Research Paper

Volume : 1 | Issue : 4 | Sep 2012 • ISSN No 2277 - 8179

A Novel Market Basket Analysis Using Adaptive Association Rule Mining Algorithm

Computer Science

KEYWORDS : Data Mining, Association Rule Mining, Adaptive Association Rule Mining.

M.Dhanabhakyam — Ph.D Scholar, Sri Ramakrishna College of Arts and Science for women, Coimbatore

Dr.M.Punithavalli — Director, Department of Computer Applications, Sri Ramakrishna Engineering College , N.G.G.O. Colony Post, Vattamalaipalayam, Coimbatore

ABSTRACT

Data mining is the process of extracting relatively useful information from a large data base. In recent business trends, it is necessary to transform the data available in a database into an informational advantage. As the technology growth is mounting, the amount of information which is stored in databases is rapidly increasing. It is very difficult to find the valuable information hidden in those databases. Many researches were done by the database community based on association rule mining and classification technique to find the related information in large databases. An effective association rule mining technique is used in this chapter for analyzing and predicting the best combination of the items which will be very useful for the users in getting items very easily without much effort. The knowledge of the buying patterns can be utilized to enhance the placement of these items in the super market or the layout of the catalog pages. This requirement has led to the establishment of techniques that automatically look for associations between items that are stored in databases. Hence, Adaptive Association Rule Mining Algorithm (AARMA) is proposed in this paper and it is evaluated using mushroom dataset accuracy, Area Under the Curve (AUC) and execution time.

Introduction

Data mining is a process for identifying past unknown and significant abstractions from the content of large databases. Moreover, Data mining is a powerful new technology which focuses on the extraction of hidden predictive information. This technique has an enormous capability to help many organizations to concentrate on the interesting information in their data warehouses. Data mining tools can effortlessly predict the future trends and behaviors of the new technologies, with which it can help the organizations to make proactive, knowledge-driven decisions. The prospective investigation provided by data mining technique is very effective compared to the analyses of previous events by retrospective tools. Data mining tools can easily resolve many business problems that are traditionally time consuming. This approach can search databases for the hidden patterns and identifies the interesting data that experts may miss as it lies outside their prospects [1].

Data mining in market basket analysis has been a hot research topic for several decades. It includes the application of data mining techniques to market basket analysis. At present, several supermarket databases are capable of providing a lot of information on customer purchasing behaviours, which can be investigated in order to find the items that are purchased together and also the items that are frequently purchased [2]. A typical example of association rule mining is market basket analysis [3]. The advancement in computing and information storage has developed huge amount of data in supermarket. The challenge is to obtain constructive information from this raw data and to increase the sales and to obtain valuable information from agriculture databases.

Physical analysis of these huge amount of information stored in modern databases is very difficult. Data mining provides tools to reveal unknown information in large databases which are stored already. A well-known data mining technique is association rule mining. It is able to discover all interesting relationships which are called as associations in a database. Association rules are very efficient in revealing all the interesting relationships in a relatively large database with huge amount of data [4]. The large quantity of information collected through the set of association rules can be used not only for illustrating the relationships in the database, but also used for distinguishing between different kinds of classes in a database. But the major difficulty in association rule mining is its complexity. Association rule mining identifies the remarkable association or relationship of a large set of data items. As huge quantity of data is constantly being obtained and stored in databases, several industries are becoming concerned in mining association rules in their databases.

Data mining techniques are very easy to handle. It is very easy to implement this technique on the existing software and hardware platforms to improve the quality of the information resources. It can also be integrated with new techniques and systems which are newly introduced in the market. When data mining techniques are applied on high performance client/server or parallel processing computers, it can examine massive databases and offer solutions to various problems [5].

A prediction model is developed using data mining technique. This prediction model can be used to predict the useful and most appropriate combination of items in a super market [6]. The reasons for using the data mining approach in the prediction process are

1. Automated prediction of trends and behaviors: Data mining approach can automatically finds the predictive information in large databases. Data mining is a less time consuming process.
2. Automated discovery of previously unknown patterns: Data mining can sweep through the whole database and identifies the previously hidden patterns in less time.

