Multidimensional Quantitative Rule Generation Algorithm for Transactional Database

R.Sridevi
Research Scholar
Manonmaniam Sundaranar University,Tirunelveli

E.Ramaraj, Ph.D
Professor,Department of Computer Science and Engineering
Alagappa University, Karaikudi

ABSTRACT
Data mining is a technology development in the present decade for guiding decision making. One of the main applications of data mining is exploration of Association Rules. The objective of the research is to find out the association rules for the sample dataset to find out the interesting and useful rules. A lot of modifications have been suggested over the last two decades for the traditional Market Basket Analysis Algorithm like Apriori, FP -Growth, E-clat etc. The proposed Multidimensional Quantitative Rule Generation (MQRG) method is to generate more number of interesting rules that satisfy minimum confidence threshold (min_conf). This paper presents the comparison results of the existing algorithm with the proposed Multidimensional Quantitative Rule generation.

Keywords
Data mining, Data Discretization, Multidimensional, Quantitative, Association rules

1. INTRODUCTION
Data Mining is a new technology and rapidly growing field which could be used in extracting valuable information from data warehouses and databases of companies and governments. Data mining is the process of discovering meaningful new correlation, patterns and trends by sifting through large amount of data stored in repositories, using pattern recognition technologies as well as statistical and mathematical techniques[1]. Data mining is implemented using tools, and the automated analysis provided, this tools go beyond evaluation of dataset to provide tangible clues that human experts would not have been able to detect due to the fact that they have never experienced or expected such.

Association rule mining has become an important data mining technique due to the descriptive and easily understandable nature of the rules. Rakesh Agrawal et al [4] introduced a common way of measuring the usefulness of association rules to use the support-confidence framework. Although association rule mining was introduced to extract associations from market basket data, it has proved useful in many other domains (e.g. microarray data analysis, recommender systems and network intrusion detection).

The paper is organized as follows: Related work is presented in section II. In section III, proposed work is presented and in Section IV experimental results and performance comparison has been made with different data mining environments [DME] [2][3] and Section V includes the discussion about the results obtained and Section VI concludes research and future works.

2. RELATED WORK
2.1 Apriori Algorithm
Apriori algorithm is one of the first few algorithms proposed, which is based on a candidate set generation logic[5]. Apriori algorithm suffers a major limitation of repeated scan limitation in developing association rules. It generates candidate itemsets and tests if they are frequent.
Steps:

1. Find the frequent itemsets, i.e., find all I with support ≥min_sup.{If I = (A,B) is frequent, then both A and B are also frequent}.
2. Find frequent itemsets with cardinality ranging from 1 to k.
3. Generate ARs from frequent itemsets.

The first pass of the algorithm simply counts item occurrences to determine the large 1itemsets.Subsequent pass, say pass k, consists of two phases. First, the large itemsets Lk-1 found in the $(k-1)^{th}$ pass are used to generate candidate itemsets Ck , using the apriori gen function[7] .Next, the database is scanned and the support of candidates in Ck is counted. For fast counting, we need to efficiently determine the candidates in Ck that are contained in a given transaction t.

2.2 FP-Growth algorithm
The FP-Growth approach is based on divide and conquers strategy for producing the frequent item sets [6]. It is a two step approach to generate frequent itemsets without candidate generation. The algorithm is based upon a tree representation of frequent itemsets. It compresses a large database of transaction into a compact frequent –pattern tree structure.

In FP-tree scan the database once to collect all frequent items and their support counts. All frequent items are sorted in descending order of support and denoted as L. After that, create the root of an FP-tree and label it as "null". Scan the database for a second time. The items of each transaction in the database are sorted according to the order of *L*.

On inserting a transaction, if the tree has the same path, then the count of each node in the path increases if the path is incomplete in the tree, then a new branch and new nodes are created. Moreover, the *node-links* of these new nodes are linked to the nodes with the same *item-name* via the node-link structure [9]. After constructing the FP-tree, the FP-growth algorithm recursively builds a conditional pattern base and conditional FP-tree. Then, it is used to generate all frequent patterns.

2.3 E- clat
It is a set intersection, depth first search algorithm [5], unlike the Apriori. It uses vertical layout database and each item use

intersection based approach for finding the support. In this way, the support of an itemset P can be easily computed by simply intersecting of any two subsets Q, R ⊆ P, such that P _ Q U R.

