IJCSE International Journal of Computer Sciences and Engineering Open Access
Research Paper Vol.-7, Issue-5, May 2019 E-ISSN: 2347-2693

Implementation of Classification Algorithms in Educational Data using Weka Tool

T.Thilagaraj[1], N.Sengottaiyan[2]

[1] Part-Time Ph.D(Category – B), R&D Centre, Bharathiar University, Coimbatore &
Assistant Professor in Computer Applications, Kongu Arts and Science College, Erode, Tamil Nadu, India

[2]Professor in Sri Shanmugha College of Engineering and Technology, Sankari, Tamil Nadu, India

DOI: https://doi.org/10.26438/ijcse/v7i5.12531257 | Available online at: www.ijcseonline.org

Accepted: 14/May/2019, Published: 31/May/2019

***Abstract*—** Extracting information from a particular dataset in various sectors and transforms it into different useful form for a particular process is called data mining. The data mining will manipulate a data to establish patterns for making decisions in needy situations. This type of process in data mining will lead the researchers to evaluate N number of process. The growth of the country lies on the background of education system. Now educational data mining deals lot of issues that may lead different form of solutions. The main objective of this paper is to compare the different classification techniques using weka tool. Using a weka tool were Navies Bayes, J48, AdaBoostM1, LMT and SMO algorithms are utilized for performing classification techniques.

***Keywords*—** Data mining, Classification, Naïve bayes, J48, AdaboostM1, LMT and SMO

I. INTRODUCTION

Data Mining is used to predict the future activities based on existing data and aims in finding the relationship between existing data. The outcome of remaining data will be predicted on the basis of relationship. Here the different methodologies like classification, clustering, sequential patterns and rule generation are available to solve various problems. These methods are also used in various sectors like marketing, finance, health care, insurance and sales etc. The major research domains of data mining are text mining, web mining and image mining [1]. The data mining and statistics are categorized independently to find predictions through statistics [2]. The Bayesian network, decision tree, fuzzy logic and neural network are the most famous supervised learning techniques for fraud detection [3]. Discovering useful patterns from educational information system is the new trend in educational data mining [4]. The education is one of the important factors to improve the status of the country. The latest techniques in data mining are needed in academics to improve the quality of education [5]. Educational data mining will provide the support to the policy makers.

II. RELATED WORK

According to Rajeshinigo and Jebamalar [6] the C4.5, Random forest, Naïve bayes, Multilayer perception and SVM classifiers are implemented to analyze the student data set using weka tool. The results were compared and approved SVM as best classifier to predict the student data results with good accuracy.

Ahmed and Elaraby [7] have used the decision tree on student database to predict student's performance. This study will show the improvement of student's performance through identifying the students who need special attention.

P. Kaur, M. Singh, and G. S. Josan [8] are used various classification techniques to predict the slow learners and also analyzed their performance. Finally the multilayer perception produces high accuracy while compare with J48, SMO, REP Tree and Naïve bayes.

According to V. Ramesh, P. Parkavi, and P. Yasodha [9] the Naive bayes simple, Multilayer perception, SMO, J48, REPTree classification algorithms are implemented in student database. The result shows that Multilayer perception algorithm haves high prediction rate which compare with other four algorithms.

III. CLASSIFICATION

Classification is one of the familiar techniques in data mining to solve difficult problems. While implementing classification algorithms a new set of data has found to

produce new results. The prediction and estimation are the two types of classification available in data mining. The various approaches are performed in classification to assume the knowledge of the data. Here the predefined classes are available to perform certain tasks which will not overlap on existing ones. Only by checking after these conditions the partition will be made.

IV. CLASSIFICATION ALGORITHMS

A. Naive Bayes

This is a very simple and strong classifier. The assumptions are independent and this classifier is based on bayes theorem. which deals with any number of independent variables that may be continuous or categorical [10]. The simple interface of Naïve Bayes will permit kernel estimator to select the numeric attributes. While converting the numeric to normal attributer it will use supervised discretization. In this Naïve Bayes classifier the output is in the form of text. This classification is done on the basis of prior probability and likelihood to a class. The Naive Bayes algorithm computes qualified probabilities of instance classes and selects the class which have highest subsequent values [11].

B. J48

It is referred as statistical classifier used to generate a decision tree and the implementation of the C4.5 decision tree algorithm makes J48. While J48 classifier executes in Java Virtual Machine the class instance are created. Here the programs which occupies large amount of size may be split into smaller one. To find the best attribute, the split criterion is implemented in decision tree. Like all other decision tree structure the J48 has a root node, intermediate node and leaf node [12]. The list of all samples must belong to the same class. Even any single information is not gained from the available features the classifier will increase the node value. If the unseen class is found, then node value will be increased by the classifier.

