Available online at www.sciencedirect.com

SciVerse ScienceDirect

Procedia - Social and Behavioral Sciences 28 (2011) 97 – 101

Procedia
Social and Behavioral Sciences

WCETR 2011

Mining significant association rules from educational data using critical relative support approach

Zailani Abdullah[a,]*, Tutut Herawan[b], Noraziah Ahmad[b], Mustafa Mat Deris[c]

[a]Department of Computer Science, Universiti Malaysia Terengganu
21030 Kuala Terengganu, Terengganu, Malaysia
[b]Faculty of Computer System and Software Engineering, Universiti Malaysia Pahang
Lebuhraya Tun Razak, 26300 Kuantan Pahang, Malaysia
[c]Faculty of Computer Science and Multimedia
Parit Raja, Batu Pahat 86400, Johor, Malaysia

Abstract

Least association rules are the association rules that consist of the least item. These rules are very important and critical since they can be used to detect the infrequent events and exceptional cases. However, the formulation of measurement to efficiently discover least association rules is quite intricate and not really straight forward. In educational domain, this information is very useful since it can be used as a base for investigating and enhancing the current educational standards and managements. Therefore, this paper proposes a new measurement called Critical Relative Support (CRS) to mine critical least association rules from educational context. Experiment with students' examination result dataset shows that this approach can be used to reveal the significant rules and also can reduce up to 98% of uninterested association rules.

Keywords: Least association rules; Educational, Critical relative support;

1. Introduction

Data mining is a process of extracting useful information from large dataset by combining statistical and artifical intelligence techniques. It aims at discovering the interesting correlations, frequent patterns, associations or casual structures among sets of items in the data repositories. One of the emerging interdisciplinary research areas (Baker and Yacef, 2009) in educational context is educational data mining. Educational data mining can be defined as the application of data mining techniques to educational data (Romero *et. al*, 2008). The main concern of educational data mining is to develop methods that can explore the interesting information extracted from educational settings, and employ those methods to better understand the students, and the settings in which they learn (http://www.educationaldatamining.org).

Until this moment, association rules mining are one of the most important issues in data mining application. One of the commonly and popular techniques used in data mining application is association rules mining. The problem of association rules mining was first introduced by Agrawal *et al.* (1993) for market-basket analysis. There are two

* Zailani Abdullah. Tel.: +609-6683536; fax: +609-6694660.
E-mail address: zailania@umt.edu.my

doi:10.1016/j.sbspro.2011.11.020

main stages involved before producing the association rules. First, find all frequent items from transactional database. Second, generate the common association rules from the frequent items. In association rules, a set of item is also defined as an itemset. The itemset is said to be frequent, if it occurs more than a predefined minimum support. The item (or itemset) support is defined as a probability of item (or itemset) occurs in the transaction. Besides that, confidence is another alternative measurement used in pair in association rules. The confidence is defined as the probability of the rule's consequent that also contain the antecedent in the transaction. The association rules are said to be strong if it meets the minimum confidence.

Least itemset is a set of item that is rarely found in the database. It is also known as non-frequent, unusual, exceptional, abnormal, in-balance or sporadic itemset. Discovering of such itemset is very important because it may reveals a valuable knowledge for certain domain applications such as detection for air pollution, network intruders, machine critical faulty (Abdullah *et al.*, 2010), abnormal learning problems (Romero *et al.*, 2010), and many more. From the previous works, most of the tradition association rules mining algorithms (Kiran *et al.*, 2009; Zhou and Yau, 2007; Koh *et al.*, 2005; Yun *et al.*, 2003; Liu *et al.*, 1999; Wang *et al.*, 1999; Tao *et al.*, 2003; Ding, 2005) are still have a limitation in term of efficiency, scalability and are rarely applied into the real datasets. The low minimum support can be set to capture the least itemset. However, the trade off is it may generate the huge number of association rules. As a result, it is enormously difficult to identify which association rules are most interesting and really significant. Furthermore, the low minimum support will also proportionally increase the computational performance and its complexity.

The association rules provide simple logical rules about associations between items but not on determining the criticality relationship among them. They are said to be highly critical if it satisfies a certain predefined minimum critical threshold. Tracing that kind of relationship is very important for certain domain applications because it will reveal some crucial and hidden information. This information can be used by users as a guideline or mechanism for prevention or enhancement the respective processes in that particular domain. The challenges are, the most influential items, which have a critical relationship with other items may appear very rare and difficult to trace using classical techniques or methods. To make the thing worst, these influential items may occur together with others frequent items and when deriving association rules, the confidence level are very high. Therefore, it is a vital to put forward an appropriate model at discovering the least and critical association rules from the given database.

