International Journal of Innovative Research in Computer Science & Technology (IJIRCST)
ISSN: 2347-5552, Volume-2, Issue-1, January 2014

Multilevel Association Rule

Monali Deshmukh, Madhu Nashipudimath

Abstract— **The problems of developing models and algorithms for multilevel association mining pose for new challenges for mathematics and computer science. These problems become more challenging when some form of uncertainty in data or relationships in data exists. In this , we present a partition technique for the multilevel association rule mining problem. Taking out association rules at multiple levels helps in discovering more specific and applicable knowledge. In multilevel association rule there are two methods Boolean matrix and Hash based method. A Boolean Matrix based approach has been employed to discover frequent itemsets, the item forming a rule come from different levels. It adopts Boolean relational calculus to discover maximum frequent itemsets at lower level. When using this algorithm first time, it scans the database once and will generate the association rules. Apriori property is used in prune the item sets. It is not necessary to scan the database again; it uses Boolean logical operation to generate the multilevel association rules and also use top-down progressive deepening method. Hash-based algorithm for the candidate set generation. Explicitly, the number of candidate 2-itemsets generated by the proposed algorithm is, in orders of magnitude, smaller than that by previous methods, thus resolving the performance bottleneck. Note that the generation of smaller candidate sets enables us to effectively trim the transaction database size at a much earlier stage of the iterations, thereby reducing the computational cost for later iterations significantly. Extensive simulation study is conducted to evaluate performance of the proposed algorithm.**

Index Terms— **Association rules, Boolean matrix,data mining ,Hash based method, itemsets, multilevel rules.**

I. INTRODUCTION

Data mining, the extraction of hidden predictive information from large databases, is a powerful new technology with great potential to help companies focus on the most important information in their data warehouses. Data mining tools predict future trends and behaviors, allowing businesses to make proactive, knowledge-driven decisions. The automated, prospective analyses offered by data mining move beyond the analyses of past events provided by retrospective tools typical of decision support systems. Data mining tools can answer business questions that traditionally were too time consuming to resolve. They scour databases for hidden patterns, finding predictive information that experts may miss because it lies outside their expectations.Data mining techniques are the result of a long process of research and product development. This evolution began when business data was first stored on computers, continued with improvements in data access, and more recently, generated technologies that allow users to navigate through their data in real time. Data mining takes this evolutionary process beyond retrospective data access and navigation to prospective and proactive information delivery.

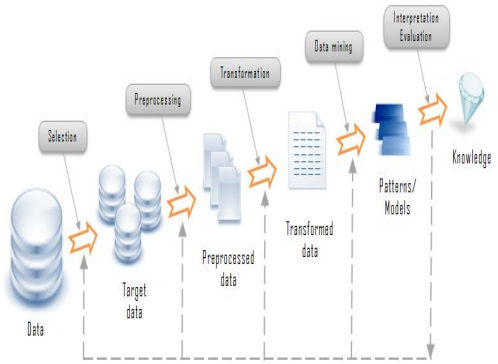

Fig 1: Knowledge Discovery Process

Data mining is a process of extraction of useful information and patterns from huge data. It is also called as knowledge discovery process, knowledge mining from data, knowledge extraction or data /pattern analysis. Data mining is a logical process that is used to search through large amount of data in order to find useful data. The goal of this technique is to find patterns that were previously unknown. Once these patterns are found they can further be used to make certain decisions for development of their businesses. The purpose of DM is to analyze and understand past trends and predict future trends. By predicting future trends, business organizations can better position their products and services for financial gain. Nonprofit organizations have also achieved significant benefits from data mining, such as in the area of scientific progress. The concept of data mining is simple yet powerful. The simplicity of the concept is deceiving, however. Traditional methods of analyzing data, involving query-and-report approaches, cannot handle tasks of such magnitude and complexity.

II. Data Mining Algorithms and Techniques

Various algorithms and techniques like Classification, Clustering, Regression, Artificial Intelligence, Neural Networks, Association Rules, Decision Trees, Genetic Algorithm, Nearest Neighbour method etc., are used for knowledge discovery from databases.

A. Classification

Classification is the most commonly applied data mining technique, which employs a set of pre-classified examples

Manuscript received January 23, 2014.
Monali Deshmukh, Computer Department, Mumbai/SCOE/ Saraswati college of Engineering , Kharghar Navimumbai, India, 7738376828., (e-mail: mona_deshmukh@rediffmail.com).
Madhu Nashipudimath, Information Technology, Mumbai/ PIIT/ Pillai's Institute of Technology, Panvel NAviMumbai, India, 9702476406.,

to develop a model that can classify the population of records at large. Fraud detection and credit risk applications are particularly well suited to this type of analysis. This approach frequently employs decision tree or neural network-based classification algorithms. The data classification process involves learning and classification. In Learning the training data are analyzed by classification algorithm. In classification test data are used to estimate the accuracy of the classification rules. If the accuracy is acceptable the rules can be applied to the new data tuples. For a fraud detection application, this would include complete records of both fraudulent and valid activities determined on a record-by-record basis. The classifier-training algorithm uses these pre-classified examples to determine the set of parameters required for proper discrimination. The algorithm then encodes these parameters into a model called a classifier.

