ISSN(Online): 2319-8753
ISSN (Print): 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(A High Impact Factor, Monthly, Peer Reviewed Journal)

Visit: www.ijirset.com

Vol. 6, Issue 10, October 2017

Prediction of Crops Methodology using Data Mining Techniques

Dr. Rakesh Poonia[1], Sonia Bhargava[2]

Assistant Professor, Department of Computer Applications, Govt. Engineering College Bikaner, Rajasthan, India[1]

M.Tech.(C.S), Suresh Gyan Vihar University, Jaipur, Rajasthan, India[2]

ABSTRACT: The agricultural system is very difficult as it is the great data that come from a series 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. We presented some of the techniques, such as the k-medium, the nearest neighbour k, the artificial neural networks and the vector support machines applied in the field of agriculture. Data extraction in agriculture is a comparatively new methodology for predicting / predicting crop / animal management. The problem has been the intricate knowledge of these raw data, which 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 numerous methods of predicting crop yields using data mining techniques.

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

I. INTRODUCTION

Agriculture is the most important 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. We also discussed the different applications of data mining in solving 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 agriculture field contains many data such as soil data, harvest data, and meteorological data, etc. Real-time weather data are difficult to analyze and handle so that several algorithms in data mining such as k -media, clustering, Apriority algorithm and other statistical methods are used to analyse agriculture data and provide useful pattern.

II. DIFFERENT METHODOLOGIES

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 three main uses of MLR analysis are in casual regression, predicting an effect, prognostic 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-Neighbour nearest K-Neighbour nearest is a classification technique in which it is assumed that samples that are similar will have a similar classification. The number of known similar samples used to assign a classification to an

ISSN(Online): 2319-8753
ISSN (Print): 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(A High Impact Factor, Monthly, Peer Reviewed Journal)

Visit: www.ijirset.com

Vol. 6, Issue 10, October 2017

unknown sample defines the parameter K. In the event that if there is no history of the samples to be classified, i.e. if the training set is not provided, the technique is used of grouping.
K-Means Approach The most important grouping technique is the K-Means grouping. This technique is used to classify data that has no previous knowledge about the data or the training set. The K parameter denotes the number of clusters required to divide the data. [9] The idea of this grouping technique is, given the number K of clusters, we can define K centers, one for each group based on all the samples belonging to a group. These centers should be placed away from each other and then associate each sample to the group having the nearest centroid. When no samples remain, the process of finding new K-centers and assigning samples to groups that have the closest centroid is performed iteratively until samples can no longer change their conglomerates.

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

SUPPORT VECTOR MACHINES: Support Vector Machines (SVMs) are binary classifiers that will classify data samples into two disjoint classes. It is a technique in which two classes are linearly separable. [7] SVM can build a model that predicts whether or not a new example falls into the category or the other. The vector backing machine is a statistical and computer science concept for a set of supervised related learning methods that analyze data and recognize the patterns used for classification and regression analysis. The SVM takes an input data set and predicts for each given input which of two possible classes forms the input making the SVM to linear non-probabilistic binary classifier.

REGRESSION MODEL: The regression model is also used in predicting crop yields. Regression is mainly used to predict future (not just crop yield). This model defines two independent variables and dependent variables. The value of the dependent variable can be predicted using the independent variable. Example: In the case of crop and soil yield, the yield depends on the type of soil. If that type of soil is suitable for the crops then the yield is high.

BICLUSTERING TECHNIQUE: The Biclustering technique is one of the techniques used in data mining when data are given in the form of rows and columns that are in the form of a matrix. The different types of bi-clustering algorithms are: Bi-cluster with constant values, Bi-cluster with constants, Bi-cluster with constant columns, and Bi-cluster with coherent values. Among these K-Means techniques, KNN techniques are adequate to predict crop yield accurately. These techniques mentioned above require statistical knowledge. The regression model is used when the investigator is having the clear image about the type of variable if it is dependent or independent variable. If there are more than two independent variables, multiple linear regressions is preferred. When the data are given as a matrix, the researcher is advised to use the bi-clustering technique. If the data are based on biological neuronal processes of the human brain, then the 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

ISSN(Online): 2319-8753
ISSN (Print): 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(A High Impact Factor, Monthly, Peer Reviewed Journal)

Visit: www.ijirset.com

Vol. 6, Issue 10, October 2017

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 Neighbour (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 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 agro meteorology, 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 so on. 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 generally been 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 than 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 crop production in the long or short term and also with sufficient and useful data to obtain an 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 computing and related fields, artificial neural networks are computational models inspired by the animal central nervous systems (in particular the brain) that are able to learn to mechanize and recognize patterns. They are usually presented as systems of interconnected "neurons" that can calculate the values of the 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 difficult to solve through regular rule-based programming, including computer vision and voice recognition.
The word network in the term "artificial neural network" refers to the inter-connections between neurons in the different layers of each system. This system contains three different layers they are the first layer has input neurons, which send data through synapses to the second layer of neurons, and then through more synapses to the third layer of the output neurons. More complex systems will have more layers of neurons with some 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:

ISSN(Online): 2319-8753
ISSN (Print): 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(A High Impact Factor, Monthly, Peer Reviewed Journal)

Visit: www.ijirset.com

Vol. 6, Issue 10, October 2017

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 below in Figure 1.

