Association Rule Mining in the field of Agriculture: A Survey

Farah Khan[*], Dr. Divakar Singh[**]

[*] UIT, BU , Bhopal, MP, India
[**] UIT, BU , Bhopal, MP, India

Abstract- Data mining is the art and science of intelligent analysis of (usually large) data sets for meaningful (and previously unknown) insights and is nowadays actively applied in a wide range of disciplines related to agriculture. Association Rule Mining is a powerful tool for generating rules from vast and diversified data such as agricultural datasets. Due to the emerging importance of data mining techniques and methodologies in the area of agriculture, this paper is a survey of some previous researches done in this field. For sustainable growth of agriculture, these methodologies need to be monitored and analyzed optimally.

Index Terms- Association rule mining, Agricultural data, Data mining.

I. Introduction

Generally, data mining [1] (sometimes called data or knowledge discovery) is the process of analyzing data from different perspectives and summarizing it into useful information. Data mining software is an analytical tool that allow users to analyze data from many different dimensions or angles, categorize it, and summarize the relationships identified. Technically, data mining is the process of finding correlations or patterns among dozens of fields in large relational databases. The patterns, associations, or relationships among all this data can provide *information*.

Information can be converted into *knowledge* about historical patterns and future trends. For example, summary information about crop production can help the farmers identify the crop losses and prevent it in future. Use of data mining techniques can provide more suitable system for the decision making. Today, data mining is used in numerous 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 behaviors, 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 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.

II. ASSOCIATION RULE MINING

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) . They reflect the usefulness and certainty of the rules . The rule P=>Q (support σ = 10%, confidence T = 80%) shows that 10% of all the transactions under analysis shows the simultaneous purchase of items P and Q by customers and 80% of confidence shows that 80% of customers who purchased item P also bought item Q [2].

Association rules relate objects 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).

Various algorithms have been proposed for mining the association rules and can be decomposed in two phases.

I. Find all the itemsets whose support and confidence are greater than the user specified minimum support (σ) and minimum confidence (T) respectively. Such items are called frequent itemsets.

II. Frequent items are used to find desired association rules. These rules must satisfy minimum support (σ) and minimum confidence (T).

The five major algorithms proposed for discovery of association rules, are as follows:

2.1. A Priori Algorithm

The process is divided in two steps (R.Aggarwal & Srikant) -

1. Minimum support is applied to find all frequent itemsets in a database.
2. These frequent itemsets and the minimum confidence constraint are used to form rules.

2.2. Partition Algorithm

Partition algorithm [3] works in two scans of the database. In one scan it generates a set of all potentially large itemsets by scanning the database once. This set is a superset of all large itemsets, i.e. it may contain false positives. But no false negatives are reported. During the second scan, counters for each of these itemsets are set up and their actual support is measured in one scan of the database.

The algorithm executes in two phases. In the first phase, the Partition algorithm logically divides the database into a number of non-overlapping partitions. The partitions are considered one at a time and all large itemsets for that partition are generated. At the end of phase I, these large itemsets are merged to generate a set of all potential large itemsets. In second phase, the actual support for these itemsets are generated and the large itemsets are identified. The partition sizes are chosen such that each partition can be accommodated in the main memory so that the partitions are read only once in each phase.

2.3. Pincer search Algorithm

The algorithm [4] begins with generating 1-itemsets as A Priori algorithm but uses top-down search to prune candidates produced in each pass. This is done with the help of MFCS set. Let MFS denote set of Maximal Frequent sets storing all maximally frequent itemsets found during the execution. So at anytime during the execution MFCS is a superset of MFS. Algorithm terminates when MFCS equals MFS. In each pass over database, in addition to counting support counts of candidates in bottom-up direction, the algorithm also counts supports of itemsets in MFC, this set is adapted for top-down search.

2.4. Dynamic Itemset Counting Algorithm

The DIC algorithm [5] works as follows :

Step 1. The empty itemset is marked with a solid box. All the 1-itemsets are marked with dashed circles. All other itemsets are unmarked.

Step 2. Read *M* transactions. Let the values of *M* range from 100 to 10,000. For each transaction, increment the respective counters for the itemsets marked with dashes.

Step 3. If a dashed circle has a count that exceeds the support threshold, turn it into a dashed square. If any immediate superset of it has all of its subsets as solid or dashed squares, add a new counter for it and make it a dashed circle.