B. Clustering

Clustering can be said as identification of similar classes of objects. By using clustering techniques we can further identify dense and sparse regions in object space and can discover overall distribution pattern and correlations among data attributes. Classification approach can also be used for effective means of distinguishing groups or classes of object but it becomes costly so clustering can be used as preprocessing approach for attribute subset selection and classification. For example, to form group of customers based on purchasing patterns, to categories genes with similar functionality.

C. Predication

Regression technique can be adapted for predication. Regression analysis can be used to model the relationship between one or more independent variables and dependent variables. In data mining independent variables are attributes already known and response variables are what we want to predict. Unfortunately, many real-world problems are not simply prediction. For instance, sales volumes, stock prices, and product failure rates are all very difficult to predict because they may depend on complex interactions of multiple predictor variables. Therefore, more complex techniques (e.g., logistic regression, decision trees, or neural nets) may be necessary to forecast future values. The same model types can often be used for both regression and classification. For example, the CART (Classification and Regression Trees) decision tree algorithm can be used to build both classification trees (to classify categorical response variables) and regression trees (to forecast continuous response variables). Neural networks too can create both classification and regression models.

D. Association rule

Association and correlation is usually to find frequent item set findings among large data sets. This type of finding helps businesses to make certain decisions, such as catalogue design, cross marketing and customer shopping behavior analysis. Association Rule algorithms need to be able to generate rules with confidence values less than one. However the number of possible Association Rules for a given dataset is generally very large and a high proportion of the rules are usually of little (if any) value.

E. Neural networks

Neural network is a set of connected input/output units and each connection has a weight present with it. During the learning phase, network learns by adjusting weights so as to be able to predict the correct class labels of the input tuples. Neural networks have the remarkable ability to derive meaning from complicated or imprecise data and can be used to extract patterns and detect trends that are too complex to be noticed by either humans or other computer techniques. These are well suited for continuous valued inputs and outputs. For example handwritten character reorganization, for training a computer to pronounce English text and many real world business problems and have already been successfully applied in many industries. Neural networks are best at identifying patterns or trends in data and well suited for prediction or forecasting needs.

III. Association Rule

Finding frequent patterns, associations, n correlations, or causal structures among sets of items or objects in transactional databases, relational databases, and other information repositories. market basket data analysis, cross-marketing, catalog design, loss-leader analysis, etc. Association rule mining can be broadly classified into categories: Boolean or quantitative associations, Single dimension or multidimensional associations, Single level or multilevel associations

A. Multi level Association Rule Mining

We can mine multilevel association rules efficiently using concept hierarchies, which defines a sequence of mappings from a set of low level concepts to higher-level, more general concepts . Data can be generalized by replacing low-level concepts within the data by their higher-level concepts or ancestors from a concept hierarchy. In a concept hierarchy, which is represented as a tree with the root as D i.e., Task-relevant data. The popular area of application for multi level association is market basket analysis , which studies the buying habits of customers by searching for sets of items that are frequently, purchased together which was presented in terms of concept hierarchy shown below. Each node indicates an item or item set that has been examined. There are various approaches for finding frequent item sets at any level of abstraction. Some of the methods which are in use are ‘using uniform minimum support for all levels’, using reduced minimum support at low levels, level-by-level independent. Multi-level databases use hierarchy-information encoded transaction table instead of the original transaction table . This is useful when we are interested in only a portion of the transaction database such as food, instead of all the items. This way we can first collect the relevant set of data and then work repeatedly on the task relevant set. Thus in the transaction table each item is encoded as a sequence of digits.

Multilevel Algorithm Based On Boolean Matrix: We propose a new multilevel association algorithm. The section is organized as follows: the correlative definition and proposition, an introduction to the MLBM algorithm details, and description of a sample execution of the MLBM algorithm.

The algorithm consists of following steps:

Step-1:Encode taxonomy using a sequence of numbers and the symbol ‘‘*’’, with the ***l*** th number representing the branch number of a certain item at levels.
Step-2:Set H = 1, where H is used to store the level number being processed whereas H ={1, 2, 3}
(as we consider up to 3-levels of hierarchies).
Step-3:Transforming the transaction database into the Boolean matrix.
Step-4:Set user defines minimum support on current level.
Step-5 Generating the set of frequent 1-itemset L1at level1.Pruning the Boolean matrix
Step-6:Perform AND operations to generate 2-itemsetsand 3- itemset at level 1.
Step-7: Generate H +1; (Increment H value by 1; i.e., H =2) itemset from Lk and go to step-4 (for repeating the whole processing for next level).

