RESEARCH ARTICLE OPEN ACCESS

Application of Data Mining in Agriculture Sector

Rajiv Senapati [1], D. Anil Kumar [2]
Department of Computer Science and Engineering
GIET, Gunupur
India

ABSTRACT
One of the fastest growing fields in India is the agriculture field. The agriculture field is generating huge amount of data every day about farmers, crops, crop prices etc. Huge amount of data is collected in daily basis from different sources, which are not mined to find out hidden information. To turns these data is into useful pattern and to predicting upcoming trends data mining approaches can be considered as an important tool. In this paper we are addressing some of the popular techniques of Data mining in agriculture domain. There are various data mining techniques such as Artificial Neural Networks (ANN),Bayesian Classifiers, decision tree and Support Vector Machines(SVM) which are used for very recent applications of Data Mining techniques.
Keywords :- Data mining, Artificial Neural Networks, decision tree, Support Vector Machines, Prediction techniques.

I. INTRODUCTION

Data mining is the method for finding interesting patterns from enormous amount of data. The process of extracting important and useful information from large sets of data is called Data Mining. In agricultural field, Data Mining is an important research mechanism for analysis and prediction. In this paper, Description and overview of data mining techniques which are applied to agriculture and their applications to agriculture related areas is described. This mechanism may leads to increases the quality of service provided to the farmers in different aspect. For example, price prediction is a very important problem for any farmer as he is the one who should know how much cost he would expect for his crops. For which past data may be collected and analyses properly by applying a suitable data mining.

The paper is organized as follows: in section II, Knowledge Discovery Process is presented. Literature on agriculture data mining is presented in section III. Agriculture data mining and prediction techniques are studied in section IV and V. Finally, conclusions are drawn in section VI.

II. KNOWLEDGE DISCOVERY

A. Knowledge Discovery Process

The terms Knowledge Discovery in Databases (KDD) and Data Mining are frequently used interchangeably. KDD is the process of changing the low-level data into high-level knowledge. Hence, KDD refers to the nontrivial removal of implicit, previously unknown and potentially useful information from data in databases. While data mining and KDD are often treated as comparable words but in real data mining is an essential step in the KDD process. The Knowledge Discovery in Databases process comprise of a few steps leading from raw data collections to some form of new information. The iterative process consists of the following steps:

- Data cleaning: During this stage noise data and unrelated data are removed from the collection.
- Data integration: During this stage, several data sources, often heterogeneous, may be shared in a common source.
- Data selection: During this stage, the data related to the analysis is decided on and retrieve from the data collection.
- Data transformation: This stage is also known as data consolidation, it is a phase in which the chosen data is transformed into forms appropriate for the mining procedure.
- Data mining: it is the essential step in which clever techniques are applied to extract patterns potentially useful.
- Pattern evaluation: During this stage, firmly interesting patterns representing knowledge are known based on given measures.
- Knowledge representation: This is the last phase in which the discovered knowledge is visually represented to the user. In this phase visualization techniques are used to help users understand and interpret the data mining results.

The process of knowledge discovery from huge amount of data is shown in figure 1.

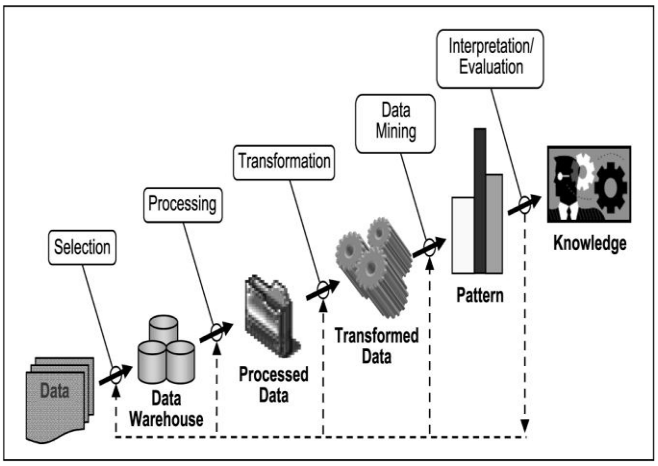

Fig. 1. Steps in KDD.

B. Data Mining Process

In the KDD process, the data mining methods are for extracting interesting hidden patterns from data. The patterns that can be exposed depend upon the data mining tasks applied. Generally, there are two types of data mining approach: these are descriptive data mining which explains the general properties of the existing data, and predictive data mining approach that attempt to do predictions based on available data.

Data mining can be done on data which are of heterogeneous types such as textual, quantitative or multimedia data. Data mining applications can use disparate kind of parameters to observe the data, which includes association, sequence or path analysis, classification and clustering. Data mining involves some of the following key steps:

- Problem definition: The first step is to discover goals. Based on the defined goal, the correct series of tools can be applied to the data to build the corresponding behavioral model.
- Data exploration: If the value of data is not suitable for an perfect model then recommendations on future data collection and storage strategies can be made at this. For analysis, all data needs to be consolidated so that it can be treated consistently.
- Data preparation: The purpose of this step is to clean and convert the data so that missing and invalid values are treated and all known valid values are made reliable for more robust analysis.
- Modeling: Based on the data and the desired outcomes, a data mining algorithm or group of algorithms algorithms is selected for analysis. These algorithms include classical techniques such as statistics, neighbourhoods and clustering but also next invention techniques such as decision trees, networks and rule based algorithms. The specific algorithm is selected based on the particular be analysed. objective to be achieved and the quality of the data to
- Evaluation and Deployment: Based on the outcome of the data mining algorithms, an analysis is conducted to find out key conclusions from the analysis and create a sequence of recommendations for consideration.

