Density Based Clustering Technique on Crop Yield Prediction

B Vishnu Vardhan and D. Ramesh
JNTUH College of Engineering, Nachupalli, Karimnagar Dist., Andhra Pradesh, India
Email: mailvishnu@yahoo.com, dantamr@yahoo.com

O Subhash Chander Goud
NIZAM College, Hyderabad, Andhra Pradesh, India
Email: organtsubhash@gmail.com

Abstract—**Over the last few decades, Information Technology become a part of our daily life. Technology breakthroughs have made in industry and services as well as in agriculture. The connection between Information Technology and Agriculture has become an interesting area of research in yield prediction subject to the available data. A farmer harvest not only crops but also growing amount of data. A farmer wants to know about the applications of recent technologies in agriculture. Such technological requirement from the farmer lead to extracting the knowledge from the available data. The knowledge extraction methods in data mining are to be explored in order to obtain the crop yield prediction. Lots of data mining techniques were used in agriculture. Some of the widely used data mining techniques over agriculture data sets are Multiple Linear Regression, Density based Clustering Technique, K-Means approach, K-Nearest Neighbor, Artificial Neural Networks, Support Vector Machines. In this context the main aim of the paper is to model and to optimize the available by means of data mining techniques to predict the crop yield. This paper presents a brief idea of the widely used data mining techniques over agriculture data sets and deal with Density based Clustering Technique.**

Index Terms—**information technology, agriculture, yield prediction, data mining, multiple linear regression, density based clustering technique, K-Means approach, K-Nearest neighbor, artificial neural networks, support vector machines**

I. Introduction

Agriculture has significant history in India. It contributes 10-15% Gross Domestic Product (GDP) to the India Economy. Agriculture as a business is unique crop production is dependent on many climate and economy factors. The agriculture yield is primarily depends on weather conditions. The volume of data is enormous in Indian agriculture. The data when become information is highly useful for many purposes. The conventional and traditional system of data analysis is agriculture is purely dependent on statistics.

Data Mining is modern data analysis technique, it has wide range of applications in the agriculture field. Data Mining techniques aim at finding the patterns or information in the data that are both valuable and interesting to the farmer. Data Mining has been used to analyze large data sets and established useful classification and patterns in the data sets. Data Mining is the processor of discovering previously unknown and potentially increasing pattern in large data sets. The mined information is used for representing as a model for prediction.

Data Mining techniques are mainly divided in two groups, classification and clustering techniques. Classification techniques are designed for classifying unknown samples using information provided by a set of classified samples. This set is usually referred to as a training set as it is used to train the classification technique how to perform its classification. Generally, Artificial Neural Networks and Support Vector Machines, these two classification techniques learn from training set how to classify unknown samples. Another classification technique, K- Nearest Neighbor, does not have any learning phase, because it uses the training set every time a classification must be performed. A training set is known, and it is used to classify samples of unknown classification.

In the event training set not available, there is no previous knowledge about the data to classify. In this case, clustering techniques can be used to split a set of unknown samples into clusters. Clustering is a one of the method in data mining, it means that process of grouping a set of physical or abstract objects into class of similar objects. Clustering is grouping of data into similar groups with respect to similarity of the data. There are many clustering methods available and each of them give different grouping of a data set.

Normally, Clustering techniques are divided in hierarchical and partitioning. While hierarchical clustering algorithms builds clusters gradually, partitioning algorithms learn clusters directly. Some of the most used clustering techniques are Density based clustering technique and K-Means approach.

Manuscript received October 1, 2013; revised January 12, 2014.

doi: 10.12720/ijeee.2.1.56-59

II. Data Mining Techniques

A. Multiple Linear Rgression

Multiple Linear Regression (MLR) is the method used to model the linear relationship between a dependent variable and one or more independent variable(s). The dependent variable is sometimes termed as predictant and independent variables are called predictors. MLR is based on least squares and probably the most widely used method in climatology for developing models to reconstruct climate variables from tree ring services.

B. K-Means Approach

The K-Means approach is one of the most used clustering in the data mining. The idea behind the K-Means approach is very simple that certain partition of the data in K clusters, the center of the cluster can be computed as the mean of the all sample belonging to a cluster. The center of the cluster can be considered as the representative of the cluster. The center is quite close to all samples in the cluster.

C. K-Nearest Neighbor

The K-Nearest Neighbor technique is one of the classification technique in data mining. It does not have learning phase because it uses the training set every time a classification is performed. Nearest Neighbor search also known as proximity search, similarity search or closest point search and it is an optimization problem for finding closest points in metric spaces.

