Educational Data Analytics using Association Rule Mining and Classification

Pornthep Rojanavasu
Department of Computer Engineering
School of Information and Communication Technology
Phayao, Thailand
pornthep.ro@up.ac.th

***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

I. Introduction

The occurrence of disrupt technology in many areas has conducted the large volume of data in various format and unprecedented speed of generated data. These data phenomenon known in term of big data [1] (volume, various, velocity). The big data required to analyses with appropriate approach for extracting useful knowledge. The data mining aims to discover pattern or useful knowledge from large collection of data [2]. The major functions of data mining are applying various algorithms to discover useful pattern and catch patterns which hidden in data. Educational Data Mining (EDM) is a new research direction in data mining. The focuses of EDM are focus on discovering useful knowledge and mining the helpful patterns from educational data, such as student profile data, student registration data, student job data and other occurrence of student data which can collect during college [3].

There are a lot of objective of EDM depend on data source and educational problem. 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. 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 [8] 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. Section II provides basic knowledge discovery process. Section III describes a brief of association rules mining and ID3 decision tree. Section IV present the design of experiments while the following section focus on results discussion. Finally, section VI contains conclusion and future works.

II. Knowledge Discovery in Database

Knowledge discovery in database (KDD) is a process to identifying valid, novel, useful and understandable pattern from large and complex datasets [8]. The KDD process consist of following step as shown in figure1.

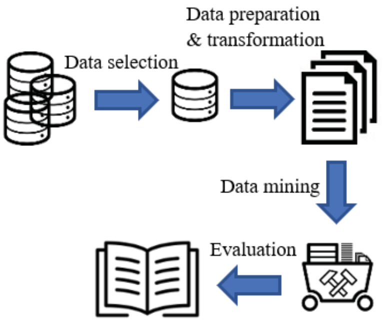

Fig. 1. The KDD process.

1. Data selections: is the process of retrieving data from multiple source which are relevant to our tasks.
2. Data preparation and transformation: is the process of collecting, cleaning, organizing data, converting data from one complex domain or format into another simple domain or format. Many data preparation and transformation method are applied in this step, such as noise filtering method, handle missing data method, data sampling method, normalization feature construction, etc.
3. Data mining: is the process of discovering useful knowledge from a large amount of data. To discovery useful knowledge data mining algorithms, such as associations rule mining, pattern classification, data clustering, are applied to data.
4. Evaluation: is the process of evaluation or scoring of the discovery pattern or knowledge to confirm useful result.

III. Association rule mining and ID3 Decesion Tree

A. *Association Rule Mining*

Association rule mining is a data mining method to discovery interesting relationship between features. The association rule mining has two important steps. First step, frequent itemset generation, is a process to find frequent itemset which have minimum support and minimum confidence value more than assigned threshold. Second step, rule generation, is a process to generate rules from frequent itemset. The generated rules are evaluated by confidence value. A widely-used algorithm for the association rules mining is the Apriori algorithm [10]. The key success of Apriori algorithm is pruning a lot of unnecessary itemset which lead to reduce computational time. The summary Apriori algorithms shown in figure 1.

```
ProcedureApriori (T, minSup){
//T is dataset and minSup is minimum support value
$C_k$: Candidate itemset of size k
$L_k$: frequent itemset of size k
$L_1$ = {frequent itemsets};
for(k=1; $L_k$ != ϕ; k++) do begin
$C_{k+1}$ = candidates generated from $L_k$;
    for each transaction t in dataset do{
            increment the count of all candidates in $C_{k+1}$ that are
contained in t
            $L_{k+1}$ = candidates in $C_{k+1}$ with min_sup
            }
end
return $U_k$, $L_k$
```

Fig. 2. Pseudocode of the Apriori algorithm [x]

B. *ID 3 Decision Tree*

ID3 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]. 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:

```
ProcedureBuildTree(S){
//S = set of not null data
repeat
    maxGain ← 0
    split A ← null
    e ← Entropy(Attributes)
        for all Attributes a in S do
            gain ← InformationGain(a, e)
                if gain > maxGain then
                    maxGain ← gain
                    splitA ← a
                end if
        end for
        Partition(S, splitA)
    until all partition processed
}
```

Fig. 3. Pseudocode of the build decision tree in ID3 [x]

IV. 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 2016 and 2017. The size of dataset is 10,342 records. Table 1 describes the attributes of dataset and their possible value

TABLE I. Attributes description of admission data set

Attributes	Description	Values
Acadyear	Acadamic year	{2016,2017}
StuCode	Student code	set of 7 number
GFacName	Group of faculty	{Technology science, Health scicence, Social science}
FacName	Faculty	{ICT, Engineering, MIS, …}
ProjName	Admission project	{Project 1, Project 2, …}

Attributes	Description	Values
	name	
EduName	School name	{School1, School2, …}
ProvName	Provice name	{Chiang rai, Phayao, … }
RegName	Region name	{North, South East, South, Central}

b) student's course grade with job dataset: This set of problems was select from student's course grade of schooll of information and communication technology between 2011-2014 which is admission only in 2011 mapped with their job after graduated. The job was collected by manual survey. The size of dataset is 106 records.