Multilevel Algorithm On Hash Based Method: In this Method, an algorithm DHP(standing for direct hashing and pruning) for efficient large itemset generation. Specifically, DHP proposed has two major features: one is efficient generation for large itemsets and the other is effective reduction on transaction database size. As will be seen later, by utilizing a hash technique, DHP is very efficient for the generation of candidate large itemsets, in particular for the large 2-itemsets, where the number of candidate large itemsets generated by DHP is, in orders of magnitude, smaller than that by previous methods, thus greatly improving the performance bottleneck of the whole process. In addition, DHP employs effective pruning techniques to progressively reduce the transaction database size. During the early iterations, tracking the candidate k-itemsets in each transaction is ineffective since the cardinality of such k-itemsets is very large, Note that the generation of smaller candidate sets by DHP enables us to effectively trim the transaction database at a much earlier stage of the iterations, i.e., right after the generation of large 2-itemsets, thereby reducing the computational cost for later iterations significantly. It will be seen that by exploiting some features of association rules, not only the number of transactions, but also the number of items in each transaction can be substantially reduced. n Extensive experiments are conducted to evaluate the performance of DHP. As shown by our experiments, with a slightly higher cost in the first iteration due to the generation of a hash table, DHP incurs significantly smaller execution times than Apriori in later iterations ,not only in the second iteration when a hash table is used by DHP to facilitate the generation of candidate2-itemsets, but also in later iterations when the same procedure for large item set generation is employed by both algorithms, showing the advantage of effective database trimming by DHP. Sensitivity analysis for various parameters is conducted. It should be noted that in , the hybrid algorithm has the option of switching from Apriori to another algorithm Apriori TID after early passes for better performance. For ease of presentation f this paper, such an option is not adopted here. Nevertheless, the benefit of Apriori TID in later passes is complementary to the focus of DHP on initial passes. The problem of mining association rules is composed of the following two steps: Discover the large item sets, i.e., all sets of item sets that have transaction support above a predetermined minimum support s. second is Use the large itemsets to generate the association rules for the database. The overall performance of mining association rules is infact determined by the first step, After the large itemsets are identified, the corresponding association rule scan be derived in a straightforward manner. In this paper, we shall develop an algorithm to deal with the first step, i.e., discovering large itemsets from the transaction database. Readers interested in more details forthe second step are referred to . As a preliminary, we shall describe the method used in the prior work Apriori For discovering the large itemsets from a transaction database given in Figure, Note that a comprehensive study on various algorithms to determine large itemsets is presented , where the Apriori algorithm is shown to provide the best performance during the initial iterations. Hence, Apriori is used as the base algorithm to compare with DHP. In Apriori, in each iteration (or each pass) it constructs a candidate set of large itemsets, counts the number of occurrences of each candidate itemset, and then determine large itemsets based on a pre-determined minimum support. In the first iteration, Apriori simply scans all the transactions to count the number of occurrences for each item. The set of candidate l-iternsets, Cl, obtained is shown in Figure. Assuming that the minimum transaction support required is 2, the set of large l-iternsets, L1, composed of candidate l-iternsets with the minimum support required, can then be determined.

TID	ITEMS
100	A C D
200	B C E
300	A B C E
400	B E

TABLE 1: An example transaction database for data

To discover the set of large 2-itemsets, in view of the fact that any subset of a large itemset must also have minimum support, Apriori uses L1 * L1 to generate a candidate set of itemsets C2 using the apriori candidate generation, where * is an operation for concatenation. C2 consists of 2-itemsets. Note that when IL1 I is large, becomes an extremely large number. Next, the four transactions in D are scanned and the support of each candidate itemset in C’2 is counted. The middle table of the second row in Figure represents the result from such counting in G’2. A hash tree is usually used for a fast counting process 3. The set of large 2-itemsets, L2, is therefore determined based on the support of each candidate 2-itemset in C2.

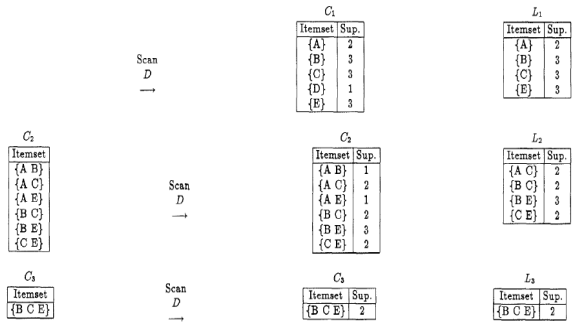

Fig 2: Generation of candidate itemsets and large itemsets

the 2-itemset {Cll},which consists of their second items, constitutes a large2-itemset or not. Since {C'E} is a large itemset by itself, we know that all the subsets of {BCE} are large and then {BCE} becomes a candidate 3-itemset. There isno other candidate 3-itemset from L2. Apriori then scans all the transactions and discovers the large 3-itemsets L3 in Figure 2. Since there is no candidate4-itemset to be constituted from L3, Apriori ends the process of discovering large itemsets. As it can be seen above, it is important to generate as small a candidate set of itemsets as possible because the support of each itemset in C; has to be counted during the scan of the entire database. As we shall see later, by exploiting this feature, algorithm DHP proposed is able to generate large itemset efficiently.