II. Related Works

Wang et al., [7] suggested a novel rule weighting approach in classification association rule mining. Classification Association Rule Mining (CARM) is a newest classification rule mining technique that built an association rule mining based classifier by using Classification Association Rules (CARs). The specific CARM algorithm which is used is not regarded, a similar set of CARs is continually produced from data, and a classifier is commonly presented as a structured CAR list, depending on a selected rule ordering approach. Several numbers of rules ordering approaches have been recognized in the recent past, which can be catego-

rized as rule weighting, support-confidence and hybrid. In this approach an alternative rule-weighting method, called CISRW (Class-Item Score based Rule Weighting) and constructs a rule-weighting based rule which orders mechanism depending on CISRW. Later on, two hybrid techniques are added and developed by merging (1) and CISRW.

Bartik [8] presented association based classification for relational data and its use in web mining. Classification according to the mining association rules is a technique with better correctness and human understandable classification scheme. The intention of the author is to force an alteration of the fundamental association based classification technique that can be useful in data gathering from Web pages. The alteration of the technique and necessary discretization of numeric characteristics are given.

Dong Liyan et al., [9] proposed a novel method of mining frequent item sets. The target of mining association rules is to decide the association relationship along with the item sets from mass data. In a number of practical applications, its responsibility is mostly to support in decision-making. The author proposed an association rule algorithm of mining frequent item sets, which introduces a new data structure and takes compressed storage tree to increase the run presentation of this algorithm. Finally, the experiment specifies that the algorithm proposed has a lot benefits in load balance and run time compared with most existing algorithms.

Lei Wen et al., [10] developed an efficient algorithm for mining frequent closed itemset. Association rule mining is a prominent field of data mining analysis. Identifying the useful and significant frequent itemset is a key step. The existed frequent itemset discovery algorithms could discover all the frequent itemset or maximal frequent itemset Pasquier et al., [11] proposed a novel method of mining frequent closed itemset. The size of frequent closed itemset was much lesser than all the frequent itemsets and did not lose any information. A new frequent closed itemset method is proposed using the directed specified itemset graph. This method can identify all the frequent closed itemset significantly through depth first search technique.

Mining frequent itemsets from secondary memory was put forth by Grahne et al., [12]. For the main memory databases it is understood as the main work of mining association rules (i.e.) mining frequent itemsets. The author reveals techniques for mining frequent itemsets when the database or the data structures used in the mining are bulk to apply in main memory. Therefore this technique decreases the required disk used by order of magnitude, and lets actual scalable data mining.

III. Methodology

This section illustrates how this algorithm to mine association rules may be used for MBA. The most important differentiation between the implementation of this association types are that the handling of different transaction data to mine the association rules, and different recommendation strategies.

2.1.Mapping Ratings to Transactions

The conversion from item ratings available for recommendation tasks to "transactions" as required for association rule mining is determined by what kind of associations and how many levels of associations want to discover. Then map the numeric ratings for an item into two categories: like and dislike according to whether the rating for the item is greater than or less than some chosen threshold value. Then convert the chosen like and dislike ratings into transactions:

i) With the intention of obtaining like associations among users, assume each user correspond to a "user" and items rated by users correspond to a "transaction". If a user likes an item, then the transaction equivalent to the item contains the item related to the user liking the item; If the user dislikes or did not rate the item, then the equivalent transaction does not include the corresponding item. The mined rules will then be of the following form: " 90% of items liked by user A and user B are also liked by user C, 30% of all articles are liked by all of them", or, in simpler notation,"[usera : like] AND [userb : like]⇒ [userc : like] with confidence 90% and support 30%".

ii) In order to mine like associations among items, assume each item corresponds to an "item" and each training user who rated the target item correspond to a "transaction". If a training user likes an item, then the transaction related to the user contains the item equivalent to the item; If the user dislikes or didn't rate the item, then the equivalent transaction does not include the related item. From here, like associations among articles can be mined as: "[item1: like] AND [item4: like] ⇒ [target-item: like]" with confidence 100% and support 40%.

2.2.Recommendation Strategy

A)User associations

For user associations, the rules mined are akin to [training-user1: like] AND [training-user2: like] ⇒ [target-user: like]. If training-user1 likes a test item and training user2 also likes this item, then it is said that this rule fires for this item. Associate each rule with a score, which are the product of the support and the confidence of the rule. Then assign a score to each item, which is the sum of the scores of all the rules that fire for that item. If the score for an itemi is greater than the threshold, then recommend the itemi to the target user.

