DATA CUBE – BASED IN QUANTITATIVE MULTIDIMENSIONAL ASSOCIATION RULES FROM AGRICULTURAL DATA WAREHOUSE

K.Rani,
Assistant professor,
Department of Computer Science,
Jayaraj Annapackiam college for women
(Autonomous), Periyakulam.

K. Renuga Devi,
Assistant professor,
Department of Mathematics,
Jayaraj Annapackiam college for women
(Autonomous), Periyakulam.

S. Irudhaya Ananthi,
Assistant professor,
Department of Computer Science,
Jayaraj Annapackiam college for women
(Autonomous), Periyakulam.

Abstract: In India Agriculture play a major role. Agriculture data is highly broadened in terms of irrigation sources, climate, soil and inputs like fertilizers and pesticides. For sustainable growth of agriculture, these resources need to be monitored, analyzed and allocated optimally. Data mining techniques may be used in agricultural data for mining the association rules among various inputs and outputs used for cropping. This paper is making an effort to study the existing data mining algorithms to mine association rules widely used in corporate sector. The paper also presents an idea for mining quantitative multidimensional association rules from Agricultural Data Warehouse based on concise data using data cubes.

Keywords: Data mining, KDD, Association rule, Multi dimensional data, Data cube, Quantitative Association Rules, Agricultural data warehouse.

1. INTRODUCTION

Data mining is the practice of automatically searching large stores of data to discover patterns and trends that go beyond simple analysis [1]. Data mining uses sophisticated mathematical algorithms to segment the data and evaluate the probability of future events [1]. Data mining can answer questions that cannot be addressed through simple query and reporting techniques [2]. Data Mining is an essential process where intelligent methods are applied for knowledge discovery in Database (KDD). It is an interactive sequence of Data Cleaning, Data Integration, Data Selection, Data Transformation, Data Mining, Pattern Evaluation and Knowledge Presentation operations [3]. Use of data mining techniques can provide more suitable system for the decision making. Today, data mining is used in numerous areas and many commercial data mining systems are available for these areas. For example financial data collected from banking and financial industries are often comparatively absolute, reliable, and of high quality, which helps methodical data analysis and data mining. Retail industry is also an important application field for data mining since it gathers huge amount of data on customer shopping history, consumption, sales etc. Retail data mining can help to identify customer buying behaviours, customer shopping patterns and trends; can help to improve the quality of customer service, achieve better customer satisfaction, enhance goods consumption ratios, design more effective goods & transportation policies and reduce the cost of business. Presently data mining is also used in many scientific applications like Biological Data Analysis, Intrusion Detection and Agricultural sector. There are many researches are going in data mining. Agricultural sector is relatively an emerging research field where lot of work is to be done. The present paper attempts to describe various algorithms for mining the association rules with their limitations for Agricultural data with some minimum specified confidence. The paper also brings out an idea for mining the quantitative association rule from Agricultural Data Warehouse by focusing on a determination of summarized data using data cubes.

II. ALGORITHMS IN DATA MINING

a) Mining the Association Rules

An Association rule is an implication of the form $A \Rightarrow B$, where $A \cap B = \Phi$ and A & B are subsets of all itemset D. There are two measures of rule interestingness i.e. Support (σ) and Confidence (T) [4]. They reflect the usefulness (worth)

IJCRCST © 2016 |All Rights Reserved

Scanned by CamScanner

d certainty (assurance) of discovered rules respectively. The le A=>B (support σ=5%, confidence T = 60%) shows that % of all the transactions under analysis shows the imultaneous purchase of items A and B by customers and 0% of confidence shows that 60% of customers who urchased item A also bought item B. Association rules express how items or objects are related to each other and how they tend to group together.

Association rules can be classified in numerous ways, based on type of values handled in rule (Boolean association rule or Quantitative association rule), based on the dimensions of data involved in the rule (Single dimension or Multidimensional) and based on level of abstractions involved (Single level association rules or Multilevel association rules). The present study focuses on quantitative association rules only.

