Jeetesh Kumar Jain, Nirupama Tiwari, Manoj Ramaiya / International Journal of Engineering Research and Applications (IJERA) ISSN: 2248-9622 www.ijera.com
Vol. 3, Issue 1, January -February 2013, pp.2065-2069

A Survey: On Association Rule Mining

***Jeetesh Kumar Jain, **Nirupama Tiwari, ***Manoj Ramaiya**
*Research Scholar, Shri Ram College of Engineering & Management
,*Asst. Professor, Shri Ram College of Engineering & Management

ABSTRACT

Data mining becomes a vast area of research in past few years. Several researches have been made in the field of data mining. The Association Rule Mining (ARM) is also a vast area of research and also a data mining technique. In this paper a survey is done on the different methods of ARM. In this paper the Apriori algorithm is defined and advantages and disadvantages of Apriori algorithm are discussed. FP- Growth algorithm is also discussed and advantages and disadvantages of FP- Growth are also discussed. In Apriori frequent itemsets are generated and then pruning on these itemsets is applied. In FP-Growth a FP-Tree is generated. The disadvantage of FP- Growth is that FP-Tree may not fit in memory. In this paper we have survey various paper based on mining of positive and negative association rules.

.

Keywords : *ARM, frequent itemset, pruning, positive association rules, negative association rules.*

1. INTRODUCTION

Data mining is a process of extracting interesting knowledge or patterns from large databases. There are several techniques that have been used to discover such kind of knowledge, most of them resulting from machine learning and statistics. The greater part of these approaches focus on the discovery of accurate knowledge. Though this knowledge may be useless if it does not offer some kind of surprisingness to the end user. The tasks performed in the data mining depend on what sort of knowledge someone needs to mine.

Data mining techniques are the result of a long process of research and product development. The main types of task performed by DM techniques are Classification, Dependence Modeling, Clustering, Regression, Prediction and Association. Classification task search for the knowledge able to calculate the value of a previously defined goal attribute based on other attributes and is often represented by IF-THEN rules. We can say the Dependence modeling as a generalization of classification. The aim of dependence modeling is to discover rules able to calculate the goal attribute value, from the values of calculated attributes. Even there are more than one goal attribute in dependence modeling. The process of partitioning the item set in a set of significant sub-classes (called clusters) is known as Clustering.

1.1 Association Rule Mining

he notion of mining association rules are as follows [1]. In the data mining, the Association rule mining is introduced in [2] to detect hidden facts in large datasets and drawing inferences on how a subset of items influences the presence of another subset. Let S= {S1, S2, S3……. Sn} be a universe of Items and T= {T1, T2, T3…….....Tn} is a set of transactions. Then expression X => Y is an association rule where X and Y are itemsets and $X \cap Y=\Phi$. Here X and Y are called antecedent and consequent of the rule respectively. This rule holds support and confidence, support is a set of transactions in set T that contain both X and Y and confidence is percentage of transactions in T containing X that also contain Y. An association rule is strong if it satisfies user-set minimum support (minsup) and minimum confidence (minconf) such as support $\geq$ minsup and confidence $\geq$ minconf. An association rule is frequent if its support is such that support $\geq$ minsup. There are two types of association rules- positive association rules and negative association rules. The forms of rules X=>¬Y, ¬X=>Y and ¬X=>¬Y are called negative association rules (NARs) [3]. In the previous researches we have seen that NARs can be discovered from both frequent and infrequent item sets.

1.2 Notation and concept of ARM

As it is discussed above that S is an universe of Items and T is a set of transactions. It is assumed to simplify a problem that every item x is purchased once in any given transaction T. generally each transaction T is assigned a number for ex.- Tid. Now it is assumed that X is an itemset then a transaction t is assumed to X iff $X \subseteq T$. hence it is clear that an association rule is an implication of the form X=>Y where X and Y are subsets of S.