In certain circumstances, it is interesting to search for least association rules which may reveal a new knowledge and interesting discovery. These rules are very significant especially when dealing with extraordinary and exceptional cases. For example in educational context, computer science student whose get a good score in programming course should be able to obtain the average performance for others programming courses. But, the studies of significant relationship among computer science subjects are rarely focused. This study is very important and crucial for computer science syllabus improvement (university level) and for the real software engineer requirement (student level). Moreover, it also can be used as a base for investigating and enhancing the current educational standards and managements (Sevindik *et al.*, 2010; Buldu *et al.*, 2010; Romero *et al.*, 2009; Romero *et al.*, 2008; Dogan *et al.*, 2008; Enclieva *et al.*, 2006). Furthermore, the current measurements to evaluate whether the least association rules are significant or not are very limited (Zhou *et al.*, 2007) and quite intricate. In fact, the algorithm for triggering such exception rules is also very complex and these rules are always mixed up with the vast number of monotonous association rules.

In this paper, we propose a novel Critical Relative Support (CRS) to deal with the criticality or significant level of least association rules. The experiment was performed based on the examination result of computer science student from Universiti Malaysia Terengganu. Due to the confidentiality, the details about the dataset will not be disclosed.

The rest of the paper is organized as follows. Section 2 explains the basic concepts and terminology of association rule mining. Section 3 discusses the proposed method. This is followed by performance analysis through five experiment tests in section 4. Finally, conclusion and future direction are reported in section 5.

2. Preliminaries

Throughout this section the set $I = \{i_1, i_2, \cdots, i_{|A|}\}$, for $|A| > 0$ refers to the set of literals called set of items and the set $D = \{t_1, t_2, \cdots, t_{|U|}\}$, for $|U| > 0$ refers to the data set of transactions, where each transaction $t \in D$ is a list of distinct items $t = \{i_1, i_2, \cdots, i_{|M|}\}$, $1 \le |M| \le |A|$ and each transaction can be identified by a distinct identifier TID.

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 $\operatorname{supp}(X)$ 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 $X \cap Y = \phi$. The sets X and Y are called antecedent and consequent, respectively.*
Definition 4. The support for an association rule $X \Rightarrow Y$, denoted $\operatorname{supp}(X \Rightarrow Y)$, 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 $\operatorname{conf}(X \Rightarrow Y)$ 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*

$$\operatorname{conf}(\mathrm{X} \Rightarrow \mathrm{Y}) = \frac{\operatorname{supp}(\mathrm{X} \Rightarrow \mathrm{Y})}{\operatorname{supp}(\mathrm{X})}.$$

3. The Proposed Model

Some definition are required prior the method proposed.

Definition 9. (Least Items). *An itemset X is called least item if $\alpha \le \operatorname{supp}(X) \le \beta$, where α and β is the lowest and highest support, respectively.*
The set of least item will be denoted as Least Items and

$$\text{Least Items} = \{X \subset I \mid \alpha \le \operatorname{supp}(X) \le \beta\}$$

Definition 11. (Frequent Items). *An itemset X is called frequent item if $\operatorname{supp}(X) > \beta$, where β is the highest support.*
The set of frequent item will be denoted as Frequent Items and

$$\text{Frequent Items} = \{X \subset I \mid \operatorname{supp}(X) > \beta\}$$

Definition 13. (Merge Least and Frequent Items). *An itemset X is called least frequent items if $\operatorname{supp}(X) \ge \alpha$, where α is the lowest support.*
The set of merging least and frequent item will be denoted as LeastFrequent Items and

$$\text{LeastFrequent Items} = \{X \subset I \mid \operatorname{supp}(X) \ge \alpha\}$$

LeastFrequent Items will be sorted in descending order and it is denoted as

$$\text{LeastFrequent Items}^{\text{desc}} = \left\{ \begin{array}{l} X_i \,|\operatorname{supp}(X_i) \ge \operatorname{supp}(X_j)\}, \ 1 \le i, j \le k, \ i \ne j, \\ k = |\text{LeastFrequent Items}|, \ x_i, \mathrm{x_j} \subset \text{LeastFrequent Items} \end{array} \right\}$$

Definition 15. (Ordered Items Transaction). *An ordered items transaction is a transaction which the items are sorted in descending order of its support and denoted as t_i^{desc}, where*

$$t_i^{\text{desc}} = \text{LeastFrequentItems}^{\text{desc}} \cap t_i, 1 \le i \le n, \left|t_i^{least}\right| > 0, \left|t_i^{frequent}\right| > 0.$$

An ordered items transaction will be used in constructing the proposed model, so-called LP-Tree.
Definition 16. (Significant Least Data). *Significant least data is one which its occurrence less than the standard minimum support but appears together in high proportion with the certain data.*

