Innovative Applications of Data Mining Tools and Techniques in Agriculture Sector to Improve the Profitability

Asst. Prof. Priyanka Thakare

PIBM Chinchwad , Savitribai Phule Pune university, Pune India

priyanka.thakare3@gmail.com

Prof. Dr. R. M. Patil

SIBACA Lonavala, Savitribai Phule Pune university, Pune India

ravidrapatil@sinhgad.edu

Abstract*—Data Mining can be defined as extracting the information from the huge set of data. We can also say that data mining is mining the knowledge from data. In Indian agriculture, large amount of data is available. Mining of such a huge amount of data in agriculture sector is a need in today's circumstance to improve the profitability. In this article, we will discuss the importance of data mining tools, techniques and its innovative applications in agriculture sector. Lots of data mining techniques will be used in agriculture sector some of these are Classification, Regression. Artificial Neural Networks, Bayesian Networks, Decisions tree, Support vector Machine etc. There are n numbers of open source data mining tools are also available but in this paper we will discuss only WEKA tool. It is a modern technique to find the solution over the traditional and predictable method which will helpful to improve the financial growth of farmer which will definitely helps to pick up the profitability. Data mining is a modern data analysis technique and it is a relatively a novel research area in agriculture sector

***Keywords*— Agriculture, Data Mining, Artificial Neural Network, WEKA Tool, Support Vector Machine, Decision Tree.**

I. Introduction

Agriculture is back bone business in India. Agriculture plays a vital role in India's economy. Over 58 per cent of the rural households depend on agriculture as their principal means of livelihood. Today, India ranks second worldwide in farm output. Rising private participation in Indian agriculture, growing organic farming and use of information technology are some of the key trends in the agriculture. In Indian agriculture, large amount of data is available. The data when become information is highly useful for many purposes. The predictable and traditional system of data analysis in agriculture is purely dependent on statistics. Data mining is a modern data analysis technique. It has broad range of applications in the field of agriculture. Generally, data mining sometimes called data or knowledge discovery is the process of analyzing data from different perspectives and summarizing it into useful information which can be used to increase revenue, cuts costs, or both. Data mining is really helpful to improve the productivity which will ultimately result in increased profitability of agriculture sector. Data mining software is one of a number of analytical tools for analyzing data. This article will discussed applications of the data mining techniques in the area of agriculture sector. Special techniques of data mining have been used in this field. This study aims to come out of the different tools and techniques being used in the agriculture. Though, there are lots of techniques available in the data mining, few methodologies such as Support vector machine, Artificial Neural Networks, k means algorithm , K nearest neighbor are popular techniques but which will use it depends on the nature of data. There are n numbers of open source data mining tools are also available. Out of that six powerful open source data mining tools are Rapid Miner (formerly known as YALE), WEKA, R-Programming, orange, KNIME (Konstanz Information miner), NLTK (Natural language tool kit). Out of this WEKA tool is very famous software for data mining and generally use for agriculture sector. For this reason in this study we will focus only on WEKA tool.

Objectives of the Study

To find how data mining technology helps to improve profitability in agriculture sector.

To study the different data mining tools and techniques which we can implement in agriculture sector.

To study the innovative applications of data mining techniques in agriculture.

To find the solution over the traditional agriculture methods and predictable method which will helpful to improve the financial growth of farmer?

Research Methodology

This research is based on secondary data that has been collected by referring various research articles, books & web-sites. The collective data has been analyzed, compiled & then the outcome of all these are presented in this research article.

II. Data Analysis

The agriculture data which has been collected for this research is processes and then this processed data is apply as input in WEKA which can be analyzed using different data mining techniques like, Classification, Clustering, Association rule mining, Artificial Neural Network, Support Vector Machine, Decision Tree ,Visualization and different algorithms etc. Then this analysis will be use for developing innovative applications in agriculture which will really helpful to improve the productivity which will ultimately result in increasing the profitability of agriculture sector. This paper shows one of the small importances of WEKA tool to utilization and analysis for agriculture data by using data mining techniques and knowledge discovery from agriculture database.

