Data Mining in Agriculture: A Review

***K Raghuveer**
**** M J Yogesh**
***** Shwetha S**

*Professor, Department of Information Science, NIE, Mysore
**Assistant Professor, Department of Computer Science and Engineering, NIE, Mysore
***Student, Department of PG Studies & CA, NIE, Mysore

Abstract- Data mining in application in agriculture is a relatively new approach for forecasting/predicting the crop yield in advance for market dynamics. In this paper an attempt has been made to review the research studies on application of data mining techniques in the field of agriculture. Some of the techniques, such as the k-means, the k nearest neighbor, Decision Tree applied in the field of agriculture were presented.

Keywords – Data Mining, Agriculture, Market Dynamics

1. Introduction

Data Mining is the process of extracting useful and important information from large sets of data. In this paper describe an overview of Data Mining techniques applied to agricultural and their applications to agricultural related areas. Yield prediction is a very important agricultural problem. Any farmer is interested in knowing how much yield he is about to expect. In the past, yield prediction was performed by considering farmers experience on particular field and crop. The production of crop growth monitoring can provide important information for government agencies, commodity firms and producers in planning transport activities, market prices etc.. Agricultural productivity is sensitive to two broad classes of climate induced effects direct effects from changes in temperature, precipitation, or carbon dioxide concentrations, and indirect effects through changes in soil moisture and the distribution and frequency of infestation by pests and diseases.

Different techniques were proposed for mining data over the years. . In this paper we present some of the most used general Data Mining techniques in the field of agriculture.

The findings of the study revealed that the decision tree analysis indicated that the productivity of soybean crop was mostly influenced by Relative humidity followed by rainfall and temperature. The decision tree analysis indicated that the productivity of paddy crop was mostly influenced by Rainfall followed by Relative humidity and Evaporation. For Wheat crop the analysis indicated that the productivity is mostly influenced by Temperature followed by Relative humidity and Rainfall. The findings of decision tree were confirmed from Bayesian classification. The decision tree in the study area fast to execute and much to be desired as representations of knowledge interpretations The rules formed from the decision tree are helpful in identifying the conditions responsible for the high or low crop productivity.

Soybean is one of the most predominant crops cultivated in the state of Madhya Pradesh, India. Over the past two decades the productivity of the crop has been in declining trend despite the area under the crop is increasing. Using the long term meteorological data it is now possible to predict the influence of different meteorological parameters on the crop yield using decision tree induction

approach. The major agricultural inputs and their effect on crop yield and also cost of cultivation affected by different inputs in selected States of India. In this study regression analysis was used in predicting the input interaction on the crop yield.

Agriculture in India has a significant history. Today, India ranks second worldwide in farm output. Agriculture and allied sectors like forestry and fisheries accounted for 16.6% of the GDP 2009, about 50% of the total workforce. The economic contribution of agriculture to India's GDP is steadily declining with the country's broad-based economic growth. Still, agriculture is demographically the broadest economic sector and plays a significant role in the overall socio-economic fabric of India. Indian agriculture is known for its diversity which is mainly result of variation in resource and climate, to topography and historical, institutional and socio economic factors. Policies followed in the country and nature of technology that became available over time has reinforced some of the variations resulting from natural factors. As a consequence, production performance of agriculture sector has followed on uneven path and large gaps have development in productivity between different geographic locations across the country.

Agriculture as a business is unique crop production is dependent on many climatic, geographical, biological political and economic factors that are mostly independent of one another. This multiple factor introduces risk. The efficient management of these risks is imperative for the successful agricultural and consistent output of food. The Agricultural yield is primarily depends on weather conditions, diseases and pests, planning of harvest operation. Effective management of these factors is necessary to estimate the probability of such unfavorable situation & to minimize the consequences. Accurate and reliable information about historical crop yield is thus vital for decisions relating to agricultural risk management.

Historical crop yield information is also important for supply chain operation of companies engaged in industries that use agricultural produce as raw material. Livestock, food, animal feed, chemical, poultry, fertilizer pesticides, seed, paper and many other industries use agricultural products as intergradient in their production processes. An accurate estimate of crop size and risk helps these companies in planning supply chain decision like production scheduling. Business such as seed, fertilizer, agrochemical and agricultural machinery industries plan production and marketing activities based on crop production estimates.[1],[13].

