A General Survey on Multidimensional And Quantitative Association Rule Mining Algorithms

R. Sridevi[1], Dr. E. Ramaraj[2]

[1]*Asst. Professor, Dept. of Information Technology Dr.Umayal Ramanathan College for Women, Karaikudi*
[2]*Professor, Dept. of Computer Science & Engineering, Alagappa University, Karaikudi.*

Abstract

Data mining is one of the significant topics of research in recent years. Association rule is a method for discovering interesting relations between variables in large databases. Support and Confidence are the two basic parameters used to study the threshold values for each database. In this paper, an overall survey of the algorithms implementing the multidimensional and quantitative data is presented. It also illustrates the various approaches in association rule. The concept of Bit Mask Search Algorithm can be identified and suggested to effectively generate the frequent Item sets from large database. Moreover, the algorithm is best suited for multidimensional and quantitative datasets. It describes the essential role of multidimensional and Quantitative data in Association rule mining algorithm.

Keywords: Data Mining, Association rules, Multidimensional rule, Quantitative rule.

I. INTRODUCTION

Data mining is a non trivial extraction of implicit, previously unknown and potentially useful information from data. Data mining technology provides a user- oriented approach to novel and hidden patterns in the data [3]. Data mining refers to extracting or “mining” knowledge from large amounts of data.

Discovery of hidden patterns and relationships often goes unexploited. Advanced data mining techniques can help to overcome this situation. Data mining technology provides a user-oriented approach to novel and hidden patterns in the data. A wide variety of areas including marketing, customer relationship management, engineering, medicine, crime analysis, expert prediction, Web mining, and mobile computing, besides others utilize Data mining.

The essential process of Knowledge Discovery is the conversion of data into knowledge in order to aid in decision making, referred to as data mining. Knowledge Discovery process consists of an iterative sequence of data cleaning, data integration, data selection, data mining, pattern recognition and knowledge presentation.

The difficulties posed by prediction problems have resulted in a variety of problem-solving techniques. For example, data mining methods comprise artificial neural networks and decision trees, and statistical techniques including linear regression and stepwise polynomial regression [8]. Data mining technology provides a user oriented approach to novel and hidden patterns in the data. Data Mining have two flavors- directed and undirected. Directed data mining attempts to explain or categorize some particular target field such as income or response. Undirected data mining attempts to find patterns of similarities among groups of records without the use of a particular target field or collection of predefined classes. Data mining is largely concerned with building models. A model is simply an algorithm or set of levels that connects a collection of inputs to a particular target or outcome.

Data analysis is a process in which raw data is prepared and structured so that valuable information can be extracted from it. The process of organizing and thinking about data is way to accept what the data does and does not contain. There are a variety of ways in which public can approach data analysis. In fact, most data mining algorithms require large amounts of data in order to build and train the models that will then be used to perform classification, prediction, estimation, or other data mining tasks.

II. DATA MINING TECHNIQUES

Basically, data mining is about processing data and identifying the patterns in that information so that it is easy to decide or judge. Data mining principles have been around for many years, but, with the advent of large amount of data, it is even more prevalent. Large data caused an explosion in the use of more extensive data mining techniques, partially because the size of the information is much larger and because the information tends to be more varied and extensive in its nature and content. With large data sets, it is no longer enough to get relatively simple and straightforward statistics out of the system.

2.1 ASSOCIATION RULES TYPES

Association rule mining, one of the most important and well researched techniques of data mining, was first introduced in [2]. Association rules are widely used in various areas such as telecommunication networks, market and risk management, inventory control etc. Association rule mining is to find out association rules that satisfy the predefined minimum support and confidence from a given database. The problem is usually decomposed into two sub-problems. One is to find those item sets whose occurrences exceed a predefined threshold in the database; those item sets are called frequent or large item sets. The second problem is to generate association rules from those large item sets with the constraints of minimal confidence [4].The various types of association rules are clearly described in the following section.

Each item can be seen as a Boolean variable presenting the absence or presence of that item in the transaction or in the row. Binary association rule in databases of items can be organized and clustered [13]. By calculating a Minimum Spanning Tree on the graph of associations, the most significant associations can be discovered and easily visualized, allowing easy understanding of existing relations.

