ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Website: www.ijircce.com

Vol. 5, Issue 3, March 2017

Educational Mining: A Comparative Study of Classification Algorithms Using WEKA

Driyani Rajeshinigo [1], J. Patricia Annie Jebamalar [2]

Research Scholar, Dept. of Computer Science, St.Xavier's College, Tirunelveli, Tamilnadu, India

Assistant Professor, Dept. of Computer Science, St.Xavier's College, Tirunelveli, Tamilnadu, India

ABSTRACT: Data mining is used to analyse large volumes of data. Classification is one of the techniques in data mining which will be used to predict the target attribute accurately from the knowledge it gained from the training data. Predicting student's performance becomes more challenging due to the large volume of data stored in the educational database. There are various classifiers such as Decision trees, Naive Bayes classifier, random forest, Multilayer Perceptron and Support vector machine can be used to predict the student's performance. This paper provides the comparative analysis of those algorithms on student's data set and suggests the best classifier for educational mining.

KEYWORDS: Classification Algorithms, Educational Data Mining (EDM), C4.5, Random Forest, Naive Bayes, Multi Layer Perceptron and SVM

I. Introduction

Educational mining is the current trend which uses data mining. There is huge amount of data stored in educational database about students but being unused. Educational institutions normally execute some queries on database to fetch past records about a student. But the data stored in educational database can predict a student's performance if used correctly. This can help the student to improve himself in future and can help the staffs to give some additional care for the students who were not performing well enough. Choosing the attributes from the data set for classification lays vital role to predict the target attribute accurately.

Data mining is used to analyse large amounts of data effectively to discover some useful information. Classification is one of the techniques of data mining which will be used to predict the target attribute accurately from the knowledge it gained from the training data.

The WEKA software is used as it contains the implementation of the classification algorithms. It is the free software tool and is widely used for research in the data mining field. Several types of classification algorithms are selected and the student dataset was applied with these algorithms. The classifiers used in this paper consist of common decision tree algorithm C4.5 (J48), Random Forest, Bayesian classifiers (Naive Bayes), Multilayer Perceptron and SVM classifiers. These classifiers will be analysed on student's data set and the results are compared and the best algorithm for predicting student's academic performance will be suggested.

II. Related work

This section summarises literature review of various surveys and comparative studies made on the classification algorithms applied on educational mining. R. Sumitha, and Vinothkumar analysed and compared Classification algorithms on students' data set and found J48 gives better accuracy of 97% [1]. Amirah mohamed Shahiri and Wahidah have done a review on predicting students performance in data mining techniques and found classification algorithms predicts the performance better than other techniques in data mining and C4.5 is highly used to by the researchers for predicting student's performance [2].Pooja Thakar and Anil Mehta broadly analysed many papers on educational mining which compared the data mining technique predicts student's performance and found the attributes which are highly correlated with the student's performance [3].C. Anuradha and T. Velmurugan selected classification algorithms and tested on students' data set and found that Classification of the students based on the attributes reveals

DOI: 10.15680/IJIRCCE.2017. 0503122

that prediction rates are not uniform among the classification algorithms and also show classification algorithms works differently depends on the selection of attributes [4]. Sumit Garg and Arvind K. Sharma analysed various classification algorithms on educational data set and explained thorough details of the implemented algorithms and suggested some good algorithms to work on students' data set to predict their future performance [8].

Sagar Nikam has done the comparative study of classification algorithms. Analysis of classification algorithm says each algorithm has its own merits and demerits and the techniques have to be selected based on the situation [5]. Bhardwaj and Pal conducted the study on the student performance and found that the factors like students' grade in senior secondary exam, living location, medium of teaching, mother's qualification, students other habit, family annual income and student's family status were highly correlated with the student academic performance [14]. Abdul Hamid and Amin analysed and compared students' enrolment approval using classification algorithms and found C4.5 gives high accuracy and lowest absolute errors [6]. Trilok Chand Sharma and Manoj Jain discussed about the classification algorithms and explained how to run those classifiers for the selected dataset and found decision tree gives better performance and high accuracy [7].

