ISSN: 2393-8528

Contents lists available at www.ijicse.in

International Journal of Innovative Computer Science & Engineering

Volume 3 Issue 4; July-August-2016; Page No. 14-17

Prototyping and Analyzing the Approach of Efficient Mining of Educational Data

Surjeet Kumar

MCA Department, VBS Purvanchal University, Jaunpur, U.P. India 222003

ARTICLE INFO

Received: 17 June 2016
Accepted 30 July 2016

Corresponding Author:

Surjeet Kumar

MCA Department, VBS Purvanchal University, Jaunpur U.P. India 222003

E-mail:
surjeet_k_yadav@yahoo.co.in

ABSTRACT

Data Mining is among the most important research domains in recent a year which focuses on the aspects of real world. To get patterns and predictions on the basis of the data, data mining or classification techniques can be used. Efficiency of the educational institutes, students performance and student employability are in the main focus of today's education system. So the efficient data mining of the educational data is become very critical. The analysis of the efficient usage of techniques, to find out the pattern between student performance, employability and the educational institute's responsibility is shown in this paper. This pattern can be found by extracting the knowledge from the educational student performance data. Using this data, different Data mining Techniques can be applied to find out the useful patterns and fill the gap between the Student Academics and Employability. This paper includes analysis on different classification techniques like decision tree algorithm, C4.5, ADTree, ID3, SVM, and compares the best performances on different aspects.

1. Introduction

Data mining is an iterative approach in which automatic or manual methods are used to define the progress. Data mining is a cooperative effort of humans and computers that can be used in an exploratory analysis in which there are no predetermined notions about what will contributes on interesting outcomes. Data mining is very useful in extracting knowledge form hidden data and the necessary information can be found according to the requirements of the process. The task of knowledge extraction is done with the help of various algorithms, techniques and tools of data mining. Then these can be implemented on training dataset. These dataset passes through various steps of data mining and finally the filtered data is used by the user according to the requirement. It plays an important role in education institute [1].

To extract knowledge from the large warehouse dataset in the field of education, Educational data mining (EDM) is used. Community of EDM defines the EDM as an emerging discipline, concerned with developing methods for exploring the unique types of data that come from educational setting and using those methods to better understand students and the setting which they learn in. It also helps academic planners to give intimation in advance to student that in which direction of studies or other activities they lack. Classification methods like rule mining, decision tree, association rule with other techniques, Bayesian network etc are uses to predict the student performance which helps in placement.

2. Related Work

EDM is a recent research field but there are many research works going on in the area of knowledge extract in educational community from many years. Considerable amounts of EDM work are done and published by various researchers working in this field around the world.

C. Romero and S. Ventura has done a survey in the field of EDM from 1995 to 2005 to show the need of student data analysis which can be further used by students, educators and administrators.
Baker and Yacef analyzed the four goals of EDM as follows

- Predicting student's future learning behaviour
- Discovering or improving domain models
- Studying the effects of educational support
- Advancing scientific knowledge about learning and learners

A case study on educational data mining for the analysis of learning behaviour of students is done in research work of El-Halees in 2008. In his analysis data mining techniques had been applied is association and classification with the help of decision trees on the student‘s data from database course and collected all available data including personal records and academic records of students, course records and data came from e-learning system.

3. Classification Methods

Weka is open source software that implements a large collection of machine leaning algorithms and is widely used in data mining applications. WEKA has a approximately 40 classifiers divided into 4 groups. With the help of its explorer tool any classifiers may be applied to available data set. The research work has chosen 8 different classifiers for comparative analysis of performance of classifiers. As given in Figure 1 LIBSVM classification accuracy is 97.3 %, RBF Network accuracy is 96.05% and Multilayer Perceptron accuracy is 95.85% accuracy. Minimum accuracy is given by Logistic, Simple Logistic and Voted Perceptron classifiers. It shows the accuracy of the predicted values. Here again LIBSVM has minimum root relative square error [2].