Quantitative Association rule

In practice, database contains much quantitative data and is not limited to categorical items only. Quantitative association rules (QARs) is an important tool in discovering association relationships among sets of attributes in business and scientific domains [13]. In QAR, attributes are not limited to being Boolean but can either be quantitative which numeric values are (eg. Age, salary) or categorical, which are enumerations (eg. Gender, name).

Fuzzy Association Rule

The new fuzzy association rule mining approach [18] emerged out of the necessity to mine quantitative data frequently present in databases efficiently. When dividing an attribute in the data into sets covering certain ranges of values, the users are confronted with the sharp boundary problem. Elements near the boundaries of a crisp set will either be ignored or overemphasized.

Multilevel Association Rule

Multilevel association rules are used to find the frequently occurring patterns and reasonably strong rule implications. A user or an expert may specify two thresholds: minimum support and minimum confidence. The multilevel association rules can be professionally mined using concept hierarchies, which defines a series of mappings from a set of low level concepts to higher-level with more general concepts[12]. On lowest levels, it strength will be that no rules may match the constraints. At highest levels, rules can be tremendously universal. Usually, a top down proceed is used where the support threshold varies from level to level.

Hegland [7] reviews the most well known algorithm for producing association rules such as Apriori and discuss variants for distributed data, inclusion of constraints and data taxonomies. The review ends with an outlook on tools which have the potential to deal with long item sets and considerably reduce the amount of (uninteresting) item sets returned.

2.2 ASSOCIATION RULE MINING

Boolean Association Rule

In this section association rule mining problem is discussed in detail. Different issues in Association Rule Mining (ARM) will be elaborated together with classic algorithms which are presented in this paper.

Basic Concepts & Basic Association Rules Algorithms

The formal statement of association rule mining problem was stated by Agrawal in [1]. Let I=I1, I2, … , Im be a set of m distinct attributes, T be transaction that contains a set of items such that T ⊆ I, D be a database with different transaction records Ts. An association rule is an implication in the form of X⇒Y, where X, Y ⊂ I are sets of items called item sets, the rule means X implies Y. There are two important basic measures for association rules, support(s) and confidence(c).

Support(s) of an association rule is defined as the percentage/fraction of records that contains X and Y (XUY) to the total number of records in the database. The count for each item is increased by one every time the item is encountered in different transaction T in database D during the scanning process. It means the support count does not take the quantity of the item into account.

$$\textbf{Support}\ (X \Rightarrow Y) = \frac{\textbf{P}\ (XUY)}{N}$$

$$\textbf{Confidence}\ (X \Rightarrow Y) = \frac{\textbf{P}\ (XUY)}{P(X)}$$

Where X, Y is the measure of Antecedent and Consequent, P is the Probability at which it can occur and N is the total number of records.

Apriori Algorithm

Apriori is a great improvement in the history of association rule mining, which was first proposed by Agrawal in [Agrawal and Srikant 1994].Initially, the candidate itemsets are generated, then the database is scanned to check the actual support count of the corresponding itemsets. During the first scanning of the database the support count of each item is calculated and the large 1 -itemsets are generated by pruning those itemsets whose supports are below the pre-defined threshold values. The candidates are generated by joining among the frequent itemsets level-wise, also candidate are pruned according the Apriori property [14].

However there are two bottlenecks in the Apriori algorithm. One is the complex candidate generation process that uses most of the time, space and memory. Another one is the multiple scan of the database. Based on Apriori algorithm, many new algorithms were designed with some modifications or improvements.

FP-Tree(Frequent Pattern Tree) Algorithm

FP-Tree [14], frequent pattern mining, is another milestone in the development of association rule mining, which breaks the two bottlenecks of the Apriori. The frequent itemsets are generated with only two passes over the database and without any candidate generation process. The process of constructing the FP-Tree is as follows.

Step I: The database is scanned for the first time, during this scanning the support count of each items are collected. As a result the frequent 1-itemsets of this process is the same as in Apriori algorithm. Those frequent itemsets are sorted in a descending order with their supports. Also the head table of ordered frequent 1 -itemsets is created.

Step II: Create the root node of the FP-Tree T with a label of Root. The database is scanned again to construct the FP-Tree with the head table, for each transaction the order of frequent items is resorted according to the head table.

Step III: If T has a child N, then the count of N is increased by 1, else a new node is created. The insertion function is called recursively until P becomes empty where P is the items list.