Surjeet Kumar Yadav and Saurabh Pal have explained Decision tree algorithms on students' data set and found C4.5 can learn effective predictive models from the student data and gives the better accuracy of classification [12].Sonali Agarwal and Pandey applied classification algorithms on educational data and found SVM classifier LIBSVM with Radial Basis Kernel has been taken as a best choice for data classification [10]. Surjeet Kumar and Brijesh Bharadwaj have done comparative analysis on the decision tree classification algorithms and found CART algorithm is classifying the First, Second, Third class and Fail students with high accuracy [11].

Pandey and Pal conducted study on the student performance based by selecting 600 students from different colleges. By means of Bayes Classification on category, language and background qualification, it was found that whether new comer students will performer or not [13]. Z. J. Kovacic presented a case study on educational data mining to identify up to what extent the enrolment data can be used to predict student's success. The algorithms CHAID and CART were applied on student enrolment data of information system to get two decision trees classifying successful and unsuccessful students. The accuracy obtained with CHAID and CART was 59.4 and 60.5 respectively [15]. V.Ramesh has done an analysis on classification algorithms for redicting student's performance and found Multi Layer Perceptron predicted the performance better than the others and also found that parents' designation plays a vital role for predicting their grades [9].

III. METHODOLOGY

3.1. *Classification Algorithms*:

Classification is one of the Data Mining techniques that are mainly used to analyse a given dataset. It is used to extract models that accurately define important data classes within the given dataset.
Classification is a twostep process.
Step 1: The model is created by applying classification algorithm on training data set
Step 2: The extracted model is tested against a predefined test dataset to measure the model trained performance and accuracy.
So classification is the process to assign class label from dataset whose class label is unknown.

3.1.1. *Decision Tree*:

A decision tree is a flow-chart-like tree structure, where each internal node is denoted by rectangles, and leaf nodes are denoted by ovals. All internal nodes have two or more child nodes. All internal nodes contain splits, which test the value of an expression of the attributes. Arcs from an internal node to its children are labelled with distinct outcomes of the test. Each leaf node has a class label associated with it.

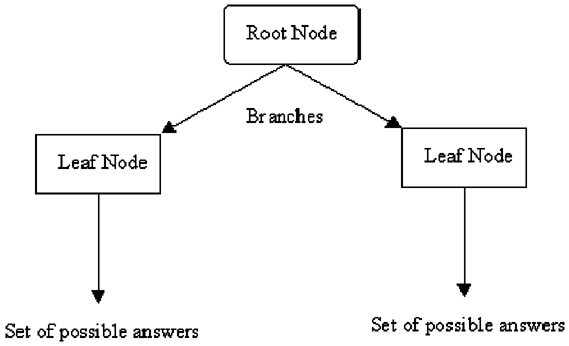


Figure 1: Decision Tree

A.*C4.5:*

This algorithm is a successor to ID3 developed by Quinlan Ross .C4.5 handles both categorical and continuous attributes to build a decision tree. In order to handle continuous attributes, C4.5 splits the attribute values into two partitions based on the selected threshold such that all the values above the threshold as one child and the remaining as another child. It also handles missing attribute values. C4.5 uses Gain Ratio as an attribute selection measure to build a decision tree. It removes the biasness of information gain when there are many outcome values of an attribute.

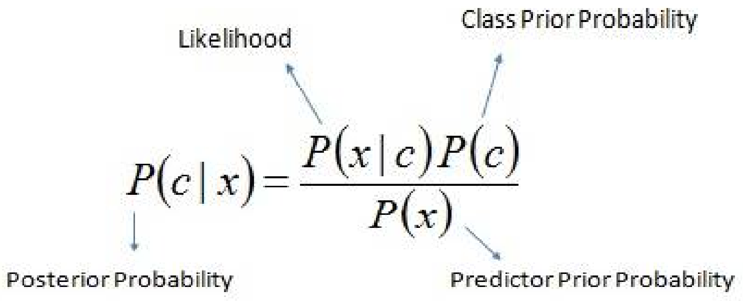
At first, calculate the gain ratio of each attribute. The root node will be the attribute whose gain ratio is maximum. C4.5 uses pessimistic pruning to remove unnecessary branches in the decision tree to improve the accuracy of classification.