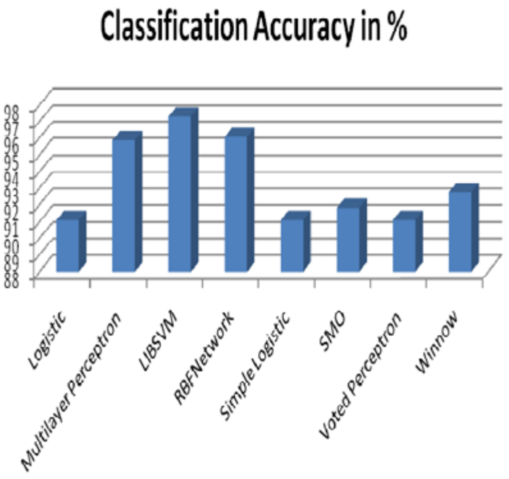


Figure 1: Classification Accuracy of WEKA Classifiers

For LIBSVM the study is further enhanced to observe the performance of various kernel types. It is observed that radial basis kernel has maximum accuracy and Sigmod type kernel as minimum accuracy.

4. Educational data Mining Process

As per the analysis in the paper, extracting the information or knowledge that is hidden or implicit in the data is the major function in the process of data mining. It is also called as Knowledge Discovery in Databases (KDD). So it is overall a knowledge discovery process. Fig 2 shown below is a block diagram of this knowledge discovery process or data mining process in the field of education. It shows the different phases of data mining [2][3]. The Knowledge Discovery from the large educational Databases process comprises of the steps raw data collections to some form of new knowledge. The process consists of the following phases:

- *Data cleaning*: also called as data cleansing. In this step noise data and irrelevant data are removed from the collection.
- *Data integration*: both homogeneous and heterogeneous multiple data sources are combined in a common source in this phase.
- *Data selection*: the decision on the relevancy of the data is done in this phase and it is retrieved from the previously integrated data collection.
- *Data transformation*: this phase is also known as data consolidation. It is a phase in which the selected data is transformed into forms appropriate for the mining procedure.
- *Data mining*: it is a very important step among all the steps, in which potentially useful patterns are extracted by applying data mining or classification techniques.
- *Pattern evaluation*: in this step, strictly interesting patterns representing knowledge are identified based on given measures.
- *Knowledge representation*: is the final phase in which the discovered knowledge is visually represented to the user. This essential step uses visualization techniques to help users understand and interpret the data mining results.

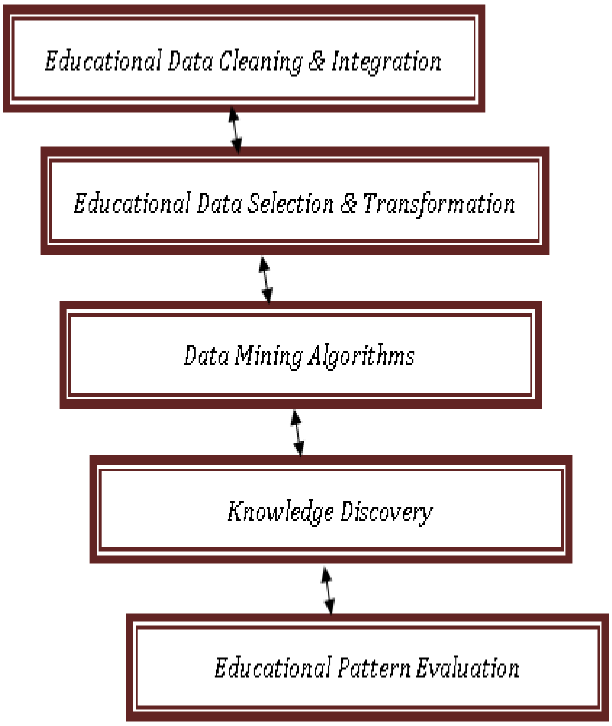

Educational Data Cleaning & Integration
Educational Data Selection & Transformation
Data Mining Algorithms
Knowledge Discovery
Educational Pattern Evaluation

Figure 2: Process of EDM

5. Applications of Algorithms in Education Mining

By implementing efficient algorithms of data mining in the field of education the improvement in education standard, high successful candidate ratio, low student drop out ratio and maximum efficiency in education system can be achieved with the increase in number of universities and students day by day [4]. Following is a detail of the algorithms used in education mining.

