Agricultural Journal 5 (2): 116-118, 2010
ISSN: 1816-9155
© Medwell Journals, 2010

Brief Survey of Application of Data Mining Techniques to Agriculture

[1]S.S. Baskar, [2]L. Arockiam, [1]V. Arul Kumar and [2]L. Jeyasimman
[1]Department of Computer Science, St. Joseph's College, Trichirappalli, India
[2]Department of Computer Applications,
JJ. College of Engineering and Technology, Trichirappalli, India

Abstract: Survey made on this area reveals the importance of data mining techniques on agriculture. Lots of data mining techniques have been used in agriculture. This survey study presents a brief idea of some of the widely used data mining techniques over agriculture data sets. It is a contemporary technique to find the solution over the traditional and conventional method. It is an infant area with respect to data mining application over agriculture.

Key words: Artificial Neural Networks, bayesian network, support vector machines and decision trees, Trichirappalli

INTRODUCTION

Agriculture is back bone business in India. It contributes 10-15% GDP to the India economy. 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 extracting vital and useful information from large sets of data (Abello *et al*., 2002; Klosgen and Zykow, 2002; Pardalos *et al*., 2007).

In this survey, applications of the data mining techniques in the area of agriculture and its allied areas are studied. Different techniques of data mining have been used in this field. The survey aims to come out of the techniques being used in the agriculture and its allied area. Though, there are lots of techniques available in the data mining, few methodologies such as Artificial Neural Networks, k means approach, K nearest neighbor are popular currently depends on the nature of the data.

Data mining: 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. Data mining techniques has put forward the guidelines for making recommendations for site specific crop management (Chosa *et al*., 2003).

MATERIALS AND METHODS

Artificial Neural Network: Artificial Neural Networks (ANN) are systems inspired by the research on human brain (Hammerstrom, 1993). Artificial Neural Networks (ANN) networks in which each node represents a neuron and each link represents the way two neurons interact. Each neuron performs very simple tasks, while the network representing of the work of all its neurons is able to perform the more complex task. A neural network is an interconnected set of input/output units where each connection has a weight associated with it. The network learns by fine tuning the weights so as able to predict the call label of input samples during testing phase. Artificial neural network is a new techniques used in flood forecast. The advantage of ANN approach in modeling the rain fall and run off 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 (SVM) is binary classifiers (Burges, 1998; Cortes and

Corresponding Author: S.S. Baskar, Department of Computer Science, St. Joseph's College, Trichirappalli, India

Vapnik, 1995). SVM is able to classify data samples in two disjoint classes. The basic idea behind is classifying the sample data into linearly separable. Support Vector Machines (SVMs) are a set of related supervised learning methods used for classification and regression. In simple words given a set of training examples, each marked as belonging to one of two categories, an SVM training algorithm builds a model that predicts whether a new example falls into one category or the other.

Decision trees: 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 Geopspatial Decision Support System (GDSS). Leisa J Armstrong states that data mining approach is one of the approach used for crop decision making.

K nearest neighbor: K nearest neighbor techniques is one of the classification techniques in data mining. It does not have any learning phase because it uses the training set every time a classification performed. Nearest Neighbor search (NN) also known as proximity search, similarity search or closest point search is an optimization problem for finding closest points in metric spaces.

Bayesian networks: A Bayesian network is a graphical model that encodes probabilistic relationships among variables of interest. When used in conjunction with statistical techniques, the graphical model has several advantages for data analysis. One, because the model encodes dependencies among all variables, it readily handles situations where some data entries are missing. Two, a Bayesian network can be used to learn causal relationships and hence can be used to gain understanding about a problem domain and to predict the consequences of intervention. Three, because the model has both a causal and probabilistic semantics, it is an ideal representation for combining prior knowledge (which often comes in causal form) and data. Four, Bayesian statistical methods in conjunction with Bayesian networks offer an efficient and principled approach for avoiding the over fitting of data

Development of a data mining application for agriculture based on Bayesian networks were studied by Huang *et al.* (2008). According to him, Baysian network is a powerful tool for dealing uncertainties and widely used in agriculture data sets. He developed the model for agriculture application based on the Bayesian network learning method. The results indicate that Bayesian Networks are a feasible and efficient.

RESULTS AND DISCUSSION

Clustering can be considered as unsupervised learning problem. It deals with finding a tructure in a collection of unlabeled data. A loose definition of clustering could be the process of organizing objects into groups whose members are similar in some way.

A cluster is therefore a collection of objects which are similar between them and are dissimilar to the objects belonging to other cluster In the event a training set is 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.

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 ≠ j and $k^{\grave{U}}C_k$ = 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.

Ahsan abdullah used the unsupervised clustering techniques of the data through Recursive Noise Removal (RNR). This study revealed that interesting patterns of farmers practices along with pesticide usage dynamics, which helps the farmers to identify the pesticide abuse.