Step 4. If a dashed itemset has been counted through all the transactions, make it solid and stop counting it.

Step 5. If we are at the end of the transaction file, rewind to the beginning.

Step 6. If any dashed itemsets remain, go to Step 2.

This way DIC starts counting just the 1-itemsets and then quickly adds counters 2,3,4,...,k-itemsets. After just a few passes over the data (usually less than two for small values of *M*) it finishes counting all the itemsets.

Since agricultural data is diversified in terms of its attributes and needs to be processed efficiently and well in time and 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.

Whereas, the frequent-pattern tree (FP-tree) is a compact structure that stores quantitative information about frequent patterns in a database. 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.

2.5 FP-Tree Growth Algorithm

This method, [6] 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 and mines each such database separately .The method considerably decreases the search cost. 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.

III. LITERATURE SURVEY

3.1. Application of Spatial Data Mining for Agriculture

D.Rajesh [7], has provided an overview of data clustering methods using cluster analysis for generating patterns and rules. Since association rule mining needs to evaluate multiple spatial relationships among a large number of spatial objects, the process could be cumbersome, so another method called progressive refinement can be used in spatial association analysis. His work is in the direction of extracting patterns from spatial database using k-means algorithm. Large data sets are mined roughly using a fast algorithm and then improves the quality of mining in a pruned data set. The above concept is applied in the area of agriculture where giving the temperature and the rainfall as the initial spatial data and then by analyzing the agricultural meteorology for the enhancement of crop yields and also reduce the crop losses.

3.2 Developing Innovative Applications in Agriculture using Data mining

Sally Jo Cunningham and Geoffrey Holmes [8] have described a WEKA (Waikato Environment for Knowledge Analysis) system which provides a complete suite of facilities for applying data mining techniques to large data sets. This WEKA-based analysis and application construction process is illustrated through a case study in the agricultural domain—mushroom grading.

The mined data is represented as a model of the semantic structure of the dataset, and it can be used for pattern discovery. For accomplishing the purpose, machine learning techniques have been used for analyzing data.

33.Data Mining for Evolution of Association rules for Droughts and Floods in India using Climate Inputs

C. T. Dhanya and D. Nagesh Kumar [9] have worked in the direction of designing an effective risk management system for tracing frequent occurrences of droughts and floods because these two factors directly affect the Indian agriculture. A data-mining algorithm is used to discover association rules between extreme rainfall events and climatic indices. Association rules are generated for the regions of India which shows strong relationships between the climatic indices chosen, i.e., Darwin sea level pressure, North Atlantic Oscillation, Nino 3.4 and sea surface temperature values, and the extreme rainfall events.

3.4. PARM -- An Efficient Algorithm to Mine Association Rules From Spatial Data

[10]Association rule mining is applied to remote sensed imagery (RSI) data composed mainly of images and ground data mainly from the field of agriculture. In most of the cases applying existing algorithms on RSI data for generating association rules can consume a reasonable amount of time . Keeping that in mind an efficient algorithm has been devised for spatial data using Peano count tree (P-tree) structure. P-tree structure provides a lossless and compressed representation of images. Based on P-trees, an efficient association rule mining algorithm PARM with fast support calculation and significant pruning techniques is introduced to improve the efficiency of the rule mining process. Experimental results showed that PARM is more efficient than FP- growth and A Priori algorithms when applied on RSI spatial data.

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

J. Bhatia and Anu Gupta [11] , have done a comparative study on various association rule mining techniques, which include A Priori Algorithm, Partition Algorithm , Pincer Search Algorithm, Dynamic Item Set(DIC) Algorithm and FP-tree growth algorithm. When these are applied in agricultural domain the best results are generated by FP-Tree growth Algorithm. Since it scans the database the least number of times from all the above algorithms, it gives the lowest time complexity. Apart from this the agricultural dataset is stored in the form of a data cube to make it concise, so that it can be monitored, analyzed and allocated optimally.

3.6. The Application of Association Rule Mining to Remotely Sensed Data

In this paper [12], we defined a new data mining problem -- Association rule mining from imagery data and its application in agricultural domain. A priori algorithm has been used as base algorithm for inventing a new data mining technique which provides effective pruning for candidate 2-item set generation. According to this novel technique a significant amount of unnecessary candidate itemsets can be pruned during the early phase of mining.

