Mining Association Rules from Student Learning Results

Van-Quyet Nguyen
Faculty of Information Technology, Hung Yen University of Technology and Education
16000 Hungyen, Vietnam
quyetict@utehy.edu.vn

Thai-Bao Mai-Hoang
Faculty of Information Technology, Hung Yen University of Technology and Education
16000 Hungyen, Vietnam
maihoangthaibao01@gmail.com

Kyungbaek Kim
Department of Artificial Intelligence Convergence, Chonnam National University
61186 Gwangju, South Korea
kyungbaekkim@jnu.ac.kr

ABSTRACT

This paper presents a method using machine learning to mine frequent association rules from student learning results. We first collect and pre-process data extracted from an online learning management system. We then convert data into a form so that it could be used in the Apriori algorithm. Finally, we use a machine learning library based on the .NET framework to implement the proposal idea. The experimental results show that obtained association rules are highly reliable. The study results can be applied in educational consultancy for students.

KEYWORDS

Association Rules, Frequent Itemset, Apriori

1 INTRODUCTION

Association rules mining is an important problem in data mining. It is applied in many fields such as finding groups of products being purchased together by customers in retail data analysis, finding groups of telecommunication services being used by clients, or finding groups of stocks bought by investors. This problem has a lot of application in reality, for example, it helps entrepreneurs/organizations develop new services or improve current ones to achieve better economic efficiency. Finding association rules is to discover relations between items in a large database. Discovering these relations is not based on the roles or the properties of items, it is based on the occurrences of items at the same time in the database.

Mining association rules has received much attention from researchers for a long time since it was first proposed by Agrawal et al. in [1]. In the first place, the model aimed to solve retail problems such as Walmart's or Amazon's whereby a set of transactions were given with a set of items usually purchased by customers, mining association rules is to discover what products are usually purchased by customers after buying several ones. The main technique used to solve this problem is finding frequent itemsets (frequent sets of products): the number of occurrences in each transaction is greater than a given threshold. Then, using these frequent itemsets to generate frequent rules by pruning method. From then on, a lot of algorithms were proposed to improve the performance of the primitive algorithm such as Apriori algorithm [2], FP-growth algorithm [3], or Eclat algorithm [4]. Moreover, there are a number of studies that improve the performance of mining association rules problems in a parallel-computing environment [5].

In recent years, the mining association rules problem has also received much attention in studies applied in education to enhance the quality of education [6][7][8]. Each study focuses on a specific aspect of education. For example, discovering the relationship between student scores in their graduation and post-graduation presented in [6]. Siahaan et al. [8] mined association rules from quiz answers of students to help improving the quality of exams.

In this paper, we present a method of mining frequent association rules from student learning results using Apriori algorithm. We first collect and process data of more than 12,000 students of Hung Yen University of Education and Technology in two years 2019 and 2020. We then convert data so that they could be used in Apriori algorithm. Finally, we use a machine learning library based on .NET framework to implement the proposal idea. The experimental results show that obtained association rules are highly reliable. The study results can be applied in educational consultancy for students.

2 BACKGROUND

2.1 Concepts and Terms

Let us assume that a given database ***D*** contains a list of courses completed by students whereby a row corresponds to the data of a student as shown in Table 1.

Table 1: Example of Itemsets

TID	Itemset
1	ABDE
2	BCE
3	ABDE
4	ABCE
5	ABCDE

6	BCD

Where the symbols are defined as follows:

A: Basic of Programming;
B: Object-Oriented Programming;
C: Database;
D: Data Structures and Algorithms;
E: Project 1.

