Educational Data Analytics using Association Rule Mining and Classification

Mrs K.Thrilochana Devi, M.Mallikarjuna Rao, CH.Mohan Gopi Krishna, CH.Rakesh, SK.Abdulla
Department of Information Technology, Vasireddy Venkatadri Institute of Technology, Guntur, Andhra Pradesh, India
Department of Information Technology, Vasireddy Venkatadri Institute of Technology, Guntur, Andhra Pradesh, India

Abstract

The education crisis is now widely spread in global in term of decreasing number of student and decreasing degree requirements for some jobs. Educational data mining (EDM) is recently interested in data mining area to discover useful knowledge in educational data to help educators improve their administration planning and student services. This paper proposes applying of two data mining technics in educational data. First, association rule was applied in admission data to find some knowledge for supporting admission planning. Second, decision tree was applied in course grades and job data of graduated student to predict job after graduated. The results of these studies give good knowledge for admission planning and job prediction.

Keywords—educational data mining, classification, association rule mining

1. INTRODUCTION

Educational Data Mining (EDM) is a new trend in the data mining and Knowledge Discovery in Databases (KDD) field which focuses in mining useful patterns and discovering useful knowledge from the educational information systems, such as, admissions systems, registrations systems, course management systems, and any other systems dealing with students at different levels of education, from schools, to colleges and universities. Researches in the field applying of two data mining technics in educational data. First, association rule was applied in admission data to find some knowledge for Supporting admission planning. Second, decision tree was applied in course grades and job data of graduated student to predict job after graduated.

There are a lot of objective of EDM depend on data source and educational problem. In 2017, F. Matsebula and E. Mnkandla [7] proposed an architecture for big data analytics in higher education. The architecture composed of five parts; data gathering device which collected student data from various data source (student's card, social networking and student information system), data storage and management system which consists of bigdata management, data analytics system which process algorithms from data, data visualization which help in decision making process, and action system for providing alerts, warning or guiding to student or administrators.

The main objective of this research is to answer two main questions using data mining algorithms. First, how data mining can help admission working process. Second, how data mining can predict the student's jobs. To answer these questions, we provide two task of data mining with two difference sources of data. First, the association rule mining is used to discover interesting relation between feature in admission data, such as school name, province, region, admission project, faculty. The result of association rule mining could be used to help for admission planning. Second, the ID3 which is a classification algorithm invented by Ross Quinlan [9] used to generate a decision tree from student's course grade dataset which is mapped each student's course grade to their job after graduated. The rule result from ID3 showed that accuracy and precision for student's job prediction.

The rest of this paper organized as follows. 2.Related Works and 3. Knowledge Discovery in Databases and 4. Presents Association rule mining and Id3 decision tree and 5. Presents Experimental Design and 6. Presents Experimental Results and Finally 7. Contains Conclusion and Future works.

2. RELATED WORK

In 2011, B.K. Baradwaj and S. Pal [4] studied on classification task of student database to predict the student's performance in end semester examination. The student' performance was divided into four stages (first, second, third and fail). It helps earlier in identifying the dropouts and student who need special attention from teacher. In 2015, N. Buniyamin et al. [5] presented the use of Neuro-Fuzzy classification in a student's academic data in an electrical engineering faculty of Malaysian public university. The study showed that the output of system can determine probability of student to achieve excellent grade even if the student achieved weak in certain course or subject.

In 2016, A. A. Saa [6] used multiple data mining tasks to create qualitative predictive models to predict the students' grades from educational dataset. Four decision tree algorithms have been implemented and Naïve Bayes algorithm. The results can motivate the university to perform data mining task on their student data, as well as student to improve their performances.

3. KNOWLEDGE DISCOVERY IN DATABSE

The term Knowledge Discovery in Databases, or KDD for short, refers to the broad process of finding knowledge in data, and emphasizes the "high-level" application of particular data mining methods. It is of interest to researchers in pattern recognition, databases, statistics, artificial intelligence, knowledge acquisition for expert systems, and data visualization.

1.Data Cleaning: to remove noise and inconsistent data.

2. Data Integration: where multiple data sources may be combined.

3.Data Selection: where data relevant to the analysis task are retrieved from the database.

4.Data Mining: an essential process where intelligent methods are applied to extract data patterns.

5.DataTransformation: where data are transformed and consolidated into forms appropriate for mining by performing summary or aggregration operations.

6.Pattren Evaluation: to identify the truly interesting patterns representing knowledge based on intrestigness measures.

7.Knowledge Representation: where visualization and knowledge representation techniques are used to present mined knowledge to users.