1.2.1 Support

Support is a fraction of transactions that contain an itemset. Frequencies of occurring patterns are indicated by support. The probability of a randomly chosen transaction T that contain both itemsets X and Y is known as support. Mathematically it is represented as-

$$P(X, Y) = \frac{\text{No.of transactions containing both X and Y}}{\text{Total No of transactions}}$$

1.2.2 Confidence

It measures how often items in Y appear in transactions that contain X. Strength of implication in the rule is denoted by confidence. Confidence is the probability of purchasing an itemset Y in a randomly chosen transaction T depend on the purchasing of an itemset X. mathematically it is represented as-

$$P(Y/X) = \frac{\text{No of transactions containing both } X \text{ and } Y}{\text{No of transactions containing } X}$$

1.3 Application of Association Rule Mining

Different application areas of association rule mining are described below

1.3.1 Market Basket Analysis

A broadly-used example of association rule mining is market basket analysis. In market basket databases consist of a large no. of records and in each record all items bought by a customer on a single purchase transaction are listed. Managers would be paying attention to know that which groups of items are constantly purchased together. This data is used by them to adjust store layouts (placing items optimally with respect to each other), to cross-sell, to promotions, to catalog design and to identify customer segments based on buying patterns [5]. For example, suppose a shop database has 200,000 point-of-sale transactions, out of which 4,0000 include both items A and B and 1600 of these include item C, the association rule "If A and B are purchased then C is purchased on the same trip" has a support of 1600 transactions (alternatively 0.8% = 1600/200,000) and a confidence of 4% (=1600/4,0000).

The probability of a randomly selected transaction from the database will contain all items in the antecedent and the consequent is known as support, whereas the conditional probability of a randomly selected transaction will include all the items in the consequent given that the transaction includes all the items in the antecedent is known as confidence.

Now a day's products are coming with bar codes. A large amount of sales data is produced by the software supporting these barcode based purchasing/ordering system which is typically captured in "baskets". Commercial organizations are interested in discovering "association rules" that identify patterns of purchases, such that the presence of one item in a basket will imply the presence of one or more additional items. This "market basket analysis" result can be used to suggest combinations of products for special promotions or sales.

1.3.2 Medical diagnosis

Association rules can be used in medical analysis for assisting physicians to cure patients. The common problem of the induction of reliable analytic rules is hard as theoretically no induction process by itself can guarantee the accuracy of induced hypotheses [5].

Basically diagnosis is not an easy process because of unreliable diagnosis tests and the presence of noise in training examples. This may result in hypotheses with insufficient prediction correctness which is too unreliable for critical medical applications [6].

A technique has been proposed by Serban [5] based on relational association rules and supervised learning methods, which helps to identify the probability of illness in a certain disease. This interface can be simply extended by adding new symptoms types for the certain disease, and by defining new relations between these symptoms.

1.3.3 Protein Sequences

Proteins are essential constituents of cellular machinery of any organism. DNA technologies have provided tools for the quick determination of DNA sequences and, by inference, the amino acid sequences of proteins from structural genes [7].

Proteins are sequences made up of 20 types of amino acids. Each protein has a unique 3-dimensional structure, which depends on amino-acid sequence. A minor change in sequence of protein may change the functioning of protein. The heavy dependence of protein functioning on its amino acid sequence has been a subject of great anxiety.

Lot of research has gone into understanding the composition and nature of proteins; still many things remain to be understood satisfactorily. Now it is usually believed that amino acid sequences of proteins are not random.

Nitin Gupta, Nitin Mangal, Kamal Tiwari, and Pabitra Mitra [8] have deciphered the nature of associations between different amino acids that are present in a protein. Such association rules are advantageous for enhancing our understanding of protein composition and hold the potential to give clues regarding the global interactions amongst some particular sets of amino acids occurring in proteins. Knowledge of these association rules or constraints is highly desirable for synthesis of artificial proteins

1.3.4 Census data

Censuses make a huge variety of general statistical information on society available to both researchers and the general community [9]. The information related to population and economic census can be forecasted in planning public services(education, health, transport, funds) as well as in public business(for setup new factories, shopping malls or banks and even marketing particular products).

The application of data mining techniques to census data and more generally to official data has great potential in supporting good community policy and in underpinning the effective functioning of a democratic society [10]. On the other hand, it is not undemanding and requires exigent methodological study, which is still in the preliminary stages.

1.3.5 CRM of credit card business

Customer Relationship Management (CRM), through which, banks expect to identify the preference of

different customer groups, products and services adapted to their liking to enhance the cohesion between credit card customers and the bank, has become a topic of great interest [11]. Shaw [4] mainly describes how to incorporate data mining into the framework of marketing knowledge management. The collective application of association rule techniques reinforces the knowledge management process and allows marketing personnel to know their customers well to provide better quality services. Song [12] proposed a method to illustrate change of customer behavior at different time snapshots from customer profiles and sales data.
The idea behind this is to discover changes from two datasets and generate rules from each dataset to carry out rule matching.

