ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 2, Issue 11, November 2014

Rule Based and Association Rule Mining On Agriculture Dataset

Dr. Rahul G. Thakkar, Dr. Manish Kayasth, Hardik Desai

Assistant Professor, ASPEE Agribusiness Management Institute, Navsari Agricultural University, Navsari, India

Principal, UCCC & SPBCBA & Udhna Academy College of Computer Application and IT, Surat, Gujarat, India

Assistant Professor, Naran Lala College of Professional & Applied Sciences, Navsari, India

ABSTRACT: The wide availability of huge amounts of agriculture data has generated an urgent need for the research of data mining. Generating rules with higher accuracy for Agriculture databases can be done using different techniques of data mining. As the analysis of agriculture dataset is usually a complex work. Two different techniques were demonstrated for mining agriculture dataset, Association rule mining and Classification technique. Both of them are successful. It is always difficult to select the appropriate data mining algorithm for the specific database, there can be many algorithms through which rules can be generated but it is always a problem to get rules with higher accuracy. Research mainly emphasizes on different algorithm for mining agriculture data.

KEYWORDS*:* Agriculture dataset, Data Mining, Association rule, Classification Techniques.

I. INTRODUCTION

Currently, databases of agriculture are becoming larger and larger in each and every sector due to the growth in IT industry. Usage of data mining is becoming more and more popular to convert the huge data into knowledge. In generating knowledge from the massive database, there is always a problem in selecting the appropriate technique of data mining. Considering agriculture atabase there can be data of two types: fundamental and technical. Both types of data has its own importance and can be used for mining purpose to discover knowledge and get huge returns on investment in stock for longer or shorter period. Different techniques becoming more and more popular in mining agriculture database

Appropriate algorithm selection for association rule or classification technique for agriculture database is critical in discovering knowledge. Rules generated from classification technique engage prediction of single attribute (class attribute) while association rule generation can engage any of the attribute. Both techniques are popular enough for mining stock market database through different algorithms like Apriori, ID3, C4.5, etc.

Association rule mining finds interesting association or correlation relationships among a large set of data items. With massive amounts of data continuously being collected and stored, many industries are becoming interested in mining association rules from their database. The discovery of interesting association relationship among huge amounts of business transaction records can help in many business decision making processes [2]. Classification and prediction are two forms of data analysis that can be used to extract models describing important data classes or to predict future data trends. Whereas classification predicts categorical labels, prediction models continuous-valued functions. Many classification and prediction methods have been proposed by researchers[2].

In this paper, most influential data mining algorithms (Apriori, ID3 and C4.5) in the research community are demonstrated that can be applied on the agriculture data to generate the effective rules.

Among number of Association Rule Mining algorithms researcher has demonstrated Apriori[3] to extract hidden knowledge from the huge agriculture database. While to perform classification tasks Decision Tree[9] and C4.5[9] algorithms were demonstrated.

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 2, Issue 11, November 2014

Present research demonstrates three popular algorithms that represent two different data mining techniques. One is Apriori of Association Rule Mining and the others are ID3 and C4.5 algorithm of Classification.

II. LITERATURE REVIEW

In [3] authors used large database of customer transactions, in which each transactions consists of items purchased by a customer in a visit. Authors have presented results of an efficient algorithm that generates all significant association rules between items in the database. In [6] a data mining solution was develop to diagnosis tuberculosis as accurate as possible and helps deciding if it is reasonable to start tuberculosis treatment on suspected patients without waiting the exact test results or not. Authors focused on three different data mining methods (Adaptive Neuro Fuzzy Inference System (ANFIS), Multilayer Perceptron and Partial Decision Trees). In [17] problem of yield prediction was solved by employing Data Mining techniques. Research paper aims at finding suitable data models that achieve a high accuracy and a high generality in terms of yield prediction capabilities. For this purpose, different types of Data Mining techniques were evaluated on different data sets. In view of [18], there is a need for an objective methodology for pre-harvest crop forecasting which may involves building up suitable forecast model(s) which has certain merits over the traditional forecasting methods.

