DATA MINING TRENDS IN AGRICULTURE : A REVIEW

***PATEL, AMIKSHA A.[1] AND KATHIRIYA, DHAVAL R.[2]**

CENTER FOR AGRICULTURAL INFORMATION AND COMMUNICATION TECHNOLOGY
SARDARKRUSHINAGAR DANTIWADA AGRICULTURAL UNIVERSITY
SARDARKRUSHINAGAR – 385 506, GUJARAT, INDIA

****E-MAIL: amiksha_patel@yahoo.com***

1.Assistant Professor (Computer Science), Centre for Agricultural Information & Computer Technology, Sardarkrushinagar Dantiwada Agricultural University, Sardarkrushinagar – 385506 Dist. Banaskantha, Gujarat.
2.Director, Information Technology, Anand Agricultural University, Anand, Gujarat

ABSTRACT

Data mining in agriculture is a relatively novel research field. Agriculture data are highly diversified in terms of nature, interdependency and use of resources for farming. The major problem of using data mining in agriculture is that to solve issues based on the available data and its meaningful outcomes. In data mining, clustering and classification technique make ingenious information in research and knowledge acquisition from integrated farming. And that produces better solution for the farmers about their cultivation (yield). Forthcoming data mining in agriculture rising research field in crop yield analysis. Different data mining techniques such as K-Means, K-Nearest Neighbor (KNN), Artificial Neural Networks (ANN) and Support Vector Machines (SVM) are used for new application research in agriculture field. In this paper, the techniques of predicting yield production of a crop are the centre of focus. Yield prediction is very important in agriculture. The problem of yield prediction can be solved by utilizing data mining techniques.

KEY WORDS: Artificial Neural Networks, Data mining, K-Means, K-Nearest Neighbour, Support Vector Machines

INTRODUCTION

The Indian agriculture is highly differentiated in terms of its climate, soil, water, crops, horticultural crops, plantation crops, medicinal crops, livestock, etc. Today, India ranks second worldwide in farm output. Agriculture is facing the problem of changing in the resources that are directly affecting to the crop yield, so the agricultural productivities in India are unpredictable. For balanced and sustainable growth of agriculture, these resources need to be evaluated, monitored and analysed, so that proper methods can construct.

Accurate and reliable information about crop yield prediction is important for taking decisions for agricultural risk management. Crop yield prediction is also important for supply chain operation of companies engaged in industries that use agricultural produce as raw material (Gleason, 1982). Livestock, food, animal feed, chemical, poultry, fertilizer, pesticides, seed, paper and many other industries uses agricultural products as intergradient in their production processes. An accurate estimate of crop yield helps these companies in planning supply chain decision like

production scheduling and it is useful for business such as seed, fertilizer, agrochemical and agricultural machinery industries and marketing activities based on crop yield.

Data mining techniques till now used widely in business and corporate sectors may be used in agriculture for data characterization, discrimination and predictive and forecasting purposes. Some use of data mining in soil characteristic evaluation has already been attempted.

Different techniques have been intended for mining data over the years. The most used data mining techniques are discussed in this paper. The application of data mining techniques like k-means, bi clustering, k nearest neighbor, Artificial Neural Networks, Support Vector Machine and Naïve Bayes Classifier in the agriculture field.

Data mining techniques can be, therefore, grouped in two different ways. An analysis of the results of crop cutting experiments in agriculture for yield of various crops (Deshpande, 2003). They can be clustering or classification techniques. Additionally, some of them provides a list of information for clustering or classification purpose, while other learns from the available data for how to perform classifications.

So we can say that "Data mining is an expressively production of previously unrevealed set of records to probably useful and materialistic data. It is the process of examining data from different perceptions and summarizing it into useful information."

Data mining techniques

Data mining techniques can be divided in two groups: Classification and Clustering techniques. 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, because, in general, it is used to train the classification technique how to perform its classification. For example, Neural Networks and Support Vector Machines exploit training sets for tuning their parameters in order to solve a particular classification problem. In other words, these two classification techniques learn from a training set how to classify unknown samples. Another classification technique, the k nearest neighbor does not have any learning phase, because it uses the training set, every time a classification must be performed. In the event, a training set is not available, there is no previous knowledge about the data to classify. In this case, clustering technique can be used to split a set of unknown samples into cluster. One of the most used clustering techniques is the k–means method. This paper mainly focuses on the most used techniques in agriculture related fields.

