STUDY & RESEARCH OF PREDICTION OF CROPS METHODOLOGY USING DATA MINING TECHNIQUES

Sharmila Choudhary[1], Dr. Devendra Nagal[2], Dr. Rakesh Poonia[3], Dr. Swati Sharma[4]
[1]PhD. Scholar, Department of Computer Science & Engineering, JNU Jodhpur
[2]Research Guide, Faculty of Engineering& Technology, JNU Jodhpur
[3]Research Guide, Assistant Prof., Dept. of Computer Applications, Govt. Engineering College,Bikaner.
[4]Faculty of Engineering& Technology, JNU Jodhpur

Abstract: Agricultural system is very difficult as it deals with the great data that come from a number of issues. Prediction of crop harvesting has been a matter of raising awareness for producers, consultants and agricultural partners. In this research, an effort has been made to review research studies on the application of data mining techniques in the field of agriculture. Some of the techniques, such as the k-medium, the nearest neighbor k, the artificial neural networks and the vector support machines applied in the field of agriculture were presented. Data extraction in agriculture is a comparatively new methodology for forecasting / predicting crop / animal management. The problem has been the intricate knowledge of this raw data, this has led to the growth of new approaches and methods such as machine learning that can be used for knowledge of the data with crop evaluation. This research aims to evaluate these innovative techniques, such as those found in the database. This research explores the applications of data mining techniques in the field of agriculture and related sciences. This research provides the numerous methods of predicting crop yield using data mining techniques.

Keywords: Crop yield, Data mining, Artificial Intelligence, K-Means, K-Nearest Neighbor (KNN), Artificial Neural Networks (ANN), Support Vector Machines (SVM)

I. DATA MINING TECHNIQUES

Agriculture is the most significant area of application particularly in developing countries such as India. The use of information technology in agriculture can change the decision-making situation and farmers can produce a better way. Data mining plays a crucial role in decision-making on server issues related to the field of agriculture. The role of data mining in perspective in the field of agriculture and also confers several data mining techniques and their work related by several authors in context to the domain of agriculture. Also discussed on different applications of data mining in the resolution of different agricultural problems. Data mining is a useful technique for finding the useful pattern of the huge data set. Therefore, an important place in agriculture was secured because field agriculture contains many data such as soil data, harvest data, and meteorological data, etc. Real-time weather data is difficult to analyze and manage so that various algorithms in data mining such as k-media, clustering, Apriority algorithm and other statistical methods are used to analyze agriculture data and provide the useful pattern.

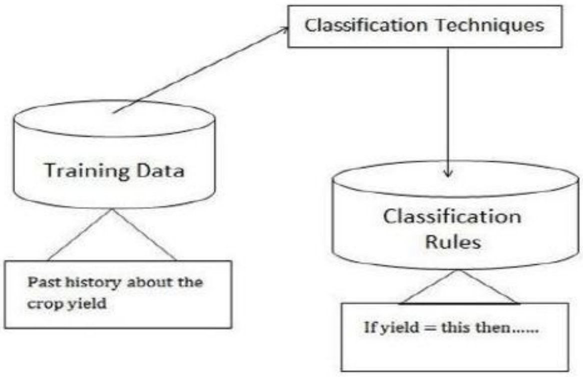

Classification Techniques
Training Data
Classification Rules
Past history about the crop yield
If yield = this then......

Fig. 1 Classification Technique

Multiple linear regressions
Multiple linear regressions are the method used to have a linear relationship between one or more independent variables and a dependent variable. The independent variable is used to estimate values for the dependent variables. MLR is a predictive analysis based on least squares and is probably the most used method in climatology. The 3 main uses of the MLR analysis are in casual regression, forecasting an effect, forecasting trend. Regression analysis focuses on the relationship between two variables, while the correlation analysis only focuses on the strength between the two or more variables.

K-Nearest Neighbor K-Nearest Neighbor is a classification technique in which it is assumed that samples that are similar will have similar classification. The number of similar known samples used for assigning a classification to an unknown sample defines the parameter K. In case if there is no history about the samples to be classified that is if the training set is not provided then clustering technique is used.

K-Means Approach The most important clustering technique is K-Means clustering. This technique is used to classify the data which have no previous knowledge about the data or the training set. The parameter K denotes the number of clusters required to partition the data. [9] The idea of this clustering technique is, given K number of clusters we can define K centers, one for each cluster based on all samples belonging to a cluster. These centers must be placed far away from each other and then associate each sample to the cluster that has the closest centroid. When no samples are left, the process of finding new K centers and assigning samples to the clusters that has the closest centroid is iteratively carried out until no longer the samples can change their clusters.