In order to work with the mining association rules problem, we describe briefly related concepts and terms as follows:

Itemset: A set of items (X) consists of a collection of items. For instance, B, AB, BC, and BCE are itemsets.
Transaction: A set of items (e.g., courses) in the database is stored with transaction identity (e.g., student identity).
Support: Support of an itemset X denoted by **sup(X)** is the frequency of transactions containing X. For instance, sup(A) = 4.
Frequent Itemset: is an itemset that occurs **frequently** in transactions and satisfies the minimum support (**minsup** – defined by users). For instance, if we choose **minsup = 3**, from the database in *Table 1,* we obtain a list of itemsets in *Table 2.*

Table 2: Example of Frequent Itemsets

sup	Frequent Itemsets
6	B
5	E,BE
4	A,C,D,AB,AE,BC,BD,ABE
3	AD,CE,DE,ABD,ADE,BCE,BDE,ABDE

K-frequent itemset denoted by $F^{(k)}$ is an itemset containing k items. For instance, for **minsup** = 3, we obtain a list of frequent itemsets as follows:

$F^{(1)}$ = {A,B,C,D,E}
$F^{(2)}$ = {AB,AD,AE,BC,BD,BE,CE,DE}
$F^{(3)}$ = {ABD,ABE,ADE,BCE,BDE }
$F^{(4)}$ = {ABDE}

Association rule: An association rule as an implication of the form X → Y, where X, Y are two itemsets satisfying X ∩ Y = ∅. For instance, AB→D, A→B are association rules.
The support of a rule X→Y, denoted by sup(X→Y) or sup(XY), is the number of transactions whereby both X and Y occur. For instance, sup(AB→D) = sup(ABD) = 3.
The confidence of a rule X→Y, denoted by conf(X→Y), is conditional probability computed by quotient sup(X→Y)/sup(X). For instance, conf(AB→D) = sup(ABD)/sup(AB) = 3/4 = 0.75. It means about 75% of students having good or higher results at course **A** (Basic of Programming) and course **B** (Object-Oriented Programming) would have good or higher results at course **D** (Data Structures and Algorithms).

Frequent association rule: A rule X→Y is said to be frequent if **sup(XY) ≥ minsup** and **conf(X→Y) ≥ minconf** (minimum confidence is defined within range (0,1] by users). For instance, if we choose **minsup** = 3, **minconf** =0.8, we obtain the results that AB→D is not a frequent association rule while AD→E is a frequent association rule.
Lift: The lift of a rule X→Y, denoted by **Lift(X->Y)**, is defined by sup(X∪Y)/(sup(X)*sup(Y)) or conf(X→Y)/sup(Y).
Then:

- If **Lift(X->Y)** = 1, X and Y are independent.
- If **Lift(X->Y)** > 1, X and Y are positively dependent on each other, rule X→Y is potentially useful.
- If **Lift(X->Y)** < 1, X and Y are negatively dependent on each other, rule X→Y is not useful.

2.2 Mining Association Rules Process

We summarize the process of mining association rules in 2 steps:
Step 1: Find all frequent itemsets from the database for support greater than or equal to **minsup**.
Step 2: Generate rules from frequent itemsets.

- For each frequent itemset Z, generate all non-empty subsets of Z.
- For each non-empty X of Z, we generate **X → Z\X** which is the association rule we need to find if **conf(X → Z\X) ≥ minconf.**

In order to solve the problem given in the two preceding steps, we can use Apriori algorithm, the idea of the algorithm is defined as follows:

- **Find Itemsets**

Step 1: Find 1-frequent itemset and put it into $\mathbf{L_k}$ (**k=1**).
Step 2: Use $\mathbf{L_k}$ to generate candidate itemset $\mathbf{C_{k+1}}$ with size (**k+1**).
Step 3: Scan the database to find itemsets that are frequent in $\mathbf{C_{k+1}}$ and put into $\mathbf{L_{k+1}}$.
Step 4: If $\mathbf{L_{k+1}}$ is not empty, then set **k=k+1** and return to step 2.