IV. ANALYSIS

These results were generated quite quickly compare to Apriori algorithm as size of candidate itemsets was reduced by hash method. However, there is still some weakness due to the collision in hash table: the processes to generate candidate item sets may be omit to filter out itemsets when they laid together in a bucket to make the entry value is greater than mins Now, the only problem we must care is the hash function of the algorithm. Hash based proposed two types: direct hash and partial hash. The direct one is simple and fast, but needed a mount of memories; the other is a bit more complex but could fit in smaller memory. Direct hash method, really simple, will associates each bucket of the hash table with unique code word – these code words are joined from two large itemsets in the prior pass, and the hash table will have buckets for all of candidates. So we can choose the function: these algorithms use hash-based approach to reduce the implement size of database and quickly process. And based on these ideas, establish a small system to find association rules in transaction database. The confidence of association rules has a specific meaning: when the antecedent of the rule is satisfied, the consequent of the rule will have *c%* (here c refers to the confidence ofthe rule) possibility of being satisfied A Boolean Matrix based approach has been employed to discover frequent itemsets, the item forming a rule come from different levels. It adopts Boolean relational calculus to discover maximum frequent itemsets at lower level.

V. CONCLUSION

The main features of this algorithm are that it only scans the transaction database once, it does not produce itemsets, and it adopts the Boolean vector "relational calculus" to discover frequent itemset. Hash-based algorithm and is especially effective for the generation of candidate n set for large 2-itemsets, where the number of candidate 2-itemsets generated is, in orders of magnitude, smaller than that by previous methods, thus resolving the performance bottleneck.

REFERENCES

[1] R. Agrawal, T. Imielinski, and A. Swami, "Mining association rules between sets of items in large databases,"Proceedings of the ACM SIGMOD Conference on Management of data, pp. 207-216, 1993.
[2]R. Agrawal and R. Srikant," Fast algorithms for mining association rules," In proceeding of the VLDBConference,1994.
[3] H. Mannila, H. Toivonen, and A, Verkamo. "Efficient algorithm for discovering association rules," AAA1 Workshop on Knowledge Discovery in Databases.
[4]Jiawei Han, Micheline Kamber, "Data Mining Concepts and Techniques," Higher Education Press 2001.
[5] Hunbing Liu and Baishenwang, "An association Rule Mining Algorithm Based On a Boolean Matrix," DataScience Journal, Vol-6, Supplement 9, S559-563, September 2007.
[6] R.S Thakur, R.C. Jain, K.R.Pardasani, "Fast Algorithm for Mining Multilevel Association Rule Mining," Journal ofComputer Science, Vol-1, pp. 76-81, 2007 .
[7] Ha and Y. Fu, "Mining Multiple-Level Association Rulesin Large Databases," IEEE TKDE. Vol-1, pp. 798-805, 1999 .
[8] R. Agrawal, C. Faloutsos, and A. Swami. EfficientSimilarity Search in Sequence Databases, Proceedingsof the 4th Intl, conf, on Foundations of DataOrganization and Algorithms, October, 1993.
[9] R. Agrawal, S. Ghosh, T. Imielinski, B. IYer, andA. Swami. An Interval Classifier for Database MiningApplications. Proceedings of the 18th InternationalConference on Very Large Data Bases, pages560-573, August 1992.
[10] AnjnaPandey and K. R. Pardasani,"Rough Set Model forDiscovering Hybrid Dimensional Association Rules,"International Journal of Computer Science and NetworkSecurity, Vol -9, no.6, pp.159-164,2009.

Prof. Monali Deshmukh, Bachelor of Engineering, Student, PIIT, Panvel Navi Mumbai. Area of Interest: Data Mining.

Prof. Madhu M Nashipudimath has Bachelors and Master's Degree in the field of Computer science and Engineering. She has keen interest in the area of data mining, software Engg, fuzzy systems, Neural networks, and has published several papers in National and International conferences and journals. Her interest is also extended to data storage and information retrieval. Prof. Madhu has more than fifteen years of experience in teaching undergraduate as well as postgraduate students. She has attended several workshops and faculty development programs. She has 20 Research Papers to her credit.
.