Artificial neural network

Artificial neural network is one of the new data mining techniques that are based on biological neural processes of human brain. According to this technique once the neural network is trained it can predict the crop yield in similar patterns even if the past data include some errors. Even if the data is complex, multivariate, nonlinear this network gives the accurate results and also without any of underlying principles the relationship between them the output is extracted.

Support Vector Machines

Support Vector Machines (SVMs) are binary classifiers that will classify data samples in two disjoint classes. It is a technique in which two classes are linearly separable which is from a simplified case. [7] SVM can build a model that predicts whether a new example falls into category or the other. A support vector machine is a concept in statistics and computer science for a set of related supervised learning methods that analyze data and recognize patterns used for classification and regression analysis. The SVM takes a set of input data and predicts for each given input which of two possible classes forms the input making the SVM a non-probabilistic binary linear classifier.

II. REGRESSION MODEL

Regression model are also used in crop yield prediction. Regression is mainly used for predicting about the future (not only crop yield). This model defines two variables independent and dependent variable. The value of the dependent variable can be predicted using the independent variable. Ex: In case of crop yield and soil, yield is dependent on the type of the soil. If that type of soil is suitable for that crop then the yield is high.

Biclustering Technique

Biclustering technique is one of the techniques used in data mining when the data is given in the form of rows and columns that is in the matrix form. Different types of bi-clustering algorithms are: Bi-cluster with constant values, Bi-cluster with constant rows, Bi-cluster with constant columns, and Bi-cluster with coherent values. Among these techniques K-Means, KNN techniques are suitable to predict the crop yield accurately. These above-mentioned techniques require knowledge of statistics. Regression model is used when the researcher is having the clear picture about the type of variable whether it is dependent or independent variable. In case if there are more than two independent variables then multiple linear regression is preferred. When the data is given in form of matrix the researcher is advised to use bi-clustering technique. If the data is based on biological neural processes of human brain then artificial neural network is advisable. 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, an association rule is a phrase of the form X => Y, where X and Y are sets of objects. The instinctive meaning of such a rule is that transactions of the database which contain X tend to contain Y.[4] 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 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.[5]

Classification and prediction are two forms of data analysis that can be used to extract models that describe 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 that can then be used to predict discrete class labels in new samples. To maximize the predictive accuracy obtained by the classification model when the classification of the examples in the test set is not seen during training is one of the main objectives of the 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), Closest Neighbor (NN), Artificial Neural Network (ANN), Support Vector Machine SVM), Rough Sets, Genetic Algorithms. [6]

In clustering, the focus is on finding a partition of data records in groups so that the points within each group are close to each other. Grouping groups data instances into subsets in such a way that similar instances are assembled together, whereas dissimilar instances belong to different groups. Since the purpose of clustering is to find a new set of categories, the latter groups are of interest in themselves, and their evaluation is intrinsic. [7] There is no prior knowledge about the data. The different clustering methods are Hierarchical Methods (HM), Partition Methods (PM), Density Based Methods (MBD), Model-Based Methods (MBCM), Grid-based Methods and Soft Computing Methods [fuzzy, neural-based networks], Square Error-Based Clustering (Vector Quantization), Network Data and Graphic Clustering [8]

Regression is the learning of a function that assigns a data element to a real value prediction variable. The different applications of regression are predicting the amount of biomass present in a forest, estimating the probability of the patient surviving or not in the set of his diagnostic tests, predicting the demand for a new product. [9] Here the model is trained to predict a continuous target. Regression tasks are often treated as classification tasks with quantitative class labels. The prediction methods are Nonlinear Regression (NLR) and Linear Regression (LR).

III. PROPOSED WORK

ANN for Crop Production Forecasting

Predicting crop yields, especially strategic crops such as wheat, maize and rice, has always been an interesting area of research for agrometeorology's, as it is important in national and international economic programming. The production of dry crops, in addition to the relationship with the genetics of the grower, adaptive terms, the effect of pests and pathology and weeds, quality of management and control during the growing season and etc. severely depend on whether events. Therefore, it is not beyond the possibility of acquiring relationships or systems that can predict the highest precision with meteorological data. Today, there are many performance prediction models, most of which have been generally classified into two groups a) statistical models, b) crop simulation models (eg CERES). Recently, the application of Artificial Intelligence (AI), such as Artificial Neural Networks (RNAs), Fuzzy Systems and Genetic Algorithm has shown more efficiency in dissolving the problem. Applying them can make models easier and more accurate from complex natural systems with many inputs. In this research, an attempt has been made to develop a prediction model of wheat yield using RNAs. If we design a network that correctly learns the relationships of effective climatic factors on crop yields, it can be used to estimate long- or short-term crop production and also with sufficient and useful data to obtain RNA model for each area. In addition, the use of RNA may find the most effective factors in crop yield. Therefore, some factors that your measures are difficult and cost-effective can be ignored. In this case only the effect of climatic factors on wheat yield has been applied.

