Extraction of Association Rules from Education Data

Khin Ei Ei Chaw, Moe Sanda Htun
Computer University(Sittway)
eiei2.chaw@gmailcom; moesdhtun@gmail.com

Abstract

Data mining concerns with developing methods for discovering knowledge from data that come from educational environment. This paper presents finding interesting patterns from educational database. Association rule mining is used to find the interesting patterns.Two important measures 'support' and 'confidence' are used to measure the interestingness of the patterns.Since support and confidence are not enough to measure the relationships of itemset.

The statistical index of the degree to which two variables are associated is the correlation coefficient and it is measured by lift ratio. Extracting the most interesting association rules can be quite tricky. One of the difficulties is that many measures of interestingness do not work effectively for all datasets. In this paper, correlation ratios of association rules are used to measure the interestingness. The lift value of rule, greater than 1, indicates a positive correlation between antecedent and consequent.

1. Introduction

Data mining is the process of extracting interesting information or patterns from large information repositories such as: relational database, data warehouses, XML repository, etc. Association rule mining is one of the most important and well researched techniques of data mining. It aims to extract interesting correlations, frequent patterns, associations or large itemsets among sets of items in the transaction databases. Discovering large itemsets is done iteratively for each large k-itemset in increasing order of k.

Association Rules represent an unsupervised learning method that attempts to capture associations between groups of items. Association Rules have also been referred to in the literature as Market Basket analysis or Affinity analysis. Support and Confidence are two important measures in association rules. An association rule must have confidence and support greater than a (user dependant) minimum confidence and minimum support.Another important concept in association rules is that of the Lift of the rule. With lift ratio, rules with strong correlation can be extracted.

Educational Data Mining, concerns with developing methods that discover knowledge from data come from educational environments. The data can be collected form historical and operational data reside in the databases of educational institutes. The student data can be personal or academic. Also it can be collected from e-learning systems which have a vast amount of information used by most institutes. This paper presents education data mining by Association rule mining algorithm. The discovered knowledge can be used to better understand students' behavior, to assist instructors, to improve teaching, to evaluate and improve e-learning systems, to improve curriculums and many other benefits.

This paper presents finding educational behavior using association rules. Lift ratio is used to measure the interestingness of frequent patterns. The organization of this paper is as follows. Section 2 is other related approach to the system. Section 3 presents association rule mining algorithms and how lift ratio is used to measure the correlation of itemsets in the frequent patterns. Section 4 is the presentation of proposed system and section 5 is the implementation of the system and system results. Section 6 is the conclusion of the system.

2. Related work

Educational data mining differs from knowledge discovery in other domains in several ways. One of them is the fact that it is difficult, or even impossible, to compare different methods or measures a posteriori and decide which is the best.

Association rules are increasingly used in educational data mining [1, 3, 6]. However, measuring the interestingness of a rule can be problematic, as explained in [2]. Two measures, support and confidence, are commonly used to extract association rules. However it is well known that even rules with a strong support and confidence may in fact be uninteresting. This is why, once the association rule $X \rightarrow Y$ has been extracted, it is wise to double check how much X and Y are related. About 20 measures have been proposed in the literature to do so. Unfortunately, no measure is better than all the others in all situations, though measures tend to agree when support is high [5].

Association is one of the fundamental tools of scientists. As one variable increases, so does the other. The statistical index of the degree to which two

variables are associated is the correlation coefficient. Developed by Karl Pearson, it is sometimes called the "Pearson correlation coefficient". The correlation coefficient summarizes the relationship between two variables.

3. Association Rule Mining

Association rule mining is one of the most important and well researched techniques of data mining. [2] It aims to extract interesting correlations, frequent patterns, associations or casual structures among sets of items in the transaction databases or other data repository The discovery of association relationship among a huge database has been known to be useful in selective marketing, decision analysis, and business management popular area of applications is the market basket analysis, which studies the buying behaviors of customers by searching for sets of items that are frequently purchased together (or in sequence).

The rule $X \Rightarrow Y$ has support s in the transaction set D if s % of transactions in D contains $X \cup Y$. Mining association rules is composed of the following two steps –

- Discover the large itemsets, i.e., all sets of itemsets that have transaction support above a predetermined minimum support s.
- Use the large itemsets to generate the association rules for the database.

