Model to Predict Price Movement of GWAR for North Gujarat Using Classification Rules: A Data Mining Approach

Dr. Rahul G. Thakkar, Hardikkumar V. Desai, Dr. Manish Kayasth

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

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

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

ABSTRACT: Research paper is based on decision tree (induction algorithm) which generates classification rules that will help in knowing next day price movement of the crop. The research paper provides a glimpse of the market and selling tips. We have used classification rule generation method of Data Mining. Research paper predicts the next day trend of Gwar based on daily price movement of the Gwar (Arrival, Minprice, Maxprice) as compared to that of the previous day price movement of the Gwar (Arrival, Minprice, Maxprice). We will considered those classification rules which have accuracy more than ninety.

KEYWORDS: Agriculture Market, Data Mining, Binning, Decision Tree, Classification Rules, Support, Confidence, Accuracy.

I. INTRODUCTION

The Agricultural Produce Market Committee is a marketing board established by the state governments of India (e.g., Gujarat and Maharashtra). The state government of Gujarat, in order to facilitate farmers to sell their produce and get reasonable price constituted APMC in many towns. Most of the APMC have market yard where traders and other marketing agents are provided go-downs and shops for purchase of agriculture produce from farmers. Farmers can sell their produce to agents or traders under supervision of APMC. We will build up our database by fetching daily data **(Arrival, Minprice, Maxprice)** from Agmark.ac.in. Data Mining will be used to discover hidden values from the database. It is a powerful technology with a great potential to focus on the most important information in data warehouses. It analyzes relationships and patterns in stored data.

Decision trees are powerful and popular tools for classification and prediction. We have used ID3 technique which searches through the attributes of the training instances and extracts the attribute that best separates the given examples. If the attribute perfectly classifies the training sets then ID3 stops; otherwise it recursively operates on the m (where m = number of possible values of an attribute) partitioned subsets to get their "best" attribute. The algorithm uses a greedy search, that is, it picks the best attribute and never looks back to reconsider earlier choices. Each discovered pattern should have measure of certainity associated with it that assesses the validity or "trustworthiness" of the pattern. A certainty measure for rules the form "A=>B" is confidence. The support of a pattern refers to the percentage of task-relevant data tuples for which the pattern is true.

II. RELATED WORK

In [1] 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. A new method for learning algorithm evaluation and selection, with empirical results based on Classification is introduced in [2]. The empirical study has been conducted among 8 algorithms/classifiers with 100 different classification problems. Research empirical results are used to generate rules,

using the rule-based learning algorithm, to describe which types of algorithms are suited to solving which types of classification problems. Most of the rules are generated with a high confidence rating. In [8] 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 [9] 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.

III.   STEPS INCLUDED IN GENERATING CALSSIFICATION RULES

A. Building Database
A basic requirement for the research is to get the historical data on the daily basis having all data regarding arrival, minprice and maxprice of Gwar in north Gujarat APMC markets, www.agmark.ac.in provides historical data. So we will use it as a main source for the data.

B. Cleaning
We will check for the missing values. If missing values will be found then past 10 trading day prices will be taken for that particular field and average will be calculated of that arrivals or prices to fill out the missing value. Following fields will be taken to build up the database:
- Market
- Date
- Arrival
- Minimum price
- Maximum price
- Modal price

C. Calculating percentage change
To generate a decision tree, we need a percentage change for arrival, minprice and maxprice as compare to that of previous day arrivals and prices.

1)   Deciding valuation: Based on the percentage change of Modalprice a valuation of previous day record will be decided, valuation can be fairly_valued, under_valued or over_valued. The valuation will be decided on the basis of the following criteria:
If percentage change of a Modalprice is
- $>=5\%$ then valuation of previous day record is "Undervalued"
- Between -5% and 5% then valuation of previous day record is "Fairlyvalued"
- $<=-5\%$ then valuation of previous day record is "Overvalued"

D. Binning
Binning will be done on each and every field of database for each Gwar record. Binning value will replace the original value which will be calculated by applying sorting on each and every attribute of each record of Gwar. Total number of values in each bin will be calculated on the basis of total number of records for a Gwar divided by ten. It means that we will allow maximum ten bins.

