Journal of Computer Science

Original Research Paper

A Classification and Prediction Model for Student's Performance in University Level

Ashraf Abazeed and Moaiad Khder

College of Art and Science, Applied Science University, East Ekir, Bahrain

Article history
Received: 05-05-2017
Revised: 19-06-2017
Accepted: 14-07-2017

Corresponding Author:
Ashraf Abazeed
College of Art and Science,
Applied Science University,
East Ekir, Bahrain
Email: support@thescipub.com

Abstract: Educational Data Mining is a new discipline, focusing on studying the methods and creating models to utilize educational data, using those methods to better understand students and their performance. We implemented two different techniques on our dataset; classification used to build a prediction model and association rules were used to find interesting hidden information in the student's records. This study will help the student's to determine their direction and improve when necessary to cope up with their studies. It also provide a great tool to predict and evaluate those students who need attention and correction actions and find out any deviation before it happen and become a decrease in performance and reduce failure rate.

Keywords: Data Mining, Educational Mining, Performance, Classification, Association Rules

Introduction

Knowledge Discovery in Database (KDD) is defined as the "extraction of implicit, unknown and potentially useful information from data". The word implicit means that we are looking for information that is contained in the database and unknown stands for a result or information we did not expect to have before. KDD consist of many steps and one of them is data mining. The knowledge discovery process takes the raw results from data mining and transforms them into useful and understandable information which can be used in different implementations and decisions making processes. Knowledge discovery is a multi-steps process, these processes include data integration, preparation and transformation, data mining as well as evaluation of the results of the data mining process, those processes can be iterative and every time the results would be enhanced or a new information could be discovered, Geist (2002).

Data mining, *the process of extracting any hidden predictive information from large databases*, is a dominating new technology with promising potential to help companies focus on the most important information in their data warehouses and decision making by decreasing time and providing new frontiers and aspect never thought of before. Data mining tools (association rules, clustering) predict future trends and behaviors, allowing businesses to make practical, knowledge-driven decisions. "Higher Educational Institution (HEI) is greatly concern on the student's enrollment data to understand the influence on student's decision to attend their institution and on their study's information to check their performance" Abu Haris *et al.* (2016).

Data mining commonly involves four classes of task:

- Classification - Arranging the data into predefined groups or classes. For example a university classifies its students into undergraduate and postgraduate. Common algorithms include nearest neighbor, Naive Bayes classifier and neural network
- Clustering - another form of classification where the arranging of items is not predefined, so the algorithm will try to group similar items together according to a center or cluster point
- Regression - Attempts to model the data into a function that represent the whole sample with least error. A common method is to use Genetic Programming
- Association rule learning - Searches for similarities among large database. For example a supermarket might collect data related to customers and using association rule learning, it can find out what products are most likely to be bought together, this kind of information can be used for marketing purposes and to improve the sales process, Agrawal *et al.* (1993)

Science Publications

Educational Data Mining is a new discipline, focusing on studying the methods and creating models to utilize the educational data, using those methods to better understand students and their performance.

Educational data mining is an emerging stream where students, academics and research analysts can use. It provides students with tools and measures to check their performance, it also provides academic with indicators and prediction for student's performance. Researchers find it very interesting basis to build applications and implementations. Provided that educational data is not used recently and it's a very promising field.

"The main Goals of educational data mining are (EDM, 2017):

- Predicting students' future learning behavior by creating student models that incorporate such detailed information as students' knowledge, motivation, metacognition and attitudes
- Discovering or improving domain models that characterize the content to be learned and optimal instructional sequences
- Studying the effects of different kinds of pedagogical support that can be provided by learning software
- Advancing scientific knowledge about learning and learners through building computational models that incorporate models of the student, the domain and the software's pedagogy"

Related Work

Educational data mining is a new and interesting area of research, recently it gained its popularity due to the vast amount of data available in the educational process(which can be mined) and the increase emphasize on quality of education in university levels. Tools and models are required nowadays to identify average and poor students, correction and detection steps can be done before it's too late.

Baradwaj and Saurabh (2011) created an ID3 classifier to classify the student's division on a 50 student's database collected from one university, 7 attributes were used to build a tree that predict and classify student's end semester mark. This study uses the classification ID3 algorithm and entropy of the data in education to help the students and teachers to improve the division of the students.

