International Journal of Information and Computation Technology.
ISSN 0974-2239 Volume 3, Number 5 (2013), pp. 411-418
© International Research Publications House
http://www. irphouse.com /ijict.htm

Educational Data Mining using Improved Apriori Algorithm

Jayshree Jha[1] and Leena Ragha[2]

*[1,2]Department of Computer Engineering, Ramarao Adik Institute of Technology,
Navi Mumbai, India.*

Abstract

Educational Data Mining (EDM) is an emerging interdisciplinary research area that deals with the development of methods to explore data originating in an educational context. EDM uses different computational approaches to analyze educational data in order to study educational questions. Different data mining techniques has been applied in this area. One of the most widely used techniques in EDM is association rules mining. Apriori algorithm is the first and best-known algorithm for association rules mining.

This paper surveys the most relevant studies carried out in EDM using Apriori algorithm. Based on the Apriori algorithm analysis and research, this paper points out the main problems on the application Apriori algorithm in EDM and presents an improved support-matrix based Apriori algorithm. The improved Apriori algorithm proposed in this research uses bottom up approach along with standard deviation functional model to mine frequent educational data pattern.

Keywords: Education Data Mining, Association rule mining, Apriori algorithm.

1. Introduction

The use of computers in learning and teaching has advanced significantly over the last decades through different e-learning systems. Nevertheless, the need for improvement has always been present. Students learn, explore content, and by doing so, they leave trail of log information. There is an important research question: what can we do with this data? These parameters could easily be recorded in e-learning system, analyzed, and as a result, teacher could adapt tests to maximize performance. Data mining (DM) techniques rise as an answer since they are used in different research areas (e.g. medicine, business, market research, etc.) with vast amount of provided data. Some similar ideas were already successfully applied in e-commerce systems, the first and most popular application of DM, in order to determine clients' interests so as to be able to increase online sales. However, until today, there has been comparatively less progress in this direction in education, although this situation is changing and there is currently an increasing interest in applying DM to the educational environment [1].

Educational Data Mining is an emerging discipline, concerned with developing methods for exploring the unique types of data that come from educational settings, and using those methods to better understand students, and the settings which they learn in [2]. Its main objective is to analyze these types of data in order to resolve educational research issues [3]. It becomes an imperative and important research topic that discovering hidden and useful knowledge from such an extensive quantity of data to guide and develop education.

The data mining techniques being widely used in teaching system is association rules mining .Association rule mining has been applied to web-based education systems from two points of view: 1) help professors to obtain detailed feedback of the e-learning process: e.g., finding out how the students learn on the web, to evaluate the students based on their navigation patterns, to classify the students into groups, to restructure the contents of the web site to personalize the courses; and 2) help students in their interaction with the e-learning system: e.g., adaptation of the course according to the apprentice's progress, e.g., by recommending to them personalized learning paths based on the previous experiences other similar students.

Apriori algorithm is the first and best-known for association rules mining. It is an influential algorithm for mining frequent itemsets for Boolean association rules. This algorithm uses a level-wise search, where k-itemsets are used to explore (k+1)-itemset, to mine frequent itemsets from transactional database.

Based on Apriori Algorithm, this paper surveys the most relevant studies carried out in EDM as a literature survey in chapter 2. This section also discusses the extent to which reviewed work overcome the limitation specified in same section. Section 3 presents the proposed algorithm which uses bottom up approach and support matrix for mining frequent data patterns.

2. Association Rule Mining

Association rule mining is one of the most well studied data mining tasks. It discovers relationships among attributes in databases, producing if-then statements concerning attribute values [4]. In association rules we find the co-occurrences among item sets through finding the large item sets. Association rule mining is usually split up into two separate steps: First, minimum support is applied to find all frequent itemsets in a database. Second, these frequent itemsets and the minimum confidence constraint are used to form rules. There are a lot of different association rule algorithms. A comparative study between the main algorithms that are currently used to discover association rules can be found in [5]: Apriori, FP-Growth, MagnumOpus, and Closet.

