International Journal of Computer Applications Technology and Research
Volume 3– Issue 11, 725 - 728, 2014, ISSN: 2319–8656

A Survey on the Classification Techniques In Educational Data Mining

Nitya Upadhyay
RITM
Lucknow, India

Vinodini Katiyar
Shri Ramswaroop Memorial University
Lucknow, India

Abstract: Due to increasing interest in data mining and educational system, educational data mining is the emerging topic for research community. educational data mining means to extract the hidden knowledge from large repositories of data with the use of technique and tools. educational data mining develops new methods to discover knowledge from educational database and used for decision making in educational system. The various techniques of data mining like classification. clustering can be applied to bring out hidden knowledge from the educational data.
In this paper, we focus on the educational data mining and classification techniques. In this study we analyze attributes for the prediction of student's behavior and academic performance by using WEKA open source data mining tool and various classification methods like decision trees, C4.5 algorithm, ID3 algorithm etc.

Keywords: Educational data mining; Classification; Analysis; WEKA,

1. INTRODUCTION:

The examination and study of student's academic performance is not a new exercise but computer based learning environment increases more interest towards student's analysis. The concepts and techniques of data mining can be implemented in education to predict the academic performance of student. On the basis of these kind of predictions the academic performance of student can be improved. EDM is applied to large amount of data accumulated by surveys and various classification techniques are implemented for better performance. The prediction of student's performance has become one of most important needs in order to improve the quality of performance. There is a need of data mining in educational system for the students as well as academic's responsible. Educational data mining is an arising regulation that promote the new techniques for extracting the new data that come from educational settings and by using those techniques, a better prediction can be done for student's behavior, academic performance, subject interest etc.

2. WHAT IS EDUCATIONAL DATA MINING?

Data mining originate a new technique known as educational data mining. In educational data mining, data mining concepts are applied to data that is related to field of education. EDM is the process of transforming the raw data aggregated by education systems.
Educational data mining means exploring hidden data that originated from educational settings by using new methods for better interpretation of students and settings they learnt. Educational data mining promote distinct tools and algorithms for analyze the data patterns. In EDM, data is accumulated during learning process and then study can be done with the techniques from statistics, machine learning and other data mining concepts. To extract the hidden knowledge from data came from educational system, the various data mining techniques like classification, clustering, rule mining etc.

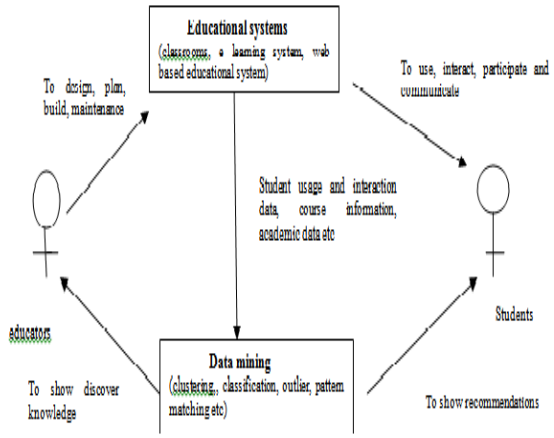

Figure 1

In figure 1 we represent the need of educational data mining. The Academics' responsible and educators worked upon the educational system to enhance the performance of students. In this diagram it is shown that educators want to design the educational system then plan to build that system and most important maintain that educational system. Educational systems include traditional classrooms and some innovative learning methods like e learning system, intelligent and adaptive web based educational system etc. The data set can be extracted from students as students are directly connected with educational system. Now the data is given as input to data mining process and in result it gives recommendations to students and to extract new knowledge to the educators by using various data mining techniques like clustering, classification, pattern matching etc.

2.1 Goals of Educational data mining:

Some of the goals of educational data mining are as follows:

1. Prediction of student's learning behavior by building student models that integrate all definite information of students like student's knowledge, behavior, academic

information etc.

2. Exploring or upgrading domain models that discriminate the content to be learnt and perfect pedagogical sequences.
3. Analysis of all the effects of various types of instructional support given by learning.
4. Advancing scientific knowledge.

2.2 Phases of Educational data mining:

Educational data mining is concerned with translation of new hidden information from the raw data collected from educational systems. EDM generally consist of four phases:

1. The first phase of educational data mining is to find the relationships between data of educational environment. The aim of establishing these relationships is to utilize these relationships in various data mining techniques like classification, clustering, regression etc.
2. The second phase of educational data mining is validation of discovered relationships between data so that over fitting can be avoided.
3. The third phase is to make predictions for future on the basis of validated relationships in learning environment.
4. The fourth phase is supporting decision making process with the help of predictions.