III. APRIORI, ID3 AND C4.5 ALGORITHMS

Study portrays the methodology to demonstrate the effectiveness of different algorithm on agriculture dataset. The emphasis is on data mining techniques applied to know the future price movement of agriculture crops. Attributes like crop, date, time, minimum market price, maximum market price, moderate price and arrivals can be used for prediction of price movement. Agriculture database can consist of past few years price data of major agriculture crops, which can be used as a raw data for analysis and to generate the rules from different algorithms of data mining techniques.

Popular techniques have been demonstrated. The algorithms under study are the Apriori, ID3 and C4.5. The choice of such methods is based on the multiplicity of strategies they use to produce the rules.

A. Apriori algorithm

Pre-processing of numerical data will improve the quality of association rules[14]. After pre-processing numerical data on the agriculture price data of major agriculture crops training data set can be obtained. The training data set can be divide day into three parts based on prices obtained in morning, afternoon and evening sessions. Training data set obtained after pre-processing steps can be used for the comparison with Day attribute indicates the day of week. The value of trend can be derived from the values of minimum price, maximum price and moderate price of major agriculture crops based on specific market/area. Finding frequent relationship between trends of different sessions on a particular day with minimum support of more than 20 is important because of its combinatorial explosion. Once frequent patterns obtained, it is straightforward to generate association rules with confidence larger than or equal to 40. Apriori is a seminal algorithm for finding frequent pattern using candidate generation. Here database can be sorted on the basis of fields like crops, day and session. Let the set of frequent trend of session on particular day for particular crop is of size k represented as Fk and their candidates be Ck. Apriori first scans the database and searches for frequent trend of session on particular day for particular crop in specific market of size 1 by accumulating the count for each and collecting those that satisfy the minimum support requirement. Then it iterates on the following three steps and extracts all the frequent trend of session.

1. Generate Ck+1, candidates of frequent trend of session on particular day for particular crop in specific market of size k +1, from the frequent trend of session on particular day for particular crop in specific market of size k.
2. Scan the database and calculate the support of each candidate of frequent trend of session on particular day for all crops in specific market.
3. Add those trends of session on particular day for particular crop in specific market that satisfies the minimum support requirement to Fk+1.

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 2, Issue 11, November 2014

The Apriori algorithm generates Ck+1 from Fk in the following two step process:
1. Join step: Generate RK+1, the initial candidates of frequent trend of session on particular day for particular crop in specific market of size k + 1 by taking the union of the two frequent trend of session on particular day for particular crop in specific market of size k, Mk, Pk and Qk that have the day and crop in specific market in common.
2. Prune step: Check if all the trends of first, second and third session on particular day for all crop in specific market of size k in Rk+1 are frequent and generate Ck+1 by removing those that do not pass this requirement from Rk+1. Apriori, then, calculates frequency only for those candidates generated this way by scanning the database.
Accuracy can be calculated using Hold-out method.

B. ID3 Algorithm

Percentage change for minimum price, maximum price and moderate price can be calculated by comparing record of second session with the record of first session and for record of third session with the record second session. On the basis of the percentage change in moderate price attribute, valuation for the trend can be decide for each session in comparison to previous session on a daily basis using day attribute for all agriculture crops on the basis of specific market.
The ID3 algorithm on each iteration calculates the entropy H(S) (or information gain IG(A)) of unused attribute, here it is Percentage change in minimum price, maximum price, moderate price. Then selects the attribute which has the smallest entropy (or largest information gain) value. The algorithm continues to recurse on each subset, considering only attributes never selected before.
Recursion on a subset may stop in one of these cases:

- every element in the subset belongs to the same class (Up or Average or Low), then the node is turned into a leaf and labelled with the class of the examples
- there are no more attributes to be selected, but the examples still do not belong to the same class (Up or Average or Low), then the node is turned into a leaf and labelled with the most common class of the examples in the subset.
- there are no examples in the subset, this happens when no example in the parent set was found to be matching a specific value of the selected attribute, for example if there was no example with Percentage change in minimum price >= 0.5, then a leaf is created, and labelled with the most common class of the examples in the parent set.

Throughout the algorithm, the decision tree is constructed with each non-terminal node representing the selected attribute (Percentage change in minimum price, Percentage change in maximum price, Percentage change in moderate price) on which the data was split, and terminal nodes representing the class (Up or Average or Low) of the final subset of this branch.