2 LITERATURE SURVEY

2.1 By Idheba Mohamad et. al.[13] ***"Mining Positive and Negative Association Rules from Interesting Frequent and Infrequent Itemsets"*** in this title authors describe that Association rule mining is one of the most important tasks in data mining. The basic idea of association rules is to mine the interesting (positive) frequent patterns from a transaction database. Though, mining the negative patterns has also attracted the attention of researchers in this area. Aim of this survey is to develop a new model for mining interesting negative and positive association rules out of a transactional data set. The model proposed in this paper is integration between two algorithms, the Positive Negative Association Rule (PNAR) algorithm and the Interesting Multiple Level Minimum Supports (IMLMS) algorithm, to propose a new approach (PNAR_IMLMS) for mining both negative and positive association rules from the interesting frequent and infrequent itemsets mined by the IMLMS model. The results show that the PNAR_IMLMS model provides significantly better results than the previous model.
As infrequent itemsets become more significant for mining the negative association rules that play an important role in decision making, this study proposes a new algorithm for efficiently mining positive and negative association rules in a transaction database. The IMLMS model adopted an effective pruning method to prune uninteresting itemsets. An interesting measure VARCC is applied that avoids generating uninteresting rules that may be discovered when mining positive and negative association rules.
2.2 By Xushan Peng, Yanyan Wu [14] "Research and Application of Algorithm for Mining Positive and Negative Association Rules" in this title an effective algorithm is presented to discover positive and negative association rules. Particularly, it analyzes the definition of the positive and negative association rules, methods of support and confidence level in affairs. It also describes the conflicted rules problems which are in the positive and negative association rules mining. The solutions of the conflicted rules problems with the related rules is obtain by it.
To the algorithm of the positive and negative association rules, this paper use linked list to implement the algorithm of the positive and negative association rules. The ways of the binary number present whether the transactions include the elements of the items of the collection. To the algorithm of the positive and negative association rules, there are more research to do.(l)How to optimize the searching space?(2)To the mine the rules, there are further to do the research and use relativity.(3)In the positive and negative association rules, Apriori algorithm displace and improvement and so on.

3 VARIOUS METHODS FOR ARM

3.1 Apriori Algorithm

Apriori is a seminal algorithm proposed by R. Agrawal and R. Srikant[1] in 1994 for mining frequent itemsets for Boolean association rules. The name of the algorithm is based on the fact that the algorithm uses *prior knowledge* of frequent itemset properties, as we shall see following. Apriori employs an iterative approach known as a *level-wise* search, where *k*-itemsets are usedtoexplore (*k*+1)-itemsets. First, the setof frequent 1-itemsets is found by scanning the database to accumulate the count for each item, and collecting those items that satisfy minimum support. The resulting set is denoted *L*1.Next, *L*1 is used to find *L*2, the set of frequent 2-itemsets, which is used to find *L*3, and so on, until no more frequent *k*-itemsets can be found. The finding of each *Lk* requires one full scan of the database. To improve the efficiency of the level-wise generation of frequent itemsets, an important property called the Apriori property, presented below, is used to reduce the search space.We will first describe this property, and then show an example illustrating its use.
Apriori property: *All nonempty subsets of a frequent itemset must also be frequent.*
The Apriori property is based on the following observation. By definition, if an itemset *I* does not satisfy the minimum support threshold, *min sup*, then *I* is not frequent; that is, *P*(*I*) < *min sup*. If an item *A* is added to the itemset *I*, then the resulting itemset (i.e., *IUA*) cannot occur more frequently than *I*. Therefore, *IUA* is not frequent either; that is, *P* (*I* UA) < *min sup.*
This property belongs to a special category of properties called antimonotone in the sense that *if a set cannot pass a test, all of its supersets will fail the same test as well*. It is called *antimonotone* because the property is monotonic in the context of failing a test.
"How is the Apriori property used in the algorithm?" To understand this, let us look at how *Lk*-1 is used to find L_k for $k \geq 2$. A two-step process is followed, consisting of join and prune actions.