Eclat

Equivalence Class Transformation (Eclat) algorithm is used to perform itemset mining by exploring the data in vertical format. Item set mining lets us to find frequent patterns in data[19]. This type of pattern is called association rules and is used in many application domains. The first scan of the database builds the TID_set of each single item. Starting with a single item (k = 1), the frequent (k+1)-item sets developed from a previous k-item set can be generated according to the Apriori property, with a depth-first computation order. The computation is done by intersection of the TID_sets of the frequent k-itemsets to compute the TID_ sets of the corresponding (k+1)-itemsets. This process repeats, until no frequent itemsets or no candidate itemsets can be found. The significant advantage of this method is that there is no need to scan the database to find the support of (k + 1)-itemsets (for k ≥ 1). This is because the TID_set of each k-itemset carries the complete information required for counting such support.

III. MULTIDIMENSIONAL RULE

To enhance the traditional model of transaction this approach can associate the records with number of attributes that helps to explain the context where the transaction happens. These attributes form a multi-dimensional space and transactions can be viewed as points in space. The dimensions can be of any kind as long as they are useful for mining. Association rules can be classified as single-dimensional association rules and Multi-dimensional association rules on the basis of dimensions. In a Single Dimensional rule or intra-dimensional association rule, items in the rule refer to only one dimension or Predicate [6].

Multidimensional association rules are rules that are found from a table structured database where each item has distinct attributes which is not found in association rules. Each item in the rule is a pair i.e., attributes and value of the attribute. In a Multidimensional rule items in the rule refer to two or more dimensions or predicates. Hence it can be say that it has no repeated predicates.

Multidimensional association rules with no repeated predicates are called inter-dimensional association rules.The main difference between association rules and multidimensional association rules is that each item in multidimensional association rules has distinct attributes, but there is no such attribute in the items of association rules. Multidimensional association rules are basically an application of general association rule algorithms to table-like databases. The table-like databases may consist of condition attributes and decision attributes. Neelu and Neeru [10] have described that multidimensional association rules have better accuracy than decision trees for most of the data sets.

3.1MULTIDIMENSIONAL APPROACHES

A. Using Static Discretization

Quantitative attributes are discretized by incorporating concept hierarchies. This process takes place before the occurrence of mining. For example, a concept hierarchy for age may be used to replace the original numeric values of this attribute by ranges, such as "1-10", "11-20", "21-40", and so on. Here, discretization is static and predetermined. The discretized numeric attributes, with their range values, can then be treated as categorical attributes. Each range is considered as a category. It is referred as mining multidimensional association rules using static discretization of quantitative attributes [16].

B. Using Dynamic Discretization

Another method to deal with numeric attributes is to perform discretizing called "bins" based on the distribution of the data. These bins may be further combined during the mining process. The discretization process is dynamic and established so as to satisfy some mining criteria, such as maximizing the confidence of the rules mined. Because this strategy treats the numeric attribute values as quantities rather than as predefined ranges or categories, association rules mined from this approach are also referred to as quantitative association rules [16].

C. Using Distance Based Discretization

It considers the values that are close to each other and grouped into the same interval. It involves two steps, firstly intervals or clusters are generated by the clustering technique and then distance based association rules are generated by searching for groups or clusters that occur frequently together. Hence, such quantitative association rules are also referred to as Distance-based association rules.[16].

D. Boolean Matrix based Approach

This approach, based on Boolean matrix [10] is used to generate the multidimensional rule which has no repetitive predicates. A Boolean Matrix based approach has been used to find the frequent itemsets, the items forming a rule originating from different dimensions. It is an algorithm for mining multidimensional association rules from relational databases. The algorithm adopts Boolean relational calculus to find out frequent predicate sets. While using algorithm for the first time, it scans the database once and will generate the association rules. Apriori property is implemented in this algorithm to prune the sets. It is not necessary to scan the database again; it uses Boolean logical operations to generate the association rules. It stores all data in the form of bits, so that it reduces memory space and can be applied to large relational databases. An algorithm to transfer the relational database in Boolean matrix is as follows:

The algorithm consists of the following steps:

1. Transforming the relational database in to Boolean matrix.
2. Generating the set of frequent 1-itemsets.
3. Pruning the Boolean matrix.
4. Perform AND operation on frequent 1-items of different dimensions to generate 2-itemsets.
5. Repeat the process to generate more number of item sets until the resultant item set is generated.