2.1 Association Rule Mining in EDM

Association rule mining has been applied to e-learning systems for traditionally association analysis (finding correlations between items in a dataset), including, the following tasks: building recommender agents for on-line learning activities or shortcuts, automatically guiding the learner’s activities and intelligently generate and recommend learning materials , identifying attributes characterizing patterns of performance disparity between various groups of students , discovering interesting relationships from a student’s usage information in order to provide feedback to the course author , finding out the relationships between each pattern of a learner’s behaviour , finding student mistakes often occurring together , guiding the search for the best fitting transfer model of student learning , optimizing the content of an e-learning portal by determining the content of most interest to the user , extracting useful patterns to help educators and web masters evaluating and interpreting on-line course activities , and personalizing e-learning based on aggregate usage profiles and a domain ontology [3].

2.2 Apriori Algorithm

Apriori algorithm is the first and best-known for association rules mining. Apriori was proposed by R.Agrawal and R.Srikant [6] in 1994. Apriori algorithm is one of the most influential Boolean association rules mining algorithm for frequent itemsets. It is an iterative algorithm to calculate the specific length of item collection of given database to produce frequent item sets. It cut down candidate item sets using the principle that all non empty subsets of frequent item sets are frequent too. Apriori algorithm basically works in two steps. In first step candidate itemset is generated using linking process and in next step frequent itemset from those candidate itemset is found based on minimum support count by scanning the database.

2.3 Limitation of Apriori Algorithm EDM

In spite of being simple and clear, Apriori algorithm has some limitation. It is costly to handle a huge number of candidate sets. For example, if there are 10^4 frequent 1-item sets, the Apriori algorithm will need to generate more than10^7 length-2 candidates and accumulate and test their occurrence frequencies. This is the inherent cost of candidate generation, no matter what implementation technique is applied. It is tedious to repeatedly scan the database and check a large set of candidates by pattern matching, which is especially true for mining long patterns. We can see from the above description of the Apriori algorithm that in general there exist two major shortcomings in Apriori algorithm [7]. Firstly, it needs to scan database repeatedly and secondly, it needs to generate large number of candidate item set. Apart from this Apriori algorithm also assumes that the weight of each item and transaction in database is same which is not true always [8].

3. Literature Survey

As basic Apriori Algorithm can't be used for education data mining due previously mentioned shortcoming, various authors have suggested variant in Apriori algorithm to address the mentioned shortcomings.

- Mixed Weighted Association Rules Mining Algorithm: Xin-hua Zhu,Ya-qiong Deng, Qing-ling Zeng [8] in their paper have taken grades of computer cultural foundation course as example. The weighted association rules algorithm have been used to analyze grades of college-wide examination course in this paper, giving the model and mining method of mixed weighted association rules on grade database. Compared to directly apply the Apriori algorithm, more valuable correlations have been obtained between the chapters, chapters and scores, colleges and chapters at the same threshold values. Thus, it is more helpful to guide teachers for their teaching and for the improvement of teaching quality.
- Clustering with Apriori algorithm: Zhiyu Zhang [9] in his paper used Clustering algorithm to first categories the students and courses and then with the help Apriori algorithm various hidden information is extracted from the large amount of education data.

Chandrani Singh, Dr. Arpita Gopal ,Santosh Mishra [10] in their paper deals with the extraction and analysis of faculty performance of management discipline from student feedback using clustering and association rule mining techniques. First the faculty members are categorised based on student feedback data and then Apriori algorithm has been implemented to undermine the hidden trends in Faculty performance and their behaviour.

- Matrix based Apriori algorithm: Hong Liu, Yuanyuan Xia [11] in their paper used matrix based Apriori algorithm to extract and analyse the Indicator-score in teaching evaluation data and found the information of Indicator-score that have high frequency, and then analyzed the strengths and weaknesses of the teaching, to provide recommendations of improving teaching quality of teachers. This method does not repeat the database scanning, thus reducing the I/O load.
- Improved Apriori algorithm based on Tid set: Qiang Yang,Yanhong Hu [12] in their paper used improved Apriori algorithm to find the correlation rules of course which provided the directive significance information for the curriculum . This algorithm need to scan the original database only once when generating candidate item set, it compute support count of the other candidate item sets through stating the count of the corresponding Tid set, not scanning the database repeatedly, which saves the visiting time greatly.