The overall performance of mining association rules is in fact determined by the first step.

3.1 Apriori Algorithm

In Apriori, in each iteration, it constructs a candidate set of large itemsets, counts the number of occurrences of each candidate itemset, and then determines large itemsets based on a pre-determined minimum support. In the first iteration, Apriori simply scans all the transactions to count the number of occurrences for each item. The meaning of various parameters is as follow:

Dk	Set of transactions for large k-item sets
|Dk|	No. of transactions in Dk
Ck	Set of candidate k-itemsets
Lk	Set of Large k-itemsets

TID	Items
100	A C D
200	B C E
300	A B C E
400	B E

Figure 1.Sample Database D

Apriori algorithm consists of two steps – join and prune actions.

Join Step – It is the candidate generation process. To find L_k, a set of candidate k-itemsets is generated by joining L_{k-1} with itself. This set of candidates is denoted Ck.

Prune Step – C_k is a superset of L_k, that is, its members may or may not be frequent, but all of the frequent k-itemsets are included in Ck. A scan of DB to determine the count of each candidate in Ck would result in the determination of L_k. if the count of candidate in L_k is less than minimum supports, it cannot be frequent and so can be removed from C.

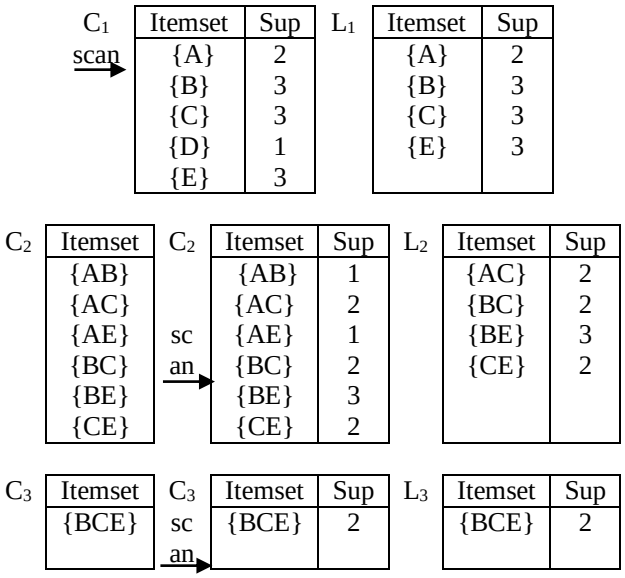

Figure 2.Sample process of Apriori Algorithm

3.2 Correlation

Lift ratio is used to represent the correlation of association rules. The lift value is the ratio of the confidence of the rule and the expected confidence of the rule. The lift is measured as the ratio of the probability of antecedent and consequent occurring together to the probability of antecedent and consequent occurring independently. The lift value of rules that greater than 1 indicates a positive correlation between antecedent and consequent. With the lift value, we can interpret the importance of a rule. The first rule, with the highest lift which means highest correlation is the most important, and so on. *lift* of a rule is defined as

$$\text{Lift}(X \Rightarrow Y) = \frac{\text{Support\_count}(X \cup Y)}{\text{Support\_count}(X) * \text{Support\_count}(Y)}$$

or the ratio of the observed confidence to that expected by chance.

4. Proposed System

This paper presents extracting interesting patterns from educational database. Educational database includes Attendance of the students, Grade Point Average (GPA), hours, resources, exercises,

homeworks, midterm, lab, final and grade information. Association mining algorithm is used to extract interesting relationships among attributes from the given dataset. It allows finding rules of the form "If antecedent then (likely) consequent", where antecedent and consequent are frequent itemsets. Itemsets are sets of one or more items. In our dataset an example of item is: attendance = good. Because, we are looking for items that characterize the final grade of students, consequent has one item which is final_grade= A where A is one value of the final grade such as excellent, very good,…etc. Association rules do not capture many interesting dependencies between items. In order to solve this problem, correlation measure (lift ratio) is to measure the dependencies between items. The overview of the proposed system is shown in the following Figure 1.