2.3 Methods of Educational data mining

There are so many promoted methods of educational data mining but all kind of methods lie in one of following specified categories:

1. **Prediction**: Ryan S. J. d. Baker has given a detail explanation of prediction in his paper. He mentioned that " In prediction, the goal is to develop a model which can infer a single aspect of data(predicted variable) from some combination of other aspects of data (predictor variables).if we study prediction extensively then we get three types of prediction: classification, regression and density estimation. In any category of prediction the input variables will be either categorical or continuous. In case of classification, the categorical or binary variables are used, but in regression continuous input variables are used. Density estimation can be done with the help of various kernel functions.

2. **Clustering**: In clustering technique, the data set is divided in various groups, known as clusters. When data set is already specified, then the clustering is more useful. As per clustering phenomenon, the data point of one cluster and should be more similar to other data points of same cluster and more dissimilar to data points of another cluster. There are two ways of initiation of clustering algorithm. Firstly, start the clustering algorithm with no prior assumption and second is to start clustering algorithm with a prior postulate.

3. **Relationship Mining**: Relationship mining generally refers to contrive new relationships between variables. It can be done on a large data set, having a no of variables. Relationship mining is an attempt to discover the variable which is most closely associated with the specified variable. There are four types of relationship mining: association rule mining, correlation mining, sequential pattern mining and causal data mining. Association data mining is based on if- then rule that is if some particular set of variable value appears then it generally have a specified value. In correlation mining, the linear correlations are discovered between variables. The aim of sequential pattern mining is to extract temporal relationships between variables.

4. **Discovery with Models:** it includes the designing of model based on some concepts like prediction, clustering and knowledge engineering etc. This newly created model's predictions are used to discover a new predicted variable.

5. **Distillation of Data for Human Judgment**: There are two objectives for human judgment for which distillation of data can be done: Identification and Classification. As per phenomenon of identification, data is represented in a way that human can easily recognize the well specified patterns.

3. LITERATURE SURVEY

3.1. Efficiency of decision trees in predicting student's academic performance

In this paper, S. Anupama Kumar et.al has suggested an approach for predicting the student's performance in examination. They have used C4.5 (J48 in WEKA) to do the prediction analysis. In data collection, a slight modification has been done in defining the nominal values for the analysis of accuracy. As per need of system, data is preprocessed, and integer values are converted into nominal values and stored in .CSV format. After that it is converted to .ARFF format that is accessible to WEKA.

In this paper, the implementation of decision trees rules can be done by dividing the data into two groups. J48 made decision trees by using a set of training data and ID3does the same with the concept of information entropy. In decision tree the attribute for splitting at each node of tree is normalized information gain. The attribute having highest normalized information gain is chosen to make decision. This paper analyzes the accuracy of algorithm in two ways, the first is by comparing the result of tree with the original marks obtained by student and the second is comparing the ID3 and C4.5 algorithm in terms of efficiency.

3.2. Classification model of prediction for placement of students

In paper Ajay Kumar Pal has presented a new approach of classification to predict the placement of students. This approach provides the relations between academic records and placement of students. In this analysis, various classification algorithms are employed by using data mining tools like WEKA for study of student's academic records. In this approach the training algorithm uses a set of predefined attributes. The most widely used classification algorithms are, naïve Bayesian classification algorithm, multilayer perceptron and C4.5 tree. For the high dimensional inputs the naïve Bayesian classification is best technique. Multilayer perceptron is most suitable for vector attribute values for more than one class. Nowadays C4.5 is most popularly used algorithms due its added features like supervising missing

values, categorization of continuous attributes, pruning of decision trees etc.
For testing, the 10 fold cross validation is selected as this evaluation approach. Here, a no of tests are regulated for estimation of input variables: chi square test, information gain test and gain ratio test. Each of the tests makes the concernment of variable in another way. According to this analysis, among three selected best algorithms, the best algorithm is Naïve Bayes classification.

3.3. Study of factors analysis affecting academic achievement of undergraduate students in international program

In this paper, Pimpa Cheewaprakobkit has done analysis to identify the weak students so that the academic performance of those weak students can be improved. In this study, WEKA open source data mining tool is used to estimate aspects for predicting the student's academic performance. In this study , data set to characterize classifier(decision tree, neural network). To predict the accuracy, a cross validation with 10 folds is used.
In this study, to explore the proposal, two classification algorithms have been accepted and distinguished: The Neural Network and C4.5 decision tree algorithm. The investigation process consists of three main steps: data preprocessing, attribute filtering and classification rules. According to this analysis, it is suggested the decision tree model is more accurate than the neural network model. It can be concluded that the decision tree technique has better efficiency data classification for this data set.