Various algorithms have been proposed for mining the quantitative association rules. All these algorithms for mining the quantitative association rules are based on support-confidence framework and can be decomposed in two ways. First phase is concerned with "Finding all sets of items whose support and confidence are greater than the user specified minimum support (σ) and minimum confidence (T) respectively" [1]. Such items are called frequent itemsets. In second phase, "Frequent items are used to find desired association rule(s). These rules must satisfy minimum support (σ) and minimum confidence (T)" [1]. Much Research has been focused on first phase for finding the frequent itemsets. There are five major algorithms proposed to identify frequent itemsets for discovery of association rules, which have been discussed as follows.

A Priori Algorithm

Apriori is a seminal algorithm proposed by R. Agarwal and R.Srikant in 1994 for mining frequent itemsets for Boolean association rule[AS94b]. It is also called level wise algorithm. This algorithm uses prior knowledge of frequent itemset properties. It explores the level wise mining apriori property that "all nonempty subsets of a frequent itemset must also be frequent. At the kth iteration (for K≥2), it forms frequent k-itemset candidates based on the frequent (k-1) itemsets, and scans the database once to find the complete set of frequent kitemset, L_k" [5]. Two-step process is followed to find L_k.- join step and prune step.

Joining step: To find L_k, a set of candidate k-itemsets is generated by joining L_{k-1} with itself. This set of candidate is denoted by C_k. The set C_k is superset of L_k, i.e. its members may or may not be frequent, but all of the frequent k-itemsets are built-in C_k. A scan of the database to determine the tally of each candidate in Ck would outcome in determination of L_k (All candidates having minimum count equal to the minimum support). Large size of C_k could involve heavy calculation.

In order to diminish the size of C_k **pruning step** is applied based on the principle that "Any (k-1) itemset that is not frequent cannot be subset of a frequent k-itemset; if any (k-1) subset of a candidate k itemset is not in L_{k-1}, then candidate cannot be frequent and can be removed from C_k"[6]. Hence this algorithm is appropriate to discover the large itemsets in transactional database, satisfying the minimum support and confidence conditions. It is an iterative method to find frequent data set by pruning many of the sets which are unlikely to be frequent sets. The limitations of the algorithm are that it may produce a larger number of candidate itemset in this process and it may require n number of data scans where n is the size of large nonempty itemset. Also the number of discovered rules is huge while most of them are non-interesting [7].

Partition Algorithm

It is based on the observation that the frequent sets are normally very few in number compared to the set of all itemsets. If set of transactions can be divided to smaller segments such that each segment can be accommodated in main memory, then set of frequent sets of each partition can be worked out. The partition algorithm executes in two phases to determine all frequent sets. Firstly, it divides the database into non-overlapping partitions. The partitions are considered one at a time and all frequent itemsets for that partition are generated. If there are n partitions, Phase I of algorithm takes n iterations. At the end of phase I, these frequent itemsets are merged to generate a set of all potential frequent itemsets. In phase II of algorithm, the actual support for these itemsets is generated and the frequent itemsets are identified. This algorithm is based on the premise that "Size of the global candidate set is significantly small than the set of all possible itemsets" [8]. In other terms it can be explained as the size of the global candidate set is bounded by n times the size of largest set of locally frequent set of any partition. For Large partition size, the number of local frequent itemsets is parallel to the number of frequent itemsets of entire database. If the data characteristics in all the partitions are uniform, then large numbers of itemsets derived for individual partitions may be common. If an itemset is not frequent in any of the segments or partition, then it is not frequent in the whole database also.

Princers Search Algorithm

It has an advantage over above two algorithms that it works in two directions simultaneously i.e. bottom-up and top-down process. It attempts to find frequent itemsets in a bottom up manner, at the same time it maintains a list of maximal frequent itemsets. While making a database pass, it also counts the support of these candidate maximal frequent itemsets to see if any itemset is actually frequent. In that situation, it can conclude that all the subsets of these frequent sets are going to be frequent and hence they are not verified for the support count in the next pass. The princers search algorithm has advantage over apriori algorithm when the largest frequent itemset is long. In each pass of database, this algorithm counts the support of the candidate in the bottom up direction. It also counts the supports of some itemsets using a top down approach. These itemsets are called the Maximal Frequent

IJCRCST © 2016 |All Rights Reserved

Scanned by CamScanner

Candidate Set (MFCS). This process helps in pruning the candidate sets very early. The performance of this algorithm is better than apriori algorithm but the concept of pruning data set remains present, which will lead not to discover proper association for less represented data.