TABLE II. 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)	{IT, Non-IT}

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

V. 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. A snapshot of rules after run on admission data set for each region are shown in Table III.

TABLE III. A SNAPSHOT OF RULES IN EACH REGION

Region	Rules	Support / Confidence
North	1. North, Faculty of Medicals, 10% direct admission => Lampang	5% / 80%
	2. North, Faculty of MIS, 10% direct admission => Phayao	4% / 60%
Central	3. Central, School of education, 10% direct admission => Sukhothai	2% / 80%
	4. Central, Faculty of Medicals, 10% direct admission => Phitsanulok	3% / 60%
South East	5. South East, School of education, 10% direct admission => Udon Thani	5% / 80%
	6. South East, Faculty of Medicals, 10% direct admission => Udon Thani	3% / 50%

After analyzing the generated association rules, it is observed as following rules:

The Rule no.1 means that the students who come from north region and admission from 10% direct project and admission in faculty of medicals have relative with Lampang province with support value 5 % and confidence value 80%.

The Rule no.2 means that the students who come from north region and admission from 10% direct project and admission in faculty of MIS have relative with Phayao province with support value 4 % and confidence value 60%.

The Rule no.3 means that the students who come from central region and admission from 10% direct project and admission in school of education have relative with Sukhothai province with support value 2 % and confidence value 80%.

The Rule no.4 means that the students who come from central region and admission from 10% direct project and admission in faculty of medicals have relative with Phitsanulok province with support value 3 % and confidence value 60%.

The Rule no.5 means that the students who come from south east region and admission from 10% direct project and admission in school of education have relative with Udon Thani province with support value 5 % and confidence value 80%.

The Rule no.6 means that the students who come from south east region and admission from 10% direct project and admission in faculty of medicals have relative with Udon Thani province with support value 4 % and confidence value 60%.

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. The configuration of important parameters is configured as above. After running ID3 decision tree with 5-fold cross validation on student's course grade with job dataset, the confusion matrix of ID3 decision tree was generated as table IV.

TABLE IV. THE CONFUSION MATRIX OF ID3 DECISION TREE

	True IT	*True Non-IT*	*Class Precision (%)*
Prediction IT	45	11	80.35%
Prediction Non-IT	17	33	66.00%
	72.58%	75.00%	

The ID3 decision tree was able to predict both true positive and true negative 78 records out of 106 records. The accuracy of ID3 decision tree is 73.58%.

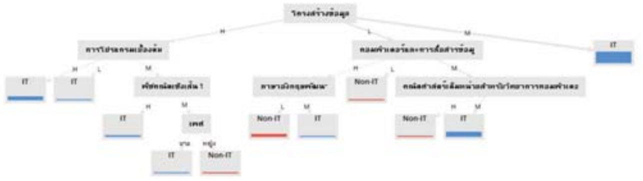

Fig. 4. The decision tree result from ID3 using Rapid Miner Studio Educational 8.1.000

The fig.4 showed the decision tree which generated from ID3 using Rapid Miner Studio. The example of IF-THEN rule could be easier to explain is shown in fig. 5.

IF Data structure = ‘M’ THEN Job = ‘IT’
IF Data structure = ‘H’ AND Intro. Programming = ‘H’ THEN Job = ‘IT’
IF Data structure = ‘H’ AND Intro. Programming = ‘M’ AND Linear Algebra = ‘H’ THEN Job = ‘IT’
IF Data structure = ‘H’ AND Intro. Programming = ‘L’ THEN Job = ‘IT’
IF Data structure = ‘L’ AND Data com. = ‘L’ THEN Job = ‘Non-IT’
IF Data structure = ‘L’ AND Data com. = ‘H’ AND Dev. Eng. = ‘L’ THEN Job = ‘Non-IT’
IF Data structure = ‘L’ AND Data com. = ‘H’ AND Dev. Eng. = ‘M’ THEN Job = ‘IT’

Fig. 5. The example of rule set generated by ID3 decisin tree

VI. 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.

VII. 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.

Acknowledgment

The dataset using in this research was supported by Division of Education Services, University of Phayao.

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’s Performance Prediction., 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.