2. Application of Data Mining Techniques in Agriculture

2.1 Prediction of problematic wine fermentations

Wine is widely produced all around the world. The fermentation process of the wine is very important, because it can impact the productivity of wine-related industries and also the quality of wine. If we were able to predict how the fermentation is going to be at the early stages of the process, we could interfere with the process in order to guarantee a regular and smooth fermentation. Fermentations are nowadays studied by using different techniques, such as, for example, the k-means algorithm, and a technique for classification based on the concept of biclustering. Note that these works are different from the ones where a classification of different kinds of wine is performed.

2.2 Detection of diseases from sounds issued by animals

The detection of animal's diseases in farms can impact positively the productivity of the farm, because sick animals can cause contaminations. Moreover, the early detection of the diseases can allow the farmer to cure the animal as soon as the disease appears. Sounds issued by pigs can be analyzed for the detection of

diseases. In particular, their coughs can be studied, because they indicate their sickness. A computational system is under development which is able to monitor pig sounds by microphones installed in the farm, and which is also able to discriminate among the different sounds that can be detected.

2.3 Sorting apples by watercores

Before going to market, apples are checked and the ones showing some defects are removed. However, there are also invisible defects that can spoil the apple flavor and look. An example of invisible defect is the watercore. This is an internalapple disorder that can affect the longevity of the fruit. Apples with slight or mild watercores are sweeter, but apples with moderate to sever degree of watercore cannot be stored for any length of time. Moreover, a few fruits with severe watercore could spoil a whole batch of apples. For this reason, a computational system is under study which takes X-ray photographs of the fruit while they run on conveyor belts, and which is also able to analyse (by data mining techniques) the taken pictures and estimate the probability that the fruit contains watercores.

2.4 Optimizing pesticide usage by data mining

Recent studies by agriculture researchers in Pakistan (one of the top four cotton producers of the world) showed that attempts of cotton crop yield maximization through pro-pesticide state policies have led to a dangerously high pesticide usage. These studies have reported a negative correlation between pesticide usage and crop yield in Pakistan. Hence excessive use (or abuse) of pesticides is harming the farmers with adverse financial, environmental and social impacts. By data mining the cotton Pest Scouting data along with the meteorological recordings it was shown that how pesticide usage can be optimized (reduced). Clustering of data revealed interesting patterns of farmer practices along with pesticide usage dynamics and hence help identify the reasons for this pesticide abuse.

2.5 Explaining pesticide abuse by data mining

To monitor cotton growth, different government departments and agencies in Pakistan have been recording pest scouting, agriculture and metrological data for decades. Coarse estimates of just the cotton pest scouting data recorded stands at around 1.5 million records, and growing. The primary agro-met data recorded has never been digitized, integrated or standardized to give a complete picture, and hence cannot support decision making, thus requiring an Agriculture Data Warehouse. Creating a novel Pilot Agriculture Extension Data Warehouse followed by analysis through querying and data mining some interesting discoveries were made, such as pesticides sprayed at the wrong time, wrong pesticides used for the right reasons and temporal relationship between pesticide usage and day of the week.

3. Data Mining Techniques

Data Mining techniques are mainly divided in two groups, classification and clustering techniques .Classification techniques are designed for classifying unknown samples using information provided by a set of classified samples. This set is usually referred to as a training set as it is used to train the classification technique how to perform its classification. Generally, Neural Networks and Support Vector Machines these two classification techniques learn from training set how to classify unknown samples. Another classification technique, K- Nearest Neighbor does not have any learning phase, because it uses the training set every time a classification must be performed. A training set is known, and it is used to classify samples of unknown classification. The basic assumption in the K Nearest Neighbor algorithm is that similar samples should have similar classification. The

parameter K shows the number of similar known samples used for assigning a classification to an unknown sample. The K-Nearest Neighbour uses the information in the training set, but it does not extract any rule for classifying the other.