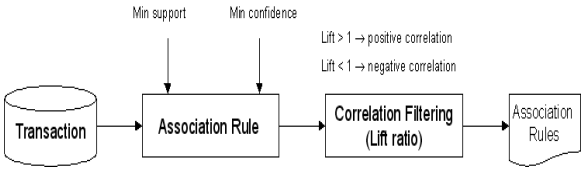

Figure 1. Overview of the System

4.1 System Implementation

This paper is implemented as a web-based educational system. It is developed using Microsoft Visual Studio 2008 and data are collected from education training center. The process flow of the system is shown in Figure 2.

4.2 Data collection

In this system, we collected the students' data from educational training center. There are 350 students attending the class. The sources of collected data are: personal records and academic records of students, course records and data came from e-learning system. This system has the following attributes Attendance, gpa, hours, resource, exercises, homeworks, midterm, lab, final, grade. We convert the numerical values of attribute to the categorical values as identifying according their range. For example,the Attendance of the student is between 100% and 85% is converted to Good, 85% and 64% is converted to Fair and less than 65% is converted to Bad. Detail data attribute and domain values are shown in Table 1.

In the database model, Student table stores the personal information of each student, and course data and section data are recorded in course table and section table respectively. Others table record academic data of students.

These data are then converted in transaction data and this transaction table is shown in Table 2. The following table 2 demonstrates the example of the educational database with 11 attributes for 9 students.

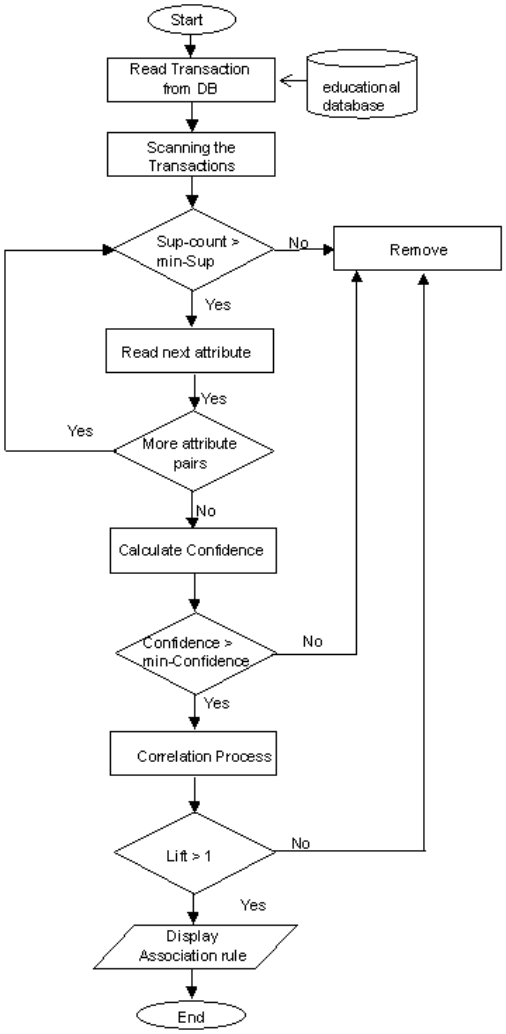

Figure 2. Process Flow of the System

Table 1. Data Attribute and Domain Values

Attribute	Domain Value
Attendance	Good (100%-85%), Fair (84%-65%), Bad (<65%)
gpa	Good, Fair, Poor
hours	Full-Time(64-50), Part-Time(32-28)
Resource	Good (100%-75%), Fair (74%-41%), Poor (<40%)
Exercises	Good (100%-85%), Fair (84%-60%), Bad (<60%)
Homework	Good (100%-75%), Fair (74%-41%), Bad (<40%)
Midterm	Distinction, Credit, Pass, Fail
Lab	Good (100%-75%), Fair (74%-41%), Bad (<40%)
Final	Distinction, Credit, Pass, Fail
Grade	A,B,C,D,E,F

Table2. Example database with 11 attributes for 9students

Student ID	Attendance	gpa	hours	Resource	Exercises	Homework	Midterm	Lab	Final	Grade
1	Good	Good	Full-time	Good	Good	Good	Distinction	Good	Distinction	A
2	Fair	Fair	Full-time	Fair	Fair	Good	Credit	Good	Credit	B
3	Bad	Fair	Full-time	Fair	Bad	Fair	Pass	Fair	Pass	D
4	Bad	Fair	Part-time	Fair	Bad	Fair	Pass	Fair	Pass	C
5	Bad	Bad	Part-time	Fair	Bad	Fair	Pass	Fair	Pass	D
6	Bad	Bad	Part-time	Fair	Bad	Fair	Pass	Fair	Fail	D
7	Bad	Bad	Part-time	Bad	Bad	Bad	Pass	Fair	Fail	D
8	Fair	Fair	Full-time	Fair	Fair	Fair	Credit	Good	Distinction	B
9	Fair	Fair	Full-time	Fair	Fair	Fair	Credit	Good	Distinction	B

4.4 System Result

The following samples rules are the result of the system after the processing. Rules are generated from the data of Table 2. In this system, attendance, gpa, hours, resources, exercises, homework, and lab are the antecedent and one of Midterm or Final or Grade should be the consequence. For example, for rule No. 1, support count of (Antecedence) Attendance=Bad, GPA=Bad, Hours=Part-Time, Resource=Fair, Exercise=Bad, Homework=Fair, Lab=Fair is 2 and support count of (Antecedence and consequence) is 2. Therefore confidence = 1. But the support count of (consequence) Midterm = Pass, is 5. Therefore lift ratio is 2 / 2 * 5 = 0.2.

Table 3.System Result

Patterns	Support	Confidence	Lift
Attendance=Bad, GPA=Bad, Hours=Part-Time, Resource=Fair, Exercise=Bad, Homework=Fair, Lab=Fair==>MidTerm=Pass	2	100.00%	0.2
Attendance=Fair, GPA=Fair, Hours=Full-Time, Resource=Fair, Exercise=Fair, Homework=Fair, Lab=Good==>MidTerm=Credit	2	100.00%	0.333333 33333333
Attendance=Fair, GPA=Fair, Hours=Full-Time, Resource=Fair, Exercise=Fair, Homework=Fair, Lab=Good==>Final=Distinction	2	100.00%	0.333333 3333333
Attendance=Bad, GPA=Bad, Hours=Part-Time, Resource=Fair, Exercise=Bad, Homework=Fair, Lab=Fair==>Grade=D	2	100.00%	1.75
Attendance=Fair, GPA=Fair, Hours=Full-Time, Resource=Fair, Exercise=Fair, Homework=Fair, Lab=Good==> Grade=B	2	100.00%	2.333333 33333333

When we put ‘2’ in the Minimum Support and’50%’in the Minimum Confidence, the support count will extract and show the amount which is greater than and equal to 2. The final result will be expressed in lift filtering.We can see whether the result is strong rule or not by means of lift filtering.

5. Conclusion

This paper presents discovering interesting relationships in educational data. Interestingness is measured by association rule algorithm's two important measures support and confidence as well as lift ratio. Because of using lift ratio value, it can measure the correlation and relevance of rules produced. It shows how useful data mining can be in higher education in particularly to improve student performance. Association Rules are sorted using lift ratio so that we got more relevant rules.

References

[1] Merceron, A., Yacef, K. Revisiting interestingness of strong symmetric association rules in educational data. *Proceedings of the*

International Workshop on Applying Data Mining in elearning

[2] Jiawei Han. Data Mining, Concepts and Techniques, Mining Association Rules in Large Databases.

[3] M. Houtsma and A. Swami, Set-Oriented Mining of Association Rules. Technical Report RJ 9567, IBM Almaden Research Laboratory, San Jose, CA, October 1993

[4] Nguyen, Son N. and Orlowska, Maria E. "A Further Study in the Data Partitioning Approach for Frequent Itemsets Mining", Proceeding at the 17th Australasian Database Conference (ADC 2006), Hobart, Australia.

[5] R. Agrawal and S. Srikant. Fast Algorithms for Mining Association Rules in Large Databases.

Proceedings of the 20th International Conference on Very Large Data Bases, September 1994.

[6] R. Agrawal, T. Imielinski, and A. Swami. Mining Association Rules between Sets of Items in Large Databases. Proceedings of ACM SIGMOD, pages207-216, May 1993

[7] R .T. Ng and J. Han. Efficient and Effective Clustering Methods for Spatial Data Mining. Proceedings on the 18th International Conference on Very Large Data Bases, pages 144-155, September 1994.