Advanced Science and Technology Letters
Vol.63 (NGCIT 2014), pp.235-239
http://dx.doi.org/10.14257/astl.2014.63.52

Educational Data Mining: an application of data mining techniques with the educational web-based system for predicting student performance

Ji Hoon Seo, Irfan Ajamal Khan, Jin Tak Choi

Incheon National Unversity, South Korea
sserz@incheon.ac.kr, manikhan@nate.com, choi@incheon.ac.kr

Abstract: Mining data in educational filed is an important and useful for educational institution to calculate and predict the performance of students. It can also help to classify the level of students and they can be guided and educated according to their level to achieve better performance during and after studies. Additionally it can help students to choose subjects those suites their interest and concern. Classification is a data mining function that assigns items in a collection to target categories or classes. The goal of classification is to accurately predict the target class for each case in the data. There are many classification algorithms because of their wide range of application. Classification Decision tree are commonly used because they are easy to understand and implement. In this paper we have explain our system, which is developed to predict and calculate the performance of the students with help of data mining techniques such as Classification and prediction. In Educational data mining the most useful classification algorithm is decision tree algorithm, because of they are easy to design and implement. The ID3, C4.5 decision tree algorithms have been applied on the data of students to predict their performance. Performance and results are compared of all algorithms and evaluation is done by already existing datasets. All the algorithms have a satisfactory performance but accuracy is more witnessed in case of C4.5 algorithm.

Keywords: Educational Data Mining, ID3, C4.5, Decision trees, Classification

1 Introduction

Twice a year, educational institutes are welcoming different students in different departments having variety of courses from different regions, educational background and with varying level of knowledge. Analyzing the past performance of these students would provide a better academic performance of students in the future. This can be achieved by using the concepts of data mining. Educational data mining is used to study the data available in the educational field and bring out the hidden knowledge from it. Classification methods like decision trees, rule mining, Bayesian network etc. can be applied on the educational data for predicting the students behavior, performance in

ISSN: 2287-1233 ASTL

examination etc. This prediction will help the tutors to pinpoint the weak students and help them to achieve better performance in future.

For this purpose, we have analyzed the data of students enrolled in first year of computer engineering department. This data was obtained from the information provided by the admission department of the institute. It includes their detailed information about education career. We have applied the ID3 and C4.5 algorithms after pruning the dataset to predict the results of these students in their first semester as precisely as possible.

2 Background

Data mining is to extract information from a data set and transform it into an understandable structure of future use. It involves methods of artificial intelligence, machine learning, statistics and database system. Now days, the amount of data a student need to get admission in any educational institute is increasing and getting complicated. On the other hand institutes have to store these all complicated information into their database. Our project is aimed to mine all that hidden knowledge from the data and make it useful and more meaningful information that institutes and students can use for better and safer decisions.

Classification is a data mining technique that assigns items in a collection to target categories or classes. The goal of classification is to accurately predict the target class for each case in the data. For example, a classification model could be used to identify students final GPA or the research area in future.

Algorithm of Decisions Tree Induction: The basic algorithm for a decision tree induction is a greedy algorithm approach. The construction of the tree is top-down recursive divide-and-conquer manner. Pseudo code for generating decision tree algorithm "gendectree(Sv,A-a)" as follow

Step1*: create a node N*
Step2*: if all samples are of the same class C the label N with C; terminate*
Step3*: if A is empty the label N with the most common class C in S(voting); terminate;*
Step4*: Select a ∈A, with highest information gain; Lagel N with a;*
Step5*: For each value v of a;*

a. *Grow a branch from N with condition a=v;*
b. *Let S_v be the subset of samples in S with a=v;*
c. *If S_v empty then attach a leaf labeled with the most common class in S;*
d. *Else attach the node generated by gendectree(S_v,A-a)*

Attribute Selection Measure: Information Gain (ID3/C4.5) Algorithm: ID3 developed by J. Ross Quinlan is based off the Concept Learning System (CLS) algorithm. ID3 improves on CLS by adding a feature selection heuristic. ID3 searches through the attributes of the training instances and extracts the attribute that best separates the given

examples. If the attribute perfectly classifies the training sets then ID3 stops; otherwise it recursively operates on the n (where n = number of possible values of an attribute) partitioned subsets to get their "best" attribute. The algorithm uses a greedy search, that is, it picks the best attribute and never looks back to reconsider earlier choices.

