International Journal of Modern Electronics and Communication Engineering (IJMECE) ISSN: 2321-2152
Volume No.-7, Issue No.-4, July, 2019

# A survey on application of data mining techniques for agricultural data analysis

Nishchal Adil
Department of Computer Science and Engineering
RSR Rungta College of Engineering and Technology
Bhilai, India
E-mail: nishchaladil@gmail.com

Somesh Dewangan
Department of Computer Science and Engineering
RSR Rungta College of Engineering and Technology
Bhilai, India
E-mail: somdev2016@gmail.com

Kusum Sharma
Department of Computer Science and Engineering
RSR Rungta College of Engineering and Technology
Bhilai, India
E-mail: prkussum@gmail.com

**Abstract:** India leads the world in production of milk, pulses and jute. In terms of agriculture production, India has gained second position all around the world. The farming sector and its associated industries such as lumbering, forestry, and fishing contribute high percentage towards the GDP of the country. As half of the manpower of India is employed in this sector, it is the largest source of livelihood in India. However, India still has many concerns related to agriculture. Use of data mining technologies in agriculture can help farmers in decision making and help them yield in a better way. In this review paper, the role of data mining and their related work by several authors in context with agriculture field has been discussed. It conjointly discusses diverse applications of data mining targeted to solve agricultural issues. To help other researchers to get information of current scenario of data mining techniques and applications in context to agriculture field, the work of various authors has been consolidated in this paper.

**Keywords:** Data mining; Clustering; Classification; Association; Regression; Agriculture

## I. INTRODUCTION

Agriculture is the most important economical sector of India; therefore one can easily assume it as backbone of the nation. It is the primary source of income and survival for more than 50% of Indian population. India has topped in production of many crops and spices all around the world. The climate of India varies from humid and dry tropical in south to temperate alpine in the northern reaches. It also has great diversity in ecosystem. However, having such importance of agriculture, there are many growing concerns. In spite of the fact that large areas in India have been brought under irrigation, only one third of the cropped part is irrigated. Stressing the water resources of the country will need realignment and rethinking. While India has achieved food sufficiency in production, this country is home for over 190 million undernourished people and quarter of world's hungry people. To overcome problems related to agricultural sector and to help farmers for better productivity certain new advancements in technologies can be employed. One of the methods that can be applied is data mining. It is a process, in which a large set of data is processed to give new patterns. Its goal is to extract information from the large data set and change it into human understandable format for further using it.

Since, the agriculture is done from the ages; a vast collection of past data of agriculture is available. Also at present scenario farmers, government and agricultural scientist have put an extra effect and technologies for increasing production. This has resulted in increase of agricultural data day-by-day.

Data mining can analyze any type of data and there is no such restriction for using any particular type of data. The data for analysis can be gathered from heterogeneous sources. It is necessary to understand appropriate technique of data mining for analyzing the data.

This paper is focused to provide details about different data mining techniques in the aspect of agriculture domain so that detailed information can be gathered by the researchers about appropriate data mining techniques related to their work area.

There are mainly two categories in which data mining tasks can be classified into: Descriptive data mining and Predictive data mining.

Descriptive data mining is used to characterize the general properties of the data in the database and provide latest information on past or recent events while predictive data mining provides future queries and predict explicit values based on patterns determined from known results [1]. Predictive approach for data mining is commonly used in agriculture, to estimate crop yield, to analyze various climatic factors, to calculate and recommend amount of fertilizers and pesticides.

The process discovering useful information from large data sets involves different steps as shown in Figure 1.