The process model of data mining is presented in Figure 2.

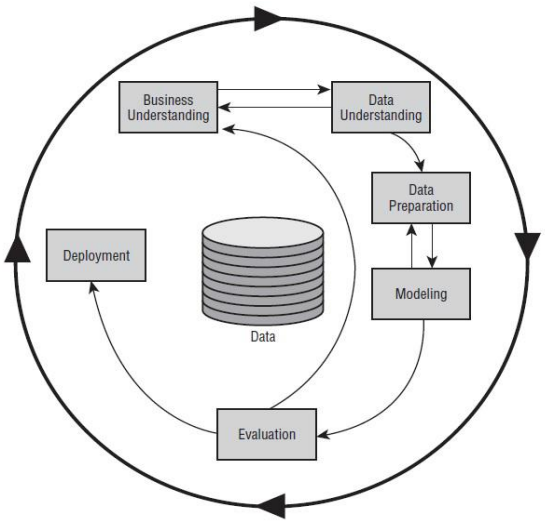

Fig. 2. Data Mining Process Representation.

III. LITERATURE REVIEW

Different techniques were used for mining data. Researchers have discussed a detailed and elaborated 10 Data Mining Techniques [7]. This paper present the most used Data Mining Techniques in agricultural field. Classification and Clustering Techniques are two types of Data Mining Techniques [4]. For classifying unknown samples in which information is provided by a set of classified samples, Classification Techniques are designed. Neural Networks [6] and Support Vector Machines [3] are two classification techniques that are used generally to classify unknown samples. The technique that does not have any learning set is the K-Nearest Neighbour (KNN)[1] but it has the training set that is used for classification and in this technique similar samples should have similar classification. The parameter K in K-Nearest Neighbour is used to show the number of similar known samples. The K-Nearest Neighbour uses the training set, in case, if any training set is not available, clustering techniques can be used to split a set of unknown samples into clusters. K-Means algorithm [2] is one of the most used clustering techniques. In a set of data with unknown classification, we will find a partition of the set in which we have same data which is grouped in same cluster. The parameter K present in K-Means algorithm specifies the number of clusters in which data is to be partitioned. The main reason behind K-Means algorithm is the centers of clusters that can be computed as means of all samples belonging to a cluster. The

representative of the cluster can be considered as the centre of cluster because the center is quite close to all samples. But one of the disadvantages in using the K-Means algorithm can be the choice of parameter K. Another important issue is the computational cost of the algorithm. Other data mining techniques such as Principle Component Analysis (PCA), Regression Model [5] and Bi-clustering techniques have some applications in agriculture or agricultural-related fields.

IV. AGRICULTURE DATA MINING

Data mining in agriculture is a very recent emerging research area. Recent technologies which are in use in this domain are able to provide a lot of information on agricultural-related activities, which can then be analysed in order to find important information and interesting hidden patterns so as to support the farmers. Some of the interesting areas like Optimizing pesticide use by data mining methods, crop price predictions etc. can be done by using data mining techniques. In the agriculture managing data mining prediction are playing vital role. The data mining prediction model is presented in Figure. 3. Some of the prediction based data mining techniques are as follows:

- Neural network.
- Bayesian Classifiers.
- Decision tree.
- Support Vector Machine.

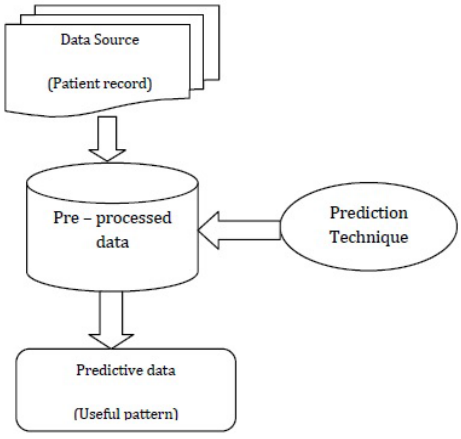

Fig. 3. Agriculture Data Mining Prediction Model.

V. PREDICTION TECHNIQUES

A. Artificial Neural Networks (ANN)

An artificial neural network is a mathematical model based on biological neural networks. It consists of an interrelated group of artificial neurons and processes information using a connectionist approach to computation. Neurons are structured into layers. The input layer consists of the original data, while the output layer nodes represent the classes. There may be several hidden layers. A main feature of neural networks is an iterative learning process in which data samples are presented to the network one at a time, and the weights are adjusted to predict the correct class label. Advantages of neural networks include their high tolerance to noisy data, as well as their ability to classify patterns on which they have not been trained. A main concern of the training phase is to focus on the interior weights of the neural network, which is used according to the transactions used in the learning process. For each training transaction, the neural network receives in addition the expected output. Figure 4 represents ANN model.