A. Support Vector Machine (SVM) : It is considered as an efficient tool to train data. If offers accurate methods among algorithms. SVM is the most worked upon algorithm for training purposes and a lot of research is still going on [5].

B. Apriori: In this technique, first of all we have to find out frequent data sets. Then the rules can be generated using these data sets those are very frequently used. These are called association rules. The steps includes generating Ak+1 the calculate support then putting items for minimum support.

C. The Expectation Maximization Algorithm: In this technique first the data is randomly observed. This is a simple mathematics based approach. The data is clustered continuously. That's why this technique is called as expectation maximization.

D. Page Ranker Page Ranker was given forth by Brin Karruy Page in 1998. On this algorithms basis they built Google, which has an excellent success ratio. It produces a static ranking of different web pages in sense that pager value is determined offline and does not depend on the online queries.

E. AD Tree (Alternating Decision Tree): This tree contains two nodes that are decision node and prediction node. The task of decision node is to specify Predicate condition. Prediction node can be as any of the leave and root. This is different from other techniques as in this the instance follows all paths in which decision nodes is true [5].

F. J-48: When there are different set of targets or it can be said that there are different set of predictors as dependent or independent variables then the technique J-48 is useful. It is a decision tree based technique. The prediction of the target variable of new dataset is allowed in this technique [5][6].

G. JRip: In this technique there are two type data sets one is growing set and the other is running set. This technique is called JRip because reduced error pruning technique is used in this algorithm. It is also a decision tree based algorithm. Heuristic method is used to form the initial set.

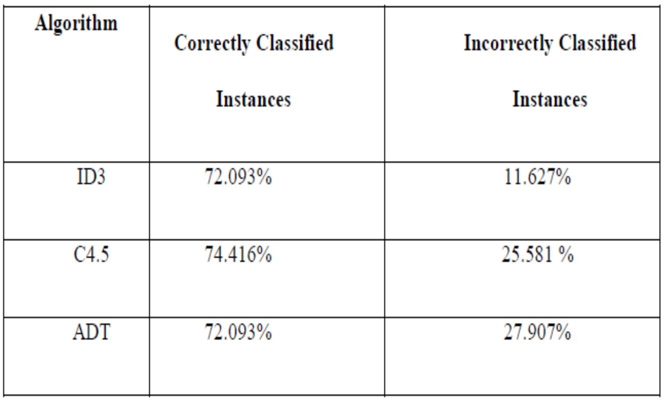
H. ID3 (Iterative Dichotomise 3): There are two phases in this technique tree building and tree pruning. If noise is there then it will not give accurate results. So before applying this technique, first we have to use some reprocessing technique to remove noise. The root node is selected with respect to the information gain and the arcs of the root not are the possible values of that attribute. Then check if all the possible outcomes are falling under the same class or not. If all instances are falling under the same class, the node is represented with single class name, otherwise choose the splitting attribute to classify the instances.

I. C4.5 and C5.0 both these algorithms are the successors of the ID3 technique, developed by Quinlan Ross based on the Hunt's algorithm. C4.5 handles both continuous and categorical attributes to construct a decision tree. In order to address continuous attributes, C4.5 separates 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. Both C4.5 and C5.0 also handles missing attribute values. C4.5 usages Gain Ratio as an attribute selection measure to develop a decision tree. It withdraws the biasness of information gain when there are many outcome values of an attribute.

6. Results and Discussion

As per the results, the analysis is done on the three decision tree based algorithms (techniques). The accuracy and execution times are measured on the basis of different set and loads of data sets to check the efficiency of the classifications techniques. Table 1 shown below is the accuracy comparison between ID3, C4.5 and ADTree classifiers.

Table 1: Classifiers Accuracy



Algorithm	Correctly Classified Instances	Incorrectly Classified Instances
ID3	72.093%	11.627%
C4.5	74.416%	25.581 %
ADT	72.093%	27.907%

Figure 3 shown below is the graphical representation of the accuracy comparison of these techniques.