International Journal of Modern Electronics and Communication Engineering (IJMECE) ISSN: 2321-2152
Volume No.-7, Issue No.-4, July, 2019

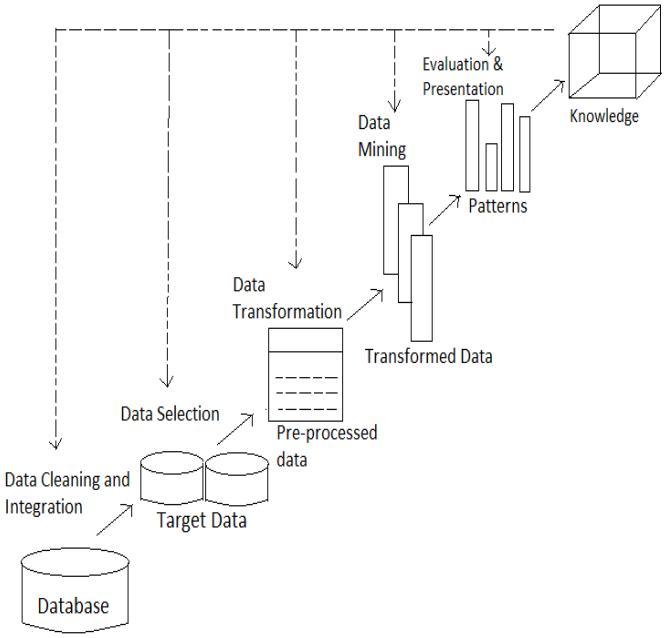


**Figure1.** Data Mining Process

## II. DATA MINING TECHNIQUES

Based on type of task tried to achieve data mining tasks can be classified into two types i.e. predictive tasks and descriptive tasks. In predictive data mining various inferences are carried out on previously stored data set in order to prognosticate how newly derived data set will behave whereas descriptive data mining finds significant patterns and information from previously stored data sets. Different data mining tasks are shown in figure 2.

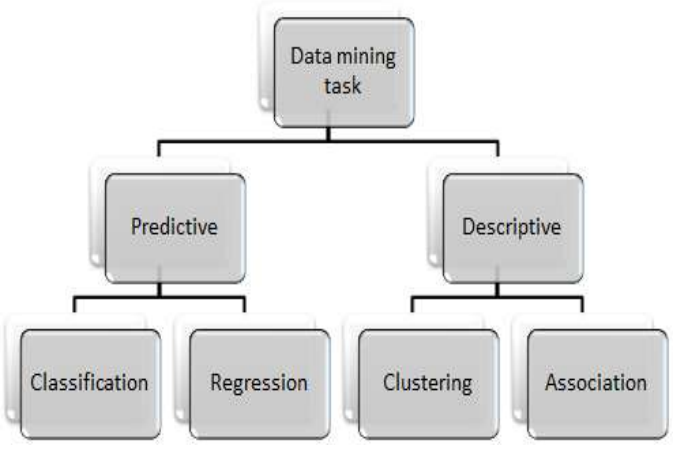


**Figure2.** Data Mining Tasks

### A. Cluster Analysis

Cluster analysis is based on identification and collection of data objects exhibiting similar behavior within a group. A cluster is formed by partitioning large data sets into groups according to their similarities, while dissimilar attributes belong to different groups. From a machine learning point of view clusters tends to bring out hidden patterns, the search for clusters is unsupervised learning, and the resulting system represents a data concept. Based on cluster models clustering algorithms can be classified as Hierarchical Clustering, Partitioning relocation Clustering, and Density Based Clustering, Grid Based methods [2].

### B. Classification Analysis

Classification analysis is a data mining technique which predicts the class of an object by building a model based on its attributes. Datasets will be available, each dataset having a group of attributes. One attribute will be class attribute and the goal of classification task is assigning a class attribute to new set of records as accurately as possible. Classification involve two step process first is learning or training phase in which classification model is constructed and various algorithms are used to build classifier, the classification model is trained using the training set. Next is classification step in which the constructed model is tested on the test data and obtain the accuracy of classification rules.

The different classification techniques for discovering knowledge are Rule Based Classifiers, Bayesian Networks (BN), Decision Tree (DT), Nearest Neighbor (NN), Artificial Neural Network (ANN), Support Vector Machine (SVM), Rough Sets, Fuzzy Logic, and Genetic Algorithms [1].