In this type of algorithm, for each frequent itemset I new database is created Di. This can be done by finding j which is frequent corresponding to I together as a set then j is also added to the created database i.e. each frequent item is added to the output set. The algorithm uses the join step as in Apriori only for generating the candidate sets but as the items are arranged in ascending order of their support thus less amount of intersection is needed between the sets. It generates the large number of candidates then Apriori because it uses only two sets at a time for intersection [5]. There is reordering step takes place at each recursion point for reducing the candidate itemsets.

In this way by using this algorithm there is no need to find the support of item sets whose count is greater than 1because Tid-set for each item carry the complete information for the corresponding support. When the database is huge and the item sets in the database is also huge then it is feasible to handle the Tid list. Thus it produce good results but for small databases its performance is not up to mark.

3. PROPOSED WORK

The proposed method generates the data set based on the attribute for the transactional database. Data Discretization and Concept hierarchy are performed to categorize the attribute value according to the constraints given. The quantitative conversion is used to simplify the given data set in which all the attributes are presented as numbers. The following Fig.1 represents the overview of the MQRG algorithm.

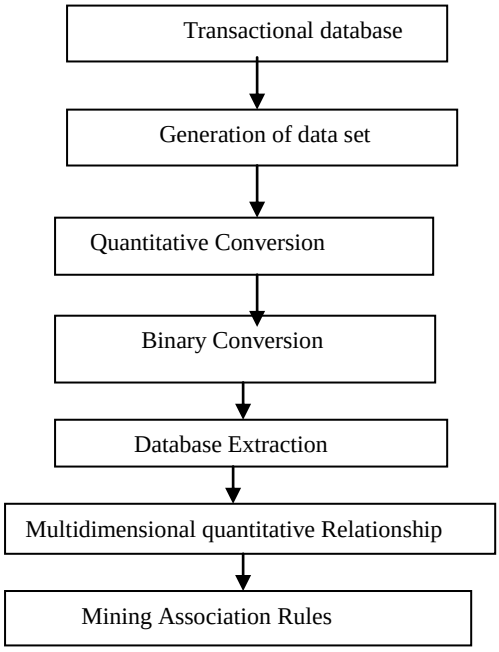

Fig.1 Overall architecture of the proposed method

The proposed methodology also extracts the database information from the given set of database based on the given problem [8].The effective association rule mining is done by making relationship between the multidimensional quantitative set of data derived from the earlier conversion.

Almost a database is a combination of categorical and numerical values. First, the categorical attributes are assigned an integer value [10]. This conversion helps to reduce the time taken to scan the database. Now, all the values in the table are numerical attributes. In this approach, the numerical values are converted to binary values according to certain constraints related to the problem. For example, Age>=20 is assigned as 0 and Age<=20 is assigned as 1. Similarly, categorical attributes are divided into two categories as per the concept hierarchy and assigned the number as 0 and 1[12].The Table 1 shows the sample dataset.

Table 1: Sample Student dataset

Tid	Place	Medium	Sex	Age	AI	Result	Subject
1	Rural	Tamil	Male	20	1,50,000	Pass	Arts
2	Rural	Tamil	Female	19	1,75,000	Pass	Arts
3	Urban	English	Male	21	90,000	Fail	Science
4	Rural	Tamil	Male	18	98,000	Pass	Science
5	Urban	English	Female	19	1,25,000	Fail	Arts
6	Urban	English	Female	21	1,50,000	Pass	Arts
7	Rural	Tamil	Male	22	95,000	Pass	Science
8	Rural	Tamil	Male	18	1,20,000	Pass	Science
9	Urban	English	Female	19	1,10,000	Fail	Arts
10	Rural	Tamil	Male	19	94,000	Pass	Arts

Concept hierarchies for numerical attributes can be constructed automatically based on data Discretization [13]. More than one concept hierarchy can be defined for the same attribute in order to accommodate the needs of various users. From the sample student database given in Table 1, the concept hierarchy is generated as shown in Fig.2.

