DOI: http://dx.doi.org/10.26483/ijarcs.v8i7.4307

ISSN No. 0976-5697

Volume 8, No. 7, July – August 2017

International Journal of Advanced Research in Computer Science

REVIEW ARTICLE

Available Online at www.ijarcs.info

# A REVIEW ON DATA MINING: SCOPE AND APPLICATIONS IN AGRICULTURE

Gurpinder Singh
Department of Computer Engineering
Punjabi University
Patiala,India

Kanwalpreet Singh Atwal (Asst. Prof.)
Department of Computer Engineering
Punjabi University
Patiala, India

***Abstract:*** Data Mining is becoming a trending topic in the field of Agriculture. Crop Yield prediction, Quality Measurement, Soil classification, Crop Disease prediction etc. are some of the recently explored topics. A proper analysis and recommendation on the basis of data mining can help farmers in better understanding and management of their crops and land. Apart from that, patterns evaluated via. Data mining can help the governments for making better policies for Agriculture. As compare to industrial data mining, education data mining, business data mining and medical data mining, agriculture data mining is a novel field. This paper emphasizes on the applications which have been developed using Data Mining and further scope of the data mining. This paper discusses about the use of data mining in some topics relating to agriculture which need immediate attention. Some tools and techniques useful for the purpose of analysis are described.

***Keywords:*** Data, Data Mining, Machine Learning, Classification, Clustering.

## 1. INTRODUCTION

Data Mining is a technique used from extracting useful information from a large dataset. Data Mining uses many techniques for evaluating different patterns from a large amount of data. Data Mining is considered to be a step in the larger process of Knowledge Discovery from Data (KDD). KDD is the process of discovering useful knowledge from data while data mining refers to a particular step in this process [2]. In data Mining large datasets relating to any subject/field are first collected and then all preprocessing is applied. Preprocessing is a process of transforming or making data appropriate for applying data mining techniques to it. Preprocessing may include: cleaning of data, summarization, transformation etc. Data is transformed into the format required for the analysis. The dataset taken into account represents the whole population. Therefore appropriate sampling is important in order to get the accurate results from the dataset. Data Warehouses are the largest storage units of data. Historical data relating to any field can be found in the data warehouse. For example; a bank ABC has many branches but has one center or headquarter. Similarly, operational data is stored in each branch's storage unit but historical data from each branch is collected and stored in one centralized unit called a data warehouse. So that in future any kind of data analysis can be applied to the data.

Data Mining incorporates many techniques like: clustering, classification, machine learning, Support Vector Machines, Regression, Association Rules etc. Further these techniques can be applied on the dataset by different algorithms. An overview of these different techniques is shown in the figure 1.1.

Data Mining in Agriculture is an emerging area and attracting many data analysts and data mining experts to focus their studies on it. Summary information about crop production can help the farmers identify the crop losses and prevent it in future [3]. Many other problems can be formulated in this field which when solved can help farmers in decision making and managing their crops efficiently. Data mining in agriculture can give farmers information about various future risks and hazards. For making more suitable systems for decision making, data mining can be used.

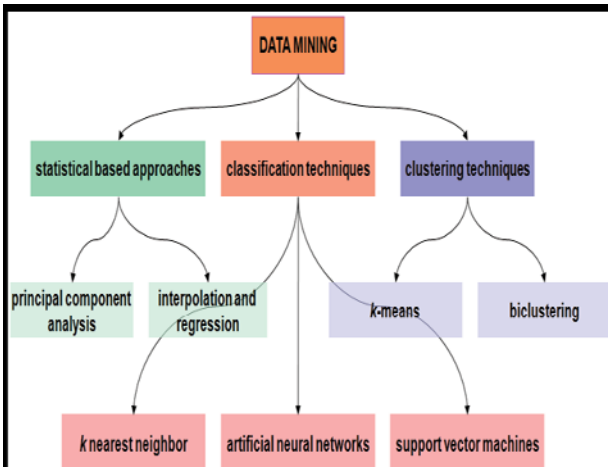


Fig. 1 A schematic representation of the classification of the data mining techniques discussed.[1]

Today, different areas are using data mining, for example financial data collected from banking and financial industries are often comparatively absolute, reliable, and of high quality, which helps methodical data analysis and data mining. It is used extensively in the retail industry because it collects huge amount of data on customer shopping trends, sales of the company etc. This helps the company to analyze both the sales data and data relating to the customer which helps them in making better business decisions. Data mining techniques bring out the customer behaviors and popular choices made by the customers from which the company can estimate that which product's sales are better than the other. Telecommunication industry also uses data mining which

has extended its application from providing telephone services to offer many add-on services like fax, Internet and cellular phone [4]. Many scientific applications including Biological Data Analysis, Intrusion Detection and Agriculture Sector also demand the use of data mining techniques. Data mining in agriculture sector however is just started to give its services in solving various problems. This paper discusses various application of data mining in agriculture.