B. *Random Forest:*

Random forest is a collection of decision trees built up with some element of random choice. Random forest works by generating a number of trees to analyse the data then it combine all the output from tree and then through the process of vote (look for the classes who have the majority) to obtain the final result.
Random forest has high robustness for large data but it consumes much cost than other techniques

3.1.2. *Naive Bayes:*

The Naive Bayes Classifier technique is based on Bayesian theorem and is particularly used when the dimensionality of the inputs is high. The Bayesian Classifier is capable of calculating the most possible output based on the input. It is also possible to add new raw data at runtime and have a better probabilistic classifier. A naive Bayes classifier considers that the presence (or absence) of a particular feature (attribute) of a class is unrelated to the presence (or absence) of any other feature when the class variable is given.



$$P(c|x) = \frac{P(x|c)P(c)}{P(x)}$$

$$P(c \mid X) = P(x_1 \mid c) \times P(x_2 \mid c) \times \cdots \times P(x_n \mid c) \times P(c)$$

Figure 2: Naive Bayes

- P(c|x) is the posterior probability of class (target) given predictor (attribute) of class.
- P(c) is called the prior probability of class.
- P(x|c) is the likelihood which is the probability of predictor of given class.
- P(x) is the prior probability of predictor of class.
- Class (*c*) is independent of the values of other predictors.

3.1.3. *Multi Layer Perceptron :*

A Multi Layer Perceptron (MLP) is a feed forward artificial neural network model that maps sets of input data onto a set of appropriate outputs. As its name suggests, it consists of multiple layers of nodes in a directed graph, with each layer fully connected to the next one. The architecture of this class of networks, besides having the input and the output layers, also have one or more intermediary layers called the hidden layers. The hidden layer does intermediate computation before directing the input to output layer.

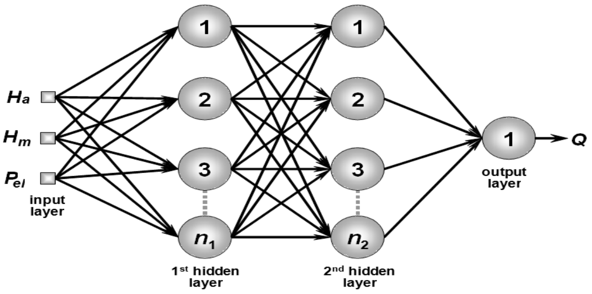


Figure 3: Multi Layer Perceptron

3.1.4. *Support Vector Machine*:

Support Vector Machines are based on the concept of decision planes that define decision boundaries. Support Vector Machine (SVM) is primarily a classier method that performs classification tasks by constructing hyper planes in a multidimensional space that separates cases of different class labels. New examples are then mapped into that same space and predicted to belong to a category based on which side of the gap they fall. An SVM model is a representation of the examples as points in space, mapped so that the examples of the separate categories are divided by a clear gap that is as wide as possible. SVM supports both regression and classification tasks and can handle multiple continuous and categorical variables. Support vector machine operator consists of kernel types including dot, radial, polynomial, neural, anova etc.

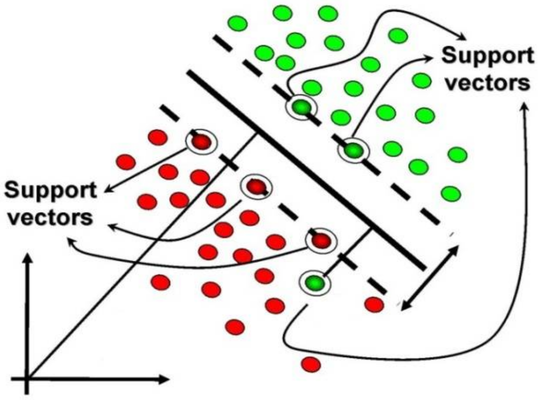


Figure 4:SVM

3.2. *Applying Classification algorithms in WEKA tool:*

3.2.1. *WEKA Tool:*

The Waikato Environment for Knowledge Analysis (WEKA) where learning algorithms were available in various languages, for use on different platforms, and operated on a variety of data formats. WEKA would not only provide a toolbox of learning algorithms, but also a framework inside which researchers could implement new algorithms without having to be concerned with supporting infrastructure for data manipulation and scheme evaluation. Nowadays, WEKA is recognized as a landmark system in data mining and machine learning. Giving users free access to the source code has enabled a thriving community to develop and facilitated the creation of many projects that incorporate or extend WEKA. In this paper WEKA tool is used to analyse the classification algorithms and predicts the students' performance.