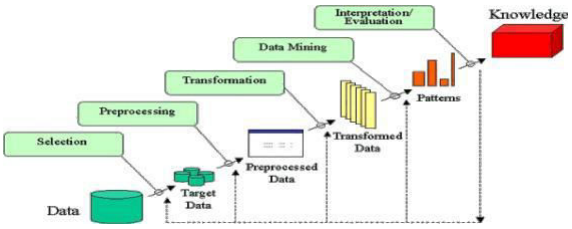

Fig. 1. The KDD process

4.ASSOCIATION RULE MINING AND ID3 DECISION TREE

a. Association Rule Mining

Association rule mining is a data mining method to discovery interesting relationship between features. Association rule generation as two-step approach:

i. Frequent Itemset Generation: Generate all itemsets whose support>=min sup

ii.Rule Generation: Generate high confidence rules from each frequent itemset,where each rule is a binary partitioning of a frequent itemset.

A widely-used algorithm for the association rules mining is the Apriori algorithm [10]. Apriori is a seminal algorithm proposed by R.Agarwal and R.Srikant in 1994 for mining frequent itemsets for Boolean association rules. The name of the algorithm is based on the fact that the algorithm uses prioir knowledge of frequent itemset properties.

```
ProcedureApriori (T, minSup){
//T is dataset and minSup is minimum
support value Ck: Candidate itemset of size
k
Lk: frequent itemset
of size k L1 =
{frequent itemsets};
for(k=1; Lk != ; k++) do
begin Ck+1 = candidates
generated from Lk;
    for each transaction t in dataset do{
        increment the count of all candidates in
Ck+1 that are contained in t
        Lk+1 = candidates in ck+1 with min_sup
        }
end
return Uk, Lk
```

Fig. 2. Pseudocode of the Apriori algorithm

b. ID3 Decision tree

ID3 (Iterative Dichotomiser) algorithm is a statistical model used for generating decision tree from a dataset. The basic idea of ID3 algorithm is to construct the decision tree by applying a top-down, greedy search through the given sets of training data to test each attribute at every node [x].It was developed by Ross Quinlan. The key of decision tree construction is root node selection which using statistical method call information gain. The information gain can evaluate how well of chosen node that might be generate smallest decision tree. The information gain equation as follows:

$$\text{Gain(S, A)} = Entropy(S) - \sum_{v \in Values(A)} \frac{|S_v|}{|S|} Entropy(S_v)$$

where *Values(A)* is the set of all possible values for attribute *A*, and S_v is the subset of *S* for which the attribute A has value *v*. The *Entropy(S)* is a measure in the information theory, which describes the diversity of an arbitrary collection of data as follows:

$$\text{Entropy(S)} = \sum_{i=1}^{c} -p_i \log_2 p_i$$

where p_i is the probability of S belonging to class *i*. The procedure for decision tree generation as follows:

```
INPUT: S, where S = set of classified instances
OUTPUT: Decision Tree
Require: S ≠ ∅, num_attributes > 0
 1: procedure BUILDTREE
 2:     repeat
 3:         maxGain ← 0
 4:         splitA ← null
 5:         e ← Entropy(Attributes)
 6:         for all Attributes a in S do
 7:             gain ← InformationGain(a, e)
 8:             if gain > maxGain then
 9:                 maxGain ← gain
10:                 splitA ← a
11:             end if
12:         end for
13:         Partition(S, splitA)
14:     until all partitions processed
15: end procedure
```

Fig. 3.Pseudocode of the build decision tree in ID3

5 EXPERIMATAL DESIGN

We define our two research questions. First, how data mining can help admission working process. Second, how data mining can predict the student's jobs.

A. Testing problems

To answer two research questions, we conduct two sets of experiments: admission data set and the mapped student's course grade with job dataset.

a) admission data set: The admission data set used in this research was collected from admission year 2018 and 2019.

Table1 Attributes descriptions of admission data set

attributes	description	values
Stu code	Student code	Set of alphabets
Acad year	Academic year	2018,2019
Faculty name	Faculty name	Ram,ram
Proj name	Project name	EDA,IT
School	School name	HS,SM
Course	Course name	Information technology,computer science

Regname	Region name	South ,east west,north

b) student's course grade with job dataset*:* This set of problems was select from student's course grade of schooll of information. The job was collected by manual survey.