C4.5 is an extension of ID3 algorithm and it is a well-known algorithm for generating decision trees. It can also be referred to as a statistical classifier. Some of the C4.5 improvements over ID3 algorithm [2] are, missing values, cost handling, Pruning the decision tree after its creation, handling attributes with discrete and continuous values and other measures, e.g., goodness of rules, such as decision tree size or compactness of classification rules.

step by step illustration as below:

Step1*: Select the attribute with the highest information gain*

Step2*: Let p_i be the probability that an arbitrary tuple in D belongs to class C_i, estimated by $|C_{i,D}|/|D|$*

Step3*: Expected information (entropy) needed to classify a tuple in D:* $Info(D) = -\sum_{i=1}^{m} p_i \log_2(p_i)$

Step4*: Information needed (after using A to split D into v) to classify D:* $Info_A(D) = \sum_{j=1}^{v} \frac{|D_j|}{|D|} \times I(D_j)$

Step5*: Information gained by branching on attribute A* $Gain(A) = Info(D) - Info_A(D)$

3 Implementing the Performance Prediction Web Application

Entire implementation of our project is as follow; *Student Database, Data preprocessing using RapidMiner and a web-based UI for student's performance prediction and guidance*. Student database contains the detailed information of each student. We then segmented the training dataset further, considering various feasible splitting attributes, i.e., the attributes which would have a higher impact on the performance of a student. Classification was done by feeding pruned database to RapidMiner with the help of classification algorithms (ID3 and C4.5).

Implementing the Performance Prediction Web Application: The predictive patterns obtained from RapidMiner for prediction were then implemented in a web-based application to get the predicted result for student(s). The web application was developed using a popular PHP framework named CodeIgniter.

Verifying Accuracy of Predicted Results: The accuracy of the algorithm was tested by the comparing obtained predicted result of evaluation with actual results obtained. The accuracy achieved was 79.315% for both ID3 and C4.5 algorithms. And the accuracy of ID3 algorithm for further research and guidance was 74.512%. C4.5 algorithm showed little bit better accuracy (75.157%) in further research and guidance prediction.

Further Evaluation: Further evaluation was done for predicting the performance and future guidance of student(s) by evaluating each student's record in a multiple and single manners. It was done by entering student's id one by one and by uploading an external file such as (MS excel file). For the purpose of accuracy some of the results were taken for graduated students and the students of 3rd and 4th grade.

The accuracy of the algorithm results was calculated and verified by comparing with predicted result of evaluation and actual results. Figure 11 shows that the accuracy achieved is 75.145% for both ID3 and C4.5 algorithms. Figure12 shows the mismatched tuples, i.e. the tuples which were predicted wrongly by the application for the current test data.

4 Future work & Conclusion

In this paper, we have explained the system that is used to predict and guide for better and safer decision in future. In this project, prediction parameters are not updated dynamically in the future we are planning to make the whole implementation. Also, we are planning to mine curriculum system and other extra activities for every semester, which we believe may have a significant impact on the overall performance and after graduation activities. Considering such parameters would result in better accuracy of prediction.

Acknowledgement. This research was supported by a grant (12-TI-C01) from Advanced Water Management Research Program funded by Ministry of Land, Infrastructure and Transport of Korean government.

References

1. Wikipeida: www.wikipedia.com
2. Han, J. and Kamber, M., (2006) Data Mining: Concepts and Techniques, Elsevier.
3. http://docs.oracle.com/cd/B28359_01/datamine.111/b28129/classify.htm#i1005746
4. Dunham, M.H., (2003) Data Mining: Introductory and Advanced Topics, Pearson Education Inc.
5. Kantardzic, M., (2011) Data Mining: Concepts, Models, Methods and Algorithms, Wiley-IEEE Press. International Journal of Data Mining & Knowledge Management Process (IJDKP) Vol.3, No.5, September 2013
6. Ming, H., Wenying, N. and Xu, L., (2009) “An improved decision tree classification algorithm based on ID3 and the application in score analysis”, Chinese Control and Decision Conference (CCDC), pp1876-1879.
7. Xiaoliang, Z., Jian, W., Hongcan Y., and Shangzhuo, W., (2009) “Research and Application of the improved Algorithm C4.5 on Decision Tree”, International Conference on Test and Measurement (ICTM), Vol. 2, pp184-187.
8. RapidMiner, http://rapid-i.com/content/view/181/190/
9. Stanley TenEyck Schuyler,"Using Problematizing Ability to Predict Student performance In A First Course In Computer Programming", Robert Morris University, Copyright © Stanley TenEyck Schuyler 2008.

10. Sebastian Nowozin, "Improved Information Gain Estimates for Decision Tree Induction", Appearing in Proceedings of the 29 th International Conference onMachine Learning, Edinburgh, Scotland, UK, 2012