A Brief survey of Data Mining Techniques Applied to Agricultural Data

Hetal Patel
Research Scholar
Charusat, Changa

Dharmendra Patel
Assistant Professor
Charusat, Changa

ABSTRACT

As with many other sectors the amount of agriculture data based are increasing on a daily basis. However, the application of data mining methods and techniques to discover new insights or knowledge is a relatively a novel research area. In this paper we provide a brief review of a variety of Data Mining techniques that have been applied to model data from or about the agricultural domain. The Data Mining techniques applied on Agricultural data include k-means, bi clustering, k nearest neighbor, Neural Networks (NN) Support Vector Machine (SVM), Naive Bayes Classifier and Fuzzy c-means. As can be seen the appropriateness of data mining techniques is to a certain extent determined by the different types of agricultural data or the problems being addressed. This survey summarize the application of data mining techniques and predictive modeling application in the agriculture field.

Key words

Agriculture, Data Mining, k-means, bi clustering, k nearest neighbor, Artificial Neural Network (ANN), Support Vector Machine, Naive Bayesian Classifier, Fuzzy c-means

1. INTRODUCTION

Agriculture is the backbone of the Indian economy. As Mahatma Gandhi said, "India lives in villages and agriculture is the soul of Indian economy". Nearly two-third of its population directly depends on agriculture for its livelihood. In spite of the fact that large areas in India have been brought under irrigation, only one-third of the cropped part is in fact irrigated. The productivity of agriculture is very low. [2] So as the demand of food is increasing, the researchers, farmers, agricultural scientists and government are trying to put extra effort and techniques for more production. And as a result, the agricultural data increases day by day. As the volume of data increases, it requires involuntary way for these data to be extracted when needed. Still today, a very few farmers are actually using the new methods, tools and technique of farming for better production. Data mining can be used for predicting the future trends of agricultural processes.

Data mining is the process to discover interesting knowledge from large amounts of data. (han, 2006) [1] Data mining is the process that results in the discovery of new patterns in large data sets. The goal of the data mining process is to extract knowledge from an existing data set and transform it into a human understandable formation for advance use. It is the process of analyzing data from different perspectives and summarizing it into useful information. There is no restriction to the type of data that can be analyzed by data mining. We can analyze data contained 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. The intention of this paper is to give details about different data mining techniques in perspective of agriculture domain so researchers can get details about appropriate data mining technique in context to their work area. Data mining tasks can be classified into two categories: Descriptive data mining and Predictive data mining.

Descriptive data mining tasks characterize the general properties of the data in the database while predictive data mining is used to predict explicit values based on patterns determined from known results. Prediction involves using some variables or fields in the database to predict unknown or future values of other variables of interest. As far as data mining technique is concern; in the most of cases predictive data mining approach is used. Predictive data mining technique is used to predict future crop, weather forecasting, pesticides and fertilizers to be used, revenue to be generated and so on.

2. METHOD:

The main techniques for data mining include Classification, Clustering, Association rules and Regression. The different data mining techniques used for solving different agricultural problem discussed by Mucherino, A., Papajorgji, P., & Pardalos, P. (2009) are shown in fig 2.1. The graphical representation of different data mining techniques is shown in Figure 2.1. [4]

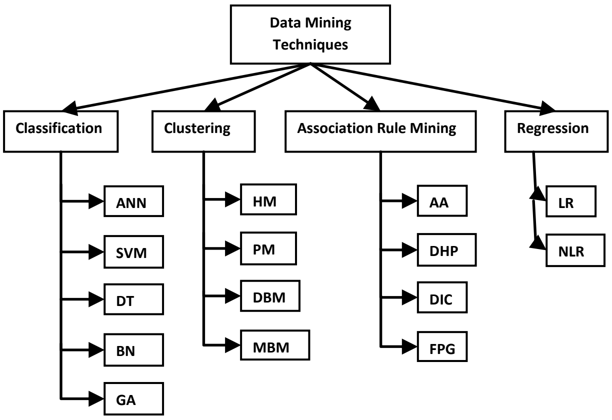

Figure 2.1: Different data mining techniques.

Classification: Classification and prediction are two forms of data analysis that can be used to extract models describing important data classes or to predict future data trends. It is a process in which a model learns to predict a class label from a set of training data which can then be used to predict discrete class labels on new samples. To maximize the predictive accuracy obtained by the classification model when classifying examples in the test set unseen during training is one of the major goals of classification algorithm. Data mining classification algorithms can follow three different learning approaches: supervised learning, unsupervised learning, or semi-supervised learning. The different classification techniques for discovering knowledge are Rule Based Classifiers, Bayesian Networks(BN), Decision Tree (DT), Nearest Neighbour(NN), Artificial Neural Network(ANN), Support Vector Machine (SVM), Rough Sets, Fuzzy Logic, Genetic Algorithms.[5]