The main techniques for data mining include association rules, classification, clustering and regression. The different data mining techniques used for solving different agricultural problem has been discussed (Mucherino *et al.* 2009).

Association rules

Association rules 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, an association rule is a phrase of the form A = > B, where A, B are sets of objects. The instinctive

meaning of such a rule is that transactions of the database which contain A tend to contain B. (Srikant *et al.,* 1997). An application of the association rules mining is the market basket analysis, customer segmentation, store layout, catalog design and telecommunication alarm prediction. The different association rules mining algorithm are Apriori Algorithm (AA), Partition, Dynamic Hashing and Pruning (DHP), Dynamic Itemset Counting (DIC), FP Growth (FPG), SEAR, Spear, Eclatand Declat, MaxEclat, etc. (Zaki, 1999).

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: semi-supervised learning, supervised learning and unsupervised 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, etc. (Beniwal and Arora, 2012).

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 assembled together, while dissimilar instances belong to diverse groups. Since the aim of clustering is to find out a new set of categories, the latest groups are of interest in themselves, and their assessment is intrinsic (Xu and Wunsch, 2005). 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), network data and Clustering graph (Fayyad *et al.,* 1996).

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 (Sawaitul *et al.,* 2012). Here the model is trained to predict a continuous target. Regression tasks are often treated as classification tasks with quantitative class tag. The methods for prediction are Nonlinear Regression (NLR) and Linear Regression (LR).

Regression analysis (Sellam and Poovammal, 2016), linear are cited. Described about various environmental factors that influence the crop yield and the relationship among these parameters is also established.

Applications of data mining techniques in agriculture

There are many studies which have been accepted on the application of data mining techniques for agricultural data sets. Naive Bayes data mining technique is used to classify soils that analyse large soil profile experimental data sets. Decision tree algorithm in data mining is used for

predicting soil fertility. By using clustering techniques based on Partitioning Algorithms and Hierarchical Algorithm, the land utilization for agriculture and non-agriculture areas for the past ten years has been determined.

As early into the growing season as possible, a farmer is always concerned with how much yield of his crop. In the past, this yield prediction has been relied on farmer's experience for particular yield, crops and climatic conditions. However, this knowledge might also be available, but not exactly for the small scale. Accurate data which can collect in seasons using a multitude of seasons. Advancement and consistence of the agricultural production at a faster in time is one of the basic necessary for agricultural development.

In India, area and yield production of different crops are the results, and reflection of the combined effect of many factors, like agro-climatic conditions, resource endowment, technology level, infrastructure, social and economic conditions. Many schemes have been invented to maximize the productivity of various crops in different agro-climate region, institution, seed, fertilizer, pesticide, fungicide companies and many other activities are actively engaged in the productivity of different crops in different regions and under different condition (Just and Weinenger, 1999; Veenadhari *et al.*, 2011).

Now-a-days, IT had grown to be more and more part of our day by day activities. With IT developments in efficiency can be made in almost any part of industry and services, and now this is true for agriculture. A farmer not harvests only crops but also emerging quantity of data. These data are accurate and in very less amount. There is a lot of data available having information about agriculture. Here soil and yield assets that should be useful in a way that farmers are beneficial. This is a general problem for which data mining is been there. Data mining techniques aim at finding that information in the data that are both important and beneficial to the farmer. A common particular problem of farmers is yield prediction.

Neural networks

Sawaitul *et al.* (2012) focuses the information about weather. The recorded parameters are used to forecast weather. If there is a change in any one of the recorded parameters like wind speed, wind direction, temperature, rainfall, humidity, then the upcoming climatic condition can be predicted using artificial neural networks, back propagation techniques. The increase in signal range will work in large areas as well.

Neural network models for predicting flowering and physiological maturity of soybean (Elizondo *et al*., 1994). Somvanshi *et al.* (2006) deliberated the modelling and prediction of rainfall using artificial neural networks and Box- Jenkins methodology along with other applications of artificial neural networks in hydrology is forecasting daily water hassle and flow forecasting.

Maier and Dandy (2000) used BP neural network and simulated the result using MATLAB. They found suitable data model for achieving high accuracy for price prediction. The prediction is mainly based on only price. Neural networks for the prediction and forecasting of water resources variables.

K-means

Data mining is the process of discovering meaningful patterns and trends by shifting through huge amount of data, using pattern detection technologies as well as statistical and mathematical techniques. Data mining techniques are often used to study soil characteristics. As an example, the K-Mean approach is used for classifying

soils in combination with GPS based techniques (Verheyen *et al.*, 2001).