- Improved Apriori algorithm based on clipping technique: Jian Wang, Zhubin Lu, Weihua Wu and Yuzhou Li [13] in their paper used improved Apriori algorithm of association rules to analyze the intrinsic link among various courses, dig out the precedence relationship and association of students' learning courses, reveal the teaching regularities and problems from large amount of data, as well as provide a strong basis for reasonable course-setting . Improved Apriori algorithm uses clipping technique to remove all candidate itemset in Ck that doesn't belong to Lk-1.
- Improved Apriori algorithm based on logo list intersection: Lanfang Lou, Qingxian Pan, Xiuqin Qiu [14] in their paper proposed a novel association rules for data mining to improve Apriori algorithm. The proposed approach uses the intersection operation to generate frequent item sets. It is different from the existing algorithm as it scans the database only one time and then uses the database to mine association rules. The proposed technique has been implemented in a teaching evaluation system, to enhance the foundation in performance evaluation for staff in teaching issues.
- Improved Apriori Algorithm based on Modified Pruning process and flag bit: Deng Jiabin, Hu JuanLi, Chi Hehua, Wu Juebo[15] in their paper put forward a kind of intelligent evaluation method based on improved Apriori, which can be used to mine different levels of association rules and evaluate the teaching quality automatically. The improvement ideological of the frequent items: Between the Lk and Ck, introducing the Lk', when one item has been validated that it is not a frequent item set, it will be inserted into Lk', but not be deleted. In order to distinguish an item set whether it is frequent item sets or non-frequent item sets flag bit is introduced into the item sets. When it is the frequent item set, we use 1, or else use 0. At the same time, the verification process and the pruning process also need to be modified: when verifying the candidate set Ck, each time, we select items from the item set Ck to verify. However, each time, we select items from Lk' in the pruning process as the pruning conditions and iteratively generate Lk +1'.

Different techniques have implemented to improve the shortcoming of classic Apriori algorithm in education data mining. Although these improved algorithms can reduce the number of candidate itemsets or improve the mining efficiency by pruning methods, but still can't completely solve the problem of which candidate itemsets appear no longer. And, what's more, facing masses of education data for mining long pattern to adopt basic association rules mining is not the solution to problem as they will be producing a large number of candidate itemsets, using lots of memory space and CPU processing time. Apart from this setting the appropriate minimum support threshold is also an issue as it may lead to too many or too few rules.

4. Proposed Algorithm

This paper presents an Improved Apriori algorithm based on Bottom up approach and Support matrix to identify frequent item set. The proposed algorithm replaces arbitrary user defined minimum support with functional model based on Standard Deviation. In proposed algorithm Minimum support value is calculated based upon Standard Deviation value of support counts of all transactions. This approach make this algorithm more comfortable for somebody non expert in data mining. Presented algorithm uses Bottom up Approach to find the frequent item set from largest frequent Item set to smallest frequent item set which help in mining long pattern easily.

This algorithm works in 2 phases, Support Matrix Generation and Bottom Up approach to mine frequent items set based upon calculated minimum support.

Phase 1:- Generation of sorted Support matrix

Steps to generate sorted Support Matrix are as follows:-

Step 1:- Scan database to generate Boolean matrix A1. Rows in matrix represent transaction. Columns in matrix represent items. Each cell will have the values either 1 or 0 for representing presence of items in the transaction. Entry value 1 indicates the corresponding item is present in transaction and value 0 indicates the corresponding item is not present in transaction.

Step 2:- Calculate the Support value of each item. Boolean Matrix A1 is transformed to Support Matrix A2, by replacing each entry value of 1 by the Support value of corresponding item and inserting two more columns to the Support Matrix A2 to hold total Support value and Count of elements in each row respectively.

Step 3:- The Support Matrix A2 will be rearranged in descending order in accordance with total Support value and non-zero entry will be replaced by 1 which leads to generation of Sorted Support Matrix A3.

Step 4:- Calculate the value of Minimum Support Count based upon standard deviation parameter.

Minimum Support, MinSup = Mean - Standard deviation

Standard Deviation, $\sigma = \sqrt{\frac{\sum_{i=1}^{N}(Support[i]-Mean)^2}{N-1}}$

Mean = $\frac{\sum_{i=1}^{N} Support[i]}{N}$

Where N = Total numbers of items.

Infrequent items are the items that have low support value that lies on the extra small region defined by Standard Deviation parameter.

Phase 2:- Bottom Up approach to mine frequent items.

Steps to mine frequent items are as follows:-

Step 1:- Simplify the matrix A3. Remove the column for items from the A3 for which support is less than Minsupport.