## 2. AGRICULTURE DATA

In India, it has been seen that problems occur in getting appropriate agricultural data for the purpose of analysis. Inappropriate Government policies, unawareness of farmers and lack of surveys are some of the major reasons for the problem of insufficient data collection in agriculture. However globally, there are various soil, weather and agricultural datasets which can be used for purpose of exploration by applying different data mining techniques. LUCAS [11], GLOBAL_MICROBIAL_BIOMASS [10], Pesticide Data Program (2015), Pesticide use in agriculture (USA) [12] etc. Although there is only a limited dataset available but still there is a lot to be done in order to analyze and explore these dataset.

Deficiency of recorded agricultural data allows the expert to collect the primary data. This process is mostly time consuming but accurate data can be recorded. Agriculture data varies according to requirement. However some vital factors relating to agricultural data are: Farmer practices (of any specific crop), specification of the crop/fruit, chemical properties of the soil, climate of the region, Government policies, MSP, Sales record of pesticide/fertilizers, Rainfall data, Temperature data etc. These are some of the factors which happen to be a part of any research in Agricultural data mining. Experts choose any attribute according to their need and then apply different data mining techniques to mine different patterns out of the data.

## 3. CLASSIFICATION

Classification is one of the major techniques used in data mining while other being the Clustering. Classification and prediction are sometimes used as synonyms but in actual there is difference between the two. Classification refers to prediction of categorical values however prediction models predict continuous values as well. In classification the class labels are already known [4]. For instance let us assume a problem which involves the prediction "whether to play or not to play" on the basis of parameters like: Rain, time, homework, Play. Here Play is the class attribute which has two categories; Play, No Play. So here classification model or decision tree will generate a model which will predict whether to play or not based on the prior information of the class label.

Classification algorithms include: k-Nearest Neighbors, Naïve Bayes, ID.3, CART (Classification and Regression Tree), CHAID (Chi-Square Automatic Interaction Detector) and MARS which extends the decision trees in order to handle numerical data more precisely. K-nearest is however the most widely used classification algorithm which has its application in Concept Search and Recommender Systems [5].

K-Nearest Neighbors [6] algorithm divides the data set into two portions which are called training set and test set. These sets are usually divided in the ration of 70:30, 70% being the training set and 30% being the test set. Then the algorithm uses the training set to train the model for accurate prediction. To check the accuracy of the developed model it is then applied to the test set and a confusion matrix is created which shows how many records belonging to a particular attribute/field have been correctly predicted.

## 4. MACHINE LEARNING

Machine learning is considered to be a part of Artificial Intelligence. Machine Learning algorithms learn on their own experience hence do not need any human to enhance their ability. Popular Machine Learning application include following: Ad placement, credit scoring, fraud detection, stock trading, Web Search, spam filters, recommender systems, computer vision and drug design. Other than these, Amazon's algorithms for book recommendation based on previous book purchase history and Netflix's algorithms for movie/show recommendation are popular examples of machine learning use in our daily life [13].

Three categories can be formed to divide machine learning algorithms: Supervised learning, unsupervised learning and enforcement learning [8]. Supervised learning is useful when the class labels are known in advanced and model is then trained to predict the class of a particular record. However, unsupervised learning is useful in cases where the challenge is to discover implicit relationships in a given *unlabeled* dataset [8]. In enforcement learning there is no proper class label or error available but some form of feedback is available at each predictive step. Ordinary least square regression, Logistic Regression and Support Vector Machines are a few machine learning algorithms which are widely used in day to day applications.