E. Bit-Mask Search Algorithm

In this approach, a new array based optimization technique called Bit stream mask and Masked Item set Processing search is used for mining complete frequent item sets [5]. Data representation with additional storage of sixteen elements in one process memory array location has been compared with the available implementation of familiar Apriori and FP-growth algorithms. This technique shows improved performance of mining frequent item sets on a number of typical datasets. It is faster than the other Apriori Tree.

IV. QUANTITATIVE RULE

Mining association rules on both categorical and numeric attributes is said to be quantitative association rules. When the domain is continuous, then the association rules are Quantitative association rule. Numeric attributes are always defined on a wide range of different values. Therefore, it is difficult to work on all possible values. A classic method to deal with numeric attributes is to divide their domains into intervals called Discretization.

Most of them are based on some form of discretization labelled as partitioning, quantization or bucketing of numeric attributes, which means dividing the attribute domain into separate ranges. Such pre-processing method results in the loss of informational value of discovered rules, or often loses significant ones. In order to find out association rules with quantitative attributes, an efficient method is required to reveal the meaningful intervals.

Many researchers proposed different methods to solve the problem of interval partition. For quantitative attributes that are not partitioned into intervals, the values are mapped to consecutive integers. These mappings lead to treat a database record as a set of (attribute, integer value) pairs, without loss of generality. Adaptation of the APRIORI algorithm for mining quantitative association rules was identified soon after the introduction of APRIORI algorithm. The necessity for quantity in mining association rules was first identified by Sujatha and Naveen [15]. It proposed rules of the form $x=q_x \Rightarrow y=q_y$ i.e. it associated a single quantity 'q' to the antecedent and the consequent. This was done by decomposition of one quantitative attribute into several binary attributes.

However, most of the approaches are designed for "market basket analysis" and operate on categorical (qualitative) data. It renders them useless for learning from many common types of data based on numeric values. Special forms of association rules for quantitative attributes may be applicable.

Quantitative approach

There are only few algorithms and methodologies to deal with quantitative associations.

Discretization

Discretization technique [9] [10] can be considered as the first comprehensive work on quantitative association rules. The basic idea of the technique is to map quantitative data to Boolean by considering a partition of the numerical attributes into sets of intervals. Then, an algorithm for finding Boolean association rules can be used to get quantitative rules.

Two main types of partitions are introduced. A fixed partition, where the set of intervals are disjoint and another type, where the boundaries of intervals are overlapped with each other. The rules are in the form $X \Rightarrow Y$, where X and Y are sets of attributes and they are either a categorical item or a range of numerical values. The main advantage of this technique is that it handles both categorical and numerical data equally.

Disjoint sets suffer from Minimum Support and Minimum Confidence whereas overlapped sets suffer from Sharp Boundary problem. Using intervals instead of the original continuous data leads to loss of information. The rules obtained will be only an approximation of the best ones. Moreover, the lack of accuracy not only affects the quality of the rules but also leads to large number of unwanted rules.

Another problem is the expansion of the attributes. The Discretization technique can be used in any kind of data that contains any combination of categorical and numerical attributes even on the case of all attributes are numerical.

APACS2 (Automatic Pattern Analysis and Classification System)

The algorithm APACS2 [9] is based on employing both adjusted difference analysis and discretization to find rules between two attributes. The two attributes can be any combination of numerical or categorical. The technique has the ability of discovering positive and negative association rules and does not need any user thresholds. The rules are in the form,

(attribute$_l$) =>attributer [type]

Where attribute$_l$ and attribute$_r$, may be categorical and quantitative in any order and type may be + or – for positive and negative rules respectively.It has the ability of finding new meaningful objective measure of the association rules both positive and negative.

Statistical Inference

In some cases numerical attributes are treated as a continuous values rather than converting it into range of numerical values based on statistical inference [9]. In order to make this algorithmically manageable, two specific association rules are defined: Categorical to Quantitative rules with an unlimited number of attributes, and Quantitative to Quantitative where both sides contain a single attribute only. The general form for the first type of association rules is

$$X \Rightarrow M_J (T_X) (M_J (T_X) \approx M_J(\cancel{D}))$$

Where X is categorical attributes, J is a set of numerical attributes, M is a statistical Measure, T_x is the set of transactions defined by X, and D is the whole database.