Dynamic Itemset Counting (DIC) Algorithm

Dynamic itemset counting techniques can be applied in a database which can be partitioned into blocks marked by start point. In this variation, new candidate itemset can be added at any start point, instead of beginning of each scan of database as in apriori algorithm. Hence this algorithm requires fewer database scans than apriori. Hence the method significantly decreases the size of candidate sets and enhances the performance. The logic behind DIC is that it works like a train running over the data with stops at some defined interval between transactions.

However, all the above methods have two main disadvantages.

(1) These methods may need to generate a huge number of candidate sets.

2) These methods may verify a large set of candidates by pattern matching and scans the database repetitively. It becomes costly to pass through each transaction in the database to find out the support of candidate itemsets.

FP - Tree Grow Algorithm

The frequent-pattern tree (FP-tree) is a compact structure that stores quantitative information about frequent patterns in a database [9]. This method adopts divide & conquer strategy. First it compresses the database representing frequent items into a frequent pattern tree or FP-tree, which retains the itemset association information. It then divides the compressed database into a set of conditional database, each associated with one frequent item or 'Pattern Fragment' and mines each such database separately.

The FP-growth method transforms the problem of developing long frequent patterns to searching for shorter ones recursively and then unite the suffix. In this method, Least Frequent Items are used as suffix for searching of frequent patterns. The method considerably decreases the search cost. When the database is huge, it is impractical to construct a main memory based FP-tree.

This algorithm has an advantage that there is no need of multiple scans of data like other algorithms, because it stores the data in a tree structure and it does not generate the candidate as in other algorithms [10]. Following table presents the comparison among various algorithms discussed above for mining the quantitative association rules.

Attribute/ **Algorithm**	DB Scans	Phases Required (2 in each)	Execution & **Searching** Direction
Apriori Algorithm	N	Join & Prune	1(Bottom Up Direction)
Partition Algorithm	2	Disjoint partitioning of data base & Generation of frequent itemsets.	1 (Depth First Search)
Princers Search Algorithm	N	Finding Frequent itemset & Maximal frequent itemset.	2 (Bi-Directional)
DIC Algorithm	< N	Partitioning of data base into blocks marked by start point & addition of new itemset at start point	1 (Drill Down Method)
FP-Tree Algorithm	1	Divide & Conquer	1 (Drill Down Method)

Table 1. Comparison of various Algorithms used for Discovering of Association Rules

III. MINING MULTIDIMENSIONAL ASSOCIATION RULES

One of the major goals of data mining is to discover association rule. Among the areas of data mining, problem of deriving association rules from data has received a great deal of attention [11].

It is referred as market basket problem. In this problem we are given a set of items and large collection of transaction which are subset (baskets) of these items. The task is to find relationship between the presences of various items within these baskets. Association Rules that involve two or more dimensions or predicates are called

Multidimensional association rules. Multidimensional association rules with no repeated predicates are called inter dimension association rules whereas multidimensional association rules with repeated predicates are called hybrid dimension association rules.

Missing data sets can be problematic and may limit the analysis and extraction of new knowledge [8]. The problem o missing values has been analyzed. R. Agrawal has proposed a fast algorithm to explore very large transactional database with association rules [9] [10]. It uses a carefully tune assessment procedure to find out itemsets that should b

Scanned by CamScanner

easured in a pass. This procedure strikes a balance between e number of passes over the data and the number of itemsets at are measured in a pass by using pruning technique. It also corporates buffer management to handle all the itemsets that eed to be measured in a pass and may not fit in memory, ven after pruning [11]. In many real world applications, data re managed in relational databases where missing values are ften inevitable "To fill the missing values, a relevant ssociation between the attributes of data is required to mine out".

Association type between the database attributes is depending on type of database attributes. Database attributes can be classified as categorical attributes or quantitative attributes [8]. Categorical attributes are also called normal attributes. It has limited number of possible values without any ordering. Whereas quantitative attributes are numeric and have embedded ordering among values. "An attribute is called discrete if it has a less (finite) number of possible values while continuous attribute is considered to have a very large number of possible values (infinite)" [13].