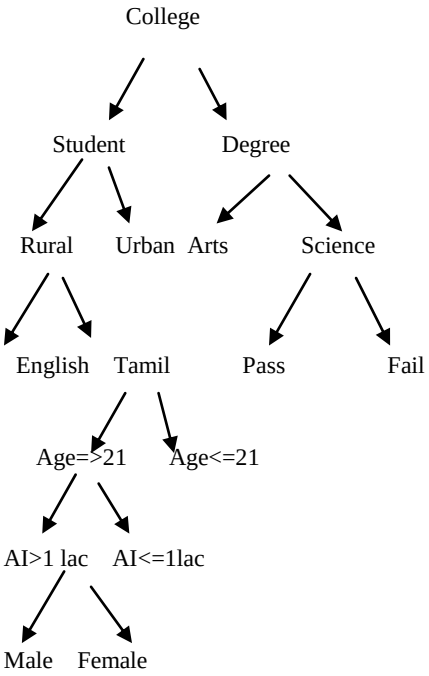

Fig: 2 Concept hierarchy for student dataset

All the attributes in the database are stored as 0 and 1.When a record is retrieved it is a series of 0 and 1.In order to integrate the record, a combined value is given for each new series. i.e.,each record is identified with an integer. This kind of transformation reduces the time complexity. For a database of m records of n attributes, assuming binary encoding of attributes in a record, the enumeration of subset of attributes requires m x 2^n computational steps.

Algorithm: Multidimensional Quantitative Rule Generation(MQRG)

Input: A dataset composed of N Tuples, MinConf, and number of attributes
Output: Quantitative and multidimensional Association Rules R
Step 1: Select a set of attributes from the given database
Step 2: Let Rt a set of constraints defined on these attributes.
Step 3: Generate concept hierarchy for the attributes
Step 4: For each record **in** the database substitute the numerical value
Step 5: Compute the binary values for numerical attributes using discretization
Step 6: Call Procedure quantitative _relation (database)
Step 7: For **e**ach tuple in N find the relationship using the combined number for the two datasets
Step 8: Generate the rules from the transformed dataset
Step 9: Filter the rules that satisfy the min-conf

4. MATERIALS AND MODELS

4.1 Data Set

4.1.1 Mushroom

This data set includes descriptions of hypothetical samples corresponding to 23 species of gilled mushrooms in the Agaricus and Lepiota Family. Each species is identified as definitely edible, definitely poisonous, or of unknown edibility and not recommended. It consists of 8124 Instances and 22 nominal attributes.

4.1.2 Adult

This data set is used to predict whether the income exceeds 50K per year based on census data. Also known as "Census Income "dataset. It exhibits Multivariate characteristics. The number of instances taken for consideration is 48842. The attributes include Categorical and Integer type and its count was 42.

4.1.3 Balloon

The Balloon data set exhibits multivariate characteristics. The number of instances used is 16. The attributes of the dataset is categorical.

4.1.4 Iris

The data set contains 3 classes of 50 instances each, where each class refers to a type of iris plant. One class is linearly separable from the other 2; the latter are NOT linearly separable from each other. The number of Instances are 150 (50 in each of three classes) and the number of Attributes includes 4 numeric, predictive attributes and the class.

5. PERFORMANCE ANALYSIS

The aim of the work is to obtain an associative model that allows studying the influence of the input variables related to the student database and the output variables related to the software product and the software process. The display includes bars, disk and colours whose meaning is given in the graph. Those experiments performed on computer with Core 2 Duo 2.00 GHZ CPU, 2.00 GB memory and hard disk 80 GB.

5.1 Comparison with Apriori and FP-Growth

For experimental study various types of datasets are used to justify the efficiency of the proposed Multidimensional Quantitative Rule Generation method. The data sets are adult, mushroom, Iris and Balloon. Some characteristics of these datasets are shown in table 2.

Table 2: Characteristics of Experiment Data Sets

Data	#items	avg. trans. length
Adult	156	51
Balloon	145	47
Iris	130	43
Mushroom	120	23

The Multidimensional Quantitative rule generation was mainly compared with two popular algorithms -Apriori and FP-growth. They were compiled in VB .Net.

Table 3: Run Time (S) For Adult data

No. of records	Apriori in msec	FP-Growth in msec	Multidimensional Quantitative method in msec
100	156	105	42
200	291	187	54
500	141	141	77
1000	197	261	83
2000	398	358	106

Table 3 shows the running time of the compared algorithms on Adult data with different number of records represented by percentage of the total transactions. Under minimum supports, Multidimensional Quantitative rule generation runs faster than Apriori and FP-Growth.

