Special Issue Published in International Journal of Trend in Research and Development (IJTRD), ISSN: 2394-9333, www.ijtrd.com

A Study of Data Mining Tools and Techniques to Agriculture with Applications

Noor Ayesha
Research Scholar, Department of Computer Science, Bharathiar University, Coimbatore, India

***Abstract*-** The commonly used data mining techniques in the field of agriculture. Some of these techniques, such as the k-means, the k nearest neighbor, Artificial Neural Networks(ANN),Support Vector Machines (SVM) and bi-clustering are discussed and an application in agriculture for each of these techniques is presented. Yield prediction is a very important agricultural problem that remains to be solved based on the available data. The problem of yield prediction can be solved and It is our opinion that efficient techniques can be developed and tailored for solving complex agricultural problems using data mining.

Keywords— *Data Mining, K-Means, K-Nearest Neighbor, Artificial Neural Networks, Support Vector Machines, Bi-clusturing.*

I. INTRODUCTION

In Indian agriculture, the volume of data is enormous. The data when become information is highly useful for many purposes. The conventional and traditional system of data analysis in agriculture is purely dependent on statistics. Data mining is a modern data analysis technique. It has wide range of applications in the field of agriculture.

Data mining is the process of discovering previously unknown and potentially interesting patterns in large datasets. The mined information is used for representing as a model for prediction or classification. Datasets from the agricultural domain appear to be significantly more complex than the datasets traditionally 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.

The data can be analyzed in a relational database, a data warehouse, a web server log or a simple text file. Analysis of data in effective way requires understanding of appropriate techniques of data mining. In this paper describe an overview of Data Mining techniques applied to agricultural and their applications to agriculturalrelated areas. Yield prediction is a very important agricultural problem. Any farmer is interested in knowing how much yield he is about to expect.The yield prediction problem can be solved by employing Data Mining techniques such as K Means, K nearest neighbor (KNN), Artificial Neural Network and support vector machine (SVM). Research paper aims at finding suitable data models that achieve a high precision and a high generality with respect to four parameters namely rainfall, year, production and area of sowing.

II. DATA MINING TECHNIQUES

Data Mining techniques are mainly divided in two groups, classification and clustering techniques [6]. 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 [3,4,5] and Support Vector Machines [7], these two classification techniques learn from training set how to classify unknown samples.

Another classification technique, K- Nearest Neighbor [8], 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. One of the most used clustering technique is the K-Means algorithm [9].

A. The k-means approach

The k-means is a data mining technique for clustering. 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 measure of similarities between data samples is provided using a suitable distance: samples that are close to each other are considered similar. The parameter k in the k-means algorithm 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 quite simple. Given a certain partition of the data in k clusters, the centers of the clusters can be computed as the mean 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. It follows that a cluster contains similar data if all its samples are closer to its center and not to the center of some other cluster. Therefore, when samples belonging to a cluster are closer to the center of a different cluster, the k means algorithm moves the corresponding data samples from their original cluster to the new cluster.

B. The k nearest neighbor approach

The k nearest neighbor (k-NN) is a technique for classification. A training set is known, and it is used to classify samples of unknown classification. The basic assumption in the k-NN algorithm is that similar samples should have similar classification. As in the k-means approach, the similarities between samples are measured using suitable distance functions.

The parameter k shows the number of similar known samples used for assigning a classification to an unknown sample. Given an unknown sample, its distances from all samples of the training set are computed, and the k nearest known samples are located. Then, the most frequent classification among known neighbor samples is assigned to the unknown sample

The k-NN method provides a very simple classification rule, but it can be quite expensive to perform. For each unknown sample, its distances from all known samples need to be computed, and this procedure can have a high computational cost. The k-NN method uses the information in the training set, but it does not extract sample needs to be classified. For this reason, many techniques have been developed with the aim of reducing the training set to the minimum indispensable number

of samples which keeps intact the quality of the classification performed by the k-NN.

C. Bi-clustering

Bi-clustering of a set of data, is actually a technique for classification. typical expressions employed when dealing with classification techniques, such as "classes of samples", and with clustering techniques, such as "partition in (bi)clusters", will be both used in the following discussion.