3.2.2. *Data Selection and Transformation :*

In this step only those fields were selected which were required for data mining. A few derived variables were selected from the database. All the predictor and response variables which were derived from the database are given in Table 1 for reference.

Table 1: Students Variables

Variable	Description	Possible Values
IAT	Internal Assessment Test	Numeric
CTG	Class Test Grade	{Poor , Average, Good}
SEM	Seminar Performance	{Poor , Average, Good}
ASS	Assignment	{Yes, No}
ATT	Attendance	{Poor , Average, Good}
LW	Lab Work	{Yes, No}
ESM	End Semester Marks	{First ≥ 60% Second ≥ 45 & <60% Third ≥ 36 & <45% Fail < 36%}

IV. SIMULATION RESULTS

The data set with the attributes mentioned in Table 1 is supplied to the WEKA tool and all the mentioned classification algorithms in this paper are executed on the data set to predict the semester results.

4.1. *Classifier accuracy:*

Accuracy of a classifier is the percentage of test set samples correctly classified by the model constructed by the classification algorithm. WEKA is supplied with the dataset contains the attributes mentioned in Table1. Classification algorithms C4.5, Random Forest, Naive Bayes, Multilayer Percetron and SVM are executed in the WEKA Explorer window on the dataset. The results obtained after execution were compared. The below mentioned Table 2 shows the percentage of correctly classified instances and incorrectly classified instances by the classifiers.

TABLE 2: CLASSIFIER ACCURACY

Algorithm	Correctly classified instances	Incorrectly classified instances
C4.5	53.1915 %	46.8085 %
Random Forest	68.0851 %	31.9149 %
Naive Bayes	65.9574 %	34.0426 %
Multilayer Perceptron	61.7021 %	38.2979 %
LibSVM	80.8511%	19.1489%

The below Figure 5 column chart shows the graphical representation of Table 2.

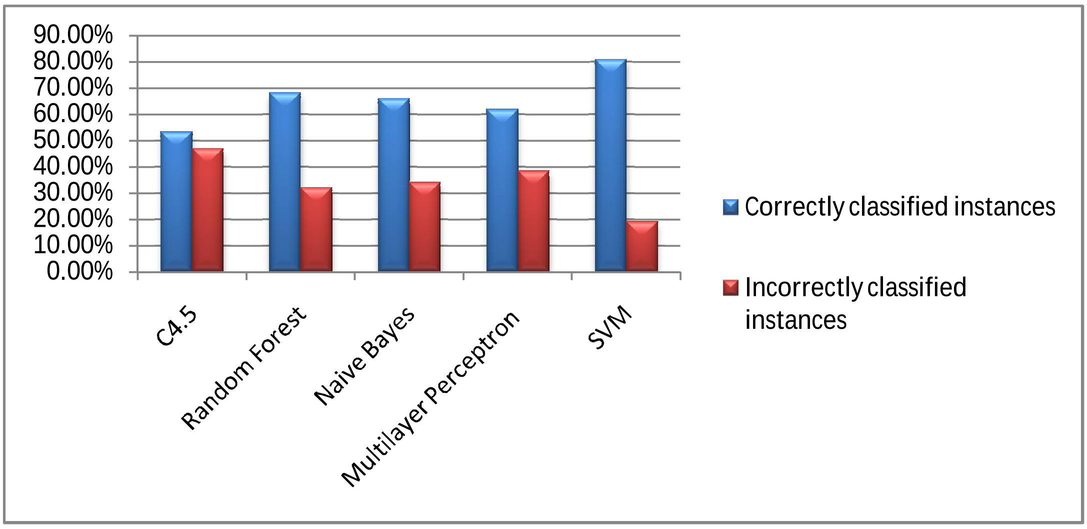


Figure 5.Classifier Accuracy

V. CONCLUSION AND FUTURE WORK

In this Paper, classification algorithms C4.5, Random Forest, Naive Bayes, Multi Layer Perceptron and SVM classifiers are analysed on the students' data set. WEKA tool is used to apply the classification algorithms on the selected data set for predicting the student's semester results. The results were compared and found SVM classifier predicts the results with high accuracy of 81% and C4.5 found to be giving lower accuracy among the algorithms compared. C4.5 algorithm provided lower accuracy because of the use of continuous data as the attribute values. As a future work, these algorithms can be applied on other data sets and techniques need to be found to handle the continuous data in C4.5 to improve the classifier accuracy.