Definition 18. (Critical Relative Support). *A Critical Relative Support (CRS) is a formulation of maximizing relative frequency between itemset and their Jaccard similarity coefficient.*

The value of Critical Relative Support denoted as CRS and

$$\mathrm{CRS}(I) = \max\left(\left(\frac{\mathrm{supp(A)}}{\mathrm{supp(B)}}\right),\left(\frac{\mathrm{supp(B)}}{\mathrm{supp(A)}}\right)\right) \times \left(\frac{\mathrm{supp}(A \Rightarrow B)}{\mathrm{supp(A)} + \mathrm{supp(B)} - \mathrm{supp}(A \Rightarrow B)}\right)$$

CRS value is between 0 and 1, and is determined by multiplying the highest value either supports of antecedent divide by consequence or in another way around with their Jaccard similarity coefficient. It is a measurement to show the level of CRS between combination of the both Least Items and Frequent Items either as antecedent or consequence, respectively. Typically, CRS value is completely depend on the range of support of antecedent and consequence. The more range of support between them, the more CRS will be reached to its maximum value.

4. Comparison Test

In this section, we do comparison tests between CRS with another two benchmarked measurements; RSAA and MSAA. All association rules have been assigned into these measurements. The details discussion and finding are only based on the CRS. These experiments have been 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 algorithm to student examination result in computer science programme for intake July 2007/2008. This programme has 30 subjects in computer science and the period of study is 3 years. There were 80 students involved and their identities were removed due to the confidentiality agreement. The data was obtained from IT Centre, Universiti Malaysia Terengganu in Microsoft excel format. The original data was given in the horizontal format which is only suitable for reporting purposes. It consists of 12 attributes: student status, metric number, name (firstname, middle name and last name), session (July or December), result (pass or fail), CGPA, GPA, course code, course name, credit hour, level (elective or compulsory) and grade. Due to the confidentiality matters, the data were converted into a new representation. A new set of data was generated based on the combinations of a new course code and a new grade code. For example, if the student obtained B+ for course code TMK3101, therefore the item 112 will be appeared in the new dataset. The first 2 number is corresponding to a new course code and the last number is for a new grade code. Association rules were generated in a form of many-to-one cardinality relationship and the maximum number of antecedents was set to four. The summary of selected association rules for analysis is shown in Figure. 1.

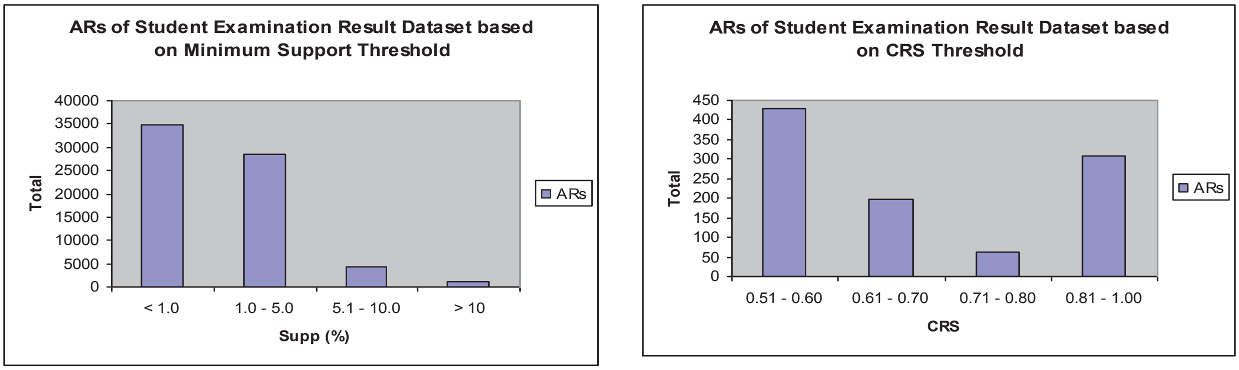

Figure 1. Total number of association rules (ARs) being produced using Minimum Support (Supp) and CRS thresholds

There were 68,855 of association rules produced from the Student Examination Result (SER) dataset. From these overloaded numbers of association rules, it is quite impossible and difficult to trace which association rules are really significant for users. However, by applying the minimum CRS of 0.5, it can dramatically reduce up to 99.55% of the existing association rules and thus providing more rationale in tracing the significant association rules.

