi Raj, S.V.S.S Lakshmi." Comparison of Different Classification Techniques Using WEKA for Diabetic Diagnosis." International Journal of Innovative Research in Computer and Communication Engineering.,Vol. 6, Issue 1, 2018.

[5] David, Satish Kumar, Amr TM Saeb, and Khalid Al Rubeaan. "Comparative Analysis of Data Mining Tools and Classification Techniques using WEKA in Medical Bioinformatics." Computer Engineering and Intelligent Systems, Vol 4,Iss 13 pp. 28-38, 2013.

[6] Ranjita kumari Dash."Selection of the best classifier from different datasets using WEKA" International Journal of Engineering Research & Technology (IJERT) , Vol. 2 Issue 3, March – 2013.

[7] Bin Othman,Mohd Fauzi and Thomas Moh Shan Yau."Comparision of different classification techniques using WEKAfor breast cancer." 3rd Kuala Lumpur International Conference on Biomedical Engineering.Springer Berlin Heidelberg, 2006.

[8] Ms S. Vijayarani , Ms M. Muthulakshmi, "Comparative Analysis of Bayes and Lazy Classification Algorithms." International Journal of Advanced Research in Computer and CommunicationEngineering,Vol.2, Issue. 8, 2013.

[9] D. Lavanya. "Ensemble Decision Tree Classifier for Breast Cancer Data," International Journal of Information Technology Convergence and Services, vol. 2, no. 1, pp. 17-24, Feb. 2012.

[10] Aized Amin Soofi and Arshad Awan." Classification Techniques in Machine Learning: Applications and Issues." Journal of Basic & Applied Sciences,Iss. 13, pp.459-465, 2017.

[11] Zia Ul Rehman ,Hira Fayyaz , Asghar Ali Shah , Numan Aslam, Muhammad Hanif , Sagheer Abbas. "Performance evaluation of MLPNN and NB: A Comparative Study on Car Evaluation Dataset," International Journal of Computer Science and Network Security, VOL.18 No.9, September 2018.

[12] S. Makki, A.Mustapha, J. M.Kassim, E. H Gharayebeh, M. Alhazmi, "Employing neural network and naive Bayesian classifier in mining data for car evaluation," In Proc. ICGST AIML-11 Conference, pp. 113-119, 2011.

Authorized licensed use limited to: Carleton University. Downloaded on October 04,2020 at 09:49:35 UTC from IEEE Xplore. Restrictions apply.