REFERENCES

[1] R. Sumitha and Vinothkumar, "Prediction of Students Outcome Using Data Mining Techniques", International Journal of Scientific Engineering and Applied Science, Volume-2, Issue-6, June 2016

[2] Amirah mohamed Shahiri, Wahidah Husain, "A Review on Predicting Student's Performance Using Data Mining Techniques", ELSEVIER, Volume 72 , Pages 414-422, 2015.

[3] Pooja Thakar, Anil Mehta, "Performance Analysis and Prediction in Educational Data Mining: A Research Travelogue", International Journal of Computer Applications, Volume 110 – No. 15, January 2015.

[4] C.Anuradha and T. Velmurugan, "A Comparative Analysis on the Evaluation of Classification Algorithms in the Prediction of Students Performance", Indian Journal of Science and Technology, Vol 8(15), 2015.

[5] Sagar Nikam, "A Comparative Study of Classification Techniques in Data Mining Algorithms", Oriental journal of computer science & technology, Vol. 8, Pgs. 13-19, 2015.

[6] Abdul Hamid and Amin, "A Comparative Analysis of Classification Algorithms for Students College Enrollment Approval Using Data Mining", Workshop on Interaction Design in Educational Environments, ISBN: 978-1-4503-3034-3, 2014

[7] Trilok Chand Sharma and Manoj Jain, "WEKA Approach for Comparative Study of Classification Algorithm", International Journal of Advanced Research in Computer and Communication Engineering, Vol. 2, Issue 4, April 2013.

[8] Sumit Garg and Arvind K. Sharma "Comparative Analysis of Data Mining Techniques on Educational Dataset", International Journal of Computer Applications (0975 – 8887) , Volume 74– No.5, July 2013.

[9] V.Ramesh, "Predicting Student Performance: A Statistical and Data Mining Approach", International Journal of Computer Applications, Volume 63– No.8, February 2013.

[10] Sonali Agarwal and Pandey, "Data Mining in Education: Data Classification and Decision Tree Approach", International Journal of e-Education, e-Business, e-Management and e-Learning, Vol. 2, No. 2, April 2012.

[11] Surjeet Kumar and Brijesh Bharadwaj, "Data Mining Applications: A comparative Study for Predicting Student's performance", international journal of innovative technology & creative engineering, Vol.1 no.12, December 2012

[12] Surjeet Kumar Yadav and Saurabh Pal, "Data Mining: A Prediction for Performance Improvement of Engineering Students using Classification", World of Computer Science and Information Technology Journal (WCSIT), Vol. 2, No. 2, 2012.

[13] U . K. Pandey, and S. Pal, "Data Mining: A prediction of performer or underperformer using classification", International Journal of Computer Science and Information Technology, Vol. 2- No.2, 2011.

[14] B.K. Bharadwaj and S. Pal. "Data Mining: A prediction for performance improvement using classification", International Journal of Computer Science and Information Security, Vol. 9, No. 4, April 2011.

[15] Z. J. Kovacic, "Early prediction of student success: Mining student enrollment data", Proceedings of Informing Science & IT Education Conference, 2010

BIOGRAPHY

Driyani Rajeshinigo is a Research Scholar in the Computer science Department, St.Xavier's College, Tirunelveli. She received Master of Computer Science (M.Sc) degree in 2007 from St.Joseph's College, Trichy. Her research interests are from Data Mining.

J. Patricia Annie Jebamalar is an Assistant professor in Department of Computer Science, St.Xavier's College, Tirunelveli. She received her Master of philosophy (M.Phil) in Computer science from Alagappa University, Karaikudi. She has published more than 9 Research Papers in Journals and Conferences. Her research interests are from Data Mining.

DOI: 10.15680/IJIRCCE.2017. 0503122