B)Item associations

For item associations, the rules are of the form: [item1: like] AND [item 2: like] ⇒ [target_ item: like]. For a test item of the target user, if the user likes item1 and item2 (which could be known from the training items of the user), then it is said that this rule fires for this item.

The recommendation strategy for item associations is dissimilar than for user associations. Here, items whose rules' supports are above a cutoff are taken into consideration. The support cutoff is adjusted during system tuning; the mining process is then restricted to rules whose support is above the cutoff. This mining process has the following advantages over algorithms such as Apriori [13] or CBA:

i. By mining item associations for one item at a time, only ratings related to the target item are used for mining, which is only a small subset of the whole rating data. The support of a rule is calculated over the small subset of the whole rating data, which enables to obtain rules for items that have only received a limited number of ratings, for example a new product.
ii. A considerable amount of runtime is saved by mining rules only over the subset of the rating data that is associated to the target item rather than over the whole data items. Systems that mine rules with unrestricted heads for instance IBM's Intelligent Miner can effortlessly take numerous days to mine item associations for all articles at once.

Volume : 1 | Issue : 4 | Sep 2012 • ISSN No 2277 - 8179

C)Combined associations mode
The following strategy is used to combine user and item associations: If a user's minimum support is greater than a threshold, then user associations is used for recommendation, otherwise item associations is used for the effective association.

2.3.Adaptive Algorithm Association Rule Mining Algorithm (AARMA)
This approach relies on information about relationships between different users' preferences in order to suggest items of potential interest to the target user. This algorithm adjusts the minimum support of the rules during mining in order to obtain an appropriate number of significant rules for the target predicate.

The new AARMA consists of two parts: AARMA-1 and AARMA-2.

AARMA-1 With the intention of mining only a specified number of most capable rules for each target item, AARMA-1 is used to control the minimum support count and discover the rules with the highest supports. The minimum support count is the smallest amount of transactions that convince a rule with the aim of making that rule frequent, specifically; it is the multiplication of the minimum support and the whole number of transactions. The overall process of AARMA-1 algorithm is shown in figure 3.1.

Input: Transactions, targetItem, minConfidence, minRulenum, maxRulenum

Output: minedRulenum

1) set initial minsuppoatCount based on targetItem's like ratio;
2) r= AARMA-2();
3) while (R,rulenum=maxRulenum) do
4) minusupportCount++;
5) R1= AARMA-2();
6) if R1.rulenum > minRlenum then R=R1;
7) else return R;
8) end
9) while(R.rulenum<minRulenum) do
10) minsupportCount--;
11) R=ARMA-2();
12) end
13) return R;

Figure 3.1: The AARMA-1 Algorithm

The working of the above algorithm is described below:

1. AARMA-1 starts the minimum support count based on the frequency of the target predicate and calls AARMA-2 to mine rules.
   When AARMA-2's output is returned, AARMA-1 will initially verify if the number of rules returned is equivalent to maxRulenum (as described below, AARMA-2 terminates the mining process when the number of rules generated is equal to maxRulenum). If it is, that means the minimum support count is small which causes above maxRulenum rules, as a result the AARMA-1 will keep raising the minimum support count and calling AARMA-2 until the number of rules is less than maxRulenum.
2. Lastly, AARMA-1 will verify if the number of rules is fewer than minRulenum; if it is, it will keep diminishing the minimum support count until the rule number is better than or equal to minRulenum.
   Within a specified support, rules with smaller bodies are mined initially. Therefore, if with minimum support count say 15 there is no rule available, but with minimum support count 16 there are at least maxRulenum rules, then AARMA-1 will return the shortest maxRulenum rules with support count of at least 16.

AARMA-2 is an alternative of CBA-RG and as a result of the Apriori algorithm also. AARMA-2 is an alternative of CBA-RG in the sense that rather than mining rules for all target classes, it only mines rules for one target item. It varies from CBA-RG in that it will simply mine a number of rules within a particular range. When it attempts to produce a new rule after having acquired maxRulenum rules previously then it just terminates its execution and returns the rules it has mined until now. AARMA-2 algorithm is presented in figure 3.2.