III. Overview of Data Mining Tools and Techniques in Agriculture

Data mining is the practice of automatically searching large stores of data to discover patterns and trends that go beyond simple analysis. Data mining uses sophisticated mathematical algorithms to segment the data and evaluate the probability of future events. Data mining is also known as Knowledge Discovery in Data (KDD).Data mining tools predict behaviors and future trends, allowing businesses to make proactive, knowledge-driven decisions. Data mining tools can answer business questions that traditionally were too time consuming to resolve. The mined information is used for representing as a model for prediction or classification. The agricultural domain datasets appear to be significantly more complex than the datasets conventionally used in machine learning. Data mining is mainly categorized as descriptive and predictive data mining. But in the agriculture area, predictive data mining is mainly used. There are two main techniques namely classification and Clustering. There are n numbers of open source data mining tools are also available but in this paper we will discuss only WEKA tool. These Data mining tools and techniques is really helpful to improve the productivity which will ultimately result in increased profitability of agriculture sector.

WEKA tool: (Waikato Environment for Knowledge Analysis) Weka is a popular suite of machine learning software written in Java, developed at the University of Waikato, New Zealand. It is a collection of machine learning algorithms for data mining tasks. The algorithms can either be applied directly to a dataset or called from your own Java code. Weka supports several standard data mining tasks, more specifically, data preprocessing, clustering, classification, regression, visualization, and feature selection. It is also well-suited for developing new machine learning schemes. WEKA is open source software issued under the GNU General Public License. Its purpose is to allow users to access a variety of machine learning techniques for the purposes of experimentation and comparison using real world data sets. Basically use to solve real word problems in particular, those arising from agricultural and horticultural domains.

Techniques: The major techniques for data mining consist of Classification and Regression. Classification is broadly categories into Neural Networks, Bayesian Networks, Decisions tree, Support vector Machine, Instance based. According to the data set different data mining techniques used for solving different agricultural problem. The graphical representation of different data mining techniques is as shown in the following figure.

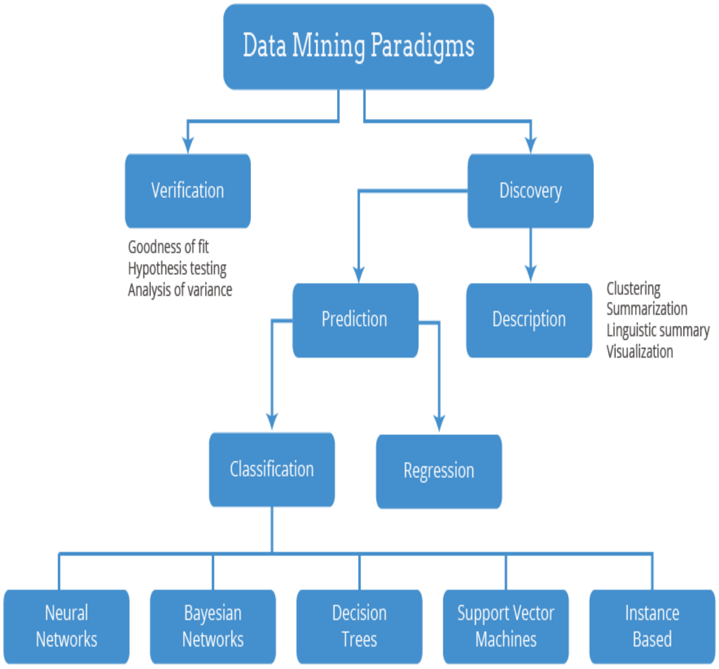

Figure 1: Different data mining techniques Paradigm.

Relationship over the conventional techniques flood forecast. Neural network has several advantages over conventional method in computing. Any problem having more time for getting solution, ANN is highly suitable states that the neural network method successfully predicts the pest attack incidences for one week in advance.