E. Rule generation
The algorithm computes the information gain of each attribute. The attribute with the highest information gain is chosen as the test attribute for the given database. A node is created and labeled with the attribute, branching are created for each value of the attribute and the samples are partitioned accordingly.
Description of algorithm is given below
Create a node N;
If samples are all of the same class, C then
        Return N as a leaf node labeled with the class C;
If attribute-list is empty then
Return N as a leaf node labeled with the most common class in samples;
Select test-attribute, the attribute among attribute-list with highest info. Gain;

Label node N with test-attribute;

For each known value ai of test-attribute

Grow a branch from node N for the condition test-attribute= ai

Let si be the set of samples in samples for which test-attribute= ai

     If si is empty then

Attach a leaf labeled with the most common class in samples;

     Else

          Attach the node returned by the algorithm

F. Tree Pruning

We have selected pre-pruning approach where a tree is "pruned" by halting its construction early (by deciding not to further split or partition the subset of training samples at a given node). Upon halting, the node becomes a leaf. The leaf holds the most frequent class among the subset samples or the probability distribution of those samples.

G. Support & Confidence

The decision tree can be converted to classification IF_THEN rules by tracing the path from the root node to each leaf node in the tree. We will calculate support and confidence for each classification rule that is that is converted into IF_THEN rule in the following manner.

Rules that satisfy both a minimum support threshold and a minimum confidence threshold are called strong.

H. Support

The rule A=>B holds in the transaction set D with support s, where s is the percentage of transactions in D that contain AnB (i.e., both A and B). This is taken to be the probability, P (AnB).

For each rule generated by ID3 Technique we will calculate support. The rule holds in the training data set with support s, where s is the percentage of transactions in training data set that contains both IF and THEN part. This is taken to be the probability that both occur. We have considered minimum support of twenty.

I. Confidence

The rule A=>B has confidence c in the transaction set D if c is the percentage of transactions in D containing A that also contain B. This is taken to be the probability,   P (B|A).

For each rule generated by ID3 Technique we will calculate confidence. The rule holds in the training data set with confidence c, where c is the percentage of transactions in training data set that contains IF part that also contains THEN part. We have considered minimum confidence of ninety.

J. Accuracy (Hold-Out method)

We will use hold-out method for determining accuracy in which two thirds of the data are allocated to the training set, and the remaining one third is allocated to the test set. The training set is used to derive the classifier, the accuracy of which is estimated with the test set. Classification rules generated having minimum support of twenty, minimum confidence of eighty and minimum accuracy of 90 will only be considered in our model.

IV.   CONCLUSION

"Technical approach" is developed for the prediction of next day trend of Gwar. For the Farmers it's interesting if they can know the next day movement before one day. Our classification rules can helps in predicting next day movement of the Gwar in north region of gujarat. So that before selling Gwar farmer may assure profit or loss percent.

REFERENCES

[1] 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.

[2] Ali, A.B.M.S., Smith, K. A. (2004). "On learning algorithm for classification". Applied Soft Computing, pp: 119-138.

[3] Ali, A.B.M.S., Wasimi, S. A. (2007). "Data mining: methods and techniques". Thomson, Victoria, Australia.

[4] 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.

[5] Cohen, W.(1995). "Fast effective rule induction". Proceedings Twelth International Conference on Machine Learning. Pp: 115-123

[6] Duda, R., Hart, P., "Pattern classification and scene analysis". Wiley, New York, 1973

[7] Han, J., Kamber, M., "Data mining concepts and techniques 2nd edithion". Morgan Kaufman, 2006, pp: 227-378.

[8] 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.

[9]Ramesh D, Vishnu Vardhan B., "Data Mining Techniques and Applications to Agricultural Yield Data", IJARCCE, Vol. 2, Issue 9, September 2013.

[9] Surya, K., Priya, K., "Exploring Frequent Patterns of Human Interaction Using Tree Based Mining". IJCST, Vol. 4, No. 4, 2013.

[10]https://www.agmarknet.nic.in/

[11]https://cacp.dacnet.nic.in