### C. Association Rule Mining

Association discovers the association or connection among a set of data items, the association rule shows how frequent the item set occur in a transaction. Association identifies the relationship between objects. In association relations between items at same transaction are considered to identify the pattern. In general association rule mining can be viewed as two steps process first find all the frequent item sets and then generate strong association rules from the frequent item sets. The different association rule mining algorithms are Apriori Algorithm, Partition, Dynamic Hashing and Pruning (DHP), Dynamic Item set Counting, FP Growth [3].

### D. Regression

Regression is a data mining technique used to map a data item into a real value prediction variable. It is used to predict a range of numeric values which is also known as continuous values. Regression involves predictor variable (the values which is known) and response variable (values to be predicted). Basic regression techniques are Linear Regression and Non-Linear Regression [3].

## III. DATA MINING TOOLS

In today's world there is number of organizations producing myriad of data and it is very important to handle this huge data to obtain useful information, data mining tools help to apply various data mining algorithms and visualizations in very less time. Given below are some the freely available and open source software which can be used for application of data mining.

### A. Rapid Miner

Rapid Miner is software developed for data analytics; it integrates all data science tasks such as data preparation, machine learning and predictive analytics. Rapid Miner Studio is a free edition; other products are Rapid Miner Auto Model, Rapid Miner Turbo Prep, Rapid Miner Server, and Rapid Miner Radoop [4].

### B. Orange

Orange is an open source toolkit for machine learning, data mining and data visualization, it's Graphical User Interface helps experts as wells as for the beginners and learners to focus on data analysis rather than coding. It features visual programming front end for explorative data analysis and interactive visualization [5].

### C. Weka

Weka stands for Waikato Environment for Knowledge Analysis. It incorporates a graphical user interface along with a collection of algorithms for machine learning and data visualization tools for application data mining and predictive modeling. It consists of tools for data preparation, classification, regression, clustering, association rules mining and visualization [6].

### D. KNIME

KNIME (Konstanz Information Miner) is open source software used for data analytics, reporting and integration platform. By using its modular data pipelining concept, it integrates various components for machine learning and data mining [7].

### E. Rattle

Rattle (R Analytical Tool To Learn Easily) is a GUI based data mining application which is written R. It was developed specifically to ease the transition for basic data mining to sophisticated data analysis using powerful statistical language [8].

## IV. LITERATURE REVIEW

In agricultural industry there are numerous decisions to be made by farmer and growers based on various environmental factors, Jharna Majumdar et al. [9] have done comprehensive study for analyzing agricultural data using data mining techniques in order to find parameters to increase production of crops. The data of districts of Karnataka has been clustered based on attributes having similar temperature, rainfall and soil type using modified approach of DBSCAN (Density Based Spatial Clustering of Applications with Noise) algorithm. PAM (Partitioning Around Medoids) and CLARA (Clustering Large Applications) algorithms are used for clustering data based on districts yielding highest crop production and derive the optimal parameters to produce maximum yield. Multiple linear regression is used to predict the annual crop production.

J. Muangprathub et al. [10] developed a system which deploys wireless sensor network that enables optimal use of water for agricultural crops. The researchers worked towards designing a control system by employing sensors in the agricultural field and manage data through smart phone and web application. System consists of three components first is the control box hardware which is connected to soil moisture sensors for the purpose of collecting the data and monitoring the crop field. Second component is a web based application which is responsible for analyzing various s details on the crop data and field information by application of data mining to predict the favorable conditions such as temperature, humidity, and soil moisture for crop growth. Last component is the mobile application in a smart phone which is responsible for controlling the crop watering. It enables two modes either manual or automatic for the user.

S. Rajeswari, K. Suthendran [11] applied C5.0: Advanced Decision Tree (ADT) classifier algorithm to predict the soil fertility level by considering the Virudhunagar District Soil information and proposed recommendations for selecting the appropriate crop according to the soil nutrient level and sowing. They developed a system called Design of Smart Information System (DSIS) which is an android based mobile phone application. The user has to send service request using this application then the system attempts to recognize the user location by activating the Global Positioning System.