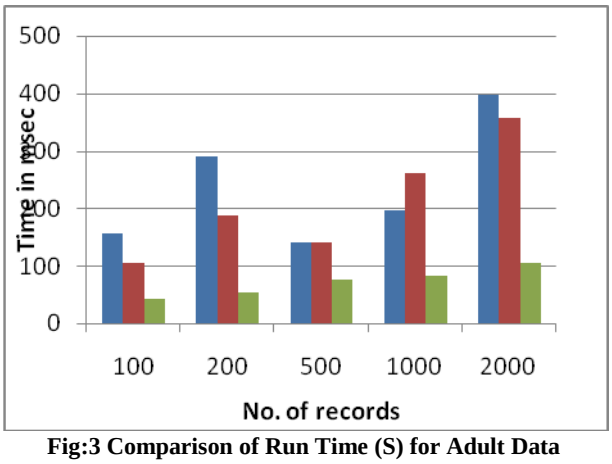

Fig:3 Comparison of Run Time (S) for Adult Data

MQRG algorithm runs faster than both algorithms under almost all support values. Fig.3 shows the relative performance of the algorithms on Adult data set.

Table 4: Run Time (S) For Balloon data

No. of records	Apriori in msec	FP-Growth in msec	Multidimensional Quantitative method in msec
100	141	72	33
200	219	101	37
500	289	175	41
1000	417	342	89
2000	634	562	199

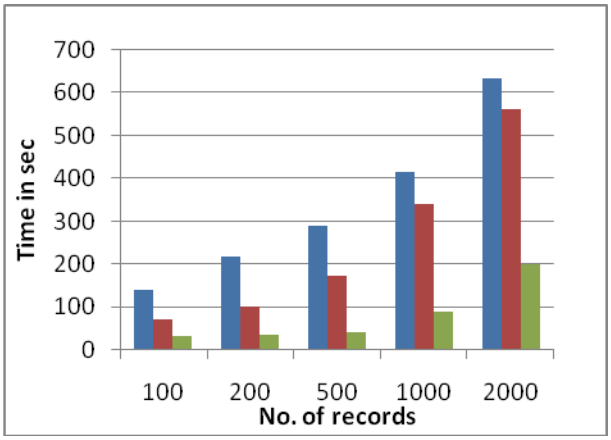

Fig: 4 Comparison of Run Time (S) for Balloon Data

Table 4 and Fig. 4 show the performance comparison of the compared algorithms on Balloon data.For this dataset, MQRG algorithm runs faster than other two algorithms.

Table 5: Run Time (S) For Iris data

No. of records	Apriori in msec	FP-Growth in msec	Multidimensional Quantitative method in msec
100	140	65	32
200	157	120	41
500	219	305	59
1000	307	413	146
2000	432	397	231

Table 6: Run Time (S) For Mushroom Data

No. of records	Apriori in msec	FP-Growth in msec	Multidimensional Quantitative method in msec
100	188	82	37
200	196	200	84
500	156	236	81
1000	379	354	152
2000	564	621	212

Table 6 shows the relative performance of the algorithms on Mushroom dataset.

The following bar chart fig.5 illustrates the comparative study of the existing and proposed work for the Iris dataset.

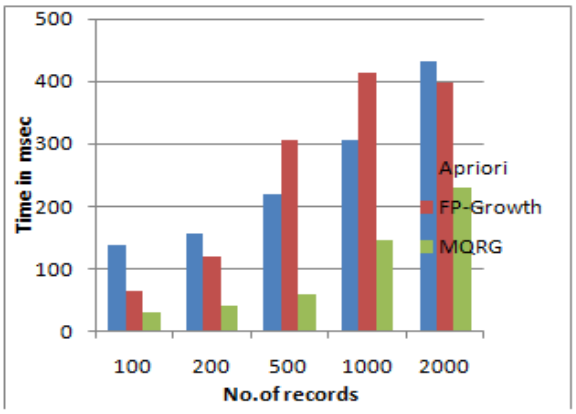

Fig: 5 Comparison of Run Time (S) for Iris Data

Fig.6 shows the relative performance of the algorithms on Mushroom dataset.

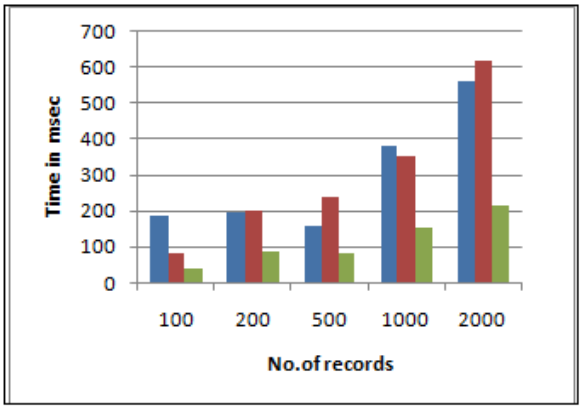

Fig: 6 Comparison of Run Time (S) for Mushroom Data

5.2 Run Time

The data sets used in the above experiments have often been used in previous research and the times shown include the time needed in all the steps. MQRG algorithm outperforms FP-growth and Apriori.