1. The join step: To find *Lk*, a set of candidate *k*-itemsets is generated by joining *Lk*-1 with itself. This set of candidates is denoted *Ck*. Let *l*1 and *l*2 be itemsets in *Lk*-1. The notation *li*[*j*] refers to the *j*th item in *li* (e.g., *l*1[*k*-2] refers to the second to the last item in *l*1). By convention, Apriori assumes that items within a transaction or itemset are sorted in lexicographic order. For the (*k*-1)-itemset, *li*, this means that the items are sorted such that *li*[1] < *li*[2] < ……< *li*[*k*-1]. The join, L_{k-1} on L_{k-1}, is performed, where members of L_{k-1} are joinable if their first (*k*-2) items are in common. That is, members *l*1 and *l*2 of L_{k-1} are joined if (*l*1[1] = *l*2[1]) ^ (*l*1[2] = *l*2[2]) ^…….^(*l*1[*k*-2] = *l*2[*k*-2]) ^(*l*1[*k*-1] < *l*2[*k*-1]). The condition *l*1[*k*-1] < *l*2[*k*-1] simply ensures that no duplicates are generated. The resulting itemset formed by joining *l*1 and *l*2 is *l*1[1], *l*1[2], …… , *l*1[*k*-2], *l*1[*k*-1], *l*2[*k*-1].
2. The prune step: *Ck* is a superset of *Lk*, that is, its members may or may not be frequent, but all of the frequent *k*-itemsets are included in *Ck*. A scan of the database to determine the count of each candidate in *Ck* would result in the determination of *Lk* (i.e., all candidates having a count no less than the minimum support count are frequent by definition, and therefore belong to *Lk*). *Ck*, however, can be huge, and so this could involve heavy computation. To reduce the size of *Ck*, the Apriori property is used as follows. Any (*k*-1)-itemset that is not frequent cannot be a subset of a frequent *k*-itemset. Hence, if any (*k*-1)-subset of a candidate *k*-itemset is not in *Lk*-1, then the candidate cannot be frequent either and so can be removed from*Ck*. This subset testing can be done quickly by maintaining a hash tree of all frequent itemsets.

2.1.1 Advantages of Apriori Algorithm

1. Uses large item set property
2. Easily parallelized
3. Easy to implement
4. The Apriori algorithm implements level-wise search using frequent item property

2.1.2 Disadvantages of Apriori Algorithm

1. There is too much database scanning to calculate frequent item (reduce performance)
2. It assumes that transaction database is memory resident
3. Generation of candidate itemsets is expensive(in both space and time)
4. Support counting is expensive
   - Subset checking (computationally expensive)
   - Multiple Database scans (I/O)

3.2 FP-Growth Algorithm

The FP-growth method described in[1] transforms the problem of finding long frequent patterns to searching for shorter ones recursively and then concatenating the suffix. It uses the least frequent items as a suffix, offering good selectivity. The method substantially reduces the search costs.

When the database is large, it is sometimes unrealistic to construct a main memory based FP-tree. An interesting alternative is to first partition the database into a set of projected databases, and then construct an FP-tree and mine it in each projected database. Such a process can be recursively applied to any projected database if its FP-tree still cannot fit in main memory. A study on the performance of the FP-growth method shows that it is efficient and scalable for mining both long and short frequent patterns, and is about an order of magnitude faster than the Apriori algorithm. It is also faster than a Tree-Projection algorithm, which recursively projects a database into a tree of projected databases.

Algorithm: FP growth. Mine frequent itemsets using an FP-tree by pattern fragment growth.
Input: *D*, a transaction database;
min sup, the minimum support count threshold.
Output: The complete set of frequent patterns.
Method:
1. The FP-tree is constructed in the following steps:
(a) Scan the transaction database *D* once. Collect *F*, the set of frequent items, and their support counts. Sort *F* in support count descending order as *L*, the *list* of frequent items.
(b) Create the root of an FP-tree, and label it as "null." For each transaction *Trans* in *D* do the following.
Select and sort the frequent items in *Trans* according to the order of *L*. Let the sorted frequent item list in *Trans* be [*pjP*], where *p* is the first element and *P* is the remaining list. Call insert tree ([*pjP*], *T*), which is performed as follows. If *T* has a child *N* such that *N.item-name=p.item-name*, then increment *N*'s count by 1; else create a new node *N*, and let its count be 1, its parent link be linked to *T*, and its node-link to the nodes with the same *item-name* via the node-link structure. If *P* is nonempty, call insert tree(*P*, *N*) recursively.
2. The FP-tree is mined by calling FP growth(*FP tree*, *null*), which is implemented as follows.
procedure FP growth(*Tree*, a) (1) If *Tree* contains a single path *P* then
(2) For each combination (denoted as b) of the nodes in the

path *P*
(3) Generate pattern b[a with *support count* = *minimum support count of nodes in* b;
(4) Else for each *ai* in the header of *Tree {*
(5) Generate pattern b = *ai* [a with *support count* = *ai*: *support count*;