A. Chougule et al. [12] developed an ontology based system which is capable recommending suitable crop and fertilizers for the agricultural field. The given predictions for crop suitability are based on consideration of particular region and type of soil. For recommending fertilizers the NPK contents of the soil is taken into account and previously stored data in ontology is used. The system uses random forest algorithm for crop recommendation and K-means clustering is used for predicting best suitable fertilizer for crop based on given available NPK content in soil.

D. Ramesh, B.Vishnu Vardhan [13] in the paper "analysis of crop yield prediction using data mining techniques" presents a brief analysis of crop yield prediction by implementing multiple linear regression and density based clustering techniques to bring out the patterns. For crop yield prediction using multiple linear regression authors used Production, Year, Rainfall, Area of Sowing, Yield and Fertilizers as input variables. Yield estimation using density based clustering involves 6-clusters approximation about East Godavari District of Andhra Pradesh.

H. Flores et al. [14] developed an integrated supply chain planning tool for fresh vegetables, in their latest research which mainly accounts the characteristics and resources of three specific states in Mexico, to suggest the most appropriate and best crops to be planted, and helps farmers to bring out the best time for planting and harvesting, and what markets to be invested so that the farmers are benefited the most; and also help the farmers to choose the suitable agricultural tools and machineries appropriate for the region and its environment taken into the reference. The first step is planning process in which climatic conditions of targeted area is analyzed and historical data like daily rainfall and temperature is collected and clusters are formed based this data. Second step is identification of products with best market and climatic requirements of targeted area. The next step is to determine the suitable agricultural technology for the crops based on environmental modifications.

S. Nath et al. [15] in their research investigated the agricultural data such as soil moisture, fertilizer, humidity, rainfall and some other attributes to improve the production of crops under varied condition using data mining techniques. The authors aimed to design smart system which is capable suggesting crops to farmers; the system consists of 4 modules i.e. Crop, Soil, Weather and Fertilizer. Results of classification for both J48 tree classifier and Naive Bayes classifier was obtained and found that J48 classifier performed better than Naive Bayes classifier.

Monali Paul et al. [16] proposed a system for analyzing soil datasets and predicting the category of soil. Based on this prediction yielding of crops is indicated. Soil datasets collected from Soil Testing Laboratory of Jabalpur, Madhya Pradesh. For prediction K-Nearest Neighbor and Naive Bayes classification algorithms are applied to the dataset.

The authors [17] in their study presented a model for analysis of agricultural data to estimate crop production and choose best crop for the agricultural field. The model includes input component which takes input from the farmer; input component consist of crop name, land area, soil type, soil pH, pest details, weather, water level, seed type; Feature selection component choose an attribute from crop details. Classification is applied on climatic data and crop parameters for crop yield forecasting.

The researchers in [18] aimed to develop information system based on agricultural data in order to enable and improve interactions between customer and farmer. The system uses cloud computing technology to provide easy and secure access to data saved on cloud at any instance of time. For crop production forecasting and bring out the patterns and knowledge from historical data, regression analysis has been implemented.

S. Jambekar et al. [19] in their study applied regression analysis for yield prediction of crops such as rice, wheat and maize based on the parameters and data of period 1950-2013. The parameters taken into consideration were rainfall, mean temperature, area under irrigation, area, production and yield. Researchers used Multiple Linear Regression, Random Forest Regression and Multivariate Adaptive Regression Splines (Earth) algorithms for regression analysis. From the results obtained they found the performance of Multivariate Adaptive Regression Splines (Earth) was better for rice and wheat dataset and performance of Multiple Linear Regression is better for Maize dataset.

P. Bhargavi, S. Jyothi [20] applied Naive Bayes classifier for classification of agricultural land soils. Soil database collected from Department of Soil Sciences and Agricultural Chemistry, S V Agricultural College, Tirupati contains measurements of soil profile data from various locations of Chandragiri Mandal, Chittoor District.

## V. CONCLUSION

In almost all developing countries agriculture plays an important role for their economy. In India also agricultural sector has significant importance. Advance technologies can help the agriculture sector to develop. Information technology can help farmers to make certain decision which can yield better results. As large data is available in context of agriculture, data mining can be used for decision making tool.