Input: Transaction, targetItem, minConfidence, maxRulenum, minsupportCount

Output: mined association rules

1) F1={frequent1-condsets};
2) R=genRules(F1);
3) if R.rulenum=maxRulenum then return R;
4) for (k=2;Fk-1^{1}Æ;k++) do Begin
5) Ck=candidateGen(Fk-1);
6) for each transaction tÎCt contained in t;
7) Ct=all candidate condsets of Ck contained in t;
8) for each candidate cÎCt do Begin
9) C.condsupCount++;
10) If t contains targetItem then c.rulesupCount++;
11) end
12) end
13) Fk={c=Ck|c.rulesupCount ^{3} minisupportCount};
14) R=R È genRules(Fk);
15) if R.rulenum=maxRulenum then return R;
16) end
17) return R;

Figure 3.2: AARMA-2 Algorithm

Here k-condset is used to indicate a set of items (or itemset) of size k which possibly will form a rule: k-condset ⇒ target-item. The support count of the k-condset called condsupCount is the amount of transactions that include the k-condset. The support count of the equivalent rule (also called rulesupCount of this k-condset) is the number of transactions that include the condset in addition to the target item.

AARMA-2 is extremely like CBA-RG as stated above. Association rules are produced by making multiple passes over the transaction data. The initial pass calculates the rulesupCounts and the condsupCounts of all the particular items and discovers the frequent 1-condsets. For pass $k > 1$, , it produces the candidate frequent k-condsets by making use of the frequent -condsets; after that it scans all transactions to count the rulesupCounts and the condsupCounts of all the candidate k-condsets; at last, it will go over the entire candidate k-condsets, choosing those whose rulesup is above the minimum support as frequent k-condsets and simultaneously generating rules k-condset ⇒ target-item, if the confidence of the rule is above the minimum confidence.

IV. Experimental Results
In order to evaluate the proposed approach, mushroom dataset is used. The performances of the algorithms were evaluated using various parameters. The performance of the proposed approach is evaluated against,

- Association Rule Mining (ARM) and
- Original Apriori Algorithm (OAA).

The performance of the proposed approach is evaluated using the parameters like

- Accuracy,
- Area Under the Curve (AUC) and
- Execution time.

There are 211 datasets available in the UCI Machine Learning Repository [University of California, Irvine (UCI)]. But mushroom dataset is more suitable for evaluating the market basket analysis approaches [12]. Mushroom records are drawn from the Audubon Society Field Guide to North American Mushrooms (1981). G. H. Lincoff (Pres.), New York: Alfred A. Knopf. This data set comprises of descriptions of hypothetical samples corresponding to 23 species of gilled mushrooms in the Agaricus and Lepiota Family. Each species is recognized as positively edible, categorically poisonous, or of indefinite edibility and not recommended. This latter class was integrated with the poisonous class.

3.1.Accuracy

Accuracy is calculated for ARM, OAA and AARMA in mushroom dataset. Table 4.1 shows the comparison of the accuracy of results for all the three approaches.

Table 4.1

Comparison of Accuracy in Mushroom Dataset

Approaches	Accuracy (%)
ARM	78.46
OAA	81.23
AARMA	94.68

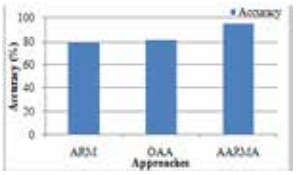

Figure 4.1: Comparison of Accuracy in Mushroom Dataset

From the figure 4.1, it can be observed that the accuracy of results using ARM and OAA is 78.46% and 81.23% respectively, and that of the proposed AARMA is 94.68% in mushroom dataset.

3.2.AUC Value

AUC value is obtained for ARM, OAA and AARMA in mushroom dataset. Figure 4.2 shows the AUC value comparison of results for all the three approaches.

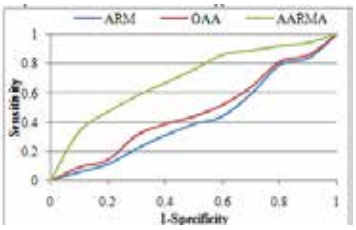

Figure 4.2: Comparison of AUC value in Mushroom DatasetFrom the figure 4.2, it can be observed that the AUC value of ARM is fairly adequate, the AUC value of OAA approach is good and that of the proposed AARMA approach is excellent in mushroom dataset.

3.3.Execution Time

Table 4.2 shows the execution time taken by the ARM, OAA and AARMA in mushroom dataset. It can be observed that the time required for execution using the proposed AARMA approach in mushroom dataset is 6.3 seconds, whereas more time is needed by the other two approaches for execution.

