Mining of Quantitative Association Rules in Agricultural Data Warehouse: A Road Map

J. Bhatia1,∗, Anu Gupta2
1Research Scholar, Panjab University, Chandigarh, India
2DCSA, Panjab University, Chandigarh, India

Received: 12 November 2013; Accepted: 15 December 2013

Abstract

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 an endeavour 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; Discretization

©Martin Science Publishing. All Rights Reserved.

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].

∗ J. Bhatia
jbhatia.nielit@yahoo.com
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. Another application of data mining is in telecommunication industry that has quickly changed from providing telephone services to offer many add-on communication services including fax, cellular phone, Internet services. This evolution creates a great demand for data mining in order to understand the business involved, identify telecommunication patterns, catch fraudulent activities, make better use of resources and improve the quality of services. Presently data mining is also used in many scientific applications like Biological Data Analysis, Intrusion Detection and Agricultural sector. 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.

2. Mining the Association Rules

An Association rule is an implication of the form P=>Q, where P ∩ Q = Φ and P & Q are subsets of all itemset I. There are two measures of rule interestingness i.e. Support (σ) and Confidence (T) [4]. They reflect the usefulness (worth) and certainty (assurance) of discovered rules respectively. The rule P=>Q (support σ=5%, confidence T = 60%) shows that 5% of all the transactions under analysis shows the simultaneous purchase of items P and Q by customers and 60% of confidence shows that 60% of customers who purchased item P also bought item Q. 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 phases. 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.

2.1 A Priori Algorithm

It is also called level wise algorithm. It is a seminal algorithm for mining frequent itemset for boolean association rules. 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 k-itemset, Lk” [5]. Two-step process is followed to find Lk. -join step and prune step. In joining step, to find Lk, a set of candidate k-itemsets is generated by joining Lk-1 with itself. This set of candidate is denoted by Ck. The set Ck is superset of Lk, i.e. its members may or may not be frequent, but all of the frequent k-itemsets are built-in Ck. A scan of the database to determine the tally of each candidate in Ck would outcome in determination of Lk (All candidates having minimum count equal to the minimum support). Large size of Ck could involve heavy calculation. In order to diminish the size of Ck 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 Lk-1, then candidate cannot be frequent and can be removed from Ck” [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].

2.2 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.

2.3 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 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.

2.4 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.

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

(b) 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.

2.5 Frequent Pattern (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.

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

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

* N refers to size of large nonempty itemset

3. 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 of missing values has been analyzed. R. Agrawal has proposed a fast algorithm to explore very large transactional databases with association rules [9] [10]. It uses a carefully tuned assessment procedure to find out itemsets that should be measured in a pass. This procedure strikes a balance between the number of passes over the data and the number of itemsets that are measured in a pass by using pruning technique. It also incorporates buffer management to handle all the itemsets that need to be measured in a pass and may not fit in memory, even after pruning [11]. In many real world applications, data are managed in relational databases where missing values are often inevitable “To fill the missing values, a relevant association 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 a continuous attribute is considered to have a very large number of possible values (infinite)” [13]. Discretization techniques are frequently used by the classification 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 quantitative attributes can be achieved by either using static discretization of quantitative attributes or using dynamic discretization of quantitative attributes.
3.1 Static Discretisation

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.

3.2 Dynamic 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]. The approach to mine association rule having two or more categorical or quantitative attributes on the left side of the rule and one quantitative attribute on the right side, is called ARCS (Association Rule Clustering System) [5][16]. In this approach, a 2-D grid for tuples satisfying a given categorical attribute condition, is formed by mapping pair of quantitative attributes. This grid is then searched for cluster of points from which the association rules are generated.

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

4. Mining Association Rule in Agricultural Data

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

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

It is not an easy task to discover the relations in the multidimensional data containing missing values of any attribute specially when data is agricultural data where outcome of the agricultural production is dependent on various inputs like seeds, fertilizers, manures, soil fertility, irrigation methods, temperature of the climate etc. Mining Association rule, searches for interesting
relationships among items in given data set so that effects of the yield of the crop can be analyzed on NPK (Nitrogen Phosphorus Potassium) composition of applied fertilizer on the crop.

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

![Lattice of cuboids](image)

**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 (Wheat(0105)) ^ 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

<table>
<thead>
<tr>
<th>Crop Code</th>
<th>Crop Name</th>
</tr>
</thead>
<tbody>
<tr>
<td>0101</td>
<td>Paddy</td>
</tr>
<tr>
<td>0102</td>
<td>Jowar</td>
</tr>
<tr>
<td>0103</td>
<td>Bajra</td>
</tr>
<tr>
<td>0104</td>
<td>Maize</td>
</tr>
<tr>
<td>0105</td>
<td>Ragi</td>
</tr>
<tr>
<td>0106</td>
<td>Wheat</td>
</tr>
</tbody>
</table>

**Table 3.** Sample fertilizer codes

<table>
<thead>
<tr>
<th>Fertilizer Code</th>
<th>Fertilizer Name</th>
<th>NPK (Nitrogen Phosphorus Potassium) Composition (In %)</th>
</tr>
</thead>
<tbody>
<tr>
<td>01</td>
<td>Ammonium Sulphate- Nitrate</td>
<td>20.6 00 00</td>
</tr>
<tr>
<td>02</td>
<td>Super Phosphate</td>
<td>00 16 00</td>
</tr>
<tr>
<td>03</td>
<td>Urea</td>
<td>46 00 00</td>
</tr>
<tr>
<td>04</td>
<td>Di-Ammonium Phosphate</td>
<td>18 46 00</td>
</tr>
</tbody>
</table>

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 trade off 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.

In given example, Crop will be considered as categorical attribute, Fertilizer will be considered
as discrete quantitative attribute and yield will be considered as continuous quantitative attribute
which will be discretized and will be partitioned in intervals, to mine association rules. By
applying the association rules, classification of new data can be accomplished so that crops can
be classified and required nutrition (Composition of NPK) can be predicated for target crop.
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.

5. 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).

References