A set of data is basically formed by samples, which are represented by a sequence of features that are considered to be relevant for the representation of the samples. Instead of considering samples only, bi-clustering aims at finding simultaneous classifications of samples and of their features. Moreover, if a training set is known, a bi-clustering can be constructed by exploiting this training set. The corresponding partition in bi-clusters is able to associate subgroups of samples to subgroups of features, so that the features causing the classification of the training set are revealed. This information can then be exploited for performing classification of samples which do not belong to the training set.

In order to perform correct classifications, it is very important that the found bi-clustering is consistet. However, reallife sets of data do not usually allow for consistent bi-clustering. This is due to the fact that some features used for representing the samples actually do not represent the data very well. Such features need therefore to be removed from the set of data, while the total number of considered features is maximized in order to preserve the information in the training set. This problem can be formulated as a 0–1 linear fractional optimization problem, which is NP-hard [14].

This technique has been used so far for analyzing gene expression data [3, 15], where samples may represent diseases, human tissues, etc., but also for studying brain dynamics in patients affected by epilepsy [16]. In this paper, we use the technique for analyzing different wine fermentations by using data experimentally measured during the first 150 hours of the fermentation process. Our main aim is to predict problematic fermentations in time for an enologist to interfere with the process and ensure that the fermentation could end regularly and smoothly. Compounds are regularly measured from diferent wine fermentations of the Cabernet sauvignon.

D. Artificial Neural Network

Artificial Neural Network is a new technique used in flood forecast. The advantage of ANN system over the other system is it can model the rainfall also it predicts the pest attack incidence for one week in advance. Data mining tools are beginning to show value in analyzing massive data sets from complicated systems and providing high-quality information (White and Frank, 2000). An artificial neural network (ANN) is an attractive alternative for building a knowledge-discovery environment for a crop production system. An ANN can use yield history with measured input factors for automatic learning and automatic generation of a system model. In the past few years, several yield simulation models have been built. Ambuel et al. (1994) used a fuzzy logic expert system to predict corn yields with promising results. The functional relationship using the fuzzy logic expert system was expressed linguistically instead of mathematically.

E. Support Vector Machine

SVM is able to classify data samples in two disjoint clusters. SVM are a set of related supervised learning method used for classification and regression. i.e. the SVM can build a model that predicts whether a new example falls into category or the other. A support vector machine is a concept is statistics and computer science for a set of related supervised learning methods that analyze data and recognize patterns used for classification and regression analysis. The SVM takes a set of input data and predicts for each given input which of two possible classes forms the input making the SVM a nonprobabilistic binary linear classifier. An SVM is used in model building which is a representation of the examples as points in space, mapped so that the examples of the separate categories are divided by a clear gap that is as wide as possible. New examples are then mapped into that same space and predicted to belong to a category based on which side of the gap they fall.[1],[2]

F. Bayesian network

Bayesian network is a powerful tool for dealing uncertainties and widely used in agriculture datasets. Bayesian network is a graphical model which encodes probabilistic relationship among variable of interest when it is used with statistical technique, the graphical model has several advantages for data analysis. This technique explicitly deals with uncertainty of data and relationships, and can include both qualitative and quantitative variable. It facilitates effective communication with stakeholders, while promoting a focus on key variables and relationships of the system, rather than being bogged down in details.[17][18]

III. APPLICATION OF DATA MINING TECHNOLOGY IN AGRICULTURE

There are several applications of Data Mining techniques in the field of agriculture. Some of the data mining techniques are related to weather conditions and forecasts. For example, the K-Means algorithm is used to perform forecast of the pollution in the atmosphere [10], the K Nearest Neighbor(KNN) is applied for simulating daily precipitations and other weather variables [11], and different possible changes of the weather scenarios are analyzed using SVMs [12]. K-Means approach to analyze color images of fruits as they run on conveyor belts. Shahin MA et al. [13] uses X-ray images of apples to monitor the presence of water cores, and a neural network is trained for discriminating between good and bad apples.