Support Vector Machines: Support Vector machines can be defined as systems which use hypothesis space of a linear functions in a high dimensional feature space, trained with a learning algorithm from optimization theory that implements a learning Support Vector Machine (SVM) was first heard in 1992, introduced by Boser, Guyon, and Vapnik in COLT-92. Support vector machines (SVMs) are a set of related supervised learning methods used for classification and regression. They belong to a family of generalized linear classifiers. In another terms, Support Vector Machine (SVM) is a classification and regression prediction tool that uses machine learning theory to maximize predictive accuracy while automatically avoiding over-fit to the data. Support bias derived from statistical learning theory. Support vector machine was initially popular with the NIPS community and now is an active part of the machine learning research around the world. SVM becomes famous when, using pixel maps as input; it gives accuracy comparable to sophisticated neural networks with elaborated features in a handwriting recognition task. It is also being used for many applications including agriculture sector also.

Decision trees: A decision tree is a decision support tool that uses a tree-like graph or model of decisions and their possible consequences, including chance event outcomes, resource costs, and utility. It is one way to display an algorithm. Decision trees are commonly used in operations research, specifically in decision analysis, to help identify a strategy most likely to reach a goal. The decision tree is one of the popular classification algorithms in current use in Data Mining and Machine Learning. Decision tree is a new field of machine learning which is involving the algorithmic acquisition of structured knowledge in forms such as concepts, decision trees and discrimination nets or production rules. Application of data mining techniques on drought related data for drought risk management shows the success on Advanced Geospatial Decision Support System (GDSS). Data mining approach is one of the approaches used for crop decision making.

K nearest neighbor: In pattern recognition, the k-Nearest Neighbors algorithm (or k-NN for short) is a non-parametric method used for classification and regression. In both cases, the input consists of the k closest training examples in the feature space. The output depends on whether k-NN is used for classification or regression:

In k-NN classification, the output is a class membership. An object is classified by a majority vote of its neighbors, with the object being assigned to the class most common among its k nearest neighbors (k is a positive integer, typically small). If k = 1, then the object is simply assigned to the class of that single nearest neighbor.

- In k-NN regression, the output is the property value for the object. This value is the average of the values of its k nearest neighbors.

k-NN is a type of instance-based learning, or lazy learning, where the function is only approximated locally and all computation is deferred until classification. The k-NN algorithm is among the simplest of all machine learning algorithms. Both for classification and regression, it can be useful to weight the contributions of the neighbors, so that the nearer neighbors contribute more to the average than the more distant ones. For example, a common weighting scheme consists in giving each neighbor a weight of 1/d, where d is the distance to the neighbor. The neighbors are taken from a set of objects for which the class (for k-NN classification) or the object property value (for k-NN regression) is known. This can be thought of as the training set for the algorithm, though no explicit training step is required. A shortcoming of the k-NN algorithm is that it is sensitive to the local structure of the data.

Bayesian networks: A Bayesian network, Bayes network, belief network, Bayesian model or probabilistic directed acyclic graphical model is a probabilistic graphical model (a type of statistical model) that represents a set of random variables and their conditional dependencies via a directed acyclic graph (DAG). For example, a Bayesian network could represent the probabilistic relationships between diseases and symptoms. Given symptoms, the network can be used to compute the probabilities of the presence of various diseases.

Formally, Bayesian networks are DAGs whose nodes represent random variables in the Bayesian sense: they may be observable quantities, latent variables, unknown parameters or hypotheses. Edges represent conditional dependencies; nodes that are not connected represent variables that are conditionally independent of each other. Each node is associated with a probability function that takes as input a particular set of values for the node's parent variables and gives the probability of the variable represented by the node. For example, if the parents are mBoolean variables then the probability function could be represented by a table of 2^mentries, one entry for each of the 2^mpossible combinations of its parents being true or false. Similar ideas may be applied to undirected, and possibly cyclic, graphs; such are called Markov networks.