C. AdaboostM1

Improving the performance of any algorithm is called Boosting. The main purpose of boosting is to reduce the faults of a weak algorithm. This algorithm is used to construct a strong classifier. It was designed to use the combination of many other algorithms to improve the performance of critical tasks. This classifier is highly sensitive while dealing with noisy data. The process here is to repeatedly run the weak algorithm by using various distributions on training data and then combining it into single classifier. This classifier is well suited in the place of defect classification. Here the accuracy is improved and also processing time is very less while comparing with other models. Another advantage of this model is that, it has a minimum level of memory usage while processing and the error rates are also less [13].

D. LMT

The Logistic Model Tree algorithm is to deal binary and multiclass target variables. It handles the numeric, nominal and missing values. All the nominal attributes are converted to binary ones before building the tree. The probabilities of the error will be less in cross-validating the logiBoost iteration.

E. SMO

The sequential minimal optimization algorithm is used to solve optimization problem in analytical manner and also it splits each problem into smaller one. This algorithm will starts randomly on the subset of the data. The algorithm is closely related to Bregman methods and also solves the convex programming problems without using cached kernel matrix. Here two lagrange multipliers are joined together to optimize the finding process [14].

Table 1: List of Attributes and its description

S.No.	Attribute Name	Values	Description
1	Teaching Methodology	Yes, No	The Method of Teaching favourable for the students and it states whether it satisfied or not
2	Attention	Yes, No	Interest is the essential factor
3	Food culture	Good, Bad	Food habits to maintain the health
4	Workouts	Yes, No	Exercise or yoga are more useful techniques to relax
5	Economic Status	High, Medium, Low	Financial status is the most important factor for Education
6	Addiction	Yes, No	It Enquires the students addiction for drugs
7	Presentation	Yes, No	Stage fear will be an obstacle to

			make effective presentation
8	Self Esteem	Yes, No	Self Esteem will create more energy
9	Affection	Yes, No	It is the fact that may divert the students
10	Panic	Yes, No	Fear on Exams will lose GPA
11	Medium	English, Tamil, Hindi	Medium of Education will have huge effect in higher Education
12	Avoid Classes	Yes, No	Avoiding of classes will lead to discontinue the course

V. WEKA

To analyze the attributes and descriptions of Table 1, we are using a tool called Waikato Environment for Knowledge Analysis (Weka). Weka tool consists various machine learning algorithms which is developed as an open source data mining tool by the University of Waikato, New Zealand. Here the workbench holds different methods to take over attribute selection and it may have different stages to generate new set of results. While getting input the control will receive the data in the form of relational table [15]. This tool will support different types of file formats like arff, csv, etc and there may be an option to convert .csv file to flat file. Weka tool is under GNU Public license written in Java and deals with machine learning problems and data mining issues. The various visualization tools and algorithms are available to furnish different tasks [16]. The above attributes are in .xls format and it will be converted to .csv format. Then it will be imported in weka tool for classification. Here we are using five various algorithms to measure different strategies. The working steps for classification techniques in weka are as follows. After loading our dataset in weka choose the option machine learning algorithm. In classify tab click "choose" button there will be a list of classifiers available. Select the required classifier according to the need and click "start" button to view the result in the classifier window.

VI. RESULTS AND DISCUSSIONS

Table 2 and Figure 1 showed the comparison of Navies Bayes, J48, AdaBoostM1, LMT and SMO Algorithms. The True Positive, False Positive, Precision, FMeasure and ROC Area are measured by using various attributes.

Table 2: Accuracy by weighted average

Algorithm	TP Rate	FP Rate	Precision	F Measure	ROC Area
Naïve Bayes	0.782	0.593	0.765	0.772	0.664
J48	0.764	0.714	0.722	0.739	0.436
AdaBoostM1	0.755	0.832	0.659	0.704	0.666
LMT	0.800	0.822	0.667	0.727	0.463
SMO	0.818	0.702	0.775	0.776	0.558