In computer science and related fields, artificial neural networks are computational models inspired by animal central nervous systems (in particular the brain) that are capable of machine learning and pattern recognition. They are usually presented as systems of interconnected "neurons" that can compute values from inputs by feeding information through the network. Like other machine learning methods, neural networks have been used to solve a wide variety of tasks that are hard to solve using ordinary rule-based programming, including computer vision and speech recognition. The word network in the term 'artificial neural network' refers to the inter–connections between the neurons in the different layers of each system. An example system has Three layers. The first layer has input neurons, which send data via synapses to the second layer of neurons, and then via more synapses to the third layer of output neurons. More complex systems will have more layers of neurons with some having increased layers of input neurons and output neurons. The synapses store parameters called "weights" that manipulate the data in the calculations.

An ANN is typically defined by three types of parameters:

1. The pattern of interconnection between different layers of neurons
2. The learning process to update the weights of the interconnections
3. The activation function that converts the weighted input of a neuron to its output activation.

One type of network sees the nodes as „artificial neurons". These are called artificial neural networks (ANNs). The back-propagation algorithm (Rumelhart and McClelland, 1986) is used in layered feed-forward ANNs.

This means that the artificial neurons are organized in layers, and send their signals forward, and then the errors propagate backwards. The network receives inputs by neurons in the input layer, and the output of the network is given by the neurons in an output layer. There may be one or more intermediate layers. In this research, we will examine one of the most common neural network architectures, the neural forward propagation network shown in the figure.

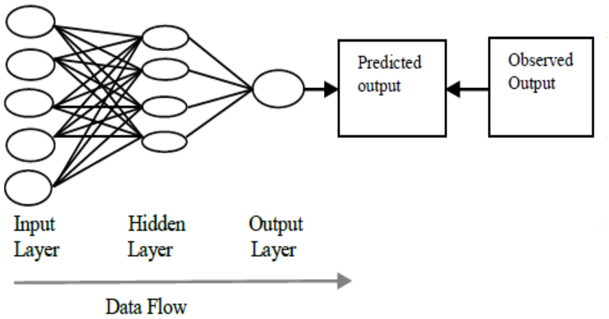


Figure: 2 Forward Back Propagation Neural Network

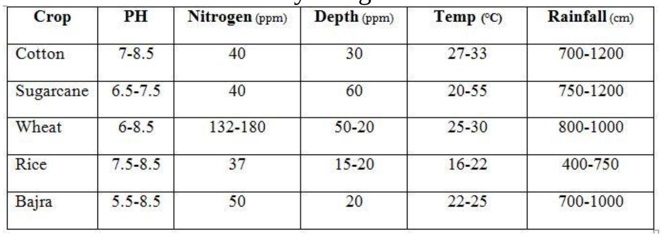
This neural network architecture is very popular because it can be applied to many different tasks. The first term, feed forward, describes how this neural network processes and remembers patterns. In a neural network of forward feeding, neurons are only connected to the prologue. Each layer of the neural network contains connections to the next layer (for example, from the input to the hidden layer), but there are no return connections [54]. The term "posterior propagation" describes how this type of neural network is formed. Post propagation is a form of supervised training. When a supervised training method is used, both input samples and anticipated outputs must be provided to the network. The expected outputs are compared to the actual outputs for a given input. Using the anticipated outputs, the posterior propagation training algorithm then takes a calculated error and adjusts the weights of the various layers backwards from the output layer to the input layer. Forward and forward propagation algorithms are often used together; however, this is not a requirement. It would be quite permissible to create a neural network that uses the feed forward algorithm to determine its output and does not use the back-propagation training algorithm. Similarly, if you choose to create a neural network that uses later propagation training methods, it will not necessarily be limited to a feed forward algorithm to determine the output of the neural network. Although these cases are less common than feeding the neural network of propagation feed. [80].

The backpropagation algorithm uses supervised learning, which means that we give the algorithm examples of inputs and outputs that we want the network to compute, and then we calculate the error (difference between actual and expected results). The idea behind the backscatter algorithm is to reduce this error, until the ANN learns training data.