Efficient algorithms exist that perform inference and learning in Bayesian networks. Bayesian networks that model sequences of variables (e.g. speech signals or protein sequences) are called dynamic Bayesian networks. Generalizations of Bayesian networks that can represent and solve decision problems under uncertainty are called influence diagrams.

Clustering: Cluster analysis or clustering is the task of grouping a set of objects in such a way that objects in the same group (called a cluster) are more similar (in some sense or another) to each other than to those in other groups (clusters). It is a main task of exploratory data mining, and a common technique for statistical data analysis, used in many fields, including machine learning, pattern recognition, image analysis, information retrieval, and bioinformatics.

Cluster analysis itself is not one specific algorithm, but the general task to be solved. It can be achieved by various algorithms that differ significantly in their notion of what constitutes a cluster and how to efficiently find them. Popular notions of clusters include groups with small distances among the cluster members, dense areas of the data space, intervals or particular statistical distributions. Clustering can therefore be formulated as a multi-objective optimization problem. The appropriate clustering algorithm and parameter settings (including values such as the distance function to use, a density threshold or the number of expected clusters) depend on the individual data set and intended use of the results. Cluster analysis as such is not an automatic task, but an iterative process of knowledge discovery or interactive multi-objective optimization that involves trial and failure. It will often be necessary to modify data preprocessing and model parameters until the result achieves the desired properties.

We can show this with a simple graphical example. Cluster is a collection of data elements that are highly similar to one another with in the cluster but weakly similar from the data elements in other clusters (Fig. 1). In mathematically, let O = {O1,O2,O3….On} be a set of n objects and let C = {C1,C2,…Ck} be a partition of O into subsets; such that Ci n Cj = Ø, I $\neq$ j and kÙCk = O. Each subset is called a cluster and C is a clustering solution. It is described as a given set of data with unknown classification to be aimed to find a partition of the set in which similar data samples are grouped in the same cluster. The similarities between two data samples are provided using a suitable distance. The samples are close to each other is considered similar.

This study exposed that interesting patterns of farmers practices along with pesticide usage dynamics, which helps the farmers to identify the pesticide abuse. So that farmer can take the decision which pesticides should be use.

K means approach: From Fig. 2, K means method is one of the most used clustering techniques in the data mining. The idea behind the k means algorithms is very simple that certain partition of the data in K clusters, the centers of the cluster can be computed as the mean of the all sample belonging to a cluster.

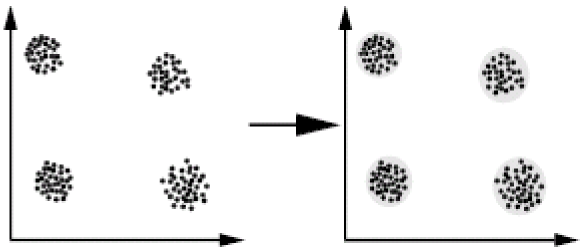

Figure 2: Clustering

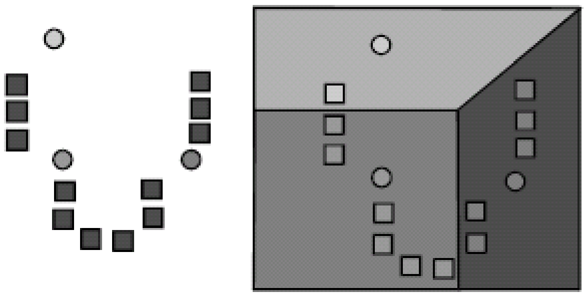

Figure 3: K means

The center of the cluster can be considered as the representative of the cluster. The center is quite close to all samples in the cluster.

IV. Innovative Applications of Data Mining in Agriculture

Several data mining techniques will be use in agriculture sector which will help to improve financial growth of the farmers. Only some of techniques are presented as follows.

- K means method is used to forecast the pollution in the atmosphere. K nearest neighbor is applied for simulating daily precipitation and other weather variables. Which will helps the farmer for crop yield.
- Different possible changes of weather are analyzed using SVM.So that, at earlier stage farmer can use specific cultivation.
- K means approach is used for classifying soil in combination with GPS readings. K Means approach was used to classify the soil and plants.
- Wine Fermentation process monitored using data mining techniques. Taste sensors are used to obtain data from the fermentation process to be classified using ANNs.