## 5. CLUSTERING

In clustering there is no training set as the class labels are unknown [5]. For example: let us assume that we have a dataset of cows belonging to different breeds and it contains the following attributes/variables; Height, width, weight and color. But we don't have the information about the breed of any of cow. So, on the basis of these four attributes we would make clusters (number of clusters can be selected with various methods) such that each cluster would contain only those records or objects which have more similarities with each other than those of other clusters. The principle which is used here is maximizing the intra-class similarity and minimizing the interclass similarity (Jiawei Han et al.).

Clustering algorithms can be divided into two categories which are unsupervised linear clustering and unsupervised non-linear clustering. The former includes the algorithms like Gaussian clustering, Hierarchical clustering, fuzzy c-means, quality threshold, k-means etc. and latter includes MST based clustering algorithms, kernel k-means clustering algorithm and density based clustering algorithm [6].

K-means algorithm which comes in handy for agricultural data mining is discussed in this paper. K-means has been used in a research about agricultural yield data [30]. Another major application of clustering was encountered in the prediction of olive production in Thassos [4]. Focus of k-

mean is to partition a dataset in which the data in a group is more similar to each other. K in k-means describes the number of clusters that should be made. Centers are marked for all the clusters in a way that they are as far from each other as possible because they can produce results if kept close.

For partitioning Euclidean distance can be used and then the objects which are near to a certain centroid will be considered a part of that cluster. The first stage is considered done after all the points are calculated. After the first stage, recalculate new k centroids and repeat the calculations of all the points. This method is performed until there is no ambiguity in the clusters and they are clearly away from each other. Other usage of k-means algorithm in the field of Agriculture includes: Forecasting pollution in the atmosphere [23], Soil classifications using GPS-based technologies [24], Classification of plant, soil, and residue regions of interest by color images [25], Predicting wine fermentation problems [26], grading apples before marketing [27], Monitoring water quality changes [28], Detecting weeds in precision agriculture [29].

## 6. APPLICATIONS OF DATA MINING IN AGRICULTURE

Many applications have been developed by the use of data mining in Agriculture. Sally Jo Cunningham and Geoffrey Holmes have discussed about the innovative techniques used in agriculture for the purpose of grading the mushrooms [7]. Three grades A, B and C was given to the mushrooms of different qualities. Dataset containing records of 282 mushrooms was used. J4.8 algorithm was used to classify the mushrooms into different grades [6].

DSSAT [8], CROPSYST [9], and GLEAMS [10] are some of the models developed for the purpose of simulating the soil dynamics. Three most used parameters are DUL, LL and PEWS. Here, DUL means drained upper limit: LL refers to lower limit of the plant and PEWS is the plant extractable soil water.

For mining spatio-temporal data, Independent component analysis technique has been used. This technique mined patterns in weather data by the use of NAO (North Atlantic Oscillation) as an example [11]. Bayesian's posteriori classifier was used to interpret distribution of the paddy in three counties of Taiwan during the year of 2000. It was done by using multi-temporal imageries together with the cadastre GIS [13].

To check the poisonous effect of pesticides on humans pesticide use on cotton crop was taken into account. Because pesticide effect can't be directly studied on human beings therefore cotton was used. For pesticide data as well as numeric data calculation, COF Clustering tool is used [12].

A fungal disease of Mango named Powdery Mildew shows devastating effects on the quality of mangos. This disease of Mangos was predicted [14] using Decision Tree induction, Rough Sets (RS) and hybridized Rough Set based Decision Tree Induction (RDT) in comparison with the standard Logistic Regression (LR) method.

To provide information to the growers of Tomato regarding various diseases and their control measures, a web based expert information system was developed. This system implements the ID3 classification algorithm [15]. Also this expert system allowed the growers to get in contact with each other to discuss various practices and control measures to use.

MLR and K-means were applied to predict the yield of the crop. The dataset used includes production, Area of sowing, rainfall and date. MLR (Multiple Linear Regression) gave better results than K-means [4]. These are only a few application of data mining but there is a bigger scope and possibility of exploration in the field of agriculture. Support Vector Machines also play a major role in exploring the datasets [16].

In order to extract regular and interesting patterns from large spatial databases of agriculture, spatial data mining methods were studied [17]. Aim of this study was to find out trends in agriculture production with reference to the availability of inputs. The Real vs. Counter and predicted graph described how closely the poly analyst prediction follows the actual value of the attribute over the range of the dataset.