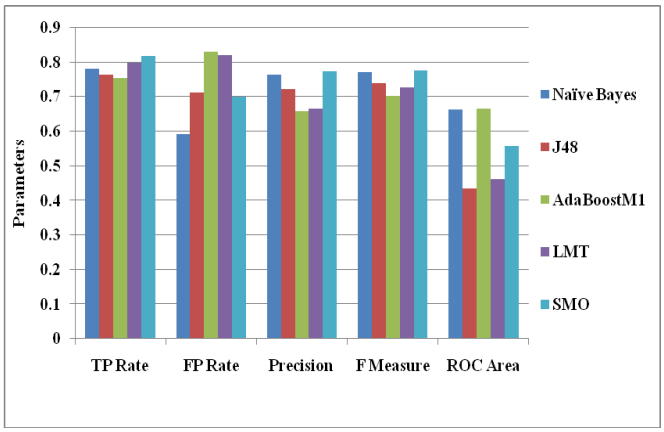

Figure 1: Weighted average of various parameters

Table 3 shows the level of correctly classified instances and incorrectly classified instances. The time measures are found to compare with the Navies Bayes, J48, AdaBoostM1, LMT and SMO Algorithms.

Table 3: Accuracy measures of Naive Bayes, J48, AdaBoost M1, LMT and SMO

Algorithm	Correctly Classified Instances	Incorrectly Classified Instances	Time Taken in Seconds
Naïve Bayes	78.18	21.89	0.02
J48	76.36	23.64	0.03

AdaBoostM1	75.45	24.55	0.2
LMT	80	20	0.05
SMO	81.82	18.18	0.09

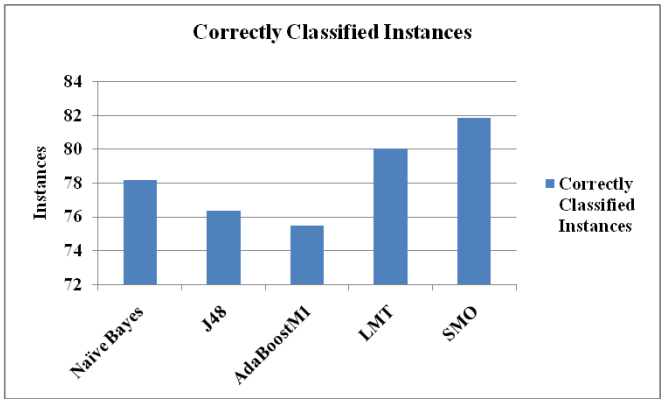

Figure 2: Comparison of correctly classified instances

Figure 2 shows the comparison of correctly classified instances with Navie Bayes, J48, AdaBoostM1, LMT and SMO. The sequential minimal optimization algorithm performs high range of correctly classified instances.

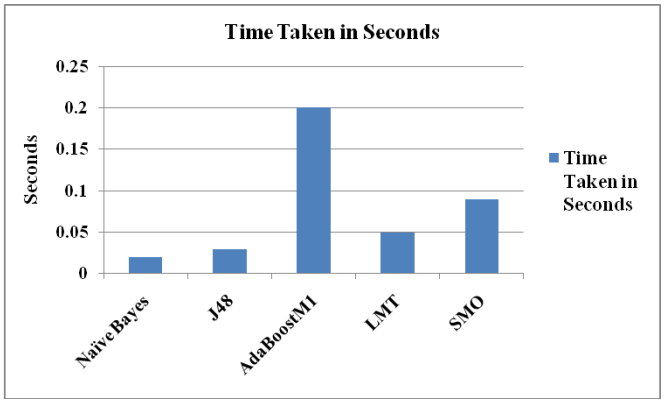

Figure 3: Comparison of time taken to complete task

Figure 3 shows the comparison of time taken to accomplish task with Navie Bayes, J48, AdaBoostM1, LMT and SMO. The Navies Bayes algorithm takes very less time to accomplish the task while compare with other models. But its accuracy level to classify the instances is less while compare with logistic model tree algorithm and sequential minimal optimization algorithm.

VII. CONCLUSION

The classification algorithms will generate a new set of results and the data accuracy is based on the training data set. Here various statistics are measured by Navies Bayes, J48, AdaBoostM1, LMT and SMO algorithms through weka tool. Each algorithm which we discussed above has its own advantages and disadvantages. Finally, the SMO algorithm is having high range of correctly classified instances but it is time consuming compared to other models. In future our work extends to minimize to the time consumption by SMO algorithm.

REFERENCES

[1] S. Vijayarani and M. Muthulakshmi, "Comparative analysis of bayes and lazy classification algorithms," International Journal of Advanced Research in Computer and Communication Engineering, vol. 2, no. 8, pp. 3118-3124, 2013.

[2] Marie Fernandes , "Data Mining: A Comparative Study of its Various Techniques and its Process", International Journal of Scientific Research in Computer Science and Engineering, Vol.5, Issue.1, pp.19-23, 2017.