A. Application of K-means to evaluate soil fertility

Weighted K-means clustering algorithm can be used to evaluate the soil fertility. The algorithm uses AHP to get the weight of soil nutrient attributes. Then combined with K-means clustering algorithm. Finally through the operational efficiency and accuracy to determine the optimal classification, that can improve the clustering algorithm of intelligent. The algorithm and the traditional K-means clustering algorithm are used in the comparison, tests showed that the weighted K-means clustering algorithm has a better accuracy, operational efficiency, significantly higher than the un-weighted clustering algorithm; Comprehensive evaluation of the changes in soil nutrients after precision fertilization that used algorithm. The soil fertility status has a significantly improvement after years of continuous precision fertilizing. The results show that the improved clustering algorithm is a good method to comprehensive evaluation of soil fertility

B. Climate forecasting

Knowing the weather a day or a week in advance is very important especially in agriculture. Weather forecast can

influence decisions, in order to avoid unwanted situations or to take advantage of favorite weather conditions. The variability of the climate is indeed one of the most important factors that seriously impacts agricultural production.While TV channels or journals are able to provide quite accurate forecasts of the weather in the next few days, it is still a big challenge forecasting the weather conditions 3 to 6 months ahead of time. These are the kinds of time intervals to deal with when working in agriculture. The uncertainty about the weather can be devastating in agriculture, because farmers may not be prepared to face the weather conditions that might occur. It can cause also poor productivity, because of the use of conservative strategies that sacrifice productivity to reduce the risk of losses. If the future weather conditions were known, this could be exploited for decreasing unwanted impacts and for taking advantage of expected favorable conditions. Most of the current climate forecasts are based on analysis on theEl Nin˜o-Southern Oscillation (ENSO). This phenomenon is characterized by three phases: warm (El Nin˜o), neutral and cool (La Nin˜a) phases.A k-NN algorithm is used for the recalibration of the precipitation outputs from the FSU-GSM (Florida State University Global Spectral Model) and FSU-RSM (Florida State University Regional Spectral Model) climate models. These climate models may not produce sufficiently accurate daily weather variable outputs to use in crop models.

C. Prediction of problematic wine fermentations

Problems occurring during the fermentation process of wine can impact the productivity of winerelated industries and also the quality of wine . Predicting how good the fermentation process is going to be may help enologists who can then take suitable steps to make corrections when necessary and to en-sure that the fermentation process concludes smoothly and successfully. Therefore, more recently, supervised bi-clustering techniques have been applied to the dataset of wine fermentations.

This technique can simultaneously solve two data mining problems. First, it is able to select the features, the compound measurements, that are actually relevant in the fermentation process, so that useless data can be discarded, and compounds that may cause problematic fermentations can be identified. Second, the information that is acquired by finding bi-clusterings of the dataset can be exploited for performing classifications of new fermentations. Therefore, we can basically perform feature selections and supervised classifications at the same time by using this technique. Each fermentation is here represented by a sample containing all compound measurements taken from the same fermentation process at different times. The technique is able to perform good-quality predictions of problematic fermentations.

D. Estimating soil water parameters

Certain soil parameters are specified. Among these parameters, the ones usually denoted by 94 4 k-Nearest Neighbor Classification the symbols LL, DUL, and PEWS are mostly used. LL is the lower limit of plant water availability; DUL is the drained upper limit; PESW is the plant extractable soil water. Unfortunately, these parameters are usually unknown. The available information about the soils usually concerns their texture, indicating the percentage of clay, silt, sand and organic carbon in the soil. If there is a relationship between the texture information and the parameters needed for the simulation models, then this relationship can be used for obtaining the needed parameters.

The k-NN method can be considered a reasonable alternative to address this category of problems The application discussed in the following and the one discussed in the previous section have some authors in common. This shows how the same methodology can be applied to different problems. Experimental observations show that soils having similar textures also have similar values for the LL, DUL and PESW parameters. Let us suppose then that a database is available where soil data are collected by their textures and LL, DUL and PESW parameters. Let us consider now another soil, whose LL, DUL and PESW parameters are unavailable. In order to find an approximation of the needed parameters, the texture of the new soil can be compared to the textures of the soils in the database. The soil under study most likely has LL, DUL and PESW parameters similar to those of the nearest soils in the database. The distances between soils are based in this case on percentages of clay, silt, sand and organic carbon in the soils. This strategy is nothing else but the k-NN method.

CONCLUSION

Several Data mining techniques used in agriculture study area. We are discussed the few techniques here. Also one technique called K means method is used to forward the pollution in atmosphere. Different changes of weather are analyzed using SVM. K means approach is used to classify the soil and plants. Wine fermentation process monitored using Data mining techniques.The applications that use the K-Means approach , utilize only the basic algorithm, while many other improvements are available.