This paper discusses the role of data mining in the field of agriculture and related work by several authors in context to agriculture domain. Various applications of data mining techniques to solve issues related to agriculture have been discussed. This paper contains the work of various authors in one place so that researchers can get knowledge of current scenario of data mining techniques and applications in context to agriculture field.

## REFERENCES

[1] M.C.S.Geetha, "A Survey on Data Mining Techniques in Agriculture," International Journal of Innovative Research in Computer and Communication Engineering Vol. 3, Issue 2, February 2015.

[2] Pavel Berkhin, "Survey of Clustering Data Mining Techniques," Accrue Software, Inc., 2002.

[3] Hetal Patel, Dharmendra Patel "A Brief survey of Data Mining Techniques Applied to Agricultural Data," International Journal of Computer Applications, vol.95 issue 9, 2014.

[4] Rapid Miner-https://rapidminer.com/ . Accessed 29 May 2019.

[5] Orange-https://en.wikipedia.org/wiki/Orange_(software). Accessed 29 May 2019.

[6] Weka- https://www.cs.waikato.ac.nz/ml/weka/ . Accessed 1 June 2019.

[7] KNIME-https://www.knime.com/knime-software/knime-analytics-platform . Accessed 1 June 2019.

[8] Graham J Williams, "Rattle: A Data Mining GUI for R," The R Journal (2019) 1(2) 45.

[9] Jharna Majumdar, Sneha Naraseeyappa, Shilpa Ankalaki, "Analysis of agriculture data using data mining techniques: application of big data", Journal of Big Data vol. 4 issue 1, 2017.

[10] J. Muangprathub, N. Boonnam, S. Kajornkasirat, N.Lekbangpong, A.Wanichsombat, P. Nillaor, "IoT and agriculture data analysis for smart farm," Computers and Electronics in Agriculture, Elsevier vol.156, pp.467-474, December 2018.

[11] S.Rajeswari, K.Suthendran, "C5.0: Advanced Decision Tree (ADT) classification model for agricultural data analysis on cloud," Computers and Electronics in Agriculture,Elsevier, 2019.

[12] Archana Chougule, Vijay Kumar Jha and Debajyoti Mukhopadhyay, "Crop Suitability and Fertilizers Recommendation Using Data Mining Techniques, Progress in Advanced Computing and Intelligent Engineering," Advances in Intelligent Systems and Computing. Springer vol. 714, pp. 205-213, 2019.

[13] D Ramesh, B Vishnu Vardhan, "Analysis of Crop Yield Prediction Using Data Mining Techniques," International Journal of Research in Engineering & Technology, vol. 4, issue 1, pp. 470-473, 2015.

[14] Hector Flores, J. Rene Villalobos, Omar Ahumada, Mark Uchanski, Cesar Meneses, Octavio Sanchez "Use of supply chain planning tools for efficiently placing small farmers into high-value, vegetable markets" Computers and Electronics in Agriculture. Elsevier. Volume -157, Pages 205-217, 2019.

[15] Suraj Nath, Debashri Debnath, Parthapratim Sarkar, Ankur Biswas, "Design of Intelligent System in Agriculture using Data Mining," Proceedings of International Conferencing on Computational Intelligence & IoT, pp. 631-637, 2018.

[16] Monali Paul, Santosh K. Vishwakarma, Ashok Verma, "Analysis of Soil Behaviour and Prediction of Crop Yield Using Data Mining Approach," International Conference on Computational Intelligence and Communication Networks, pp. 766-771, 2015.

[17] R.Sujatha, Dr.P.Isakki, "A study on Crop Yield Forecasting using Classification Techniques," IEEE, vol. 7, 2016.

[18] Pallavi V. Jirapure, Prarthana A. Deshkar, "Qualitative data analysis using regression method for agricultural data," IEEE World Conference on Futuristic Trends in Research and Innovation for Social Welfare 2016.

[19] Suvidha Jambekar, Shikha Nema, Zia Saquib, "Prediction of Crop Production in India Using Data Mining Techniques," IEEE, 2018.

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