EFFICIENT CLASSIFIER FOR PREDICTING STUDENTS KNOWLEDGE LEVEL USING DATA MINING TECHNIQUES

S.Visalaxi[1], S.Usha[2], S.Poonkuzhali[3]

[1,2]Assistant Professor,[3]Professor, Department of Information Technology

Rajalakshmi Engineering College, Chennai (India)

ABSTRACT

The enormous growth of academic data size in higher education institutions makes learning process tedious to analyze the student's knowledge level. The main motto of all educational institutions is to provide high quality of education. This can be achieved by predicting the student's knowledge level of a particular subject. As traditional modelling approaches are unable to make predictions regarding knowledge, data mining methods was adopted to predict the performance of students. In this research work various classification techniques were applied on the Students Modelling Dataset taken from UCI Machine Learning Repository for predicting the student knowledge level. This work mainly focuses on finding best classifiers for predicting the user knowledge level in particular domain of interest. The results of this study indicate the level of accuracy and other performance measures of the algorithms in predicting the performance of student's knowledge level. The results revealed that Rnd Tree and IBK Classifiers are considered as the best classification algorithms which yields 100 % accuracy on this dataset.

Keywords*: *Accuracy, Classifiers, Student Knowledge Level, Predictors, Random Classifier

I.INTRODUCTION

Data mining techniques plays a vital role in all the application areas of education research and development. The objective of Data Mining in each application area is different. These goals are sometimes difficult to quantify and require their own special set of measurement techniques [5]. In conjunction with the increase of huge volume of daily data collection, data mining has served as the tool for analyzing large amounts of data. Data mining has been expanded from not only to analyze financial data, retail industries data, recommender system data, and intrusion detection data, but also to analyze data in higher education [8][9].

Educational Data Mining refers to techniques, tools, and research designed for automatically extracting meaning from large repositories of data generated by or related to people's learning activities in educational settings. Quite often, this data is extensive, fine-grained, and precise. For example, several learning management systems (LMSs) track information such as when each student accessed each learning object, how many times they accessed it, and how many minutes the learning object was displayed on the user's computer screen. As another example, Intelligent tutoring systems record data every time a learner submits a solution to a problem; they may collect the time of the submission, whether or not the solution matches the expected solution, the amount of time that has passed since the last submission, the order in which solution components were entered into the interface, etc. The precision of this data is such that even a fairly short session with a computer-based learning environment (*e.g.*, 30 minutes) may produce a large amount of process data for analysis.

Classification as a supervised learning technique has been used in predicting new data to be classified based on training dataset. Model resulted from classification can be utilized to predict future data trends. Data classification is defined as a predictive methods in data mining that is used to classify unseen data. There are two main steps in data classification, namely learning step and classification step. In learning step, a classification model is built using an algorithm on a training set. Training set used for learning step must have class labels for given data. After a classifier model is built, it is utilized for predicting class labels for unseen data. Furthermore, traditional student modeling approaches are unable to make predictions regarding knowledge and skill changes under various future training schedules or to prescribe how much training will be required to achieve specific levels of readiness at a specific future time.

Student modeling poses several challenges. The first is that student knowledge is inherently latent – in other words, the goal is to assess a quantity that is not directly measured. Instead, knowledge must be assessed from performance, which has a noisy relationship to knowledge: students often guess and get correct answers without knowledge, and students also often make simple errors (“slips”) on material they know. However, performance can be used to validate models of knowledge – a successful knowledge model should be more successful at predicting future correctness than an unsuccessful knowledge model. Student Knowledge level analysis is one of the most powerful mechanism, which helps understanding the learning interest of the user in particular domain of interest comparing to all the other areas considered. It also focus on how particular class of users can be categorized in a particular subject of interest based on the attributes such as goal object materials(STG), related goal objects(STR), repetition number of user for goal object materials(SCG), and performance of user for goal objects(PEG).