In conclusion, it is our opinion there is a lot of work to be done on this emerging and interesting research field. The multidisciplinary of this research field is very important, because mathematicians and computer scientists can help agronomists in finding complex solutions to these complex problems. We believe that the use of more complex techniques, such as bi-clustering and high-performing computational systems, such as parallel computers, in data mining, will help solving complex problems in agriculture-related fields.

References

[1] Bernhard E. Boser, IsabelleM. Guyon, and Vladimir N. Vapnik. A training algorithm for optimal margin classifiers. In Proceedings of the 5th Annual ACM Workshop on Computational Learning Theory, pages 144–152. ACM Press, 1992

[2] Ronan Collobert, Samy Bengio, and C. Williamson. Svmtorch: Support vector machines for large-scale regression problems. Journal of Machine Learning Research, 1:143–160, 2001

[3] Georg Ruß, Rudolf Kruse, Martin Schneider, and Peter Wagner. Estimation of neural network parameters for wheat yield prediction. In Max Bramer, editor, Artificial Intelligence in Theory and Practice II, volume 276 of IFIP International Federation for Information Processing, 109–118. Springer, July 2008.

[4] Georg Ruß, Rudolf Kruse, Martin Schneider, and Peter Wagner. Optimizing wheat yield prediction using different topologies of neural networks. In Jos´e Luis Verdegay, Manuel Ojeda-Aciego, and Luis Magdalena, editors, Proceedings of IPMU-08, 576– 582. University of M´alaga, June 2008.

[5] Georg Ruß, Rudolf Kruse, Peter Wagner, and Martin Schneider. Data Mining with neural networks for wheat yield prediction. In Petra Perner, editor, Advances in Data Mining (Proc ICDM 2008), 47–56, Berlin, Heidelberg, July 2008. Springer Verlag.

[6] A. Mucherino, P. Papajorgji, P.M. Pardalos, A Survey of Data Mining Techniques Applied to Agriculture, Operational Research: An International Journal 9(2), 121–140, 2009.

[7] M. Kovacevic, B. Bajat, B. Gajic, Soil Type Classification and Estimation of Soil Properties using Support Vector Machines, Geoderma 154(3–4), 340–347, 2010.

[8] Cover TM, Hart PE, Nearest Neighbor pattern classification. IEEE Trans Info Theory 13(1) : 21-27, 1967.

[9] J. Hartigan, Clustering Algorithms, John Wiles & Sons, New York, 1975.

[10] Jorquera H, Perez R, Cipriano A, Acuna G Short term forecasting of air pollution episodes. In: Zannetti P (eds) Environmental modeling 4. WIT Press, UK, 2001.

[11] Rajagopalan B, Lall U, A K-Nearest Neighbor simulator for daily precipitation and other weather variables. Wat Res Res 35(10) : 3089–3101, 1999.

[12] Tripathi S, Srinivas VV, Nanjundiah RS Downscaling of precipitation for climate change scenarios: a Support Vector Machine approach. J Hydrol 330:621–640, 2006.

[13] Shahin MA, Tollner EW, McClendon RW Artificial intelligence classifiers for sorting apples based on watercore. J Agric Eng Res 79(3):265–274, 2001.

[14] A. Mucherino, S. Cafieri, A New Heuristic for Feature Selection by Consistent Biclustering, arXiv e-print, arXiv:1003.3279v1, March 2010.

[15] S. Busygin, N. Boyko, P.M. Pardalos, M. Bewernitz and G. Ghacibeh, Biclustering EEG Data from Epileptic Patients Treated with Vagus Nerve Stimulation, AIP Conference Proceedings 953, Data Mining, System Analysis and Optimization in Biomedicine, 162–173, 2007

[16] XIA ZJ, LI PP, HU YG. The application of data mining in the prediction of crop growth in the greenhouse [J].Journal of Jiangsu University (Natural Science Edition),2003,24(2):20 22,31.(in Chinese).

[17] J. R. Quinlan. Induction of decision trees. Machine Learning, 1(1):81–106, March 1986.

[18] Applying Naive Bayes Data Mining Technique for Classification of Agricultural Land Soils P.Bhargavi, Dr.S.Jyothi, IJCSNS International Journal of Computer Science and Network Security, VOL.9 No.8, August 2009 117