C. C4.5 Algorithm

C4.5 is a program used for creating classification rules using decision trees from a set of given data. C4.5 algorithm is an extension of the basic ID3 algorithm and it was designed by Quinlan. C4.5 algorithm builds decision trees from a set of training data similar to the ID3algorithm, using the concept of information entropy[16]. C4.5 is a program that inputs a same set of data that is used for ID3 algorithm and generates a decision tree as output. This resultant decision tree is then tested against testing agriculture dataset

IV. CONCLUSIONS

This paper demonstrates/investigates the associations rule mining algorithm and classification techniques rule mining. The main contributions are:

- Theoretical survey on association rule mining algorithm like Apriori and also on classification techniques algorithms like ID3 and C4.5 on agriculture crops.
- Comparison of association rule mining like Apriori algorithm and classification techniques algorithms like ID3 and C4.5. This includes the different strategies they employ to extract the rules from data sets.

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 2, Issue 11, November 2014

REFERENCES

[1]Alor-Hernandez, G., Gomez-Berbis, J. M., Jimenez-Domingo, E., Rodríguez-González, A., Torres-Niño, J., "AKNOBAS: A Knowledge-based Segmentation Recommender System based on Intelligent Data Mining Techniques". Computer Science and Information Systems, Vol. 9, No. 2, 2012, pp: 713-740.
[2]Han, J., Kamber, M., "Data mining concepts and techniques 2nd edithion". Morgan Kaufman, 2006, pp: 227-378.
[3]Agrawal R., Imielinski T. "Mining associations between sets of items in large databases". Proceedings of the ACM SIGMOD International Conference on Management of Data. pp: 207-216.
[4]Konda, S., "Web Data Mining Based Business Intelligence and Its Applications". IJCST, Vol. 4, No. 4, 2013, pp: 112-116.
[5]Aher, B., "Association Rule Mining in Data Mining". IJCST, Vol. 4, No. 3, 2013.
[6]Nagabhushanam, D., Naresh., N., "Prediction of Tuberculosis Using Data Mining Techniques on Indian Patient's Data". IJCST, Vol. 4, No. 4, 2013, pp: 262-265.
[7]Surya, K., Priya, K., "Exploring Frequent Patterns of Human Interaction Using Tree Based Mining". IJCST, Vol. 4, No. 4, 2013.
[8]Rajnikanth, J., "Database Primitives, Algorithms and Implementation Procedure: A Study on Spatial Data Mining". IJCST, Vol. 4, No. 2, 2013.
[9]Quinlan, J. R., "C4.5: Program for machine learning. CA". Morgan Kaufmann, San Francisco, 1992
[10]Clark, P., Niblitt, T., "The CN2 induction algorithm. Machine Learning". Vol. 3, No. 4, 1989, pp: 261-283.
[11]Cohen, W.(1995). "Fast effective rule induction". Proceedings Twelth International Conference on Machine Learning. Pp: 115-123
[12]Duda, R., Hart, P., "Pattern classification and scene analysis". Wiley, New York, 1973
[13]Ardakani, H. D., Hajizadeh, E., Shahrabi, J., "Application of Data Mining Techniques in stock markets: A survey". Journal of Economics and International Finance, Vol. 2, No. 7, 2010, pp: 109-118.
[14]Lopez, V. F., Moreno, M. N., Polo, M. J., Segrera, S., "Improving the Quality of Association Rules by Preprocessing Numerical Data". II Congreso Espanola Informatics, 2007, pp: 223-230.
[15]Chung, F. L., Fu, T. C., Ting, J. (2006), "Mining of Stock Data: Intra-Stock and Inter-Stock Pattern Associative Classification". Proceedings International Conference on Data Mining, pp: 30-36.
[16]Frank, E., Witten, I. H., (1998). "Generating accurate rule sets without global optimization". Proceedings Fifteenth International Conference on Machine Learning, pp: 141-151.
[17]Ramesh D, Vishnu Vardhan B., "Data Mining Techniques and Applications to Agricultural Yield Data", IJARCCE, Vol. 2, Issue 9, September 2013.
[18]Raorane A.A., Kulkarni R.V, "Review- Role of Data Mining in Agriculture", IJCSIT, Vol. 4, No. 2, pp: 270-272, 2013
[19]https://www.agmarknet.nic.in/
[20]https://cacp.dacnet.ni