In this paper Student knowledge Modeling dataset is from UCI Machine learning repository[6] is taken for analyzing various classification techniques using Weka[1]data mining tool. In this evaluation process 5 different classification algorithms are chosen. Finally performance evaluation is done to analyze the various classification algorithms to select the best classifier for discovering knowledge level of the user in more efficient and effective manner.

II. RELATED WORKS

P.V.Praveen Sundar et al stated that Educational Data Mining (EDM) is a field that exploits statistical, machine-learning, and data-mining algorithms over the different types of educational data. Its main objective is to analyze these types of data in order to resolve educational research issues. EDM is concerned with developing methods to explore the unique types of data in educational settings and, using these methods, to better understand students and the settings in which they learn[11]. Romero .C et al has proposed 10 common tasks in education that have been tackled using data mining techniques and predicting students‘ performance is one of them. Predicting students‘ performance using data mining methods has been performed at various levels: at a tutoring system level to predict whether some specific knowledge or skills are mastered, at a course level or degree level to predict whether a student will pass a course or a degree, or to predict her/his mark[12].

S.Poonkuzhali et al. analyzed various classification algorithm for predicting efficient classifier for T53 Mutants[2]. R.Kishore Kumar et al. performed comparative analysis for predicting best classifiers for emails spam[1]. Tamizharasi et al, compared three classification algorithms, namely K- Nearest Neighbour classifier, Decision tree and Bayesian network algorithms. The authors mentioned that results were validated by a twenty four month data analysis conducted on mock basis [4]. Chandra.E and Nandhini.K et al used k-means clustering algorithm to predict student's learning activities [10]. Zachary. A [13] explored models with varying levels of skill generality (1, 5, 39 and 106 skill models) and measured the accuracy of these models by predicting student performance within the tutoring system called ASSISTment as well as their performance on a state standardized test. Samad kardan et al [14] proposed knowledge level of a student by using knowledge domain scaffolding, which is dividing the domain into separate sub-domains called learning objectives. These learning objectives may be decomposed to sub-objectives, the dividing process continues until a single unit of knowledge or skill is reached. Ryan S.J.D investigated how well the Contextual-Guess-and-Slip model can predict student learning outside of the tutoring software, comparing it both to the canonical four-parameter version of Bayesian Knowledge Tracing, and to the Individual Difference Weights version of Bayesian Knowledge Tracing [6]. The Individual Difference Weights version finds student-level differences in the four parameters, and has been shown to improve the prediction of post-test performance for students who have reached mastery within the tutor [15]. Ryan et al examined whether ensemble methods, which integrate multiple models, can produce prediction results comparable to or better than the best of nine student modelling frameworks, taken individually [16].

III. FRAMEWORK OF THE PROPOSED SYSTEM

The overall design of the proposed system is given in Fig. 1 and each of these components is addressed in the following sections briefly.

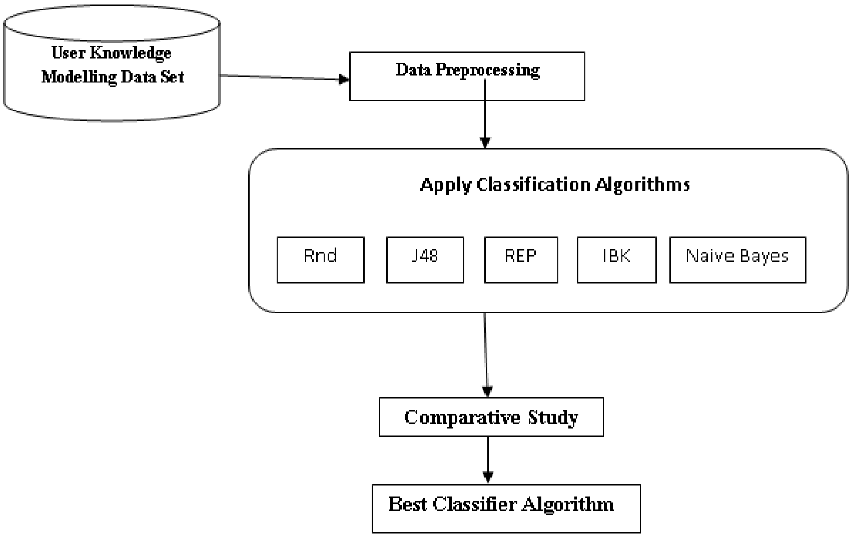
User Knowledge Modelling Data Set
Data Preprocessing
Apply Classification Algorithms
Rnd
J48
REP
IBK
Naive Bayes
Comparative Study
Best Classifier Algorithm