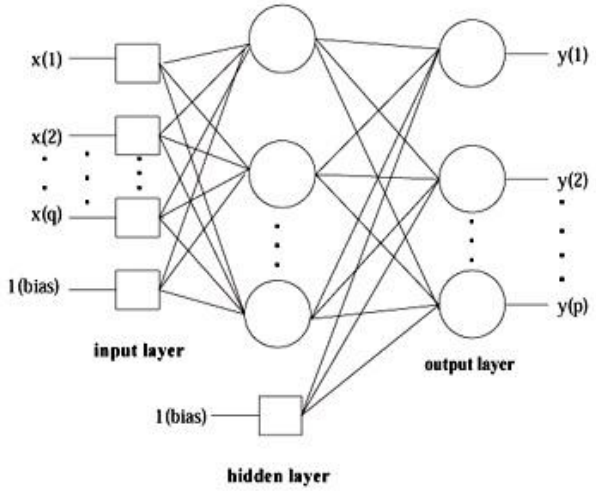

Fig. 4. A neural network Model.

B. Bayesian Classifier

Bayesian Classifier Bayesian classifier is a statistical classification approach based on the Bayes theorem. To calculate probability of A given B, P (B given A) = P (A and B)/P (A) the algorithm counts the number of cases where A and B occurs simultaneously and divides it by the number of cases where A alone occurs. Let X be a data tuple, X is considered "Evidence", in Bayesian terms. Let H be some hypothesis, such that the data tuple X belongs to class C. P (H|X) is posterior probability, of H conditioned on X. P (H) is the prior probability of H in contract.

C. Decision Tree

Decision tree uses the simple divide-and conquer algorithm. In these tree structures, leaves represent classes and branches signify conjunctions of features that lead to those classes. The attribute that most effectively splits samples into different classes is chosen, at each node of the tree. A path to a leaf from the root is found depending on the assessment of the predicate at each node that is visited, to predict the class label of an input. Decision tree is fast and easy method since it does not require any domain information. In the decision tree inputs are divided into two or more groups continue the steps till to

complete the tree as shown on Figure 5. Various decision tree algorithms as follows:

- CART (Classification & Regression Tree)
- C4.5 (Successor of ID3)
- ID3 (Iterative Dichotomise 3)
- CHAID (CHI-squared Automatic Interaction Detector)

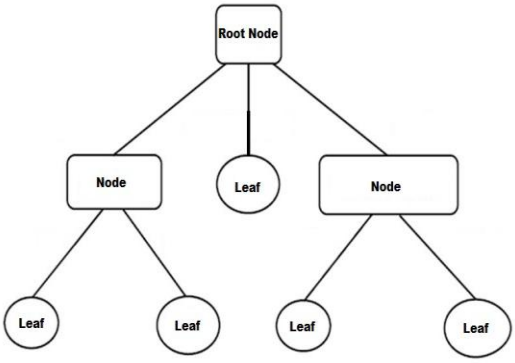

Fig. 5. Decision tree Structure.

D. Support Vector Machine(SVM)

Normally SVM is the classification technique. Initially it is developed for binary type classification later extended to multiple classifications. This SVM creates the hyper plane on the original inputs for effective separation of data points.

VI. CONCLUSION

This paper is presented to study about the various data mining techniques in the agriculture domain to discover hidden interesting patterns. There is variety of data mining tools and techniques that can be applied in agriculture domain. An appropriate data mining technique can be applied for a particular agriculture problem to predict or forecast the future trends and interesting patterns, which can be helpful for the farmers and agriculture department for decision making.

REFERENCES

[1] Thomas Cover and Peter Hart. Nearest neighbor pattern classification. IEEE transactions on information theory, 13(1):21–27, 1967.

[2] JA Hartigan. Clustering aloritms, 1975. [3] Milovs Kovacevic, Branislav Bajat, and Bovsko Gajic. Soil type classification and estimation of soil properties using support vector machines. Geoderma, 154(3):340–347, 2010.

[4] Antonio Mucherino, Petraq Papajorgji, and Panos M Pardalos. A survey of data mining techniques applied to agriculture. Operational Research, 9(2):121–140, 2009.

[5] Antonio Mucherino and Alejandra Urtubia. Consistent biclustering and applications to agriculture. In Industrial Conference on Data Mining-Workshops, pages 105–113, 2010.

[6] Georg Ruß, Rudolf Kruse, Martin Schneider, and Peter Wagner. Estimation of neural network parameters for wheat yield prediction. In IFIP International Conference on Artificial Intelligence in Theory and Practice, pages 109–118. Springer, 2008.

[7] Xindong Wu, Vipin Kumar, J Ross Quinlan, Joydeep Ghosh, Qiang Yang, Hiroshi Motoda, Geoffrey J McLachlan, Angus Ng, Bing Liu, S Yu Philip, et al. Top 10 algorithms in data mining. Knowledge and information systems, 14(1):1–37, 2008.