In the event training set not available, there is no previous knowledge about the data to classify. In this case, clustering techniques can be used to split a set of unknown samples into clusters. One of the most used clustering techniques is the KMeans algorithm. Given a set of data with unknown classification, the aim is to find a partition of the set in which similar data are grouped in the same cluster. The parameter K plays an important role as it specifies the number of clusters in which the data must be partitioned. The idea behind the K-Means algorithm is, given a certain partition of the data in K clusters, the centers of the clusters can be computed as the means of all samples belonging to a cluster. The center of the cluster can be considered as the representative of the cluster, because the center is quite close to all samples in the cluster, and therefore it is similar to all of them. There are some disadvantages in using K-Means method. One of the disadvantages could be the choice of the parameter K. Another issue that needs attention is the computational cost of the algorithm.

There are other Data Mining techniques statistical based techniques, such as Principle Component Analysis (PCA), Regression Model and Biclustering Techniques have some applications in agriculture or agricultural - related fields.

4. Methods and Materials

4.1 Influence of climatic parameters on soybean productivity using decision tree Induction technique. Decision tree induction technique is adopted in the present study to develop innovative approaches to predict the influence of climatic parameters on the predominant crop (soybean) productivity of Bhopal district. A decision tree is a flow-chart-like tree structure, where each internal node denotes a test on an attribute, each branch represents an outcome of the test, and leaf nodes represent classes or class distributions. The top most nodes in a tree is the root node. In order to classify an unknown sample, the attribute values of the sample are tested against the decision tree. A path is traced from the root to a leaf node that holds the class prediction for that sample. Decision trees were then converted to classification rules using IF-THEN-ELSE. Interactive Dichotomizer 3 (ID3) is one of the decision tree algorithm adopted in this study which is information based method that depends on two assumptions. Let C contain p objects of class P and n of class N. The assumptions are:

(1) Any correct decision tree for C will classify objects in the same proportion as their representation in C. An arbitrary object will be determined to belong to class P with probability p/(p + n) and to class N with probability n/(p + n).

(2) When a decision tree is used to classify an object, it returns a class. A decision tree can thus be regarded as a source of a message 'P' or 'N', with the expected information needed to generate this message given by

$$I(p,n)=-\frac{p}{p+n}\log_2\frac{p}{p+n}-\frac{n}{p+n}\log_2\frac{n}{p+n}$$

If attribute A with values {A1, A2, ... Av) is used for the root of the decision tree, it will partition C into (C1, C2, ... CV} where Ci contains those objects in C that have value Ai of A. Let Ci contain pi objects of class P and ni of class N. The expected

information required for the sub tree for Ci is I(pi , ni). The expected information required for the tree with A as root is then obtained as the weighted average

$$E(A)= \sum_{i=1}^{v} \frac{p_i + n_i}{p + n} I(p_i , n_i)$$

Where the weight for the ith branch is the proportion of the objects in C that belong to Ci

The information gained by branching on A is therefore gain(A)= I(p,n)- E(A)

A good rule of thumb would seem to be to choose that attribute to branch on which gains the most information.

Decision Tree Induction: The basic algorithm for decision tree induction is a greedy algorithm that constructs decision trees in a top-down recursive divide-and-conquer manner. The basic strategy for the algorithm is as follows:

 The tree starts as a single node representing the training sample

 If the samples are all of the same class, then the node becomes a leaf and is labeled with that class.

 Otherwise, the algorithm uses an entropy-based measure known as information gain as a heuristic for selecting the attribute that will best separate the samples into individual classes. This attribute becomes the test or decision attribute at the node. In this version of the algorithm individual classes. all attributes are categorical, that is, discrete valued. Continuous-valued attributes must be discretized.

 A branch is created for each known value of the test attribute, and the samples are partitioned accordingly.

 The algorithm uses the same process recursively to form a decision tree for the samples at each partition. Once an attribute has occurred at a node, it need not be considered .

 The recursive partitioning stops only when any one of the following condition is true:

a) All samples for a given node belong to the same class

b) There are no remaining attributes on which the samples may be further partitioned. In this case, majority voting is employed. This involves converting the given node into a leaf and labeling it with the class in majority among samples.

c) There are no samples for the branch test attribute = ai In this case, a leaf is created with the majority class in samples.

K-Means Algorithm

The k-means is a well-known and commonly used partitioning method. Objects are classified as belonging to one of k groups. K-means algorithm has been chosen because of the popularity due to following reasons. Its time complexity is O (nkl), where n is the number of patterns, k is the number of clusters, and l is the number of iterations taken by the algorithm. The k-means algorithm time complexity is minimum as compared to other clustering algorithms.