Fig 1. Architectural Design of the Proposed System

3.1 Input Dataset

The User Knowledge Modelling Dataset was taken from UCI machine learning repository and was created by Hamdi Tolga Kahraman. Faculty of Technology ,Department of Software Engineering, Karadeniz Technical University, Trabzon, Turkiye.This Dataset contains 6 attributes(5 continuous input attributes and 1 discrete target attribute) and 258 examples. The attribute description [6] are given in Table 1.

Table 1. Attributes of User Knowledge Modelling Dataset

Attributes	Description
STG	The degree of study time for goal object materials
SCG	The degree of repetition number of user for goal object materails
STR	The degree of study time of user for related objects with goal object
LPR	The exam performance of user for related objects with goal object
PEG	The exam performance of user for goal objects
UNS	The knowledge level of user) - Target Attribute

3.2 Preprocessing

Today, most of the data in the real world are incomplete containing aggregate, noisy and missing values. As the quality decision depends on quality mining which is based on quality data, pre-processing becomes a very important tasks to be done before performing any mining process .Major tasks in data pre-processing are data cleaning, data integration, data transformation and data reduction. In this dataset data normalization is done before applying classification algorithms [1].

3.3 Classification Algorithms

For conducting this analysis five classifiers were namely Random Tree classifier, J48, REP Tree (ID3), Naive Bayes and IBK algorithm.

3.4 Rnd Tree (Random Tree)

Random forests are an ensemble learning method for classification, regression and other tasks, that operate by constructing a multitude of decision trees at training time and outputting the class that is the mode of the classes (classification) or mean prediction (regression) of the individual trees. Random forests correct for decision trees' habit ofoverfitting to their training set.The algorithm for inducing a random forest was developed by Leo Breiman and Adele Cutler, and "Random Forests" is their trademark.

3.5 J48

J48 is an algorithm used to generate a decision tree developed by Ross Quinlan. J 48 is a jave version of the C4.5 algorithm. The decision trees generated by J 48 can be used for classification, and for this reason, J48 is often referred to as a statistical classifier[7].

3.6 REP Tree (ID3 algorithm)

ID3 algorithm begins with the original set as the root node. On each iteration of the algorithm, it iterates through every unused attribute of the set and calculates the entropy (or information gain IG(A)) of that attribute. Then selects the attribute which has the smallest entropy (or largest information gain) value[7].

3.7 Naive Bayes

In machine learning, naive Bayes classifiers are a family of simple probabilistic classifiers based on applying Bayes' theorem with strong (naive) independence assumptions between the features.Naive Bayes has been studied extensively since the 1950s. It was introduced under a different name into the text retrieval community in the early 1960s and remains a popular (baseline) method for text categorization, the problem of judging documents as belonging to one category or the other (such as spam or legitimate, sports or politics, etc.) with word frequencies as the features[7].

3.8 IBK (K-Nearest Neighbor):

K-Nearest Neighbor classifier that uses that same distance metric for classification. It is an instance based classifier in which the class of the test instance is based on the class of those training instances similar to it as determined by the similarity function based on distance. This algorithm uses normalized distances for all attributes. An object is classified by a majority vote of its neighbor, with the object being assigned to the class most common among its K nearest neighbors.