Urtubia *et al.* (2007) stated that 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 jammed or be slow can help the enologist to correct it and ensure a good fermentation process. The K-Means algorithm is used in performing atmosphere pollution forecast (Jorquera *et al.*, 2001). Verheyen *et al.* (2001) used K-Means approach to classify soils and plants and Camps-Valls *et al.* (2003) used SVMs to classify crops. Apples are checked using different approaches before sending them to the market (Breiman *et al.*, 1984).

Fathima, G. N. and Geetha (2014) used k means and Appriori algorithm, crop type and irrigation parameters and focused on the policies that government could frame by the cropping practices of farmers.

Fuzzy set

Jagielska *et al.* (1999) described the applications to agricultural related areas such as yield prediction is a very important agricultural problem. Any farmer might be interested in knowing how much yield is expected. In the past, yield prediction was achieved by considering farmer's experience on particular field, crop and climate condition. They have discussed additional information about data like probability in probability theory, grade of membership in fuzzy set theory.

Tellaeche *et al.* (2007) summarized an automatic computer vision system for the detection and differential spraying of *Avena sterilis*, a toxic weed growing in cereal crops. With such purpose, it have been designed a hybrid decision making system based on the Bayesian and Fuzzy k-Means classifiers, where the a priori probability required by the Bayes framework is supplied by the Fuzzy k-Means. To classifying plant, soil and residue regions of interest from colour images using fuzzy clusters (Meyer *et al.*, 2004).

Naive Bayes, J 48, random forests, support vector machines, artificial neural networks were implemented (Sujatha and Isakki, 2016). Bhargavi and Jyothi (2009) used climate data and crop parameters for crop yield prediction. Predicted yield using Naive Bayes, Aprioritzy algorithm, the main focus was on various soil parameters like pH, nitrogen, moisture, etc. and comparison accuracy is also presented. Only 77 per cent of accuracy is achieved (Hemageetha, 2016).

Decision tree and Bayesian classification

Veenadhari (2007) considered the influence of climatic factors on major *kharif* and rabi crops production in Bhopal District of Madhya Pradesh State. The findings of the study revealed that the decision tree analysis indicated that the productivity of soybean crop was mostly influenced by comparative humidity followed by temperature and rainfall. The decision tree analysis shows that the productivity of paddy crop was mostly inclined by rainfall followed by comparative evaporation and humidity. For wheat crop, the analysis showed that the productivity is mostly influenced by temperature followed by relative humidity and rainfall. The results of decision tree were confirmed from Bayesian classification. The rules formed from the decision tree are useful for identifying the conditions intended for high or low crop productivity.

Shalvi and De Claris (1998) stated that Bayesian network is a powerful tool and broadly used in agriculture datasets. The model developed for agriculture application based on the Bayesian network learning method. The results showed that Bayesian Networks are feasible and efficient. Bayesian approach improves hydro geological site characterization even when using low-resolution resistivity surveys.

K-nearest neighbour

The k-nearest neighbor classification algorithmic rule may be divided into two phases: coaching section and testing section. Bermejo associated Cabestany urged a reconciling learning algorithmic rule to permit fewer information points to be utilized in coaching information set. Several different techniques are projected to scale back procedure burden of k-nearest neighbour algorithms (Chinchulunn *et al.,* 2010).

A number of studies have been carried out on the application of data mining techniques for agricultural data sets. For example, the K-Nearest Neighbor is applied for simulating daily precipitations and other weather variables (Rajagopalan and Lall, 1999).

K-Nearest Neighbor approach was used to analyse and estimate forest variables for analyzing satellite imagery (Holmgren and Thuresson, 1998). A K–Nearest-Neighbor approach is simulator for daily precipitation and other weather variables (Rajagopalan, and Lall, 1999).

Support vector machine

The main plan of Support Vector Machine (SVM) is to classify information samples into two disjoint categories. The essential plan behind is classifying the sample information into linearly severable. Support Vector Machine (SVM) area unit a group of connected supervised learning ways used for classification and regression (Veenadhari *et al.,* 2011).

The SVM-based data mining is applied to future climate predictions from the second generation Coupled Global Climate Model (CGCM2) to obtain future projections of precipitation. The results are then analysed to assess the crash of climate change on rainfall over India. It is shown that SVMs provide a promising alternative to conventional artificial neural networks for statistical downscaling and are appropriate for conducting climate impact studies (Tripathi *et al.,* 2006).