The training begins with random weights, and the goal is to adjust them so the error is minimal. Since posterior propagation uses the gradient descent method, it is necessary to calculate the derivative of the quadratic error function with respect to the weights of the network. The square error function is:
E= (t-y)2,
E=square error,
t= target output,
y=actual output of output neuron.

$$y = \sum_{i=1}^{n} w_i x_i,$$

n=the number of input units to the neuron,
wi=the ith weight,
xi=the ith input value to the neuron.
In this research crop prediction methodology is used to predict the suitable crop by sensing various parameter of soil and also parameter related to atmosphere. For that purpose, we are used artificial neural network (ANN). This research shows the ability of artificial neural network technology to be used for the approximation and prediction of crop yields at rural district. It is verified by using ANN is shown below-:



Crop	**PH**	**Nitrogen** (ppm)	**Depth** (ppm)	**Temp** (°C)	**Rainfall** (cm)
Cotton	7-8.5	40	30	27-33	700-1200
Sugarcane	6.5-7.5	40	60	20-55	750-1200
Wheat	6-8.5	132-180	50-20	25-30	800-1000
Rice	7.5-8.5	37	15-20	16-22	400-750
Bajra	5.5-8.5	50	20	22-25	700-1000

Table: 4.4 Classification of various crops verified by ANN

Back Propagation

Back propagation nets are the most common kind of ANN. The basic topology is that layers of neurons are connected to each other. Patterns cause information to flow in one direction, then the errors "back-propagate" in the other direction, changing the strength of the interconnections between layers[60]. A very simple example of Neural Networks using back propagation this program is a simple example of Neural Networks using back propagation. Back Propagation network learns by example to give weights that used to get output from provided input. Learning steps:
1. Initialized by setting up all its weights to be small random numbers.
2. Apply input and calculate output that different form target one.
3. Calculate error rate as target – Actual.
4. Used error rate to modify weights to get small error rate.
5. Apply step 2, 3, 4 more and more to get final trusted model.
There are different methods to calculate the error rate based on how training process work like:
1. If we used threshold in training the error function = target – actual.
2. No threshold the error function expresses as outα (1 - outα) (Targetα - outα).
Whatever how error calculate the new weight value will be calculated as:
W (new) = w (old) + n * error rate * output.
Based on this equation, we calculate the new weight for each neuron, test the error function and check how the new weight reduces the difference between the output and the target and work the same way until reaching the goal of the model.
Forecasting the behavior of a complex system has been a broad application domain for neural networks. In particular, we have studied extensively, such as the prediction of electric charge [1], [2], economic forecast [3], prediction of natural physical phenomena [4], forecasting river flows [5] and foresight of students' income in schools [18]. In addition to predictions based on neural networks, diffuse time series predictions emerged as a noble approach to predict future values in a situation in which neither a trend nor a pattern of time series variations is visualized and the information is imprecise and vague. Song and Chissom successfully used the concept of fuzzy sets with linguistic variables presented by Zadeh [11,12] and the application of fuzzy logic to approximate reasoning by Mamdani [13] to develop the basis for the prediction of diffuse time series. Song and Chissom [13, 14] implemented their invariant developed time model and time variants in the historical time series data of the University of Alabama student enrollments. Chen [14] presented a simplified method of invariant time for the prediction of time series by using the arithmetic operations instead of the max-min operation used by Song and Chissom [7]. In addition, Chen [14] applied the high-order diffuse time series model for the prediction of enrollments and found some points of ambiguity in forecast trends and suggested using the high order fuzzy logic relation group to deal with ambiguity. MR. Singh [10] presented an improved and versatile method for predicting diffuse time series using a difference parameter as a fuzzy relation for prediction. Rajesh Joshi [17] used a diffuse time series model for agricultural forecasting production consisting of the development and implementation of diffuse series models using metrological parameters as indicators for forecasting. In this research to achieve objectivity on the subjectivity of methods based on diffuse time series has been proposed a method based on neural networks. The proposed method has been applied to the historical data and parameters of influence (temperature, sun and rain) of the crop production (wheat) of the Pant Nagar farm, G.B. Pant Nagar (India) [17] The agricultural production system is one of the real-life problems that fall into the category that has uncertainty in known parameters and some unknown, making it a natural option for the implementation of diffuse time series predictive models in their production system. Uncertainty lies in crop production due to some uncontrolled parameters of which "time", "agrometeorological" variables are the key contents. In addition, crop production is concerned with field data, accuracy of data is always a cause for concern. Past experience shows that the crop production system can observe the large variation in production data as the system is affected by many uncertain production parameters and the uncertain occurrence of natural calamities. The proposed method produces better results than the previously discussed fuzzy time series methods.