Santillan *et al*. (2016) built an incremental interaction system to predict student's performance, data was gathered from two different years, three different classification algorithms were used to classify the same data and a comparison was made between the three output. In our study we used classification and association rules mining, it tends to be more accurate and efficient to use more than one data mining aspect in order to analyze the data.

Widyahastuti *et al*. (2017) used a different technique; using linear regression to predict the student's performance by monitoring and using of an online discussion forum, in their proposed prediction model the user is the main input of the dataset, the dataset was divided into three different parts: Online discussion, forums and course assessment. 11 features were used as an input to the data mining tool. A correlation analysis was conducted and the result was discussed and explained.

Ahadi *et al*. (2017) did an interesting study on student's performance during the whole semester (a week by week and assignment by assignment basis) and compared the student's performance towards the end of the semester. Students were clustered based on their weekly performance. The study showed that the student's performance declines when it comes towards the end of the semester and that was due to the nature of studying programming subjects which increase in its difficulty week by week.

Daud *et al*. (2017) prepared a study that tackled the student's performance prediction problem, data has been collected from graduate and undergraduate universities which made up a 3000 record, after preprocessing the number of records was reduced to around 700 records. The study tried to answer the question of will the student complete his study or not. A feature analysis was conducted and feature spaces was selected from relevant attributes. Four categories of attributes were introduce and the influence of those categories on the student's performance was the result of that study.

Maja *et al*. (2015) used association rule mining in education, using the data extracted from a learning and management system (Moodle), taking one subject as a testing and studying case using 77 records. Five attributes were selected for the mining process, the grade of the student was the class label.

After the mining algorithm was implemented on the dataset, a number of rules were generated with a specified support and confidence. The rules generated can help to give an indication of student's performance in general.

Data Preparation and Data Mining

Educational Data Mining is concerned with developing models for exploring and analyzing the vast amount of (unused) data that come from educational institutions and using hose models to better understand students, predict and help them to perform better in their study.

Nowadays, student's performance is measured by internal assessment methods such as midterm exams, quizzes, final exams and assignments, student's results should be above a certain mark to pass the subject.

Academic institutions (Schools and universities) generate huge data on students, courses, faculty, staff that includes managerial systems, organizational personnel, lectures details and so on. This data is the base for any data mining application, the input to any academic institution for improving the quality of education process, Ranjan and Malik (2007).

The data set of 242 students used in this study was obtained from the college of art and science of the applied science university (Bahrain) from the session 2015/2016, the dataset was obtained from the registration department according the student's latest information and recorded into one main table.

Feature Selection

This is a fundamental filtering step to discard any of the irrelevant attributes and reduce the dimensionality of data while improving accuracy.

Using the Gain Ratio Attribute Evaluation, which evaluates the worth of an attribute by measuring the gain ratio with respect to the class. It rank all the attributes according to the importance. We choose 12 attributes to implement in our model as explained in Table 1.

Other attributes such as (parent's background knowledge, student's age, student's distance from college and long Vs short semester attributes among others were discarded as it was in the bottom of the ranked list.

Data Selection

Only those relevant fields were selected which were suitable for the data mining process. While some of the information for the fields were extracted from the database. The dataset was stored in a nominal format to suit the classification and association rules mining process.

Table 1. Dataset descriptions

Field name	Description	Options
Gender	student's Gender	M: Male F: Female
High School Grade	Represent student's high school grade.	90z: 90-100 80z: 80-89 70z: 70-79 60z: 60- 69
Major in high school	Represent student's major in high school	Science Commercial Industry
Previous GPA	Represent student's previous GPA	Excellent Very Good Good Poor
Number of courses registered	A number represent the current semester registered course.	Four Five Six
Sponsor	Specify whether the students is sponsor or not	Yes No
Advisory visit	Specify the frequency of visiting the academic advisor.	Frequent Average Poor
English score	The student's level in English	Excellent Good Poor
Attendance	The student's attendance in the whole semester overall	Excellent Good Poor
Core Vs elective	Specify if the number of core subjects is more than elective courses and vice versa	Core Elective
Student time	Specify the average study time per day for all the subjects	Two Three Four
Performance	The class field. Specify if the performance is getting better or stable or decreasing.	Better Stable Decrease

The Random Tree Classifier

The random trees classifier is a powerful technique for classification in general which is resistant to over fitting and can work with segmented fields it perform the Random Trees classification on a field basis, based on the input training feature file.