Various changes of the weather scenarios are analysed using SVMs. Data mining techniques are also applied to study sound recognition problems. Fagerlund (2007) used SVMs for classification of the sound of birds and other different sounds. Camps-Valls *et al.* (2003) used Support vector machines for crop classification using hyper spectral data in Pattern recognition and image analysis.

In the field of Agriculture two or more Data Mining techniques can be applied. Some are related to weather conditions or forecasts

To predict the rainfall, used data mining over two techniques and compares yield prediction based on rainfall between MLR Technique and K-Means. The estimation of average production was 98 per cent using MLR Technique and 96 per cent using K-Means algorithm was given as accuracy (Ramesh and Vishnu Vardhan, 2013).

In the agriculture k-means, ID3 algorithms, the k nearest neighbor, support vector machines, artificial neural networks presented the purpose of data mining techniques and were detailed discussed (Veenadhari *et al.,* 2014).

For weather forecasting, Bendre *et al.* (2015) used Map Reduce and Linear Regression algorithm. The effective model to improve the accuracy of rainfall forecasting is investigated. The forecasting is done based on only a weather data.

CONCLUSION

Agriculture is the most important application area mainly in the developing countries like India. Use of information technology in agriculture can change the condition of decision making and farmers can yield in better way. Data mining plays a crucial role for decision making on several issues related to agriculture field. It

discusses about the role of data mining in the agriculture field and their related work by several authors in context to agriculture domain. There are growing applications for data mining techniques in agriculture. This is relatively a new research field and it is expected to grow in the future. Using data mining techniques in agriculture can take a revolution the current condition of decision making and farmers yield in an advanced way. Several data mining techniques related to agriculture domain is useful for researchers to get information of current scenario of data mining techniques and applications in context to agriculture field.

REFERENCES

Bendre, M. R.; Thool, R. C. and Thool, V. R. (2015). Big data in precision agriculture: weather forecasting for future farming. *1st International Conference on Next Generation Computing Technologies,* pp.744-750.

Beniwal, S. and Arora, J. (2012). Classification and feature selection techniques in data mining, *Int. J. Engg. Res. Tech.* **1**(6): 1-6.

Bhargavi, P, and Jyothi, S. (2009). Applying Naive Bayes data mining technique for classification of agricultural land soils. *Int. J. Compt. Sci. Network Security*, **9**(8): 117-122.

Breiman, L.; Friedman, J. H.; Olshen, A. R. and Stone, C. J. (1984). *Classification and Regression Trees. Monterey,* Calif., U.S.A.: Wadsworth, Inc.

Camps-Valls, G.; Gómez-Chova, L.; Calpe-Maravilla, J.; Soria-Olivas, E.; Martín-Guerrero, J. D. and Moreno, J. (2003). Support vector machines for crop classification using hyper spectral data. In: *Iberian Conference on Pattern Recognition and Image Analysis Pattern*. pp. 134-141.

Chinchulunn, A.; Xanthopoulos, P.; Tomaino, V. and Pardalos, P. M. (2010). Data Mining Techniques in Agricultural and Environmental Sciences. *Int. J. Agril. Environ. Info. Syst*., **1**(1): 26-40.

Deshpande, R. S. (2003). An analysis of the results of crop cutting experiments. Agricultural Development and Rural Transformation Unit, Institute for Social and Economic Change Nagarbhavi, Bangalore, p. 6.

Elizondo, D. A.; McClendon, R. W. And Hoogenboom, G. (1994). Neural network models for predicting flowering and physiological maturity of soybean. *Transactions of the American Aoc. Agric. Engineers (USA),* **37**(3): 981-988.

Fagerlund, S. (2007). Bird species recognition using Support Vector Machines. EURASIP J. Adv. Signal Processing, Article ID 8637: 1-8.

Fathima, G. N. and Geetha, R. (2014). Agriculture crop pattern using data mining techniques. *Int. J. Adv. Res. Computer Sci. Engg.*, 4(5): 781-786.

Fayyad, U.; Piatetsky-Shapiro, G. and Smyth, P. (1996). From data mining to knowledge discovery in databases. *AI Magazine*, **17**(3): 37-54.

Gleason, C. P. (1982). Large area yield estimation/forecasting using plant process models. Presentation at the Winter Meeting American Society of Agricultural Engineers, Palmer House, Chicago, Illinois. December 14-17, 1982.