Clustering: In clustering, the focus is on finding a partition of data records into clusters such that the points within each cluster are close to one another. Clustering groups the data instances into subsets in such a manner that similar instances are grouped together, while different instances belong to different groups. Since the goal of clustering is to discover a new set of categories, the new groups are of interest in themselves, and their assessment is intrinsic. [6] There is no prior knowledge about data. The different clustering methods are Hierarchical Methods(HM), Partitioning Methods (PM), Density-based Methods(DBM), Model-based Clustering Methods(MBCM), Grid-based Methods and Soft-computing Methods [fuzzy, neural network based], Squared Error—Based Clustering (Vector Quantization), Clustering graph and network data etc.. [1][7][8]

Association Rule Mining: The technique of discovering association rules was originated by Agrawal, Imielinski, & Swami in 1993.[10] Association rule mining technique is one of the most efficient techniques of data mining to search unseen or desired pattern among the vast amount of data. In this method, the focus is on finding relationships between the different items in a transactional database. Association rules are used to find out elements that co-occur repeatedly within a dataset consisting of many independent selections of elements (such as purchasing transactions), and to discover rules. The simple problem statement is: Given a set of transactions, where each transaction is a set of literals (called items), an association rule is an expression of the form X => Y, where X and Y are sets of items. The intuitive meaning of such a rule is that transactions of the database which contain X tend to contain Y.[9] An application of the association rules mining is the market basket analysis, customer segmentation, catalog design, store layout and telecommunication alarm prediction.[11] The different association rule mining algorithm are Apriori Algorithm(AA), Partition, Dynamic Hashing and Pruning(DHP), Dynamic Itemset Counting(DIC), FP Growth(FPG), SEAR, Spear, Eclat & Declat, MaxEclat.[11]

Regression: Regression is learning a function that maps a data item to a real-valued prediction variable. The different applications of regression are predicting the amount of biomass present in a forest, estimating the probability of patient will survive or not on the set of his diagnostic tests, predicting consumer demand for a new product.[3] Here the model is trained to predict a continuous target. Regression tasks are often treated as classification tasks with quantitative class labels. The methods for prediction are Linear Regression (LR) and Nonlinear Regression(NLR).

3. APPLICATION OF DATA MINING TECHNIQUES/ALGORITHMS IN AGRICULTURE

There are number of studies which have been carried out on the application of data mining techniques for agricultural data sets. Naive Bayes Data Mining Technique is used to classify soils that analyze large soil profile experimental datasets. [12] Decision tree algorithm in data mining is used for predicting soil fertility. [13] By using clustering techniques (Based on Partitioning Algorithms and Hierarchical Algorithms) author examines the current usage and details of agriculture land vanished in the past seven years. The overall aim of the research was to determine the land utilization for agriculture and non-agriculture areas for the past ten years.[14] D Ramesh [15] used k-means approach to estimate the crop yield analysis. Some data mining methodology which are used in agricultural domain are reviewed by author Vamanan, R, & Ramar, K [16] in his paper shown in table 3.1 and the outcome of his research is soil classification using Naïve Bayes classifier.

Table: 3.1 Data mining methodologies and its use in Agriculture domain

Methodology	Applications
K-means	Forecasts of pollution in atmosphere Classifying soil in combination with GPS
k-nearest Neighbor	Simulating daily precipitations and other weather variable
Support Vector Machine	Analysis of different possible change of the weather scenario
Decision Tree Analysis	Prediction soil dept
Unsupervised Clustering	Generate cluster and determine any existence of pattern
WEKA Tool	Classification system for sorting and grading mushrooms.

The application of k-means algorithm in the field of agriculture:
The k-means algorithm is used for soil classifications using GPS-based technologies. [17] , Classification of plant, soil, and residue regions of interest by color images [18], Grading apples before marketing [19], Monitoring water quality changes [20] , Detecting weeds in precision agriculture [21], The prediction of wine fermentation problems can be performed by using a k-means approach. Knowing in advance that the wine fermentation process could get stuck or be slow can help the enologist to correct it and ensure a good fermentation process. [22].

The k-nearest neighbor application in the field of agriculture:
The k-nearest algorithm is used in simulating daily precipitations and other weather variables [23] and Estimating soil water parameters and Climate forecasting [4].

The applications of neural networks in the field of agriculture:
The neural network is used in Prediction of flowering and maturity dates of soybean [24] and in forecasting of water resources variables [25].

The applications of SVMs in the field of agriculture:
The application of support vector machine is the crop Classification [26] and in the analysis of the climate change scenarios [27].

4. CONCLUSION

Agriculture is the most important application area particularly in the developing countries like India. Use of information technology in agriculture can change the scenario of decision making and farmers can yield in better way. For decision making on several issues related to agriculture field; data mining plays a vital role. In this paper we have discussed about the role of data mining in perspective of agriculture field. We have also discussed several data mining techniques and their related work by several authors in context to agriculture domain. This paper also focuses on different data mining applications in solving the different agricultural problems. This paper integrates the work of various authors in one place so it is useful for researchers to get information of current scenario of data mining techniques and applications in context to agriculture field.

5. REFERENCES

[1] Han, J, Kamber, M., & Pei, J. (2006). Data mining: concepts and techniques. Morgan kaufmann.