In addition to this following are some major applications of data mining in agriculture are as follows.

➢ K-means algorithm and classification techniques use for 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 bi-clustering. Note that these works are different from the ones where a classification of different kinds of wine is performed.

- 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.

- Analyzing techniques use for Sorting apples by water cores

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 water core. This is an internal apple disorder that can affect the longevity of the fruit. Apples with slight or mild water cores are sweeter, but apples with moderate to severe degree of water core cannot be stored for any length of time. Moreover, a few fruits with severe water core 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 analyze (by data mining techniques) the taken pictures and estimate the probability that the fruit contains water cores.

- Optimizing pesticide use 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 use. These studies have reported a negative correlation between pesticide use 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 use can be optimized (reduced). Clustering of data revealed interesting patterns of farmer practices along with pesticide use dynamics and hence help identify the reasons for this pesticide abuse.

- Explaining pesticide abuse by data mining techniques

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, which will really helpful to improve the productivity which will ultimately result in increasing the profitability of agriculture sector.

V. Conclusion

This research is foundation in the direction of applying data mining tools and techniques in the agricultural domain. Several data mining techniques are applicable to agriculture sector such as SVM, ANN, and k means are generally used. In addition to this, current research tells us that decision tree can be help for crop decision making to the farmer. This work helps to provide efficient techniques for agriculture sector which will helps the farmer to use the new technology over the traditional one to improve the profitability. This research study is based on how to use data mining technology in the agriculture sector which can focus on the mission of improving lives and can help in meeting the aggressive future growth of the farmer and this field. Which will really helpful to improve the productivity which will ultimately result in increasing the profitability of agriculture sector.. This research shows that how data mining techniques can be used as a tool for knowledge management in agriculture. In conclusion, such current data mining tools and techniques will provide a cheaper, faster and a greener service to agriculture sector.

VI. REFERENCES

1. Mucherino, A.; Papajorgji, P.J.; Pardalos, P. (2009). Data Mining in Agriculture, Springer.
2. Schatzki, T.F.; Haff, R.P.; Young, R.; Can, I.; Le, L-C.; Toyofuku, N. (1997). "Defect Detection in Apples by Means of X-ray Imaging". Transactions of the American Society of Agricultural Engineers 40(5): 1407–1415.
3. Jyotshna Solanki, Prof. (Dr.) Yusuf Mulge,Different Techniques Used in Data Mining in Agriculture,International Journal of Advanced Research in Computer Science and Software Engineering , Volume 5, Issue 5, May 2015 ISSN: 2277 128X
4. “Naïve Bayes”, ”C4.5 (J48)”, Wikipedia.
5. Palace, B. (1996). Data Mining: What is Data Mining? http://www.anderson.ucla.edu/faculty/jason.frand/teacher/technologies/palace/datamining.htm
6. Agrawal, R., Imielinski, T. and Swami, A.N. (1993) “Database mining: a performanceperspective.” IEEE Transactions on Knowledge and Data Engineering, Vol. 5,914–925.

7. Latika Sharma, Nitu Mehta,Data Mining Techniques: A Tool For Knowledge Management System In Agriculture, International journal of scientific & technology research, volume 1, issue 5, June 2012 ISSN 2277-8616 67.
8. Jeysenthil.KMS1, Manikandan.T2, Murali.E3, “Third Generation Agricultural Support System Development Using Data Mining”, International Journal of Innovative Research in Science, Engineering and Technology ISSN: 2319-8753 Vol. 3, Issue 3, March 2014.
9. www.cs.waikato.ac.nz/ml/weka/
10. Hetal Patel,Dharmendra Patel, “A Brief survey of Data Mining Techniques Applied to Agricultural Data”International Journal of Computer Applications (0975 – 8887) Volume 95– No. 9, June 2014.