Hemageetha, N. (2016). A survey on application of data mining techniques to analyse the soil for agriculturalpurpose, *3rd International Conference on Computing for Sustainable Global Development (INDIACom),* pp.3112-3117.

Holmgren, P. and Thuresson, T. (1998). Satellite remote sensing for forestry planning: a review. Scand. J. For. Res., **13**(1): 90– 110.

Jagielska, L.; Mattehews, C. and Whitfort, T. 1999. An investigation into the application of neural networks, fuzzy logic, genetic algorithms, and rough sets to automated knowledge acquisition for classification problems, *Neurocomputing*, 24: 37-54.

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

Just R. E. and Weinenger Q. (1999). Are crop yields normally distributed? *American J. Agric. Econ.,* **81**: 287-304.

Maier, H. R. and Dandy, G. C. (2000). Neural networks for the prediction and forecasting of water resources variables: a review of modelling issues and applications. *Environ. Modeling Software*, **15**(1): 101-124.

Meyer, G. E.; Camargo Neto, J.; Jones, D. D. And Hindman, T. W. (2004). Intensified fuzzy clusters for classifying plant, soil, and residue regions of interest from color images. *Comput. Electronics Agric*., **42**(3), 161-180.

Mucherino, A.; Papajorgji, P. and Pardalos, P. M. (2009). A survey of data mining technique applied to agriculture. *Operational Res.*, **9**(2): 121-140.

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

Ramesh, D. and Vishnu Vardhan, B. (2013). Data mining techniques and applications to agricultural yield data. *Int. J. Adv. Res. Compu. Communi. Engg.*, **2**(9): 3477-3480.

Sawaitul, S. D.; Wagh, K. P. and Chatur, P. N. (2012). Classification and prediction of future weather by using back propagation algorithm - An approach. *Int. J. Emerging Tech. Adv. Engg.*, **2**(1): 110-113.

Sellam, V. and Poovammal, E. (2016). Prediction of crop yield using regression analysis. *Indian J. Sci. Tech.,* **9**(38): 1-5.

Shalvi, D. and De Claris, N. (1998). Unsupervised neural network approach to medical data mining techniques. In: *Proceedings of IEEE International Joint Conference on Neural Networks, (Alaska)*, pp. 171-176.

Somvanshi, V. K.; Pandey, O. P.; Agrawal, P. K.; Kalanker, N. V.; Ravi Prakash, M. and Chand, R. (2006). Modeling and predicaion of rainfall using artificial neural and ARIMA techniques. *J. Indian Geophys Union*, **10**(2): 141-151.

Srikant, R.; Quoc, V. and Agrawal, R. (1997). Mining association rules with item constraints. *Proceedings of the Third International Conference on Knowledge Discovery and Data Mining.* pp. 67-73.

Sujatha, R. and Isakki, P. (2016). A study on crop yield forecasting using classification techniques, *International Conference on Computing Technologies and Intelligent Data Engineering (ICCTIDE)*, pp.1-4.

Tellaeche, A.; BurgosArtizzu, X. P.; Pajares, G. And 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. Springer Berlin Heidelberg.

Tripathi, S.; Srinivas, V. V. and Nanjundiah, R. S. (2006). Downscaling of precipitation for climate change scenarios: a support vector machine approach. *J. Hydro.,* **330**(3): 621-640.

Urtubia, A.; Pérez-Correa, J. R.; Soto, A., and Pszczolkowski, P. (2007). Using data mining techniques to predict industrial wine problem fermentations. *Food Control,* **18**(12): 1512-1517.

Veenadhari, S. (2007). Crop productivity mapping based on decision tree and Bayesian classification. M. Tech Thesis (Unpublished) submitted to Makhanlal Chaturvedi National University of Journalism and Communication, Bhopal.

Veenadhari, S.; Misra, B. and Singh, C. D. (2011). Data mining techniques for predicting crop productivity – A review article. *Int. J. Comp. Sci. Tech*., **2**(1): 98-100.

Veenadhari, S.; Misra, B. and Singh, C. D. (2014). Machine learning approach for forecasting crop yield based on climatic parameters. *International Conference on Computer Communication and Informatics*, pp.1-5.

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

Xu, R. and Wunsch, D. (2005). Survey of clustering algorithms. *IEEE Transactions on Neural Networks* , **16**(3): 645-678.

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

[MS received : September 22, 2017] ***[MS accepted : October 07, 2017]***