Table 2 ATTRIBUTES DESCRIPTION OF STUDENT'S COURSE GRADE WITH JOB DATASET

Attributes	Description	Values
Gender	Gender	{Male, Female}
GradSubject 1	Grade of course 1	{High(A,B+,B), Medium(C+,C), Low (D+,D)}
GradSubject 2	Grade of course 2	{High(A,B+,B), Medium(C+,C), Low (D+,D)}
…	…	…
GradSubject N	Grade of course N	{High(A,B+,B), Medium(C+,C), Low (D+,D)}
Job	Job of graduated student (Label or Class)	{networking, hardware engineer, networking engineer}

B. Experimatal setup

a) Association Rule Mining*:* In our study we used APRIORI algorithms to analyze all association rule that have support and confidence higher than a given minimal support threshold (minsup=0.01) and a minimal confidence threshold (minconf=0.5)

b) ID 3 algorithm: the following setting used with ID3 operator to produce the decision tree:

- Splitting criterion = information gain ratio
- Minimal size of split = 4
- Minimal leaf size = 2
- Minimal gain = 0.1

6 EXPERIMATAL RESULTS

A. First Research question

For the first research question, how data mining can help admission working process. To answer this question, we selected association rule mining as a mining tool to discover interesting relation between features. The threshold of minsup and minconf are configured as above.

Fig: Admission plan

B. Second research question

For the second question, how data mining can predict the student's jobs. To answer this question, we selected ID3 decision tree as a mining tool to predict student's job.

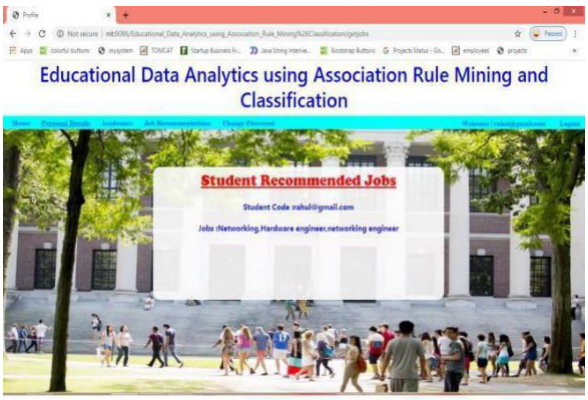

Fig: View Job Recommendations

7. CONCLUSION

In this paper, we applied two data mining algorithms to discover useful knowledge from educational dataset. First, the association rule mining is used on admission dataset to answer the question "how data mining can help admission working process". The result shown the significant relationship between region, admission project name, faculty and province. This result might be help educators who response for admission working process to plan their admission

promotion. Second, ID3 decision tree is applied to student's course grade with job dataset to answer the question "how data mining can predict the student's jobs". The result rule show that the significant subject which student should be important for future career.

FUTURE WORKS

Future work will be aimed at collect more career data from graduated student. It would be also applied other classification algorithms such as neural network, SVM, etc. to improve classification accuracy.

REFERENCES

[1] N. Elgendy and A. Elragal. Big Data Analytics: A Literature Review Paper, Industrial Conference on Data Mining (ICDM), 2014, pp214- 227.

[2] Heikki, Mannila, Data mining: machine learning statistics, and database, IEEE, 1996.

[3] Jiawei Han and Micheline Kamber (2011) Data Mining: Concepts and Techniques. 3 editions. Morgan Kaufmann.

[4] N. Elgendy and A. Elragal. Big Data Analytics: A Literature Review Paper, Industrial Conference on Data Mining (ICDM), 2014, pp214- 227.

[5] Heikki, Mannila. Data mining: machine learning statistics, and database, IEEE, 1996.

[6] AlejandroPeña-Ayala. Educational data mining: A survey and a data mining-based analysis of recent works, Expert Systems with Applications, March, 2014, Vol. 41(4).

[7] B.K. Baradwaj, S.Pal. Mining Educatinal Data to analyze Students' Performance, International Journal of Advanced Computer Science and Applications (IJACSA), 2011, Vol. 2(6).

[8] N. Buniyamin, U.B. Mat, P.M. Arshad. Educational data mining for prediction and classification of engineering students achievement, The 7 th International Conference on Engineering Education (ICEED), 2015, Kanazawa, Japan, pp. 49-53.

[9] A. A. Saa., Educational Data Mining & Student'sPerformancePrediction., International Journaal of Advanced Computer Science and Application (IJACSA), 2016, Vol.7(5).

[10] F. Matsebula, E. Mnkandla., A big data architecture for learning analytics in higher education, IEEE Africon, 2017, pp.951-956.

[11] J. Han, M. Kamber, J. Pei,. Data Mining: Concepts and Techniques, Morgan Kaufmann , 2011.

[12] Quinlan, J. R. Induction of Decision Trees. Mach. Learn. 1, 1 ,Mar. 1986, pp. 81–106.

[13] S. Rao, R. Gupta,. Implementing Improved Algorithm Over APRIORI Data Mining Association Rule Algorithm., International Journal of Computer Science And Technology, Mar. 2012, pp. 489- 493.