(6) Construct b's conditional pattern base and then b's conditional FP tree *Tree*b;
(7) if *Tree*b 6= /0 then
(8) call FP growth(*Tree* s, □); }

2.2.1 Advantages of FP- Growth Algorithm

1. Only 2 passes over data-set
2. "Compresses" data-set
3. No candidate generation
4. Much faster than Apriori

2.2.2 Disadvantages of FP- Growth Algorithm

1. FP-Tree may not fit in memory!!
2. FP-Tree is expensive to build

- Trade-off: takes time to build, but once it is built, frequent itemsets are read off easily.
- Time is wasted (especially if support threshold is high), as the only pruning that can be done is on single items.
- Support can only be calculated once the entire data-set is added to the FP-Tree.

4. CONCLUSION

This paper is a survey paper on association rule mining. In this paper a survey on association rule mining and methods of association rule mining is described. In this paper two classical mining algorithms- apriori algorithm and FP- Growth algorithm is described. The negative and positive association rules are also described in this paper. Our future work is to reduce the negative association rules from frequent itemsets

REFERENCES

[1] Han and M. Kamber, Data Mining: Concepts and Techniques, San Francisco: Morgan Kaufmann Publishers,2001.
[2] R. Agrawal, T. Imielinski, and A. Swami, "Mining association rules between sets of items in large databases," in Proceedings of the 1993 ACM SIGMOD International Conference on Management of Data, Washington, D.C., 1993, pp. 26–28.
[3] Wu, X., Zhang, C., Zhang, S.: Efficient Mining of both Positive and Negative Association Rules. ACM Transactions on Information Systems 22(3):381– 405(2004)
[4] M. J. Shaw, C. Subramaniam , G. W. Tan and M. E. Welge, "Knowledge management and data mining for marketing", Decision Support Systems, v.31 n.1, pages 127-137, 2001
[5] G. Serban, I. G. Czibula, and A. Campan, "A Programming Interface For Medical diagnosis Prediction", Studia Universitatis, "Babes-Bolyai", Informatica, LI(1), pages 21-30, 2006 Applications, 2001.
[6] D. Gamberger, N. Lavrac, and V. Jovanoski, "High confidence association rules for medical diagnosis", *In Proceedings of IDAMAP99*, pages 42-51.
[7] C. Branden and J. Tooze, "Introduction to Protein Structure", Garland Publishing inc, New York and London, 1991
[8] N. Gupta, N. Mangal, K. Tiwari and P. Mitra, "Mining Quantitative Association Rules in Protein Sequences", *In Proceedings of* Australasian Conference on Knowledge Discovery and Data Mining – AUSDM, 2006
[9] D. Malerba, F. Esposito and F.A. Lisi, "Mining spatial association rules in census data", *In Proceedings of Joint Conf. on "New Techniques and Technologies for* Statistcs and Exchange of Technology and Know-how", 2001
[10] G. Saporta, "Data mining and official statistics", *In Proceedings of Quinta Conferenza Nationale di Statistica*, ISTAT, Roma, 15 Nov. 2000.
[11] R. S. Chen, R. C. Wu and J. Y. Chen, "Data Mining Application in Customer Relationship Management Of Credit Card Business", *In Proceedings of 29th Annual International Computer Software and Applications Conference (COMPSAC'05)*, Volume 2, pages 39-40.
[12] H. S. Song, J. K. Kim and S. H. Kim, "Mining the change of customer behavior in an internet shopping mall", Expert Systems with Applications, 2001
[13] Idheba Mohamad Ali O. Swesi, Azuraliza Abu Bakar, Anis Suhailis Abdul Kadir, "Mining Positive and Negative Association Rules from Interesting Frequent and Infrequent Itemsets " in proceeding of 9th International Conference on Fuzzy Systems and Knowledge Discovery, 2012.
[14] Xushan Peng, Yanyan Wu"Research and Application of Algorithm for Mining Positive and Negative Association Rules" in proceeding of International Conference on Electronic & Mechanical Engineering and Information Technology 2011.