Step 2:- Select first transaction from A3 and compare its total Support value and count with next transaction total support and count respectively. If the next transaction total support value and count is greater than or equal to the first transaction then do the BITWISE AND operation between the transaction, if the resultant of AND operation is equal to first transaction structure then increase the support count of first transaction item set by 1. Continue this process with remaining transaction. At the end check the total support count of first transaction if it is greater than or equal to the Minimum support count extract the item set of that transaction and all its subse1t and move it to frequent Item set. The same process will be repeated for remaining transaction.

The Major advantage of this algorithm is that it avoids comparison of currently chosen transaction with other transaction to mine the frequent item set if the total Support value or count of the other transactions on which comparison needs to be done is lesser than the chosen transaction. Since the lesser support value in next transaction indicates that transaction does not contain all items or item set of transaction under scanning process. Another feature of this approach is that once the largest frequent item set is found all its subsets will be identified and moved to frequent items set. While considering next transaction to find next largest frequent item set first it checks whether item set of transaction under scanning process is already in frequent items set, if it's already in frequent item set, it avoids another set of comparison required to find the support of item set. These two striking features leads to reduction in number of scans and time required to mine the frequent item set. Moreover it also replaces user defined arbitrary minimum support threshold value of standard Apriori with functional model based on standard deviation which means that this algorithm can be well used by a non data mining expert.

5. Conclusion

The paper addresses the importance of knowledge mining from education dataset and overview of existing algorithm used in education data mining and its flaws and innovative solution with a new algorithm for data mining from the education dataset. This paper propose a new improved Apriori algorithm with a main motive of reducing time and number of scans required to identify the frequent itemset and association rules among education data using bottom up approach. Moreover proposed algorithm also replaces the user-defined minimum threshold with standard deviation based functional model as mentioning minimum support value in advance may lead to either too many or too few rules which can negatively impact the performance of entire model.

6. Acknowledgments

My sincere thanks to all the people who have contributed in anyway in carrying out this work.

References

[1] Divna Krpan, Slavomir Stankov, " Educational Data Mining for Grouping Students in E-learning System", In Proceeding of ITI 2012 34th Int. Conf. on Information Technology Interfaces, June 2012

[2] www.educationaldatamining.org

[3] Cristobal Romero,"Educational Data Mining: A Review of the State of the Art", IEEE transaction on systems, MAN, and Cybernetics-part c: Application and Reviews, VOL. 40, NO. 6, November 2010

[4] Agrawal, R., Imielinski, T. and Swami, A.N., Mining Association Rules between Sets of Items in Large Databases. In Proceedings of SIGMOD, 207-16, 1993.

[5] Zheng, Z., R. Kohavi and Mason, L., Real world performance of association rules. Sixth ACM SIGKDD International Conference on Knowledge Discovery & Data Mining. 2(2),86-98, 2001.

[6] Agrawal R., Srikant R.,"Fast algorithms for mining association rules", In Proceedings 20th International Conference on VeryLarge Data Bases (VLDB' 94), pp. 487-499, 1994.

[7] Deng Jiabin, Hu JuanLi, Chi Hehua, Wu Juebo, "An Apriori-based Approach for Teaching Evaluation", IEEE, 2010.

[8] Xin-hua Zhu, Ya-qiong Deng, Qing-ling Zeng," The Analysis on Course Grade of College-wide Examination Based on Mixed Weighted Association Rules Mining Algorithm", ICCASM 2010.

[9] ZhiyuZhang,"Study and Analysis of Data Mining Technology in College Courses Students Failed", IEEE, 2010.

[10] Chandrani Singh Dr. Arpita Gopal Santosh Mishra "Extraction and Analysis of Faculty Performance of Management Discipline from Student Feedback using Clustering and Association Rule Mining Techniques", IEEE 2011.

[11] Hong Liu Yuanyuan Xia,"Teaching Evaluation System Based on Association Rule Mining",IEEE 2011.

[12] Qiang Yang, Yanhong Hu, "Application of Improved Apriori Algorithm on Educational Information",IEEE 2011.

[13] Jian Wang, Zhubin Lu, Weihua Wu and Yuzhou Li," The Application of Data Mining Technology based on Teaching Information", ICCSE 2012.

[14] Lanfang Lou, Qingxian Pan, Xiuqin Qiu," New Application of Association Rules in Teaching Evaluation System",IEEE 2010.

[15] Deng Jiabin, Hu JuanLi, Chi Hehua, Wu JueboAn, "Apriori-based Approach for Teaching Evaluation,IEEE 2010.