3.7. Data Mining Techniques and Applications to Agricultural Yield Data

In this paper by D Ramesh and B Vishnu Vardhan [13], different Data Mining techniques , such as K-Means, K-Nearest Neighbour(KNN), Artificial Neural Networks(ANN) and Support Vector Machines(SVM) were adopted to estimate crop yield analysis with existing data. Multiple Linear Regression (MLR) is used to model the linear relationship between a dependent variable and one or more independent variable(s). The dependent variable is rainfall and independent variables are year, area of sowing and production .By adopting K-Mean clustering approach four clusters are formed .

3.8. Data Mining Techniques: A Tool for Knowledge Management System in Agriculture

Latika Sharma and Nitu Mehta [14] have attempted to bring out the computational needs of agriculture data and how data mining techniques can be used as a tool for knowledge management in agriculture. Data warehouses can be prepared to hold agriculture data, which makes transaction management, information retrieval and data analysis much easier. On Line Analytical Processing (OLAP) can easily answer multi-dimensional queries it can be used for applications such as forecasting or prediction in agriculture. It also provides an opportunity of viewing agriculture data from different points of view to discover data characterization, data discrimination and association analysis.

3.9. Classification of Agricultural Land Soils -- A Data Mining Approach

Ramesh Vamanan and K.Ramar [15] have surveyed the various classification techniques of data mining to be applied on soil database to establish meaningful relationships. The characteristics used to classify soil are as follows- soil moisture regimes, soil temperature regimes and physical and chemical properties of soil. Furthermore, the survey showed that soil can be divided into eight classes depending upon its agricultural productivity.

IV. CONCLUSION

This paper is an Endeavour to provide an overview of some previous researches and studies done in the direction of applying data mining and specifically, association rule mining techniques in the agricultural domain. We have also tried to evaluate the current status and possible future trends in this area. The theories behind data mining and association rules are presented at the beginning and a survey of different techniques applied is provided as part of the evolution.

REFERENCES

[1] Jaiwei Han and Micheline Kamber, “Data Mining Concepts and Techniques”, Second Edition , Morgan Kaufmann Publishers.

[2] “Association Rules Mining” ,Sotiris Kotsiantis, Dimitris Kanellopoulos.

[3] Ashok Savasere and Shamkant Navathe. “An Efficient Algorithm for Mining Association Rules in Large Databases.”

[4] Pincer-Search Algorithm for Discovering Maximum Frequent Sets:Notes by Akash Saxena.

[5] Sergey Brin, Rajeev Motwani, Jeffrey D. Ullman ,Dynamic Itemset Counting and Implication Rules for Market Basket Data.

[6] Jiawei Han, Jian Pei and Yiwen Yin, “ Mining Frequent Patterns without Candidate Generation.”

[7] D.Rajesh ,International Journal of Computer Applications ,Volume 15, February 2011.

[8] Sally Jo Cunningham and Geoffrey Holmes, Department of Computer Science University of Waikato Hamilton, New Zealand.

[9] C. T. Dhanya and D. Nagesh Kumar, Journal of Geophysical Research, Vol. 114.

[10] Qin Ding Dept of Comp. Sci. East Carolina Univ,Greenville,NC,Systems,Man and Cybemetics, IEEE Transactions(Volume 38, Issue :6)

[11] J.Bhatia and Anu Gupta,International Journal of Information Science and Intelligent System,2014.

[12] Jianning Dong, William Perrizo, Qin Ding and Jingkai Zhou.

[13] D Ramesh and B Vishnu Vardhan, International Journal of Advanced Research in Computer and Communication Engineering ,Vol. 2, Issue 9, September 2013.

[14] Latika Sharma and Nitu Mehta, INTERNATIONAL JOURNAL OF SCIENTIFIC & TECHNOLOGY RESEARCH VOLUME 1, ISSUE 5, JUNE 2012.

[15] Ramesh Vamanan and K.Ramar, International Journal on Computer Science and Engineering (IJCSE).

AUTHORS

First Author – Farah Khan, UIT, BU , Bhopal, MP, India, Email: farah86khan@gmail.com
Second Author – Dr. Divakar Singh, UIT, BU , Bhopal, MP, India, Email: divakar_singh@rediffmail.com