Table 4.2

Comparison of Execution Time in Mushroom Dataset

Approaches	Execution Time (Sec)
ARM	14.6
OAA	11.3
AARMA	6.3

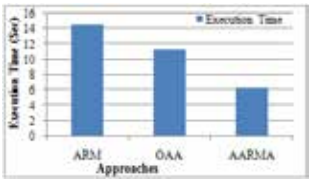

Figure 4.3: Comparison of Execution Time in Mushroom Dataset

From the figure 4.3, it is observed that the proposed AARMA approach takes very low execution time when compared with the ARM and OAA which takes 14.3 and 11.3 seconds respectively.

V.Conclusion

A novel effective approach for market basket analysis based on the Adaptive Association Rule Mining Algorithm has been introduced. There are various techniques available for selecting the best combination of itemsets in the super market for better sales. Most of the association rule mining algorithms suffer from the problems of too much execution time and generating too many association rules. Moreover, it is difficult to choose a proper minimum confidence and support for each item before the mining process because users' interests. If the minimum confidence and support for mining are set too high, enough rules for accurate combination will not be obtained. It increases the easiness of the customers which in turn increases the sales rate of the super market. This approach mainly analyzes the process of discovering association rules in this kind of big repositories. Therefore, this approach is very significant for effective market basket analysis and it helps the customers in purchasing their items with more comfort which in turn increases the sales rate of the markets.

REFERENCE

[1]Gurjit Kaur and Lolita Singh, "Data Mining: An Overview", International Journal of Computer Science and Telecommunications, Vol. 2, No. 2, Pp. 336-339, 2011. | [2]Yong Yin, Ikou Kaku, Jiafu Tang and JianMing Zhu, "Association Rules Mining in Inventory Database", Data Mining, Decision Engineering, Pp. 9-23, 2011. | [3]A. Trnka, "Market Basket Analysis with Data Mining Methods", International Conference on Networking and Information Technology (ICNIT), Pp. 446 - 450, 2010. | [4]Pei-ji Wang, Lin Shi, Jin-niu Bai and Yu-lin Zhao, "Mining Association Rules Based on Apriori Algorithm and Application", International Forum on Computer Science-Technology and Applications, Vol. 1, Pp. 141-143, 2009. | [5]Reena Hooda and Nasib S. Gill, "Applications and Issues of Data Mining", International Journal of Research in IT & Management, Vol. 2, No. 3, Pp. 11-17, 2012. | [6]Chris Rygielski, Jyun-Cheng Wang and David C. Yen, "Data mining techniques for customer relationship management", Technology in Society, Vol. 24, Pp. 483–502, 2002. | [7] Y.J. Wang, Qin Xin, F. Coenen, "A Novel Rule Weighting Approach in Classification Association Rule Mining," ICDM Workshops 2007, Seventh IEEE International Conference on Data Mining Workshops, Pp. 271 – 276, 2007. | [8]V. Bartik, "Association based Classification for Relational Data and its Use in Web Mining," CIDM '09, IEEE Symposium on Computational Intelligence and Data Mining, Pp. 252 – 258, 2009. | [9]Dong Liyan, Liu Zhaojun, Shi Mo, Yan Pengfei, Tian Zhuo and Li Zhen, "A novel method of mining frequent item sets," IEEE International Conference on Information and Automation (ICIA), pp. 173-178 , 2010. | [10]Lei Wen, "An efficient algorithm for mining frequent closed itemset," Fifth World Congress on Intelligent Control and Automation (WCICA 2004), Vol. 5, pp. 4296 – 4299, 2004. | [11]Nicolas Pasquier , Yves Bastide , Rafik Taouil and Lotfi Lakhal, "Discovering Frequent Closed Itemsets for Association Rules", in Proceedings of the ICDT International Conference on Database Theory (ICDT'1999), vol. 1540, pages 398-416, Jerusalem, Israel, 1999. | [12]G. Grahne and Jianfei Zhu, "Mining frequent itemsets from secondary memory," Fourth IEEE International Conference on Data Mining (ICDM '04), pp. 91 – 98, 2004. | [13]Abhijit Raorane and R.V. Kulkarni, "Data Mining Techniques: A Source for Consumer Behavior Analysis", International Journal of Database Management Systems (IJDMS), Vol. 3, No. 3, Pp. 45-56, 2011. |