Effects of climatic factors on major kharif and rabi crops production were studied [18]. The dataset taken in account focused the crop records from Bhopal District of Madhya Pradesh. The results of the study showed that productivity of the soybean crop was mostly influenced by Rainfall, Humidity and temperature. These finding were illustrated in the form of a decision tree. However production of paddy crop was influenced mostly by Rainfall and further by Relative humidity and Evaporation which was also analyzed in the form of a decision tree. On the other hand Temperature was the main factor which influenced the production of wheat. Bayesian classification algorithm was used for this study.

A real-time grading method for classifying apples is proposed [19]. Properties like color of fruit, shape of fruit, length of fruit and soil etc. can be used to extract and identify target features by the use of machine vision. In order to assure the quality of the apple or to classify them into defected or good, first pictures of the surface of the apple are taken by a camera. Camera is located above the conveyer belt and when apples pass on the belt it take pictures. In the next step segmentation is applied. For the purpose of segmentation different supervised or unsupervised techniques are used.

Neural Networks has been also used with success in the field of agriculture. For example an approach has been proposed to [20] to evaluate sugar and acid content of oranges by the use of machine vision. The proposed model described that low Height, reddish, medium size and glossy orange fruits are relatively sweet.

The coughing sound of pigs has been monitored in order to find out any possible health problems in pigs [21]. Neural network approach is used and a chamber was built. This chamber was made of metal and covered with transparent plastic for performing experiments. It was 2m long, 0.80 m wide and 0.95 m high. The pigs enter the chamber and their coughing sounds are recorded by a microphone. 354 sounds were recorded for purpose of training a model to predict the health problem accurately [22].

## 7. CONCLUSION

Various problems in agriculture not only relating to crop growth, quality assurance but also the condition of a farmer can be dealt using appropriate data mining techniques.

Proper actions should be taken by the Governments in order to collect the appropriate agricultural data for the sole reason of applying data mining techniques on it. Agriculture field promises a great deal of work to be done and new applications to be developed in order to enhance the knowledge about certain behaviors of the crops, animal etc. Some problems which can be taken into account include: analysis of farmer conditions, mining of crop data in order to recommend various pesticide/fertilizers. Recommendation to the farmer can be given on the basis of growth potential of his/her soil. Applications can be developed to manage risks, quality assurance etc. Bi-clustering techniques can be applied to more complex agricultural data. Also data mining techniques can be applied in a parallel environment which is still unexplored. Mathematicians and computer scientists have to come together to bring out the best results in agricultural data mining with the help of agronomists.

## 8.REFERENCES


[1] Mucherino A., Papajorgji P.J and Pardalos P.M(2009). Data Mining in Agriculture. Springer.

[2] Kamber, J. H. (2000). Data Mining Concepts and Techniques. UrbanaChampaign: Morgan Kaufmann.

[3] N. Gandhi and L. J. Armstrong, "A review of the application of data mining techniques for decision making in agriculture," **IEEE Conference Publications** 2016- 2nd International Conference on Contemporary Computing and Informatics (IC3I), Greater Noida, India, 2016, pp. 1-6.

[4] Ramar, V. R. (2011). Classification Agricultural Land Soils: A Data Mining Approach. Agricultural Journal , 8286.

[5] A. Mucherino, P. P. (2009). A survey of Data Mining Techniques applied to Agriculture. SpringerVerlag.

[6] Holmes, S. J. (1999). Developing Innovative Applications in Agriculture Using Data Mining.

[7] Cunningham S.J, Holmes Geoffrey,(1999). Developing Innovative applications in Agriculture using Data Mining.New Zealand Foundation for Reseachr and technology.

[8] Jain Rajni, Minz, S., V. Rama Subramaniam. (2009). Machine learning for forewarning cropdiseases. J. Ind. Soc. Agri. Stat. 63(1): pp. 97-107.

[9] Meyer GE, Neto JC, Jones DD, Hindman TW, (2004), Intensified fuzzy clusters for classifying plant, soil, and residue regions of interest from color images. Computer Electronics Agric Vol. 42: pp. 161–180.

[10] Jones JW, Tsuji GY, Hoogenboom G, Hunt LA, Thornton PK, Wilkens PW, Imamura DT, Bowen WT, Singh U., (1998), Decision support system for agrotechnology transfer: DSSAT v3. In: Tsuji GY, Hoogenboom G, Thornton PK (eds) , Understanding options for agricultural production. Kluwer Academic Publishers, Dordrecht, pp 157–177.