6. CONCLUSION

Mining frequent item sets for the association rule mining from the large transactional database is a very crucial task. There are many approaches that have been discussed. Nearly all of the previous studies were using Apriori approach and FP-Tree approach for extracting the frequent itemsets, which have scope for improvement. Thus the goal of this research was to find a scheme for pulling the rules out of the transactional data sets considering the time and the memory consumption. The MQRG method can also be used for Medical database, Bank database and other databases to bring out the interesting relationship between the itemsets.

7. REFERENCES

[1] Anbalagan Pakkirisamy, Chandrasekaran Ramasamy, Saravanan Subramanian . Performance Comparison Of Rule Generation Algorithm In Open Source Data Mining Environments,Advances in Engineering and Technology Convergence, 28th April, 2013, Bangkok, Thailand, ISBN NO: 978-93-82208-89-1.

[2] Lei Wangt, Xing-Juan Fan2, Xing-Long Lwt, Huan Zha Mining data association based on a revised FP-growth Algorithm Proceedings of the 2012 International Conference on Machine Learning and Cybernetics, Xian, 15- 17 July,

[3] Bo Wu, Defu Zhang, Qihua Lan, Jiemin Zheng ,An Efficient Frequent Patterns Mining Algorithm based on Apriori Algorithm and the FP-tree Structure Department of Computer Science, Xiamen University, Xiamen.

[4] R. Agrawal, T. Imielinski, and A. Swami. Mining association rules between sets of items in large databases. In Proc.1993 ACM-SIGMOD Int. Conf. Management of Data, Washington, D.C., May 1993, pp 207–216.

[5] R. Agrawal and R. Srikant. Fast algorithms for mining association rules. In VLDBY94, pp. 487-499.

[6] C.Borgelt. "Efficient Implementations of Apriori and Eclat". In Proc. 1st IEEE ICDM Workshop on Frequent Item Set Mining Implementations, CEUR Workshop Proceedings 90, Aachen, Germany 200.

[7] J.S .Park, M.S.Chen and P.S.Yu. An effective hash based algorithm for mining association rules. In SIGMOD1995, pp 175-186.

[8] J. Han, J. Pei, and Y. Yin. Mining Frequent Patterns without Candidate Generation (PDF), (Slides), Proc. 2000 ACM-SIGMOD Int. May 2000.

[9] E.Ramaraj and R.Sridevi A general Survey on multidimensional and Quantitative Association Rule mining Algorithms. International Journal of Engineering Research and Applications(IJERA) 2013.

[10] A.B.M.Rezbaul Islam, Tae-Sun Chung An Improved Frequent Pattern Tree Based Association Rule Mining Techniques Department of Computer Engineering Ajou University Suwon, Republic of Korea.

[11] E. Ramaraj and N. Venkatesan, ― Bit Stream Mask Search Algorithm in Frequent Itemset Mining,‖ European Journal of Scientific Reasearch,‖ Vol. 27 No.2 (2009),

[12] G. Grahne, J. Zhu, Fast algorithms for frequent itemset mining using FP-Trees, IEEE Transactions on Knowledge and Data Engineering 17 (10) (2005) 1347–1362.

[13] E.Ramaraj and R.Sridevi Finding frequent patterns Based on Quantitative Binary Attributes Using FP-Growth Algorithm,International Journal of Engineering Research and Applications(IJERA) 2013.

[14] J. Han, H. Cheng, D. Xin, X. Yan, Frequent pattern mining: current status and future directions, Data Mining and Knowledge Discovery (2007). 10th Anniversary Issue.

[15] J. Han, J. Pei, Y. Yin, Mining frequent patterns without candidate generation, in: Proceedings of the 2000 ACM SIGMOD International Conference on Management of Data, 2000, pp. 1–12.

[16] T.-P. Hong, C.-W. Lin, Y.-L. Wu, Incrementally fast updated frequent pattern trees, Expert Systems with Applications 34 (4) (2008) 2424–2435.

[17] H. Huang, X. Wu, R. Relue, Association analysis with one scan of databases, in: Proceedings of the IEEE International Conference on Data Mining, 2002,pp. 629–632.