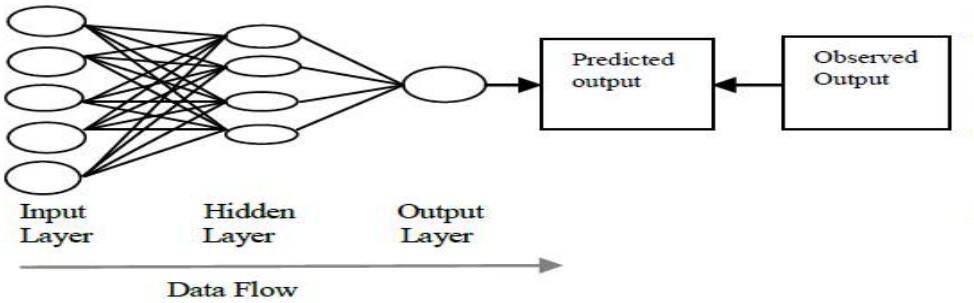

Figure: 1. Forward Back Propagation Neural Network

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 forward feeding network, 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. 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. Direct and direct 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 subsequent 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 propagation feed.

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.

ISSN(Online): 2319-8753
ISSN (Print): 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(A High Impact Factor, Monthly, Peer Reviewed Journal)

Visit: www.ijirset.com

Vol. 6, Issue 10, October 2017

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

n=the number of input units to the neuron,

wi=the i^{th} weight,

xi=the i^{th} 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 using 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: 1 Classification of various crops verified by ANN

BACK PROPAGATION: Post-propagation networks are the most common type of ANN. The basic topology is that the layers of neurons are connected to each other. Patterns cause information to flow in one direction, then errors "retro-propagate" in the other direction, changing the strength of interconnections between layers. A very simple example of Neural Networks that uses the later propagation of this program is a simple example of Neural Networks that uses the later backpropagation. The propagation network learns for example to give weights that is used to obtain the output of the input provided.

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 in the same way until reaching the goal of the model. Predicting the behavior of a complex system has been a broad application domain for neural networks. In particular, we have studied extensively, such as prediction of electric charge [1][2], economic forecasting [3],

ISSN(Online): 2319-8753
ISSN (Print): 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(A High Impact Factor, Monthly, Peer Reviewed Journal)

Visit: www.ijirset.com

Vol. 6, Issue 10, October 2017

prediction of natural physical phenomena [4], forecasting river flows [5] and predicting student incomes in schools. In addition to predictions based on neural networks, diffuse time series predictions emerged as a noble approach to predicting future values in a situation where neither a trend nor a time series pattern 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] and the application of fuzzy logic to approximate Mamdani's reasoning to develop the basis for the prediction of diffuse time series. Song and Chissom implemented their invariant time-developed model and time variants in the historical time series data from the University of Alabama student inscriptions. Chen presented a simplified invariant time method for predicting time series by using arithmetic operations rather than the max-min operation used by Song and Chissom [7]. In addition, Chen applied the high-order diffuse time series model for the prediction of enrolments 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 relationship for prediction. Rajesh Joshi used a diffuse time series model for the production of agricultural forecasts, consisting of the development and implementation of diffuse series models using metrological parameters as indicators for prediction. 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. The agricultural production system is one of the real-life problems that fall into the category that has uncertainty in known parameters and some unknowns, which makes it a natural option for the implementation of predictive models of diffuse time series its system of production. The uncertainty lies in the production of crops due to some uncontrolled parameters of which the variables "time", "agrometeorological" are the key contents. In addition, crop production takes care of the field data, the accuracy of the data is always causing for concern. Past experience shows that the crop production system can observe 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 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 the future values of agricultural processes. Data mining of this process results in the discovery of new patterns in large datasets. 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 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.

ISSN(Online): 2319-8753
ISSN (Print): 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(A High Impact Factor, Monthly, Peer Reviewed Journal)

Visit: www.ijirset.com

Vol. 6, Issue 10, October 2017

[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.