D. Artificial Neural Networks

An Artificial Neural Network (ANN) is an attractive alternative for building a knowledge-discovery environment for a crop production system. An ANN can use yield history with measured input factors for automatic learning and automatic generation of a system model. A Multilayer Perceptron (MLP) is a feed forward Artificial Neural Network model that maps sets of input data into a set of appropriate output. The MLP consists of an input and an output layer with one or more hidden layers of non linearly activating nodes. Each node in one layer connects with a certain weight to every node in the following layer.

E. Support Vector Machines

Support Vector Machines (SVMs) are binary classifiers able to classify data samples in two disjoint classes. The basic idea behind this technique comes from the simplified case in which the two classes are linearly separable. SVM are a set of related supervised learning method used for classification and regression. i.e. the SVM can build a model that predicts whether a new example falls into category or the other. A support vector machine is a concept is 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.

III. Literature Survey

Clustering techniques are widely used in data compression in image processing, it is otherwise known as vector quantization [1]. Clustering in data mining was brought to routine by greatest development in information retrieval and text mining [2], [3]. This techniques has been used in different areas such as spatial data base applications GIS or astronomical data [4]. This techniques can also been studied in sequence and heterogeneous data analysis [5] and web applications [6].

In the agricultural science, Data Mining clustering techniques are found in Grading apples before marketing [7], detecting weeds on precision agriculture [8].

Data Mining is the Process of discovering meaningful new correlation, patterns and trends by shifting through large amount of data, using pattern recognition technologies as well as statistical and mathematical techniques. Data Mining techniques are often used to studied soil characteristics. As an example, the K-Mean approach is used for classifying soils in combination with GPS based techniques [9] and K-Means approach is used to perform forecasts of the pollution in the atmosphere [10].

The researchers worked on Rainfall variability analysis and its impact on crop productivity [11]. In this case study collected the weekly rainfall data and number of rainy days recorded from 1958 to 1996 (39 years) at the main Dry farming research station. The correlation and regression studies were worked out using rainfall(x) as independent variable and yield(y) as dependent variable to derive information on rainfall-yield relationship and to develop yield prediction model for important crops.

A number of studies have been carried out on the application of data mining techniques for agricultural data sets. For example, the K-Nearest Neighbor is applied for simulating daily precipitations and other weather variables [12] and the different possible changes of the weather scenarios are analyzed using SVMs [13].

From the research article [14], the researcher express that large amount of data which is collected and stored for analysis. Making appropriate use of these data often leads to considerable gains in efficiency and therefore economic advantages.

The researchers [15] explain comparison of different classifiers and the outcome of research could improve the management and systems of soil uses throughout large fields that include agriculture, horticulture, environmental and land use management.

IV. Proposed Model

Density based clustering technique tries to divide the data into non-equal clusters, based on the mathematical model “Euclidean distance”. According to the model the process is done in the following steps:

1. Picking the no of parameters for the no of independent factors.
2. Picking the no of clusters to be divided up on the data.

3. Now pick up the first ‘n’ no of points in the given data, where n is ”no of clusters” and with ‘m’ co-ordinates where ‘m’ is no of parameters as (x,y,…m).
4. Now these choosen points are calculating the distance from one to another in the complete data set as shown below.

   i. A={ $(x_1,y_1,\ldots)$, $(x_2\ ,y_2\ ,\ldots)$, $(x_3, y_3\ ,\ldots)$, ……….. $(x_n , y_n , \ldots.)$}

   ii. Now if m,n are given as m=4 and n=3 then the first points are picked in the format n=3 as co-ordinates $(x_1\ ,\ y_1\ ,\ z_1\ ,\ l_1)$ so we now have 3 points of as a=(x_1 , y_1 , z_1 , l_1) b=(x_2 , y_2 , z_2 , l_2) c=(x_3 , y_3 , z_3 , l_3). Let set A= {a,b,c………..k}.