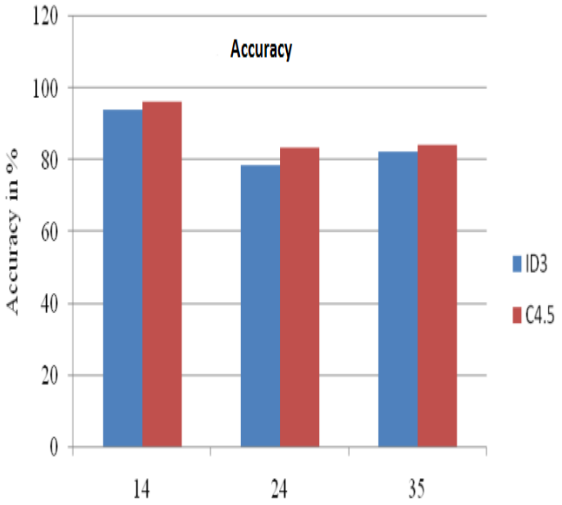


Figure 3: Classifiers Accuracy

Table 2 shows the time complexity in seconds of various classifiers to build the model for training data.

Table 2: Execution time bound

Algorithm	Execution Time (Sec)
ID3	0.12
C4.5	0.08
ADT	0.06

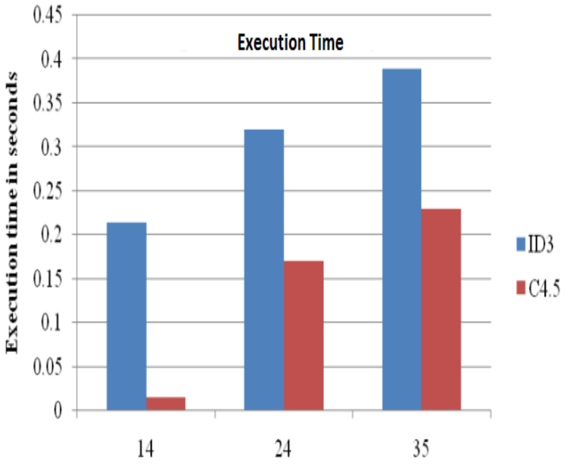


Figure 4: Classifiers Time Complexity

7. Conclusion

In this paper, study about the education data mining and the rules to extract useful information from the large educational databases is carried out. The importance of implementing data mining rules on educational databases and extracting knowledge from them is that student performance improvement and finding the path for placements can be made more efficient. The analysis in the paper is based on the implementation of the advanced data mining techniques like Graph structure, Naive Bayes, C4.5, Ripper and Machine learning algorithm SVM. These are some of the predictive algorithms to apply on educational data to generate the rules and check how many these yields correct results.

REFERENCES

1. J. Han and M. Kamber: Data Mining: Concepts and Techniques, Morgan Kaufmann, 2000. [5] Witten, I. H., Frank, E., Hall, M. A., Data Mining: Practical Machine Learning Tools and Techniques , 3rd Ed. Morgan Kaufmann, 2011.
2. Tinto, V: Dropout from Higher Education: A theatrical synthesis of recent research . Review of Education Research, 45, 89-125, 1975.
3. Kember, D: Open Learning Courses for Adults: A model of student progress . Englewood Cliffs, NJ. Educational Technology Publications, 1995.
4. Mosima Anna Masethe, Hlaudi Daniel Masethe : Prediction of Work Integrated Learning Placement Using Data Mining Algorithms Proceedings of the World Congress on Engineering and Computer Science 2014 Vol 1 WCECS 2014, 22-24 October, 2014, San Francisco, USA
5. Agrawal Bhawana, Gaurav Bharti: “Review on data mining techniques used for educational system”, IJETAE, Vol 4 Issue 11, 2014
6. May, P., Ehrlich, H.C., Steinke, T. ZIB Structure Prediction Pipeline: Composing a Complex Biological Workflow through Web Services. In: Nagel, W.E., Walter, W.V., Lehner, W. (eds.) Euro-Par 2006. LNCS, vol. 4128, pp. 1148--1158. Springer, Heidelberg 2006
7. Chandra, E. and Nandhini, K: Knowledge Mining from Student Data, European Journal of Scientific Research, vol. 47, no. 1, pp. 156-163, 2010
8. Z. N. Khan : Scholastic achievement of higher secondary students in science stream , Journal of Social Sciences, Vol. 1, No. 2, pp. 84-87, 2005