Discretization techniques are frequently used by the lassification algorithms but their applications are not limited to these algorithms. Discretization can also be used by instance-based learning and heritable algorithms. "The goal of discretization is to find a set of cut points to partition the range into a small number of intervals that have good class coherence, which is usually measured by an evaluation function" [14]. Many data mining techniques require that the attributes of the data sets are not continuous but are discrete. If attributes of the data sets are continuous, one has either to opt a different algorithm or to discover a method to discretize the continuous data attributes before applying the desired algorithm. Most of the experimental data are not discrete but are continuous; the discretization of the continuous attributes is key issue. At the same time, some machine learning algorithms that can handle both discrete and continuous attributes perform better with the discrete-valued attributes [15]. Mining of multidimensional association rules for uantitative attributes can be achieved by either using static discretization of quantitative attributes or using dynamic iscretization of uantitative attributes.

In static discretization method, quantitative attributes are discretized prior to mining using concept hierarchies, where as numerical values are replaced by ranges (through classification process of data mining). If resultant data are stored in relational table, slight modification in apriori algorithm will be sufficient to discover all frequent predicate sets rather than finding frequent data sets. To find all frequent K predicate sets, K+1 scans of table are required. Sample data may be used to reduce the number of scans i.e. to improve the performance.

a) Dybamic Discretisation

In dynamic discretization method, during the mining process, quantitative attributes are discretized into bins to satisfy the mining criteria, such as maximizing the confidence [6]. approach to mine association rule having two or m categorical or quantitative attributes on the left side of the and one quantitative attribute on the right side, is called AR (Association Rule Clustering System) [5][16]. In approach, a 2-D grid for tuples satisfying a given categor attribute condition, is formed by mapping pair of quantita attributes.. This grid is then searched for cluster of points f which the association rules are generated.

Discretization of numerical values of attributes for mining quantitative association rule, helps to reduce the tally values. Discretization changes scattered values of quantita attribute to determined values. It helps for machine lear algorithm to perform better for mining the quantita association rules.

IV. MINING ASSOCIATION RULE AGRICULTURAL DATA

Knowledge acquisition and prediction of effective sustainable agriculture has become an important issue agricultural sector, data mining technology can play powerful role. Correct predictions are dependent on accuracy of mined association rules. Traditionally associ rules applied on transactional data that is generally determ on a single dimension
or predicate. However, it is not sufficient for agricultural which involves more than one dimensions or predicates.

An agricultural data warehouse is modelled multidimensional database structure, where each cell of e dimension corresponds to an attribute or set of attributes i schema stores the value of some aggregate measure. actual structure of agricultural data warehouse car represented as multidimensional data cube as shown in fi 1. Data cubes consist of lattice of cuboids that multidimensional data structures and are designed with concept of hierarchies [17]. These structures can hold rele information for each dimension as well as informatio groups.

It is not an easy task to discover the relations in multidimensional data containing missing values of attribute specially when data is agricultural data w outcome of the agricultural production is dependent on va inputs like seeds, fertilizers, manures, soil fertility, irrig methods, temperature of the climate etc. Mining Associ rule, searches for interesting relationships among item given data set so that effects of the yield of the crop ca analyzed on NPK (Nitrogen Phosphorus Potass composition of applied fertilizer on the crop.

Apriori algorithm can't be used to mine the association r multidimensional quantitative data due to the limitation algorithm as discussed earlier in section II that pruning o sets will lead not to discover proper association for represented data and will generate large number of cand and need huge number of data scans. Similarly, the prem small size of frequent set considered in partition algo

Scanned by CamScanner

cannot be accepted for huge multidimensional agricultural data base.

However, if the resultant data are stored in data cube which are well suited for mining of multidimensional association rules then association rules can be mined by single scan only. Following figure shows the lattice of cuboids that defines a data cube for the dimensions Crop, Fertilizer and Yield, assuming other dimensions like irrigation method, soil quality etc. as static. This multidimensional structure can be implemented by either Star Schema or Snowflakes Schema or Galaxy Schema.