   iii. So now the Euclidean distance between the points
   a → a , a → b, a → c, ……..…,a → k_a
   d_{11} , d_{12} , d_{13} ………….. , d_{1i}
   b → a , b → b, b → c, ………. , b → k_a
   d_{21}, d_{22}, d_{23} ,……………, d_{2i}
   c → a , c → b, c → c,………..., c → k_a
   d_{31} , d_{32}, d_{33},………….. , d_{3i}

   iv. Now these distances are compared with each other and then the least distance points are grouped to one section i.e. if $d_{11}<d_{21}$ and $d_{11}<d_{31}$ then d_{11} point belongs to Cluster1, if $d_{22}<d_{12}$ and $d_{22}<d_{32}$ then d_{22} point (x_2 , y_2 , z_2 , l_2) belongs to Cluster 2 and so on this comparison is done for set A.
5. Now the step iv grouped points Cluster 1,2 and 3 are reconsidered and average of the Cluster belonging points are taken as the points from where the steps iii and iv are repeated. This is shown below:
   Let Cluster 1={ $(x_1\ ,\ y_1\ ,\ z_1\ ,\ l_1)$,($x_4\ ,\ y_4\ ,\ z_4\ ,\ l_4$), ……. , (x_n , y_n , z_n , l_n)}
   Now a point a1 is generated as
   a1=$(x_1+x_4+x_i+\ldots..+x_n/m_1,y_1+y_4+y_i+\ldots\ldots+y_n/m_1$, $z_1+z_4+z_i+\ldots..+z_n/m_1$, $l_1+l_4+l_i+\ldots.l_n/m_1)$
   So in same fashion b1, c1, points are generated and after which these a1, b1, c1 points are calculated with step iii Euclidean distances as
   a1 → a , a1 → b, a1 → c, ……..…,a1 → k_a
   d_{11} , d_{12} , d_{13} ………….. , d_{1i}
   b1 → a , b1 → b, b1 → c, ………. , b1 → k_a
   d_{21}, d_{22}, d_{23} ,……………, d_{2i}
   c1 → a , c1 → b, c1 → c,………..., c1 → k_a
   d_{31} , d_{32}, d_{33},………….. , d_{3i}

And later step iv repeats again until the no of iterations are completed.

V. Results

The data available in this paper has been obtained for the years from 1955 to 2009 in East Godavari district of Andhra Pradesh in India. The data is taken in four input variables. They are year, area of sowing in hectares, rainfall in centimeters and production in metric tons.

Table I describes the estimation of crop yield where the minimum and maximum production defines the lowest and approximated highest crop yield for the specific year for East Godavari District region in Andhra Pradesh, India using density based clustering technique.

TABLE I. Minimum and Maximum Production for East Godavari District Using Density Based Clustering

Year	Minimum Production	Maximum Production	Exact Production
1966	362576	459285	385410
1970	470964	596589	245740
1974	416741	527903	454349
1978	391954	496504	532630
1982	430684	545565	539308
1986	596451	755549	298452
1990	827286	1047956	443676
1994	453923	575002	544155
1998	549975	696675	386639
2002	333083	421930	551115
2006	497301	629951	547716

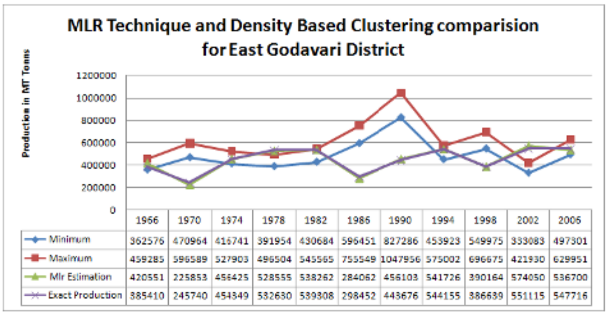
The co-relational Model MLR bring a co-relation between dependent variables and independent variables, here independent variables are rainfall, area of sowing, production per acre and time duration of crop grown. Table II shows exact production comparison with MLR estimations for East Godavari District region in Andhra Pradesh, India.

TABLE II. Estimation of the Production for East Godavari District Using MLR Technique

Year	Area of Sowing	Rainfall	Yield / Acre	Estimated Production	Exact Production
1966	626046	2.52	1379	420551	385410
1970	258499	3.04	1178	225853	245740
1974	244274	2.69	1860	456425	454349
1978	268436	2.53	2058	528555	532630
1982	242821	2.78	2318	538262	539308
1986	253222	3.85	1229	284062	298452
1990	274919	5.34	1682	456103	443676
1994	242472	2.93	2340	541726	544155
1998	256332	3.55	1573	390164	386639
2002	194809	2.15	2829	574050	551115
2006	227930	3.21	2403	536700	547716

The models MLR and density based clustering when compared, give us the best mechanism to utilize in order to predict the crop yield and regulate the production by controlling the independent factors.

In Fig. 1 the estimated results by MLR and Density based clustering technique are compared with respect to exact production to give out the best co-relational model to fit the prediction analysis.