Table 2. Results of the Various Classifiers

Classifier	Accuracy	Precision	Recall
Rnd	100%	100%	100%
J48	98.8%	98.8%	98.8%
REP Tree	95.7%	96.1%	95.7%
Naive Bayes	89.5%	90.1%	89.5%
IBK	100%	100%	100%

IV. RESULTS

Student Knowledge Modelling Dataset is taken from UCI Machine Learning Repository which is created by Hamdi Tolga Kahraman, Ilhami Colak, Seref Sagiroglu consist of 403 with 6 attributes (5-input attributes,1-target attribute) and are listed in Table 1. This dataset is loaded into the WEKA data mining tool after preprocessing. Then 5 classification algorithms namely REP Tree , J48, Naïve Bayes, IBK and Random tree are applied. The results of these 5 classification algorithms are depicted in Table 2. The results portrayed exhibit accuracy, precision and recall. The performance of all these classifiers is analyzed based on the accuracy to predict the best classifier. Here, IBK an instance based classifier and Random tree an ensemble classifier have been identified as a best classifier for this knowledge modelling dataset as they classified with an accuracy of 100%. The algorithm and sample rules for the best classifier Random Tree is given below.

4.1 Pseudo for Random Tree Algorithm

```
1: For 1 to N do (N -Number of records in User Knowledge Modelling dataset D)
2: Select 'm' input attributes at random from the 'n' total number of attributes in dataset D
3: Find the best spilt point among the 'm' attributes according to a purity measure based on Gini index.
```

$$G = \frac{2\sum_{i=1}^{n} i y_i}{n\sum_{i=1}^{n} y_i} - \frac{n+1}{n}$$

```
4: Spilt the node into two different nodes on the basis of split point.
5: Repeat the above steps for different set of records to construct possible decision trees.
```

6: Ensemble all constructed trees into single forest for classifying student knowledge about the particular subject.

4.2 Sample Rnd Tree Classifier Rules

```
PEG <= 0.35
|  PEG <= 0.13
|  |  LPR <= 0.62: very_low
|  |  LPR > 0.62
|  |  |  PEG <= 0.09
|  |  |  |  STG <= 0.17: Low
|  |  |  |  STG > 0.17: very_low
|  |  |  PEG > 0.09: Low

PEG > 0.35
|  PEG <= 0.67
|  |  LPR <= 0.83: Middle
|  |  LPR > 0.83: High
|  PEG > 0.67: High
```

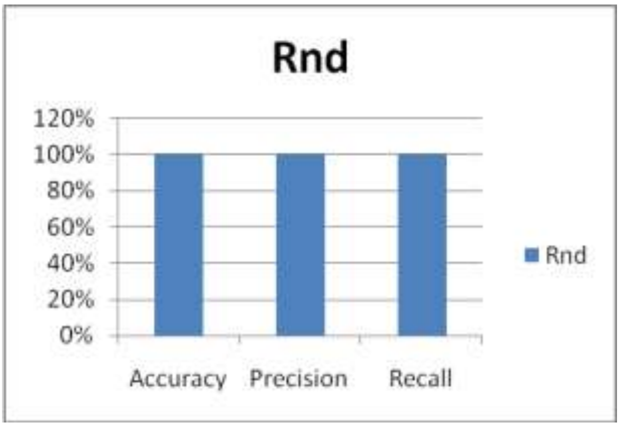

Fig 2. Accuracy, Precision and Recall of Random tree Classifier Algorithms

Performance analysis in terms of accuracy, precision and recall of the random tree classifier is depicted Fig. 2.

4.2.1 Accuracy

Accuracy of a classifier was defined as the percentage of the dataset correctly classified by the method.

$$Acuracy = \frac{No. of\ correctly\ classified\ samples}{Total\ no. of\ samples\ in\ the\ class}$$

4.2.2 Recall

Recall of the classifier was defined as the percentage of errors correctly predicted out of all the errors that actually occurred.

$$Recall = \frac{True\ Positive}{True\ Positive\ + False\ Negative}$$

4.2.3 Precision

Precision of the classifier was defined as the percentage of the actual errors among all the encounters that were classified as errors.

$$Precision = \frac{True\ Positive}{True\ Positive\ + False\ Positive}$$

V. CONCLUSION

Student knowledge level prediction helps us to identify the learning behaviour of students as well as for acquiring the potential knowledge level of student on particular domain interest. After analyzing various classifiers the Random tree classification algorithm is considered as a best classifier for predicting students knowledge level as it produced 100% accuracy for this user knowledge modelling dataset.