Random Trees is a collection of individual decision trees where each tree is generated from different samples and subsets of the training data. The idea behind calling these decision trees is that for every field that is classified, a number of decisions are made in rank order of importance.

When you graph these out for a field, it looks like a branch. When you classify the entire dataset, the branches form a tree. This method is called random trees because you are actually classifying the dataset a number of times based on a random sub selection of training fields, thus resulting in many decision trees, Ali *et al*. (2012).

To make a final decision, each tree has a vote. This process works to mitigate over fitting. Random Trees is a supervised machine-learning classifier based on constructing a multitude of decision trees, choosing random subsets of variables for each tree and using the most frequent tree output as the overall classification.

The Apriori Algorithm

When association rule mining was first introduced by Agrawal *et al*. (1993) an algorithm called AIS, Agrawal and Ramakrishnan (1994) was given for discovering the large itemsets. However, the AIS algorithm is not efficient, since it generates too many unnecessary candidates.

In the following year, the Apriori algorithm was proposed, which improves the performance from AIS by reducing the number of unnecessary candidates. Also, an OCD algorithm with a similar approach was proposed by Hipp *et al*. (2000) concurrently.

Results and Discussions

The data set of 242 students used in this study was obtained from the computer science department of the applied science university (Bahrain) from the session 2015/2016.

To better understand the data used in out experiment, we chose the fields according to students and professors surveys, field relevance selections and according to the factors affecting student's performance reviews. Only useful and relevant field which we found out that it affect student's performance were used.

We first implemented the Random Tree Classifier on the dataset and come out with the Model in Table 2, which can be used to predict and evaluate any student against his/her performance during their study in the university.

Table 2. Prediction model

```
studytime = four
| gender = f
| | numberofcoursesregistered = five: better
| | numberofcoursesregistered = four: stable
| | numberofcoursesregistered = six: better
| gender = m: better
studytime = one
| numberofcoursesregistered = five
| | previousgpa = Excellent: decrease
| | previousgpa = Good: decrease
| | previousgpa = Poor: decrease
| | previousgpa = VeryGood
| | | majorinhighschool = Commercial: better
| | | majorinhighschool = Industry: better
| | | majorinhighschool = science: decrease
| | previousgpa = poor: decrease
| numberofcoursesregistered = four: stable
| numberofcoursesregistered = six: decrease
studytime = three
| sponsor = no
| | corevselective = Core: better
| | corevselective = Elective
| | | advisoryvisits = Poor: decrease
| | | advisoryvisits = average: better
| | | advisoryvisits = frequent
| | | | englishscore = Excellent: better
| | | | englishscore = Good: decrease
| | | | englishscore = Poor: better
| | corevselective = c: better
| | corevselective = core: better
| sponsor = yes
| | englishscore = Excellent: better
| | englishscore = Good: stable
| | englishscore = Poor
| | | numberofcoursesregistered = five: better
| | | numberofcoursesregistered = four: stable
| | | numberofcoursesregistered = six: better
studytime = two
| gender = f
| | majorinhighschool = Commercial
| | | corevselective = Core
| | | | englishscore = Excellent: stable
| | | | englishscore = Good: better
| | | | englishscore = Poor: better
| | | corevselective = Elective: decrease
| | | corevselective = c: better
| | | corevselective = core: better
| | majorinhighschool = Industry: stable
| | majorinhighschool = science
| | | previousgpa = Excellent: better
| | | previousgpa = Good: stable
| | | previousgpa = Poor: stable
| | | previousgpa = VeryGood: decrease
| | | previousgpa = poor: stable
| gender = m
| | englishscore = Excellent: better
| | englishscore = Good
| | | corevselective = Core: decrease
| | | corevselective = Elective: better
| | | corevselective = c: decrease
| | | corevselective = core: decrease
| | englishscore = Poor
| | | attendace = Excellent: better
| | | attendace = Good
| | | | highschoolgrade = 60z
| | | | | advisoryvisits = Poor: decrease
| | | | | advisoryvisits = average: decrease
| | | | | advisoryvisits = frequent: better
| | | | highschoolgrade = 70z: stable
| | | | highschoolgrade = 80z: stable
| | | | highschoolgrade = 90z: decrease
| | | attendace = Poor: decrease
```

The Size of the tree was 73 and we choose a Maximum depth of tree: 3, the number of the tree's depth can be modified, it can be reduced to 2 or expand to more than 3.