5. Conclusion

Least association rules mining is one of the fundamental topic in data mining. The least rules that may consist of least and frequent items are sometimes very useful and critical for certain domain such as in educational. However, the research in least association rules and specifically in educational domain is nearly none. In fact, the special and scalable measurement is also required to deal with this type of association rules. In educational context, these unseen rules are very valuable since it can be used as a basis to improve the current educational standards and managements. In this paper, we have proposed a novel measurement, Critical Relative Support (CRS) to discover the significant and critical least association rules. Students' examination result dataset was employed and applied with CRS. The results show that CRS can easily discover the significant and least association rules. Furthermore, CRS can dramatically reduce the number of unwanted rules up to 98% as compared to the tradition minimum support threshold.

In a near future, we are going to evaluate the performance of CRS against several real datasets. We also believed that the CRS is also suitable to others real domain applications.

References

Abdullah, Z., Herawan, T., & Deris, M.M. (2010). Mining significant least association rules using fast slp-growth algorithm. *Lecture Notes in Computer Science*, 6059, Springer, Berlin Heidelberg New York, 324–336.

Agrawal, R., Imielinski, T., & Swami, A. (1993). Database mining: a performance perspective. *IEEE Transactions on Knowledge and Data Engineering*, 5(6), 914–925.

Baker, R., & Yacef, K. (2009). The state of educational data mining in 2009: a review and future visions, *Journal of Education Data Mining*, Vol. 1, Issue 1, 3–17.

Buldu, A., & Ucgun, K. (2010). Data mining application on students' data. *Procedia Social and Behavioral Sciences*, 2 (2010), 5251-5259.

Bas, G. (2010). Effects of multiple intelligences instruction strategy on students achievement levels and attitudes towards English Lesson. *Cypriot Journal Of Educational Sciences, 5*(3), 167-180.

Ding, J. (2005). Efficient association rule mining among infrequent items. Ph.D. Thesis, University of Illinois at Chicago

Enclieva, S., & Tumin, S. (2006). Application of association rules for efficient learning work-flow. In Z. Shi, Shimohara K., & Feng D. (Eds.), FIP International Federation for Information Processing, Vol. 228, *Intelligent Information Processing III*, (pp. 499-504). Boston: Springer.

Kay, J., Maisonneuve, N., Yacef, K., & Reimann, P. (2006). The big five and visualization of team work activity. In Ikeda, M., Ashley, K.D., & Chan, T-W. (Eds.), *Intellegent Tutoring Systems* (pp. 197-206). Springer-Verlag, Taiwan.

Kiran, R.U., & Reddy, P. K. (2009). An improved multiple minimum support based approach to mine rare association rules. In *Proceeding of IEEE Symposium on Computational Intelligence and Data Mining*, 340–347.

Koh, Y.S., and Rountree, N. (2005). Finding sporadic rules using apriori-Inverse. *Lecture Notes in Computer Science*, 3518, Springer, Berlin Heidelberg New York, 97–106.

Liu, B., Hsu, W., & Ma, Y. (1999). mining association rules with multiple minimum supports, *In Proceeding of ACM SIGKDD'07*, 337–341.

Romero, C., Romero, J.R., Luna, J. M., & Ventura S. (2010). Mining rare association rules from e-Learning data, *In Proceeding of The Third International Conference of Education Data Mining*, Pittsburgh, USA, 171-180

Romero, C., Venturaa, S., & Garcíaa, E. (2008). Data mining in course management systems: Moodle case study and tutorial. *Journal of Computers & Education*, Vol. 51, Issue 1, 368–384.

Romero, C., Venturaa, S., Zafraa, A., & Brab., P. (2009). Applying web usage mining for personalizing hyperlinks in web-based adaptive educational systems. *Journal of Computers & Education*, Vol. 53, Issue 3, 828-840

Sevindik, T., Demirkeser, N., & Cömert, Z. (2010). Virtual education environments and web mining. *Procedia Social and Behavioral Sciences*, 2 5120–5124

Tanimoto, S.L. (2007). Improving the prospects for educational data mining. *In Proceedings of the complete on-line proceedings of the workshop on data mining for user modelling, 11th International conference on user modeling (UM 2007)*, 106-110

Tao, F., Murtagh, F., & Farid, M. (2003). Weighted association rule mining using weighted support and significant framework. *In Proceeding of ACM SIGKDD '03*, 661–666.

Wang, K., Hee, Y., & Han, J. (2003). Pushing support constraints into association rules mining. *IEEE Transactions on Knowledge and Data Engineering*, 15(3), 642–658.

Yun, H., Ha, D., Hwang, B., & Ryu, K.H. (2003). Mining association rules on significant rare data using relative support. *The Journal of Systems and Software*, 67 (3), 181–19.

Zhou L., & Yau, S. (2007). Assocation rule and quantative association rule mining among infrequent items. *In Proceeding of ACM SIGKDD'07*, Article No. 9, 156-167

http://www.educationaldatamining.org/ Retrieved on 01-03-2011