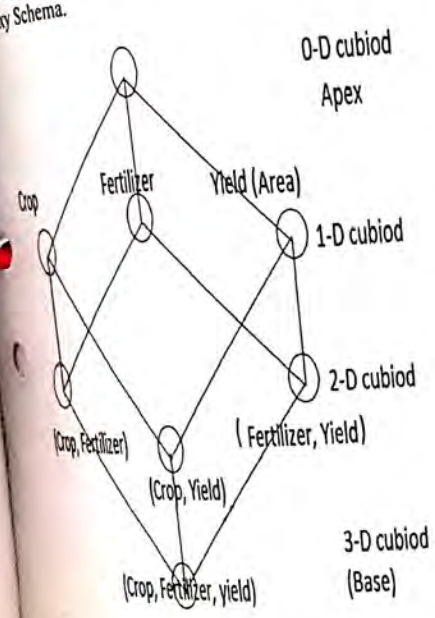

Figure 1. Lattice of cuboids, making up of 3-D data Cube [17]

Following example shows the method to represent the two dimensional quantitative association rules for agricultural data

Crop (Red Gram (0104)) ^ Fertilizer(DAP with NPK (18:46:0))=>Yield(A > 2500 Kg/Hact)

Value of A can be extracted from historical data from data warehouse and based on the pattern, quantitative association rules can be formed to predicate the yield for further decision making.

Table 2. Sample crop codes

Crop Code	Crop Name
0101	Paddy
0102	Maize
103	Red Gram
0104	Bengal Gram
0105	Plantain Tree
0106	Millet

Table 3. Sample fertilizer codes

Fertilizer Code	Fertilizer Name	NPK (Nitrogen Phosphorus Potassium) Composition		
		N	P	K
01	Ammonium Sulphate-Nitrate	20.6	00	00
02	Super Phosphate	00	16	00
03	Urea	46	00	00
04	Di-Ammonium Phosphate	18	46	00

An efficient algorithm is required to mine the quantitative association rules for above mentioned agricultural data. History of previously sown crops along with applied fertilizers and yield is required to define the support (σ) and confidence (T) to identify frequent itemsets for discovery of association rules. Attributes used for mining the association rules will be classified in categorical attributes (Like Crop) or the quantitative attributes (Fertilizer where NPK composition is given, yield i.e. Agricultural Output in Kg/Hectare). If the domain of values for a quantitative attribute is large then before mapping each pair of attribute and interval to Boolean attribute, we first partition the values into intervals. Then we can find the boolean association rule.

But if the number of values (if attribute is not partitioned) or intervals for a quantitative attributes (if attributes are partitioned) are large, the support for any single value/ interval can be low. Hence, some rules involving this attribute may not be found without using larger intervals, because of lack of minimum support(σ).

Similarly, when we partition values into intervals, it is possible that some information may lose.

In that case if an item in the ancestor consists of a unique value or a small interval, such rules may have minimum confidence (T). This loss of information increases as the interval size becomes larger. This is catch-22 situation created by these two problems [18]. If the intervals are too large, some rules may not have minimum confidence; if they are too small, some rules may not have minimum support. To come out from lack of minimum support situation, we can consider all possible continuous ranges over the values of the quantitative attributes, or over the partitioned intervals combine adjacent intervals/values. The minimum confidence problem i.e. information loss can be reduced by increasing the number of intervals, without touching the minimum support situation. When we increase the number of intervals and combine the adjacent intervals simultaneously, it introduces problems of increased execution time and ambiguity of association rules. There is a tradeoff between faster execution time with fewer intervals and reducing information loss with more intervals. We can reduce the information loss by increasing the number of intervals, at the cost of increasing the execution time and potentially generating many uninteresting rules. These uninteresting rules will be pruned out in the last which will increase the execution time.

IJCRCST © 2016 |All Rights Reserved

Scanned by CamScanner

given example, Crop will be considered as categorical ribute, Fertilizer will be considered as discrete quantitative ribute and yield will be considered as continuous antitative attribute which will be discretized and will be artitioned in intervals, to mine association rules. By applying e association rules, classification of new data can be ccomplished so that crops can be classified and required utrition (Composition of NPK) can be predicated for target rop.

Healthy yield of the crop can be harvested, after applying the resultant NPK composition. It will also be helpful to intact the soil fertility for next cropping session.