The model clearly show the importance of the amount of study time the student spend as it is the root node and it was used to classify the model according to its attributes and as a general note it shows that the student's previous GPA is not an indication of what might happen in the following semesters.

The academic advisory visits and the follow up with the supervisor plays a major role in student's performance as it clearly shows that students who have frequent visits to their academic supervisor tend to perform well in their academic results.

The major of the high school plays an interesting role in this model, students majoring in science stream tends to perform well and students majoring in industry stream tends to perform less than expected and need more time to study to follow up with other student's streams.

The complete model can be transformed into an if-then rule where the current students can be checked for their performance and any new student's information can be used to predict his performance using this model.

Prediction and Evaluation Model

Association Rules Mining

The same dataset then have been used in the process of association rules mining, where large item sets are created and intersected in order to generate an interesting rules for student's performance.

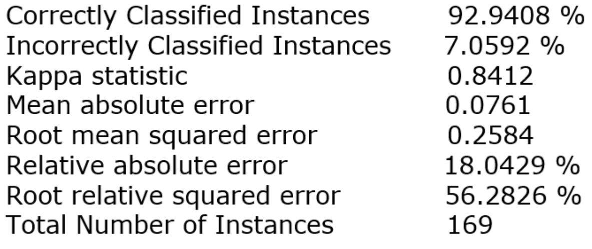
The Best rules found according to our minimum support of 100% and a confidence of 95%. We reduced the minimum support so we can generate a vast amount of rules. Then we increase the confidence to only consider those rules with high impact and relevance. A sample of association rules is presented below.

- majorinhighschool = science sponsor = no corevselective = Core studytime = three 2 ==> performace = better conf:(100)
- majorinhighschool = science sponsor = no studytime = three ==> performace = better conf:(100)
- highschoolgrade = 80z majorinhighschool = science attendace = Good ==> performace = better conf:(0.99)
- gender = m englishscore = Excellent ==> performace = better conf:(0.99)
- numberofcoursesregistered = five sponsor = no studytime = three ==> performace = better conf:(0.98)
- highschoolgrade = 80z majorinhighschool = science englishscore = Good attendace = Good ==> performace = better conf:(0.98)
- majorinhighschool = science corevselective = Core studytime = three ==> performace = better conf:(0.98)
- advisoryvisits = average studytime = three ==> performace = better conf:(0.97)
- gender = f highschoolgrade = 80z englishscore = Good ==> performace = better conf:(0.97)
- majorinhighschool = science studytime = three ==> performace = better conf:(0.97)
- highschoolgrade = 80z corevselective=Core studytime = three ==> performace = better conf:(0.96)
- majorinhighschool = science advisoryvisits = frequent studytime = three ==> performace = better conf:(0.96)
- attendace = Excellent corevselective = Core studytime = three ==> performace = better conf:(0.95)

since association rules mining finds the interesting and hidden relations between fields, so academic advisors and supervisors can use this information to detect the variation of student's performance and try to fix it before any unwanted change occur. The number of rules can vary according to minimum support and minimum confidence. We chose some of the interesting rules to present in this study. As we analyzed the two results we found out that the two models present many interesting rules that can be used to analyze and predict student's performance.

Cross Validation Analysis

10-fold Cross validation is used to validate the robustness of the mining model. 70% of the data were used as a training data set and the remaining 30% were used as a testing data set. The results in Fig. 1 shows an interesting results for the model validation process.



```
Correctly Classified Instances        92.9408 %
Incorrectly Classified Instances      7.0592 %
Kappa statistic                       0.8412
Mean absolute error                   0.0761
Root mean squared error               0.2584
Relative absolute error              18.0429 %
Root relative squared error          56.2826 %
Total Number of Instances             169
```

Fig. 1. Cross validation results

Conclusion

In this study, we presented a model for predicting and evaluating student's performance in education. We implemented two different techniques on our dataset, classification used to build a prediction model and association rules were used to find interesting hidden information in the student's records.