- **Generate association rules**

Input: A set of all frequent items **F,** minimum confidence **minconf.**
Output: A set of frequent association rules.
Main Idea:

- For each frequent itemset **Z** in **F**, generate all non-empty subsets of **Z**.
- For each non-empty **X** of **Z**, we generate **X → Z\X** which is the association rule we need to find if **conf(X → Z\X) ≥ minconf.**

3 MINING STUDENT LEARNING RESULTS

3.1 Data Collecting

Student learning results used in this paper are collected from the Educational Quality-Survey System of Hung Yen University of Education and Technology (UTEHY-Eclass). In which, data of the

UTEHY-Eclass system are stored in tables in the database management system (MS SQL Server). A student can learn and have results from several courses in a semester. Grade data of students are retrieved in the structure shown in Table 3.

Table 3: Example of Data Structure Extracted from Student Learning Results

Student Id	Course Id	Course Name	Grade	Academic Year	Semester
220213	111125	Linear Algebra	9.0	2020-2021	1
220213	321104	Digital Engineering	8.0	2020-2021	1
220088	321800	Basic of Electronics	9.0	2020-2021	2
…	…	…	…	…	…

In which, course grade ranges from 0 to 10. A student is assessed to be passed if his/her course grade is greater or equal 5.0, otherwise, he/she fail.

3.2 Data Pre-processing

In order to mine association rules from the student learning result database above, we implement a program that converts data structure at the preceding Table 3 into data form used in mining association rules having the following structure:

Table 4: Data Structure after Normalization

Student ID	List of Course Identity (with conditional)
220213	111125,321104
220088	321800
…	…

Accordingly, a student identity plays a role as a "transaction identifier" and each course identity plays a role as an "item". In which, course identities in a row represent courses which that student learned and already be graded, course identities are separated by a comma (,).

3.3 Experimental Evaluation

Environment. We implement the problem using C# programming language and machine learning library **Accord.MachineLearning**.
Dataset. We run experiments on two databases. The first contains a list of courses which corresponding students fail and the second contains a list of courses which student has good or higher results (course grade is greater or equal 8.0). Details will be described in the following.

Experiment 1: Mining association rules between courses which students fail.

In this experiment, we run the mining association rules algorithm on 7,404 students fail among 538 courses with 24,132 times that students failed. We set **minsup** = 60 and **minconf** = 0.8. The result is presented in Table 5.

Table 5: Frequent Association Rules between Courses which Students Fail

#	Frequent Association Rule	Conf	Sup	Lift
1	[Linear Algebra] [Digital Technology] -> [Basic of Electronics]	0.908	69	1.17
2	[Theory of circuit 1] [Electronic Capacity (2+1*)] -> [Electric drive]	0.907	88	1.50
3	[Theory of circuit 1] [Microprocessor engineering practice] -> [Electric drive]	0.890	73	1.48
4	[Electric – Electronic engineering mathematics 2] [Electronic Capacity (2+1*)] -> [Electric drive]	0.888	79	1.47
5	[Electric – Electronic engineering mathematics 1][Theory of circuit 1] -> [Electric drive]	0.837	87	1.39
6	[Digital Technology] -> [Basic of Electronics]	0.827	91	1.06
7	[Electric – Electronic engineering mathematics 2] [Theory of circuit 1] -> [Electric drive]	0.826	109	1.37
8	[Electronic Capacity (2+1*)] -> [Electric drive]	0.822	134	1.36
9	[Microprocessor engineering practice] -> [Electric drive]	0.808	97	1.34
10	[Engineering Physics (3+1*)] [Theory of circuit 1] -> [Electric drive]	0.802	73	1.33

We obtain exactly 10 rules in this experiment. From Confidence and Lift indicators, we find that obtained rules are highly reliable. This result can be used to advise students to concentrate on the course **Y** in case they fail the course **X**. It can be also applied to help students choose suitable courses (in registering credit-based courses).

Experiment 2: Mining association rules between courses in which students have good results.

In this experiment, we implement mining association rules algorithm on the data of 10,494 students which contains good results among 781 courses with 56,576 times that students have good results. We set **minsup** = 100 and **minconf** = 0.8. After running the algorithm, we obtain 149 association rules.