IV. CONCLUSION

The Indian economy depends mainly on the agricultural sector. Seventy-five percent of the population depends on agriculture for subsistence. Now a day very few farmers are using the various methods, tools and techniques of agriculture for a good production. Data mining can be used to predict future values of agricultural processes. Data mining this process on results in the discovery of new patterns in large data sets. The main purpose of the data mining process is to extract knowledge from the previous data set. This process examines data from different perspectives and describes in useful information. There is no restriction on the type of data that can be scanned by data mining. The objective of this research is to provide information on different techniques of data mining in the perspective of the domain of agriculture.

REFERENCES

[1] "Data Mining Techniques and Applications to Agricultural Yield Data" by D Ramesh, B Vishnu Vardhan, Associate Professor, Department of CSE, JNTUH College of Engineering, Telangana State, India, Professor, Department of CSE, JNTUH College of Engineering, Telangana State, India.Vol 2, Issue 9, September 2013.

[2] "Analysis of Data Mining Techniques for Agriculture Data" E.Manjula, S.Djodiltachoumy, Vol.4, Issue.2, Page.1311-1313, (2016).

[3] "Analysis Of Crop Yield Prediction Using Data Mining Techniques", D Ramesh, B Vishnu Vardhan, Associate Professor, Department of CSE, JNTUH College of Engineering, Telangana State, India Professor, Department of CSE, JNTUH College of Engineering, Telangana State, India. Volume: 04 Issue: 01| Jan-2015.

[4] "Data Mining: An effective tool for yield estimation in the agricultural sector" Raorane A.A.-Department of computer science, Vivekanand College, Tarabai park Kolhapur INDIA. Kulkarni R.V.-Head of the Department, Chh. Shahu Institute of business Education and Research Centre Kolhapur 416006 INDIA, Volume 1, Issue 2, July – August 2012.

[5] "Agriculture Crop Pattern Using Data Mining Techniques" G. NasrinFathima, Research Scholar, Dept. of computer Science, Jamal Mohamed College, Trichirapalli, TN, India, R. Geetha , Assistant Professor , Dept. of computer Science, Jamal Mohamed College, Trichirapalli, TN, India, Volume 4, Issue 5, May 2014.

[6] "Data Mining Technique to Predict Annual Yield for Major Crops", Rajshekhar Borat, Rahul Ombale, SagarAhire, ManojDhawade, P.S. Kulkarni, Department of Computer Engineering , NBN Sinhgad School of Engineering, Pune-411041, Vol. 4, Issue 03, 2016.

[7] "Density Based Clustering Technique on Crop Yield Prediction", B Vishnu Vardhan and D. Ramesh, JNTUHCollegeofEngineering,Nachupalli,KarimnagarDist.,AndhraPradesh,India.O SubhashChanderGoud, NIZAMCollege,Hyderabad,AndhraPradesh, India.Vol.2, No.1, March, 2014.

[8] "An Approach for Mining Accumulated Crop Cultivation Problems and their Solutions" Samhaa R. El-Beltagy, Ahmed Rafea , Said Mabrouk and Mahmoud Rafea".Cairo University, The American University in Cairo, The Central Lab for Agricultural Expert SystemsFaculty of Computers and Information, 5 Dr. Ahmed Zewail Street, 12613, Orman, Giza, Egypt, Computer Science Department, AUC Avenue, P.O. Box 74, New Cairo 11835, Egypt.Ministry of Agriculture and Land Reclamation, Giza, Egypt. 2009

[9] "Agriculture Crop Pattern Using Data Mining Techniques" G.NasrinFathima, Research Scholar, Dept. of computer Science,Jamal Mohamed College, Trichirapalli, TN, India. R. Geetha , Assistant Professor , Dept. of computer Science, Jamal Mohamed College, Trichirapalli, TN, India. Volume 4, Issue 5, May 2014

[10] "A survey on Data Mining Techniques for Crop Yield Prediction" Ramesh A. Medar Dept. of Computer Science & Engineering Gogte Institute of Technology Belgaum, Karnataka, India Vijay. S. Rajpurohit Dept. of Computer Science & Engineering Gogte Institute of Technology Belgaum, Karnataka, India. Volume 2, Issue 9, September 2014

[11] "Data mining Techniques for Predicting Crop Productivity – A review article"1S.Veenadhari,2 Dr. Bharat Misra, 3Dr. CD Singh1,2 Mahatma Gandhi GramodayaVishwavidyalaya, Chitrakoot, Satna, India 3Central Institute of Agricultural Engineering, Bhopal, India. IJCST Vol. 2, Issue 1, March 2011.