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

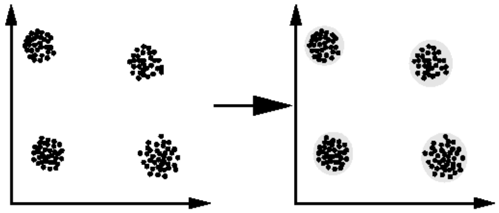

Fig. 1: Clustering

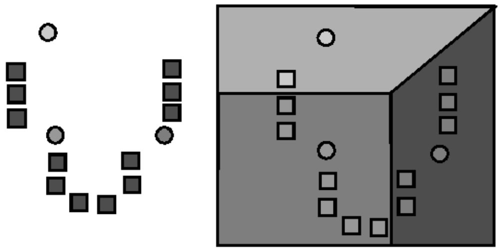

Fig. 2: K means

of the cluster can be computed as the mean of the all sample belonging to a cluster. The center of the cluster can be considered as the representatative of the cluster. The center is quite close to all samples in the cluster.

DM application in agriculture: Several data mining techniques are in agriculture. Few of techniques are elaborated here. K means method is used to forecast the pollution in the atmosphere (Jorquera *et al*., 2001). K nearest neighbor is applied for simulating daily precipitation and other weather variables (Rajagopalan and Lall, 1999). Different possible changes of weather are analyzed using SVM (Tripathi *et al*., 2006). K means approach is used for classifying soil in combination with GPS readings (Verheyen *et al*., 2001). K Means approach was used to classify the soil and plants (Camps-Valls *et al*., 2003). Wine Fermentation process monitored using data mining techniques. Taste sensors are used to obtain data from the fermentation process to be classified using ANNs (Riul *et al*., 2004).

CONCLUSION

It is opinioned that efficient techniques can be developed and tailored to solve complex agricultural problems using data mining.

RECOMMENDATION

In future, lot more advanced techniques can be tailored to agriculture area to solve the complicated problems. Some of mining techniques have not been applied to agricultural problems. It is the opinion that more techniques and algorithms to be studied related agricultural problems will give good result in agricultural growth.

REFERENCES

Abello, J., P.M. Pardalos and M. Resende, 2002. Hand Book of Massive Data Sets. Kluwer Academic Publishers, Norwell, MA, USA., pp: 1236.

Burges, C.J.C., 1998. A tutorial on support vector machines for pattern recognition. Data Min. Knowledge Discovery, 2: 121-167.

Camps-Valls, G., L. Gomez-Chova, C.J. Maravilla, E. Soria-Olivas, J.D. Martin-Guerrero and J. Moreno, 2003. Support vector machines for crop classification using hypersctral data. Lect Notes Comput. Sci., 2652: 134-141.

Chosa, T., Y. Shibata, M. Omine, K. Kobayashi, K. Toriyama and R. Sasaki, 2003. Map based variable control system for granule applicator. J. Jap. Soc. Agric. Machinery, 65: 128-135.

Cortes, C. and V. Vapnik, 1995. Support vector networks. Machine Learning, 20: 273-297.

Hammerstrom, D., 1993. Neural networks at work. IEEE Spectrum, 30: 26-32.

Huang, J., Y. Yuan, W. Cui and Y. Zhan, 2008. IFIP international federation for information processing. http://en.wikipedia.org/wiki/International_Federation_for_Information_Processing.

Jorquera, H., R. Perez, A. Cipriano and G. Acuna, 2001. Short Term Forecasting of Air Pollution Episodes. In: Environmental Modeling, Zannetti, P. (Ed.). WIT Press, UK.

Klosgen, M. and J.M. Zykow, 2002. Hand Book of Data Mining and Knowledge Discovery. Oxford University Press, Oxford.

Pardalos, P.M., L.V. Boginiski and A. Vazacopoulos, 2007. Data Mining in Biomedicine. Springer, New York.

Rajagopalan, B. and U. Lall, 1999. A K-nearest neighbor simulator for daily precipitation and other weather variables. Water Resour. Res., 35: 3089-3101.

Riul, Jr. A., H. de Souse, R.R. Malmegrim, D.S. Jr. dos Santos and A.C.P.L.F. Carvalho *et al*., 2004. Wine classification by taste sensors made from ultra-thin films and using neural networks. Sensors Actuators B Chem., 98: 77-82.

Tripathi, S., V.V. Srinivas and R.S. Najundiah, 2006. Downscaling of precipitation for climate change scenarios: A support vector machine approach. J. Hydrol., 330: 621-640.

Verheyen, K., D. Adriaens, M. Hermy and S. Deckers, 2001. High resolution continuous soil classification using morphological soil profile descriptions. Geoderma, 101: 31-48.