[11] Abdullah, A., Bulbul.R., Tahir Mehmood. (2005). Mapping nominal values to numbers by data mining spectral properties of leaves. Proc. of 3rd International Symposiumm on Intelligent Information Technology in Agriculture. Beijing, China. Oct, 2005.

[12] McQueen Robert J, Garner S.R.,Nevill- Manning C.G. , Ian H. Witten, (1995). Applying machine learning to agricultural data. Comptuers and Electronics in Agriculture. Vol. 12:pp. 275- 293.

[13] Basak J., Sudharshan, A., Trivedi D., M.S.Santhanam. (2004). Weather Data Mining Using Independent Component Analysis. J. of Machine Learning Research 5: pp. 239-253.

[14] Chi-Chung LAU, Kuo-Hsin HSIAO, (2005). Bayesian Classification For Rice Paddy interpretation. Paper presented in Conference on data mining held at China Tapei. December, 2005

[15] Using Data Mining to Discover Patterns in Autonomic Storage Systems. Zhenmin Li, Sudarshan M. Srinivasan, Zhifeng Chen, Yuanyuan Zhou, Peter Tzvetkov, Xifeng Yan, and Jiawei Han. 1st Workshop on Algorithms and Architectures for Self- Managing Systems in conjunction with ISCA and SIGMETRICS, June 2003.

[16] CampsValls G, G.C. L.M. O. G. (2003). Support Vector Machine for crop classification using hyperspectral data. Lect Notes Comp Sci , 134141.

[17] Shahin MA, T. E. (2001). Artificial Intelligence classifiers for sorting apples based on watercore. J Agric Eng , 265274.

[18] Miller, D., J. McCarthy and A. Zakzeski,(2009). A Fresh Approach to Agricultural Statistics: Data Mining and Remote Sensing. Section on Government Statistics – JSM 2009, pp. 3144-3155.

[19] V. Leemans, M.F. Destain, A Real Time Grading Method of Apples based on Features Extracted from Defects, Journal of Food Engineering **61**, 83–89, 2004.

[20] N. Kondo, U. Ahmad, M. Monta, H. Murase, Machine Vision based Quality Evaluation of Iyokan Orange Fruit using Neural Networks, Computers and Electronics in Agriculture **29**, 135–147, 2000.

[21] A. Chedad, D. Moshou, J.M. Aerts, A. Van Hirtum, H. Ramon, D. Berckmans, Recognition System for Pig Cough based on Probabilistic Neural Networks, Journal of Agricultural Enginnering Research **79** (4), 449–457, 2001.

[22] B. Moreaux, D. Beerens and P. Gustin, Development of a Cough Induction Test in Pigs: Effects of SR 48968 and Enalapril, Journal of Veterinary Pharmacology and Therapeutics **22**, 387–389, 1999.

[23] H. Jorquera, R. Perez, A. Cipriano, and G. Acuna, Short Term Forecasting of Air Pollution Episodes, In: Environmental Modeling **4**, P. Zannetti (Ed.), WIT Press, UK, 2001.

[24] V.N. Vapnik, Statistical Learning Theory, JohnWiley & Sons, 1998.

[25] G. E. Meyer, J. C. Neto, D. D. Jones, T.W. Hindman, Intensified Fuzzy Clusters for Classifying Plant, Soil, and Residue Regions of Interest from Color Images, Computers and Electronics in Agriculture **42**, 161–180, 2004.

[26] A. Urtubia, J. R. Perez-Correa, A. Soto, P. Pszczolkowski, Using Data Mining Techniques to Predict Industrial Wine Problem Fermentations, Food Control **18**, 1512–1517, 2007.

[27] V. Leemans, M.F. Destain, A Real Time Grading Method of Apples based on Features Extracted from Defects, Journal of Food Engineering **61**, 83–89, 2004.

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

[29] A. Tellaeche, X.-P. Burgos-Artizzu, G. Pajares and A. Ribeiro, A Vision-Based Hybrid Classifier for Weeds Detection in Precision Agriculture Through the Bayesian and Fuzzy k-Means Paradigms, Advances in Soft Computing 44, 72–79, 2008.

[30] D Ramesh, B. V. (2013). Data Mining Techniques and Applications to Agricultural Yield Data. International Journal of Advanced Research in Computer and Communication Engineering , 34773480.