V. CONCLUSION

In this paper, an attempt is made to summarize all the major techniques of discovering quantitative association rules for large databases. It is very much clear from the discussion that said major techniques are not efficient for multidimensional data like agricultural data stored in data warehouse. The discussion also includes two variants of discretization for multidimensional database. Further, model of agricultural data warehouse has been proposed for multidimensional database structure, where each cell of every dimension corresponds to an attribute or set of attributes in the schema that stores the value of some aggregate measure. Using this 3-D data cube, association rules can be formed for multidimensional data warehouse. The future work will be carried out considering the challenges to collect the historical data of previously sown crops along with applied fertilizers and yield of crops to find the frequent itemsets and defining the minimum support (σ) and minimum confidence (T).

VI. REFERENCES

[1] http://docs.oracle.com/cd/B28359_01/datamine.111/b28129 process.htm, "Oracle® Data Mining Concepts 11g Release 1 (11.1)", 2013.

[2] Jasmine K S, "Data Mining: A Process to Discover Data Patterns and Relationships for Valid Predictions", *Publishing of CSI communications*, vol 34, no 9.

[3] D. Rajesh, "Application of Spatial Data Mining for Agriculture", *International Journal of Computer Applications*, vol. 15, no 2, pp 7-9, 2011.

[4] Waleed A. Aljandal, William H. Hsu, Vikas Bahirwani, Doina Caragea, Tim Weninger, "Validation Based Normalization and Selection of Interestingness Measures for Association Rules", Department of Computing and Information Science, Kansas State University, Manhattan, KS, USA.

[5] Jiawei Han, Micheline Kamber, "Data Mining Concepts and Techniques", New York: Morgan Kaufmann Publishers, 2011.

[6] Guofeng Wang, Xiu yu, Dongbiao Peng, Yinhu Cui, Qiming Li, "Research of Data Mining Based on Apriori Algorithm in Cutting Database", *International Conference on Machanic Automation and Control Engineering*, pp 3765-3768, 2010.

[7] Sachin Sharma, Vidushi Singhal, Seema Sharma, "A Systematic Approach and Algorithm for Frequent Data litemsets", *Journal of Global Research in Computer Science*, vol. 3, no. 11, 2012.

[8] Arun K Pujari, "Data Mining Techniques", University Press, 2009.

[9] B. Santhosh Kumar, K.V.Rukmani, "Implementation of Web Usage Mining Using APRIORI and FP Growth Algorithms", *International Journal of Advanced Networking and applications*, vol. 1, no. 6, pp. 400-404 ,2010.

[10] J. Han, H. Pei, Y. Yin, "Mining Frequent Patterns without Candidate Genration", I*n: Proc. Conf. on the Management of Data* (SIGMOD'00, Dallas, TX). ACM Press, New York, NY, USA,2000.

[11] R. Agrawal, T. Imielinski and A. Swami. "Mining Association Rules Between Sets of Items in Large Databases", *In Proc. of the ACM SIGMOD Conference on Management of Data*, Washington, D.C., pp 207-216, 1993.

[12] R. Agrawal, H. Mannila, R.Srikant, H. Toivonen and A.I. Verkamo, "Fast Discovery of Association Rules in Knowledge Discovery and Data".

[13] Cios. K., Pedrycz, W. Swiniarki, R., Kurgan, L, "Data Mining A knowledge Discovery Approach", Springer, 2007.

[14] Sotiris Kotsiantis, Dimitris kanellopous "DiscretizationTechniques: A recent survey", Educationa Software Development Laboratory, University of patras Greece.

[15] Daniela Joita, "Unsupervised Static Discretizatio Methods in Data Mining", Titu Maiorescu University Bucharest, Romiania.

[16] Brain Lent, Arun Swami, Jennifer widom, "Clustering Association Rules", DCS, Standard University.

[17] Ramakrishnan Srikant, Rakesh Agrawal, "Minin Quantitative Association Rules in Large Relational Tables" ACM SIGMOD Conference, 1996.

[18] Farah Khan, Dr. Divakar Singh, "Association Rul Mining in the Survey" International Journal of Scientific an Research Publications, Volume 4, Issue 7, July 2014.

Scanned by CamScanner