This rule is interesting because it brings out the value of certain group of attributes significantly lower than the threshold. Obviously, the technique has a good advantage, which is the use of numerical data as it is. Also, it introduces a new, indeed an interesting measure which is finding set of frequent behaviour that is different from overall behaviour.

BAR mining Algorithm

Mining Quantitative Association Rules (QAR) by a generic Boolean Association Rule (BAR) mining algorithm, [17] however, is infeasible in most cases for the following reasons. First, QAR mining suffers from the problem of a combinatorial explosion of attribute sets as does BAR mining. In practice, the number of distinct attributes in a QAR mining problem may not be as large as that in a BAR mining problem.

However, as shown by Srikant and Agrawal [1], it is necessary to combine the consecutive intervals of a quantitative attribute in order to gain sufficient support and more meaningful intervals. This leads to another combinatorial explosion problem if the domain of a quantitative attribute is partitioned into n intervals; the total number of intervals of the attribute grows to O (n^2) afer combining the consecutive intervals. When we join the attributes in the mining process, the number of itemsets can become prohibitively large if the number of intervals associated with an attribute is large then the attributes are joined. Effective techniques to prune the large search space of QAR mining are necessary.

V. ISSUES AND CHALLENGES OF MUTIDIMENSIONAL AND QUANTITATIVE ASSOCIATION RULES

Association rule mining on high dimensional data, now a days is a high topic of research interest in many fields of data mining tasks. There are numerous data repositories which stores the data in different dimensions. Mining association rules on these databases is a challenging issue. There is a need to find association rules on high dimensional data a more effectively for different applications and decision making. There has been work on quantifying the “usefulness" or “interestingness" of a rule.

The mining of all association rules existing in the database with respect to certain user interesting measures. Database in this case consisting of a data set with varying attributes or dimensions in the average out a total set. Although the association mining is still a topic of further research, during recent years many algorithms for specialized tasks have been developed. First of all, there are the approaches that enhance the association rules itself. E.g. quantitative association rules, generalized association rules and to some extent the work on sequential patterns.

The emergence of various new application domains, such as bioinformatics and e-commerce, emphasize the need for analyzing high dimensional data. Many organizations have enormous amounts of data containing valuable information for running and building a decision making system. Extracting the value of that data is a big challenge.

First and foremost is to understand and analyze the large amount data for effective decision making. For example, generally, in a gene expression microarray data set, there could be tens or hundreds of dimensions, each of which corresponds to an experimental condition. In a customer purchase behavior data set, there may be up to hundreds of thousands of merchandizes, each of which is mapped to a dimension. Researchers and practitioners are very eager in analyzing these types of data sets.

However, before analyzing the data mining models, the researcher will analyze the challenges of attribute selection, the curse of dimensionality, the specification of similarity in high dimensional space for analyzing high dimensional data set. Association Rule Mining in high dimensional spaces presents tremendous difficulty in generating the rules, much better than in predictive mining. Attribute selection is a one which reduces the impact of high dimensional data space at the time of generating rules.

Dimensionality Curse is a way of speaking about lack of data separation in high dimensional space. The complexity of many existing data mining algorithms is exponential with respect to the number of dimensions. With increasing dimensionality, these algorithms soon become computationally intractable and therefore inapplicable in many real applications.

Multi dimensional association rule mining carried on multi dimensional datasets which stores the data in more than one dimensions or predicates. Multidimensional association rules are used to mine when the data or transactions are located in multidimensional space, such as in a relational database or data warehouse. Multiple dimensional association rule mining is to discover the correlation between different predicates or attributes. Each attribute or predicate is called a dimension. The data may contain different types of data such as categorical, Boolean, numeric. The attributes are also called as quantitative attributes. Association rule mining on these attributes can be carried both in static and dynamic discretization methods.

More research work is carried under Dynamic discretization method. In this the numeric attributes are dynamically discretized during the mining process so as to satisfy the mining criteria. Predicate attributes or values in the dimensions are converted into Boolean values so as to carry the mining process very effectively. It is carried on the following steps.

1) Determine the Number of Partitions for each predicate or attribute.
2) Generation of Frequent Itemsets.
3) Generating Interesting rules using the Minimum confidence.