This study will help the student's to determine their direction and improve when necessary to cope up with their studies. It also provide a great tool to predict and evaluate those students who need attention and correction actions and find out any deviation before it happen and become a decrease in performance and reduce failure rate.

Acknowledgment

The authors wishes to thank the deanship of art and science and the deanship of admission and registration in the Applied Science University Bahrain.

Author's Contributions

Ashraf Abazeed: Participated in the data preparation, feature selection, prediction model, testing and discussions.

Moaiad Khder: Participated in the related work, literature review, association rules generation, results, cross validation analysis and proof reading.

Ethics

The authors declares that there is no ethical issues that may arise after the publication of this manuscript.

References

Abu Haris, N., A. Munaisyah, H. Nurdatillah and A. Fauziah, 2016. A study on students enrollment prediction using data mining. Proceedings of the 10th International Conference on Ubiquitous Information Management and Communication, Jan. 04-06, ACM, Danang, Viet Nam. DOI: 10.1145/2857546.2857592

Agrawal, R., T. Imielinski, N. Arun and Z. Swami, 1993. Mining association rules between sets of items in large databases. Proceedings of the ACM SIGMOD International Conference on Management of Data, May 25-28, ACM, Washington, D.C., pp: 207-216. DOI: 10.1145/170035.170072

Agrawal, R. and S. Ramakrishnan, 1994. Fast algorithms for mining association rules in large databases. Proceedings of the 20th International Conference on Very Large Data Bases, Sept. 12-15, Morgan Kaufmann, USA, pp: 487-499.

Ahadi, A., S. Lal, J. Leinonen, A. Hellas and R. Lister, 2017. Performance and consistency in learning to program. Proceedings of the 19th Australasian Computing Education Conference, Jan. 31-Feb. 03, ACM, Australia, pp: 11-16. DOI: 10.1145/3013499.3013503

Ali, J., R., Khan, N. Ahmad and I. Maqsood, 2012. Random forests and decision trees. Int. J. Comput. Sci., 9: 272-278.

Baradwaj, B. and P. Saurabh, 2011. Mining educational data to analyze students performance. Int. J. Adv. Comput. Sci. Applic., 2: 63-69.

Daud, A., N. Aljohani, R. Abbasi, M. Lytras and F. Abbas *et al*., 2017. Predicting student performance using advanced learning analytics. Proceedings of the 26th International Conference on World Wide Web Companion, Apr. 03-07, International World Wide Web Conferences Steering Committee, Perth, Australia, pp: 415-421. DOI: 10.1145/3041021.3054164

EDM, 2017. What is Educational Data Mining (EDM).

Geist, I., 2002. A framework for data mining and KDD. Proceedings of the ACM Symposium on Applied Computing, Mar. 11-14, ACM, Madrid, Spain, pp: 508-513. DOI: 10.1145/508791.508887

Hipp, J., G. Ulrich and N. Gholamreza, 2000. Algorithms for association rule mining-a general survey and comparison. ACM SIGKDD Explorat. Newslett., 2: 58-64. DOI: 10.1145/360402.360421

Maja, M., M. Bakaric and S. Sisovic, 2015. Association rule mining and visualization of introductory programming course activities. Proceedings of the 16th International Conference on Computer Systems and Technologies, Jun. 25-26, ACM, Dublin, Ireland, pp: 374-381. DOI: 10.1145/2812428.2812438

Ranjan, J. and K. Malik, 2007. Effective educational process: A data mining approach. VINE J. Inform. Knowl. Manage. Syste., 37: 502-515. DOI: 10.1108/03055720710838551

Santillan, M., M. Paule-Ruiz, R. Cerezo and J. Nuñez, 2016. Predicting students' performance: Incremental interaction classifiers. Proceedings of the 3rd ACM Conference on Learning @ Scale, Apr. 25-26, ACM, Edinburgh, Scotland, UK, pp: 217-220. DOI: 10.1145/2876034.2893418

Widyahastuti, F., Y. Riady and Z. Wanlei, 2017. Prediction model students' performance in online discussion forum. Proceedings of the 5th International Conference on Information and Education Technology, Jan. 10-12, ACM, Tokyo, Japan, pp: 6-10. DOI: 10.1145/3029387.3029393