[3] Namrata Ghuse, Pranali Pawar, Amol Potgantwar, "An Improved Approch For Fraud Detection In Health Insurance Using Data Mining Techniques", International Journal of Scientific Research in Network Security and Communication, Vol.5, Issue.3, pp.27-33, 2017.

[4] Himanshi, Komal Kumar Bhatia, "Prediction Model for Under-Graduating Student's Salary Using Data Mining Techniques", International Journal of Scientific Research in Network Security and Communication, Vol.6, Issue.2, pp.50-53, 2018.

[5] M. F. Uddin and J. Lee, "Predicting good fit students by correlating relevant personality traits with academic/career data," in Proceedings of the 2016 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining, 2016: IEEE Press, pp. 968-975.

[6] D. Rajeshinigo and J. P. A. Jebamalar, "Educational Mining: A Comparative Study of Classification Algorithms Using Weka," *Innovative Res. Comput. Commun. Eng*, 2017.

[7] A. B. E. D. Ahmed and I. S. Elaraby, "Data mining: A prediction for student's performance using classification method," *World Journal of Computer Application and Technology*, vol. 2, no. 2, pp. 43-47, 2014.

[8] P. Kaur, M. Singh, and G. S. Josan, "Classification and prediction based data mining algorithms to predict slow learners in education sector," *Procedia Computer Science*, vol. 57, pp. 500-508, 2015.

[9] V. Ramesh, P. Parkavi, and P. Yasodha, "Performance analysis of data mining techniques for placement chance prediction," *International Journal of Scientific & Engineering Research*, vol. 2, no. 8, p. 1, 2011.

[10] D. K. Tiwary, "A Comparative study of classification algorithms for credit card approval using weka," GALAXY International Interdisciplinary Research Journal, GIIRJ, vol. 2, no. 3, pp. 165-174, 2014.

[11] Deepika Mallampati, "An Efficient Spam Filtering using Supervised Machine Learning Techniques", International Journal of Scientific Research in Computer Science and Engineering, Vol.6, Issue.2, pp.33-37, 2018.

[12] M. N. Amin and M. A. Habib, "Comparison of different classification techniques using WEKA for hematological data,"

American Journal of Engineering Research, vol. 4, no. 3, pp. 55-61, 2015.
[13] R. Kaur and V. Chopra, "Implementing AdaBoost and enhanced AdaBoost algorithm in web mining," International Journal of Advanced Research in Computer and Communication Engineering, vol. 4, no. 7, pp. 306-311, 2015.
[14] G. Taneja and A. Sethi, "Comparison of classifiers in data mining," International Journal of Computer Science and Mobile Computing, vol. 3, no. 11, pp. 102-115, 2014.
[15] F. Alam and S. Pachauri, "Detection using weka," Advances in Computational Sciences and Technology, vol. 10, no. 6, pp. 1731-1743, 2017.
[16] I. Charalampopoulos and I. Anagnostopoulos, "A comparable study employing weka clustering/classification algorithms for web page classification," in 2011 15th Panhellenic Conference on Informatics, 2011: IEEE, pp. 235-239.

Authors Profile

Mr. T.Thilagaraj pursed Bachelor of Computer Science from Bharathiar University, Coimbatore in 2006, Master of Computer Applications from Bharathiar University in 2009 and Master of Philosophy from Bharathiar University, Coimbatore in 2014. He is currently pursuing his Ph.D in Bharathiar University, Coimbatore, His main research work focuses on Data Mining. He has 10 years of teaching experience and 4 years of Research Experience. He is currently working as Assistant Professor in Department of Computer Applications, Kongu Arts and Science College, Erode.

Dr. N.Sengottaiyan pursed Bachelor of Engineering from Bangalore University, Bangalore in 1986, Master of Engineering from Annamalai University, Chidambaram in 2004 and Ph.D. from Anna University of Technology, Coimbatore in 2011. He is a Member in Computer Society of India, Life Member in Indian Society for Technical Education and Institute of Engineers India, Associate Member in IEEE, India. He has published more than 20 research papers in reputed international journals including IEEE, Springer and it's also available online. He has published 3 books. His main research work focuses on Networks, Wireless Sensor Networks, Networks Security and Steganography, Communication Systems: Space Time Block Coding, Artificial Neural Network and Ad-hoc Networks. He has 30 years of Teaching and Research Experience. He is currently working as Director, Professor in Sri Shanmugha College of Engineering and Technology, Sankari.