Fabulous research work is carried under Quantitative association rule ranging from low-level predicates to multi dimensional predicate and used for different range of applications.

VI. CONCLUSION

This research survey motivates us to find the issues in multidimensional and quantitative association rules. This paper is organized to categorize the survey process as follows:

1) Rule Mining algorithm categorization
2) Multidimensional approach to generate association rules
3) Quantitative association rules review process
4) Existing applications which are categorized with respect to fields which it applied.

The Bit search is one of the greatest techniques that provide to search all the elements from the dataset. Bit Stream Mask Search is used to find the frequent itemsets from the dataset.

Multidimensional and quantitative association rule generation is done with the help of frequent and infrequent itemsets. The fuzzy based multidimensional and Quantitative association rule generation with bit mask search will definitely identify the frequent and infrequent itemset by reducing the scanning process significantly. This new procedure will yield the execution time must be reduced.

References

[1] Agrawal, R. and Srikant, R. 1994, " Fast algorithms for mining association rules", In Proc. 20th Int. Conf. Very Large Data Bases, VLDB, J. B. Bocca, M. Jarke, and C. Zaniolo, Eds. Morgan Kaufmann, 487-499.

[2] Agrawal, R., Imielinski, T., and Swami, A. N. 1993. "Mining association rules between sets of items in large databases". In Proceedings of the 1993 ACM SIGMOD International Conference on Management of Data, 207-216.

[3] N.Deepika* et al., (IJAEST)" International Journal of Advanced Engineering Sciences and Technologies", Vol No. 11, Issue No. 2, 253 – 257.

[4] Dimitris Kanellopoulos et.al., GESTS, "International Transactions on Computer Science and Engineering ", Vol.32 (1), 2006, pp. 71-82, Association Rules Mining: A Recent Overview.

[5] E.Ramaraj and N.Venkatesan," Bit Stream Mask-search algorithm in Frequent Item set Mining", in European Journal of scientific research. Vol.27 No.2 (2009), pp.286-297.

[6] Han, J. 1995. Mining knowledge at multiple concept levels. In CIKM. 19-24.

[7] Hegland, M., "Algorithms for Association Rules, Lecture Notes in Computer Science", Volume 2600, Jan 2003, Pages 226 – 234.

[8] Milan Kumari et.al ., "Comparative Study of Data Mining Classification Methods in Cardiovascular Disease Prediction " in " IJCST Vol. 2, Iss ue 2, June 2011 I S S N : 2 2 2 9 - 4 3 3 3 (P r i n t) | I S S N : 0 9 7 6 - 8 4 9 1 (On l i n e),.

[9] Mohamed Amin Eldesouky Elmetwaly Elkhoul Further remarks on Quantitative Association Rules.

[10] An Algorithm for Mining Multidimensional Association Rules using Boolean Matrix by Neelu Khare , Neeru Adlakha and K.R.Pardasani.

[11] Mining Multi-Dimensional Quantitative Associations Michaª Okoniewski, .ukasz Gancarz, Piotr Gawryiak

[12] An Efficient Algorithm for Mining Multilevel Association Rule Based on Pincer Search Dr. Pratima Gautam, Rahul Shukla.

[13] Mining Quantitative Association Rules in Large Relational Tables by Ramakrishnan Srikant and Rakesh Agarwal.

[14] Association Rule Mining :A Survey Qiankun Zhoa and Sourav Bhowmick

[15] Quantitative Association Rule Mining on Weighted Transactional Data D. Sujatha and Naveen CH International Journal of Information and Education Technology, Vol. 1, No. 3, August 2011.

[16] International Journal of Emerging Science and Engineering (IJESE) ISSN: 2319–6378, Volume- 1, Issue-4, February 2013 Survey on Multidimensional and Conditional Hybrid Dimensional Association Rule Mining.

[17] An Information – Theoretic Approach to Quantitative Association Rule Mining by Yiping Ke James and Cheng Wilfred Ng.

[18] An Algorithm for Mining multidimensional Fuzzy Association Rules by Neelu Khare , Neeru Adlakha and K.R.Pardasani

[19] Frequent pattern mining:current status and future directions.Jiawei Han,Hong Cheng,Dong Xin,Xifeng Yan.