Table 6 represents 10 association rules which have the highest confidences. From the results with Confidence and Lift indicators, we find that obtained association rules are highly reliable. This result can be used to encourage students to register for courses (in credit-based learning) or predict the ratio of good-result students of courses as well as of the university.

#	Frequent Association Rule	Conf	Sup	Lift
1	[Advanced major 1] -> [Graduation internship]	1.000	110	1.05
2	[Advanced major 2] -> [Graduation internship]	1.000	109	1.05
3	[Internship at plants] [Soft skill 4] -> [Graduation internship]	1.000	106	1.05
4	[Advanced major 1] [Advanced major 2] -> [Graduation internship]	1.000	101	1.05
5	[Internal combustion engine internship] [Soft skill 4] -> [Soft skill 3]	1.000	109	1.18
6	[Introduction to scientific research method] [Graduation internship (9T)] -> [Soft skill 4]	0.993	150	1.15
7	[Introduction to scientific research methods] [Manufacturing garment product using high technology] -> [Soft skill 4]	0.993	109	1.15
8	[Introduction to scientific research methods] [Graduation internship (9T)] [Manufacturing garment product using high technology] -> [Soft skill 4]	0.992	128	1.15
9	[Special processing technology internship at plants] -> [Technical skill internship at plants]	0.990	103	1.02
10	[Technical skill internship at plants] -> [Special processing technology internship at plants]	0.990	103	1.03

Table 6: Top 10 frequent association rules between courses in which students have good results

4 CONCLUSION

This paper presented a method using machine learning techniques to mine frequent association rules from student learning results. We collected and processed data from student learning results, normalized data into a form which could be used in Apriori algorithm. The experimental results showed that obtained association rules are highly reliable. The results can be used in an educational consultant for students to help them adjust as well as register courses that are suitable to achieve high learning results.

ACKNOWLEDGMENTS

This research was supported by the MSIT(Ministry of Science and ICT), Korea, under the ITRC(Information Technology Research Center) support program(IITP-2021-2016-0-00314) supervised by the IITP(Institute for Information & communications Technology Promotion). This research was supported by the Bio & Medical Technology Development Program of the National Research Foundation (NRF)& funded by the Korean government (MSIT) (NRF-2019M3E5D1A02067961)

REFERENCES

[1] Agrawal, Rakesh, Tomasz Imieliński, and Arun Swami. "Mining association rules between sets of items in large databases." Proceedings of the 1993 ACM SIGMOD international conference on Management of data. 1993.

[2] Agrawal, Rakesh, and Ramakrishnan Srikant. "Fast algorithms for mining association rules." Proc. 20th int. conf. very large data bases, VLDB. Vol. 1215. 1994.

[3] Han, Jiawei, Jian Pei, and Yiwen Yin. "Mining frequent patterns without candidate generation." ACM sigmod record 29.2 (2000): 1-12.

[4] Zaki, Mohammed Javeed. "Scalable algorithms for association mining." IEEE transactions on knowledge and data engineering 12.3 (2000): 372-390.

[5] Nguyen, Giang-Truong, et al. "Efficient Association Rule Mining based SON Algorithm for a Bigdata Platform." Journal of Digital Contents Society 18.8 (2017): 1593-1601.

[6] Kumar, Varun, and Anupama Chadha. "Mining association rules in student's assessment data." International Journal of Computer Science Issues (IJCSI) 9.5 (2012): 211.

[7] Siahaan, Andysah Putera Utama, Ali Ikhwan, and Solly Aryza. "A Novelty of Data Mining for Promoting Education based on FP-Growth Algorithm." (2018).

[8] Shannaq, Boumedyen. "Machine Learning Model-Ensuring Electronic Exam Quality using Mining Association Rules." Machine Learning 29.03 (2020): 12136-12146.