The k-means algorithm is suitable for the huge amount of dataset and cluster quality is good. While hierarchical clustering algorithms are suggested for small dataset and achieve good result. Performance of k-means are better than other hierarchical clustering algorithms.

5. Conclusion

The decision trees suggested there exists a correlation between climatic factors and soybean crop productivity and these variables influence on the soybean crop productivity were confirmed from the rule accuracy and Bayesian classification.

The salient conclusions of the present study are:

i) The decision tree analysis indicated that the productivity of soybean crop was mostly influenced by Relative humidity followed by temperature and rainfall.

ii) The decision tree from the present study are fast to execute and much to be desired as representations of knowledge interpretations.

iii) The rules formed from the decision tree are helpful in predicting the conditions responsible for the high or low soybean crop productivity under given climatic parameters.

References

[1] Abdullah, A., Brobst, S., M.Umer M. 2004. "The case for an agri data ware house: Enabling analytical exploration of integrated agricultural data". Proc. of IASTED International Conference on Databases and Applications. Austria. Feb

[2] Abdullah, A., Brobst, S, Pervaiz.I., Umer M.,A.Nisar. 2004. "Learning dynamics of pesticide abuse through data mining". Proc. of Australian Workshop on Data Mining and Web Intelligence, New Zealand, January.

[3] Abdullah, A., Bulbul.R., Tahir Mehmood. 2005. "Mapping nominal values to numbers by data mining spectral properties of leaves". Proc. of 3rd International Symposium on Intelligent Information Technology in Agriculture. Beijing, China. Oct, 2005.

[4] Data mining Techniques for Predicting Crop Productivity – A review article S.Veenadhari, Dr. Bharat Misra, Dr. CD Singh IJCST Vol. 2, Issue 1, March 2011

[5] V. Ramesh and K. Ramr Classification of agricultural land soils: A data mining approach" International Journal on Computer Science and Engineering (IJCSE) ISSN: 0975-3397 Vol. 3 No. 1 Jan 2011 379

[6] Chi-Chung LAU, Kuo-Hsin HSIAO, 2005. "Bayesian Classification For Rice Paddy interpretation". Paper presented in Conference on data mining held at China Tapei. December, 2005

[7] Cunningham S.J., G. Holmes. 2005. "Developing innovative applications in agriculture using data mining". Proc. Of 3rd International Symposium on Intelligent Information Technology in Agriculture. Beijing, China. Oct, 2005

[8] "Risk in Agriculture: A study of crop yield distribution and crop insurance" by Narsi Reddy Gayam Thesis (M. Eng. In Logistics)-- Massachusetts Institute of Technology, Engineering Systems Division, 2006. Includes bibliographical references (leaves 52-53).

[9]. Abdullah, A., S. Brobst, and M. Umer. 2004.The Case study for an Agri Data Warehouse: Enabling Analytical Exploration Integrated Agricultural Data. Proceeding of the IASTED International Conference on Databases and Applications, Innsbruck, Austria.

[10]. Chandel, K.P.S., G. Shukla and N. Sharma. 1996. Biodiversity in Medicinal and Aromatic Plants in India: Conservation and Utilization, 239.

[11]. Cunningham, S. J., and G. Holmes. 1999. Developing innovative applications in agriculture using data mining. Proceeding of Southeast Asia Regional Computer Confederation Conference

[12]. Data mining Techniques for Predicting Crop Productivity – A review article S.Veenadhari, Dr. Bharat Misra, Dr. CD Singh IJCST Vol. 2, Issue 1, March 2011.

[13]. Rainfall variability analysis and its impact on crop productivity Indian agriculture research journal 2002 29,33.,8) SPRS Archives XXXVI-8/W48 Workshop proceedings: Remote sensing support to crop yield forecast and area estimates GENERALIZED SOFTWARE TOOLS FOR CROP AREA ESTIMATES AND YIELD FORECAST by Roberto Benedetti A, Remo Catenaro A, Federica Piersimoni B

[14]. R S Deshpande AN ANALYSIS OF THE RESULTS OF CROP CUTTING EXPERIMENTS Agricultural Development and Rural Transformation Unit Institute for Social and Economic Change February 2003.