	1966	1970	1974	1978	1982	1986	1990	1994	1998	2002	2006
Minimum	362576	470964	416741	391954	430684	596451	827286	453923	549975	333083	497301
Maximum	459285	596589	527903	496504	545565	755549	1047956	575002	696675	421930	629951
Mlr Estimation	420551	225853	456425	528555	538262	284062	456103	541726	390164	574050	536700
Exact Production	385410	245740	454349	532630	539308	298452	443676	544155	386639	551115	547716

Figure 1. Plotted estimation of the MLR technique and density based clustering approach.

VI. Conclusions

In this paper Density based clustering technique was presented in order to find and evaluate the agricultural yield data. While evaluating with this model an attempt is made to predict the crop production over three specific regions of Andhra Pradesh in India. This model estimates the crop yield by taking several parameters which were explained earlier. Though there are different clustering techniques available in data mining. Density based clustering techniques is found suitable for the approximate prediction. In the subsequent work a comparison of the yield prediction can be made by applying different data mining techniques such as K-Means approach, K-Nearest Neighbor, Support Vector Machine.

Acknowledgment

The authors wish to thank the Principal, Jawaharlal Nehru Technological University Hyderabad College of Engineering, Nachupalli, Kondagattu, Karimnagar Dist., Andhra Pradesh, India for his continuous support and encouragement. The authors are highly motivated by the University authorities for providing the provision in order to publish the papers in International research domain.

References

[1] A. Gersho and R. M. Gray, "Vector quantization and signal compression," in *Communications and Information Theory*, Norwell, MA, USA: Kluwer Academic Publishers, 1992.

[2] I. Dhillon, J. Fan, and Y. Guan, "Efficient clustering of very large document collection," in *Data Mining for Scientific and Engineering Applications*, USA: Kluwer Academic Publishers, 2001.

[3] M. Steinbach, G. Karypis, and V. Kumar, "A comparison of document clustering techniques," in *Proc. 6th ACM SIGKDD World Text Mining Conference*, Boston, MA, 2000, pp. 1-2.

[4] M. Easter, A. Frommelt, H. P. Kreigel, and J, Sander, "Spatial data mining: Database primitives, algorithms and efficient DBMS support," *Data Mining Knowledge Discovery*, pp. 193-216, 2000.

[5] I. Cadez, P. Smyth, and H. Mannila, "Probabilistic modeling of transaction data with applications to profiling, visualizaion and prediction," in *Proc. 7th ACM SIGKDD International Conference on Knowledge Discovery nad Data Mining*, 2001, pp. 37-46.

[6] I. Heer and E. Chi, "Identification of web user traffic composition using multi-model clustering and information scent," in *Proc. 1st SIAM ICDM Workshop on Web Mining*, 2001, pp. 51-58.

[7] V. Leemans and M. F. Destain, "A real time grading method of apples based on features extracted from defects," *J. Jood Eng*., vol. 61, pp. 83-89, 2004.

[8] A. Tellaeche, X. P. B. Artizzu, G. Pajares, and A. Ribeiro, "A vision-based classifier for weeds detection in precision agriculture through the bayesian and fuzzy K-Means paradigms," *Adv.Soft. Comp.*, vol. 44, pp. 72-79, 2008.

[9] K. Verheyen, D. Adriaens, M. Hermy, and S. Deckers, "High resolution continuous soil classification using morphological soil profile descriptions," *Geoderma,* vol. 101, pp. 31-48, 2001.

[10] H. Jorquera, R. Perez, A. Copriano, and G. Acuna, "Short term forcasting of air pollution episoides," in *Environmental Modeling*, UK: WIT Press, 2001.

[11] D. R. Mehta, A. D. Kalola, D. A. Saradava, and A. S. Yusufzai, "Rainfall variability analysis and its impact on crop productivity - A case study," *Indian Journal of Agricultural Research*, vol. 36, no. 1, pp. 29-33, 2002.

[12] B. Rajagopalan and U. Lal, "A K-nearest neighbor simulator for daily precipitation and other weather variable," *Water Resources*, vol. 35, pp. 3089-3101, 1999.

[13] S. Tripathi, V. V. Srinivas, and R. S. Najundiah, "Downscaling of precipitation for climate change scenarios: A support vector machine approach," *J Hydrol*, vol. 330, pp. 621-640, 2006.

[14] G. Ruß, "Data mining of agricultural yield data: A comparison of regression models," in *Conference Proceedings, Advances in Data Mining–Applications and Theoretical Aspects*, P. Perner Ed. Lecture Notes in Artificial Intelligence 6171, