Mining Least Association Rules of Degree Level Programs Selected by Students

Tutut Herawan¹, Zailani Abdullah², Wan Maseri Wan Mohd¹, A Noraziah¹

¹Faculty of Computer Systems and Software Engineering
Universiti Malaysia Pahang, Malaysia
Lebuhraya Tun Razak, 26300 Kuantan Pahang, Malaysia
²Department of Computer Science
Universiti Malaysia Terengganu
21030 Kuala Terengganu, Terengganu, Malaysia

{tutut,maseri,noraziah}@ump.edu.my, zailania@umt.edu.my

Abstract. One of the most popular and important studies in data mining is association rules mining. Generally, association rules can be divided into two categories called frequent and least. However, finding the least association rules is more complex and time consuming as compared to the frequent one. These rules are very useful in certain application domain such as determining the exceptional association between university’s programs being selected by students. Therefore in this paper, we apply our novel measure called Definite Factors (DF) to determine the significant least association rules from undergraduate’s program selection database. The dataset of computer science student for July 2008/2009 intake from Universiti Malaysia Terengganu was employed in the experiment. The result shows that our measurement can mine these rules and it is at par with the existing benchmarked Relative Support Apriori (RSA) measurement.

Keywords: Data Mining; Association rules; Significant Least, Measure; Educational Data.

1 Introduction

Data mining can be defined as the process of extracting hidden and useful information from large data repositories [1]. One of the emerging interdisciplinary research areas in data mining is educational data mining [2]. By definition, educational data mining is an application of suitable data mining techniques to analyze the educational data [3]. It aims at developing new methods that can discover the interesting information from educational settings, and used those methods to better understand the students, and their learning settings (http://www.educationaldatamining.org). The problem of association rules mining was first coined by [4] in an attempt for market-basket analysis. The classification of frequent or least items is based on the mechanism of support threshold. A set of items (itemset) is said to be frequent, if it appears more than minimum support count. The item (or itemset) support count is defined as a probability of item (or itemset) appears in the transaction. In addition, confidence is
another measure that always used together with support count. The confidence is defined as the probability of the rule’s consequent (right side) that also contain the antecedent (left side) in the transaction. The association rule is said to be strong if it meets the minimum confidence threshold.

Least itemset is a set of item that is rarely occurred in the transactional database. It is also known as non-frequent, unusual, exceptional, abnormal, in-balance or sporadic itemset. In some applications domain, these itemsets are very important and in fact it can provide significant information such as air pollution level [5], relationship management [6], image processing [7], abnormal learning problems [8], educational data mining [9-11], text mining [12-13], information visualization [14-15], business process management [16] and many more. From the past literature, most of the tradition association rules mining algorithms [17-27] suffer in term of efficiency and evaluating the real datasets.

Educational data is one of the potential resources in discovering the significant least association rules. These rules can be very useful for higher authority personnel in assisting them to make right decision. For instance, in every July semester, our university receives approximately 160 students to enroll in computer science program. There are always the cases that the students are uncertain and taken for granted by combining with the various fields of interests. The research question is how to justify the student interests since there is no such field to be specified in the online application system. At the moment, if their choices are not selected, they will be offered to any program in the university according to programs availability.

Therefore, in this paper, we apply our novel measure called Definite Factors (DF) to detect the abnormal relationship among university’s programs that have been selected by students. Indeed, DF will take into consideration the combination of both frequent and least university’s program for generating the desired least association rules. We also employed our LP-Tree and LP-Growth algorithms [9] prior to produce the rules. In this study, the experiment was performed based on the 2008/2009 intake students’ that have been offered in Bachelor of Information Technology (Software Engineering) at Universiti Malaysia Terengganu (UMT).

The rest of the paper is organized as follows. Section 2 describes the related work. Section 3 discusses the proposed method. This is followed by experiment tests in section 4. Finally, conclusion and future direction are reported in section 5.

2 Related Works

Nowadays, varieties of data mining method methods have been proposed in educational data mining. Romero et al. [28] suggested two categories of education data mining. The first category contains both statistics and visualization. The second one is web mining which can be divided into three parts. The first part covers clustering, classification, and outlier detection. The second part consists of association rule mining and sequential pattern mining. Finally, the third part is associated with text mining. It can be conclude that, the initial educational data mining is come into sight by analyzing the interaction between student and computer based on detailed logs of all their activities.
Baker et al. [2] proposed educational data mining into five different categories named prediction, clustering, relationship mining, distillation of data for human judge and discovery with model. Recently, discovery with models category is now become one of the most popular methods in educational data mining research. It deals with the sophisticated analysis such as discovering which learning materials sub-categories of students are the most beneficial [29], finding how different type of student behavior contribute to student’s learning in different way [30] and revealing how variety of designing the intelligent tutor influence student’s behavior over time [31].

Nowadays, only few attentions have been paid to extract least association rules from educational data. To the best of our knowledge, only one paper [8] is specifically discussed about least association rules. They applied the existing Rare Association Rules Mining (Apriori-like) algorithms to extract association rules from e-learning data. Their objective is to discover the information about infrequent student behavior. Four Apriori-based algorithms were employed to extract these rules named Apriori-Frequent [4], Apriori-Infrequent, Apriori-Inverse [19] and Apriori-Rare [32]. From the experiments, Apriori-Inverse and Apriori-Rare are proven more suitable in finding the least association rules.

In term of measurement least association rules, one of the popular measurement is Relative Support Apriori (RSA) proposed by [20]. RSA requires three (3) predefined measurements called 1st support, 2nd support and relative support (1st support > 2nd support). An item is said a least item if its support is less than 1st support and greater or more than 2nd support. A frequent item is an item having a support which equal or greater than 1st support. The least association rules are those rules that satisfied all the predefined supports. The main constrain of this algorithm is it increases the computational cost if the minimum relative support is set close to zero. In addition, determination of three predefined measurements is also another issue for this algorithm. Besides RSA, the others approach to capture least association rules are Multiple Support Apriori [17], Matrix-based Scheme [18], Collective Support Apriori [33], etc.

3 Proposed Method

Throughout this section the set \( I = \{i_1, i_2, \ldots, i_M\} \), for \(|I| > 0\) refers to the set of literals called set of items, \( W = \{w_1, w_2, \ldots, w_M\} \), refers to the set of literals called set of weights with a non-negative real numbers, and the set \( D = \{t_1, t_2, \ldots, t_M\} \), for \(|D| > 0\) refers to the data set of transactions, where each transaction \( t \in D \) is a list of distinct items \( t = \{i_1, i_2, \ldots, i_M\} \), \( 1 \leq |M| \leq |A| \) and each transaction can be identified by a distinct identifier TID.

3.1 Definition

In order to easily comprehend our measurement, some required definitions together with a sample transactional data are presented.
Definition 1. A set \( X \subseteq I \) is called an itemset. An itemset with \( k \)-items is called a \( k \)-itemset.

Definition 2. The support of an itemset \( X \subseteq I \), denoted \( \text{supp}(X) \), is defined as a number of transactions contain \( X \).

Definition 3. Let \( X, Y \subseteq I \) be itemset. An association rule between sets \( X \) and \( Y \) is an implication of the form \( X \Rightarrow Y \), where \( X \cap Y = \emptyset \). The sets \( X \) and \( Y \) are called antecedent and consequent, respectively.

Definition 4. The support for an association rule \( X \Rightarrow Y \), denoted \( \text{supp}(X \Rightarrow Y) \), is defined as a number of transactions in \( D \) contain \( X \cup Y \).

Definition 5. The confidence for an association rule \( X \Rightarrow Y \), denoted \( \text{conf}(X \Rightarrow Y) \) is defined as a ratio of the numbers of transactions in \( D \) contain \( X \cup Y \) to the number of transactions in \( D \) contain \( X \). Thus

\[
\text{conf}(X \Rightarrow Y) = \frac{\text{supp}(X \Rightarrow Y)}{\text{supp}(X)}.
\]

Definition 6. (Definite Factor). Definite Factor is a formulation of exploiting the support difference between itemsets with the frequency of an itemset against a baseline frequency. The baseline frequency of itemset is presumed as statistically independence. The Definite Factor denoted as DF and

\[
DF(I) = \frac{P(X \cup Y) - P(Y) \times \frac{P(X \cup Y)}{P(X)P(Y)}}{\text{supp}(X) - \text{supp}(Y) \times \frac{\text{supp}(X \Rightarrow Y)}{\text{supp}(X) \times \text{supp}(Y)}}.
\]

3.2 Construct Definite Least Association Rules

Rule is classified as Definite Least Association Rules (DLAR) if it fulfilled two conditions. First, DF of association rule must be greater than the predefined minimum DF. The range of min-DF is in between 0 and 1. Second, the antecedent and consequence of association rule must represent either Least Items or Frequent Items, respectively. The computation of DF of each association rule is employed from Definition 6. The complete procedure to construct the DLAR algorithm is as follows.

<table>
<thead>
<tr>
<th>DLAR Algorithm</th>
</tr>
</thead>
<tbody>
<tr>
<td>1: Specify ( DF^m )</td>
</tr>
<tr>
<td>2: for ( {DI, \in \text{DefiniteItemset}} ) do</td>
</tr>
<tr>
<td>3: for ( {DFI, \in DI, \cap \text{FrequentItems}} ) do</td>
</tr>
<tr>
<td>4: for ( {DLI, \in DI, \cap \text{LeastItems}} ) do</td>
</tr>
<tr>
<td>5: Compute ( DF(DFI, DLI) )</td>
</tr>
<tr>
<td>6: if ( DF(DFI, DLI) &gt; DF^m ) do</td>
</tr>
</tbody>
</table>
4 Experimental Results

In this section, we do experiment tests with DF measurements. The weight of all association rules were assigned according to this measurement. These experiments were conducted on Intel® Core™ 2 Quad CPU at 2.33GHz speed with 4GB main memory, running on Microsoft Windows Vista. All algorithms have been developed using C# as a programming language. We evaluate the proposed measurement to 2008/2009 intake students in computer science program. The data was obtained from Division of Academic, Universiti Malaysia Terengganu in a text file and Microsoft excel format. There were 160 students involved and their identities were removed due to the confidentiality agreement. In the original set of data, it consists of 35 attributes and the detail information were explained in 10 tables in Microsoft excel format.

Here, 8 chosen university programs by the students are extracted according to the fix location in the original flat file. The actual location for each programs are based on the fix column. There were in total of 822 bachelors programs offered in Malaysian public universities for July 2008/2009 students’ intake. From this figure, 342 bachelor programs were selected by our 160 students and it can be generalized into 47 unique general fields. In addition, LP-Tree and LP-Growth algorithm with DF measure (Abdullah, et al, 2010) are employed in the experiment. The total of 4,177 association rules was successfully extracted. Table 1 depicts top 10 of association rules based on the 3% of minimum support. Table 2 illustrates the meaning of association rules based on the Table 1.

<table>
<thead>
<tr>
<th>No</th>
<th>Association rules</th>
<th>Supp of Antc</th>
<th>Supp of Consq</th>
<th>Supp of Itemset</th>
<th>Jaccard</th>
<th>Corr</th>
<th>RSA</th>
<th>DF</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>25→9</td>
<td>4.38</td>
<td>90.63</td>
<td>4.38</td>
<td>0.048</td>
<td>22.86</td>
<td>1.00</td>
<td>0.95</td>
</tr>
<tr>
<td>2</td>
<td>25→34</td>
<td>6.25</td>
<td>90.63</td>
<td>6.25</td>
<td>0.07</td>
<td>16.00</td>
<td>1.00</td>
<td>0.93</td>
</tr>
<tr>
<td>3</td>
<td>25→8</td>
<td>12.50</td>
<td>90.63</td>
<td>12.50</td>
<td>0.14</td>
<td>8.00</td>
<td>1.00</td>
<td>0.86</td>
</tr>
<tr>
<td>4</td>
<td>25→43</td>
<td>9.38</td>
<td>90.63</td>
<td>8.75</td>
<td>0.10</td>
<td>10.67</td>
<td>0.93</td>
<td>0.84</td>
</tr>
<tr>
<td>5</td>
<td>28→40</td>
<td>3.75</td>
<td>68.75</td>
<td>3.12</td>
<td>0.05</td>
<td>26.67</td>
<td>0.83</td>
<td>0.79</td>
</tr>
<tr>
<td>6</td>
<td>25→41</td>
<td>8.13</td>
<td>90.63</td>
<td>6.88</td>
<td>0.07</td>
<td>12.31</td>
<td>0.85</td>
<td>0.77</td>
</tr>
<tr>
<td>7</td>
<td>25→38</td>
<td>15.00</td>
<td>90.63</td>
<td>13.75</td>
<td>0.15</td>
<td>6.67</td>
<td>0.92</td>
<td>0.76</td>
</tr>
<tr>
<td>8</td>
<td>25→31</td>
<td>5.63</td>
<td>90.63</td>
<td>4.38</td>
<td>0.05</td>
<td>17.78</td>
<td>0.78</td>
<td>0.73</td>
</tr>
<tr>
<td>9</td>
<td>28→34</td>
<td>6.25</td>
<td>68.75</td>
<td>5.00</td>
<td>0.07</td>
<td>16.00</td>
<td>0.80</td>
<td>0.73</td>
</tr>
<tr>
<td>10</td>
<td>25→28→34</td>
<td>6.25</td>
<td>60.23</td>
<td>5.00</td>
<td>0.08</td>
<td>16.00</td>
<td>0.80</td>
<td>0.72</td>
</tr>
</tbody>
</table>

Fig. 1: DLAR Algorithm
### Table 2. Explanation of top 10 of positive association rules

<table>
<thead>
<tr>
<th>No</th>
<th>Association rules</th>
<th>Explanation</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>25→9</td>
<td>The student chose Forestry program also chose Banking program</td>
</tr>
<tr>
<td>2</td>
<td>25→34</td>
<td>The student chose Forestry program also chose Nursing program</td>
</tr>
<tr>
<td>3</td>
<td>25→8</td>
<td>The student chose Forestry program also chose Art Design</td>
</tr>
<tr>
<td>4</td>
<td>25→43</td>
<td>The student chose Forestry program also chose Radiotherapy program</td>
</tr>
<tr>
<td>5</td>
<td>28→40</td>
<td>The student chose IT program also chose Psychology</td>
</tr>
<tr>
<td>6</td>
<td>25→41</td>
<td>The student chose Forestry program also chose Pure Sciences</td>
</tr>
<tr>
<td>7</td>
<td>25→38</td>
<td>The student chose Forestry program also chose Physiotherapy</td>
</tr>
<tr>
<td>8</td>
<td>25→31</td>
<td>The student chose Forestry program also chose Management</td>
</tr>
<tr>
<td>9</td>
<td>28→34</td>
<td>The student chose IT program also chose Nursing</td>
</tr>
<tr>
<td>10</td>
<td>25→28→34</td>
<td>The student chose Forestry and IT program also chose Nursing</td>
</tr>
</tbody>
</table>

The link of interest between the antecedent and consequence for the first rule until fifth rule is quite strange due to the contradiction in the field of study among the respective programs. The sixth rule is very realistic since both programs have a similarity in term of basic requirements, link of interest and nature of study. For the sixth until tenth rules, it is very hard and confused to explains, since there is no link of interest between the programs. From here we can see that the students have mixed up with several interests during choosing their preferred university’s programs. Moreover, most of them had chosen Forestry program. In summary, there are existed exceptional association rules in the university’s program selection database. This information is very important to give an overall idea about the student interests and how to channel them to a more appropriate university’s program.

### 5 Conclusion

Mining least association rules is very useful to help the organization in making a right decision. In educational context, identifying the suitable program for prospect students is very troublesome and usually ends up with programs availability. Therefore, this paper employed the Definite Factors measure to the students’ enrolment data of computer science program (intake 2008/2009) at University Malaysia Terengganu. The result shows that the applied measure can discover the significant least association rules. From the generated rules, 32% of the students that have been offered in computer science program are not within their program interests. Thus, effective monitoring process and analysis of these students are very important in helping them to adapt and finally enjoy with the current program.
Acknowledgement. This research is supported by Fundamental Research Grant Scheme (FRGS) from Ministry of Higher Education of Malaysia Vote RDU 110104.

References