REFERENCES

[1] Kishore Kumar. R, Poonkuzhali. G, Sudhakar. P, "Comparative Study on Email Spam Classifier using Data Mining Techniques", Proceedings of International Multiconference on Engineers and Computer Scientist, Vol.1, 2012.

[2] Poonkuzhali. S, Geetha Ramani, Kishore Kumar.R, "Efficient Classifier for TP53 Mutants using Feature Relevance Analysis", Proceedings of International Multiconference on Engineers and Computer Scientist, Vol.1, 2012.

[3] Poonkuzhali. S, Kishore Kumar R and Ciddarth Viswanathan, "Law Reckoner for Indian Judiciary: An Android Application for Retrieving Law Information Using Data Mining Methods", Advanced Computer and Communication Engineering Technology, Springer International Publishing Switzerland, 2015, Chapter 55, Page 585-593.

[4] Tamizharasi. K, Dr. UmaRani, "Employee Turnover Analysis with Application of Data Mining Methods", International Journal of Computer Science and Information Technologies, Vol. 5, No. 1, 2014, pp. 562-566.

[5] Crist´obal Romero , and Sebasti´an Ventura "Educational Data Mining: A Review of the State of the Art" Vol. 40, No. 6, 2010

[6] UCI Machine Learning Repository – UserModelingData Dataset https://archive.ics.uci.edu/ml/datasets/UserKnowledgeModelingData

[7] Archana.S,Dr.Elangovan.K,,"Survey of Classification Techniques in Data Mining", International Journal of Computer Science and Mobile Applications, Vol.2 Issue. 2, pg. 65-71.

[8] J. Han, M. Kamber, and J. Pei. "Data Mining Concepts and Techniques,3rd ed. Waltham: Elsevier Inc, 2012".

[9] C. Vialardi, J. Bravo, L. Shafti, A. Ortigosa. "Recommendation in higher education using data mining techniques. Available: eric.ed.gov/?id=ED539088, Retrieved September 7, 2014

[10] Chandra, E. and Nandhini, K. (2010) ,"Knowledge Mining from Student Data‟, European Journal of Scientific Research, vol. 47, no. 1, pp. 156-163.

[11] P.V.Praveen Sundar Iosr Journal Of Engineering ,"A Comparative Study For Predicting Student's Academic Performance Using Bayesian Network Classifiers (Iosrjen)", E-Issn: 2250-3021, P-Issn: 2278-8719

[12] C. Romero, and S. Ventura, ―"Educational Data Mining: A Review of the State of the Art",‖ IEEE transactions on Systems, Man and Cybernetics, vol. 40(6), pp.601-618, 2010.

[13] Zachary A. Pardos, Neil T. Heffernan, Brigham Anderson, Cristina L. Heffernan "The Effect of Model Granularity on StudentPerformance Prediction Using Bayesian Networks".

[14] Samad Kardan and Ahmad Kardan,"Towards a More Accurate Knowledge Level Estimation".

[15] Ryan S.J.d. Baker1 , Albert T. Corbett2, Sujith M. Gowda1, Angela Z. Wagner2, Benjamin A. MacLaren2, Linda R. Kauffman3, Aaron P. Mitchell3, Stephen Giguere1 "Contextual Slip and Prediction of Student Performance After Use of an Intelligent Tutor".

[16] Ryan S. J. d. Baker, Zachary A. Pardos, Sujith M. Gowda, Bahador B. Nooraei, Neil T. Heffernan, "Ensembling Predictions of Student Knowledge within Intelligent Tutoring Systems" .