3.4. Predicting student's performance using modified ID3 algorithm

Ramanathan L. has overcome the shortcoming of famous algorithm ID3. This algorithm is used to generate the decision trees. In this analysis, instead of information gain, the gain ratio is used. One additional aspect of this study is assignment of weights to each attribute at every decision point. In this paper, in place of traditional ID3 algorithm, a modified ID3 algorithm is used. This modified ID3 algorithm is known as weighted ID3 algorithm. To enhance the normalization, gain ration is more beneficial as compared to information gain. To get a new value, gain ratio is multiplied with the weight and among these new values, the attribute having maximum gain ratio will be elected as node of the tree. Here, WEKA tool is used to analyze the J48 and naïve Bayes algorithm. The modified weighted ID3 algorithm is based on gain ratio and the attributes should be converted by accounting its weight. As per analysis, it is concluded that WID3 algorithm is more efficient than other two algorithms J48 and Naïve Bayes algorithm.

3.5. Analysis and predictions on student's behavior using decision trees in WEKA environment

In this paper, Vasile Paul Bresfelean has worked on data accumulated by different surveys. it is necessary to identify the different conducts of the student's belonging to different specializations. In this paper, the author develops a progression of decision trees based on WEKA's implemented J48 algorithm. In this effort, to discriminate and predict the student's choice in continuing their education. WEKA workbenches applied in this research two of the most common decision tree algorithms are implemented: ID3 and C4.5 (called version J48). In this study, author used J48 because as compared to ID3, J48 gives better result in any circumstances.

Comparison Table

Paper	Author	Technology used	Accuracy	Advantage	Disadvantage
Analysis and predictions on students behavior using decision trees in WEKA envirnment	Vasile Paul Bresfelean	Decision tree construction algorithm: ID3 and C4.5	In IE: 88.68% In CIG: 71.74%	-	-
Classification model of prediction for placement of students	Ajay kumar pal	Classification algorithm: 1.Naive Bayesian classification 2.Multilayer perceptron 3. C4.5 tree Tool: WEKA	1.Naïve Bayesian classification: 86.15 2.Multilayer perceptron: 80.00 J48: 75.38	-	-
Predicting student's performance using modified ID3 algorithm	Ramnathan L., Sakhsham Dhandha, Suresh kumar	J48 and Naïve Bayesian classification algorithms Tool:WEKA	ID3: 93% J48: 78.6% Naïve bayes classifiers: 75%	Shortcoming of ID3 is removed. Gain ration is used instead of information gain	It is inclined towards the attributes with more vales.
Efficiency of decision trees in predicting student's academic performance	S. Anupama kumar Dr. vijaylaxmi	ID3 and C4.5 algorithm J48	ID3: PASS: 103 FAIL: 12 J48:	-	-

			PASS: 103 FAIL: 13		
Study of factors analysis affecting academic achievement of undergraduate students in international program	Pimpa cheewaprakobkit	Classifiers: Decision tree Neural network	Decision tree model: 85.188% Neural network model: 83.875%	-	-

5. CONCLUSION:

This paper described about the Educational data mining, goals of educational data mining and phases of educational data mining and existing classification techniques. Various classification techniques can be implemented on the data set but which classification technique will be applied on the data to improve the academic performance of students, it is important. In this paper, we made a comparison analysis on different existing approaches and methods of classification of data sets. We did the comparative analysis on the basis of accuracy percentage on the application of various classification techniques like Naïve Bayesian Classification, Multilayer Perceptron, J48 and ID3 etc. we also analyzed the advantages and shortcomings of each algorithm applied to data set. So we can say that this paper will provide a beneficial glance of existing solution for classification with their advantages and shortcomings.

6. REFERENCES

[1] Kumar S. Anupama and N. vijaylaxmi M.2011 Efficiency of Decision trees in predicting Student's Acedemic performance.

[2] L. Ramanathan, Dhanda S. and D. S. kumar 2013 Predicting Student's Performance using Modified ID3 Algorithm

[3] Pal A. kumar and Pal S. 2013 Classification Model of Prediction for Placement of Students

[4] Cheewaprakobkit P.2013 Study of Factors Analysis Affecting Academic Achievement of Undergraduate Students in International Program

[5] Bresfelean V. Paul 2007 Analysis and Predictions on Student's Behaviour using Decision Trees in Weka Environment Babes Bolyai University

[6] Baker Ryan S.J.d. Data mining for education Carnegie Mellon University