[2] http://www.publishyourarticles.net/knowledge-hub/essay/essay-on-the-importance-of-agriculture-in-the-indian-economy.html

[3] Fayyad, U., Piatetsky-Shapiro, G., & Smyth, P. (1996). From data mining to knowledge discovery in databases. AI magazine, 17(3), 37.

[4] Mucherino, A., Papajorgji, P., & Pardalos, P. (2009). Data mining in agriculture (Vol. 34). Springer.

[5] Beniwal, S., & Arora, J. (2012). Classification and feature selection techniques in data mining. International Journal of Engineering Research & Technology (IJERT), 1(6).

[6] Lior Rokach, Oded Maimon. Clustering Methods. Chap-15

[7] Xu, R & Wunsch, D (2005). Survey of clustering algorithms. Neural Networks, IEEE Transactions on, 16(3), 645-678.

[8] Periklis Andritsos Data Clustering Techniques. University of Toronto, Department of Computer Science. ftp://ftp.cs.toronto.edu/csrg-technical-reports/443/depth.pdf

[9] Srikant, R V Q & Agrawal, R (1997, August). Mining Association Rules with Item Constraints. In KDD (Vol. 97, pp. 67-73).

[10] Agrawal, R., Imieliński, T., & Swami, A. (1993, June). Mining association rules between sets of items in large databases. In ACM SIGMOD Record (Vol. 22, No. 2, pp. 207-216). ACM.

[11] Zaki, M J (1999). Parallel and distributed association mining: A survey. IEEE concurrency, 7(4), 14-25.

[12] Bhargavi, P, & Jyothi, S. (2009). Applying Naive Bayes data mining technique for classification of agricultural land soils. International journal of computer science and network security, 9(8), 117-122.

[13] Jay Gholap. (2012). Performance tuning of j48 algorithm for prediction of soil fertility. Asian Journal of Computer Science And Information Technology 2: 8 (2012) 251– 252.

[14] Megala, S., & Hemalatha, M. (2011). A Novel Datamining Approach to Determine the Vanished Agricultural Land in Tamilnadu. International Journal of Computer Applications, 23.

[15] D Ramesh, B Vishnu Vardhan, (2013). Data Mining Techniques and Applications to Agricultural Yield Data. International Journal of Advanced Research in Computer and Communication Engineering 2(9).

[16] V. Ramesh and K. Ramar, 2011. Classification of Agricultural Land Soils: A Data Mining Approach. Agricultural Journal, 6: 82-86.

[17] Verheyen, K., Adriaens, D., Hermy, M., & Deckers, S. (2001). High-resolution continuous soil classification using morphological soil profile descriptions. Geoderma, 101(3), 31-48.

[18] Meyer, G. E., Camargo Neto, J., Jones, D. D., & Hindman, T. W. (2004). Intensified fuzzy clusters for classifying plant, soil, and residue regions of interest from color images. Computers and electronics in agriculture, 42(3), 161-180.

[19] Leemans, V., & Destain, M. F. (2004). A real-time grading method of apples based on features extracted from defects. Journal of Food Engineering, 61(1), 83-89.

[20] K.A. Klise and S.A. McKenna.(2006). Water Quality Change Detection: Multivariate Algorithms. Proceedings of SPIE 6203, Optics and Photonics in Global Homeland Security II, T.T. Saito,D. Lehrfeld (Eds.)

[21] Tellaeche, A., BurgosArtizzu, X. P., Pajares, G., & Ribeiro, A. (2007). A vision-based hybrid classifier for weeds detection in precision agriculture through the Bayesian and Fuzzy k-Means paradigms. In Innovations in Hybrid Intelligent Systems (pp. 72-79). Springer Berlin Heidelberg.

[22] Urtubia, A., Pérez-Correa, J. R., Soto, A., & Pszczolkowski, P. (2007). Using data mining techniques to predict industrial wine problem fermentations. Food Control,18(12), 1512-1517.

[23] Rajagopalan, B., & Lall, U. (1999). A k–nearest-neighbor simulator for daily precipitation and other weather variables. WATER RESOURCES RESEARCH,35(10), 3089-3101.

[24] Elizondo, D. A., McClendon, R. W., & Hoogenboom, G. (1994). Neural network models for predicting flowering and physiological maturity of soybean. Transactions of the ASAE (USA).

[25] Maier, H. R., & Dandy, G. C. (2000). Neural networks for the prediction and forecasting of water resources variables: a review of modelling issues and applications. Environmental modelling & software, 15(1), 101-124.

[26] Camps-Valls, G., Gómez-Chova, L., Calpe-Maravilla, J., Soria-Olivas, E., Martín-Guerrero, J. D., & Moreno, J. (2003). Support vector machines for crop classification using hyperspectral data. In Pattern recognition and image analysis(pp. 134-141). Springer Berlin Heidelberg.

[27] Tripathi, S., Srinivas, V. V., & Nanjundiah, R. S. (2006). Downscaling of precipitation for climate change scenarios: a support vector machine approach.Journal of Hydrology, 330(3), 621-640.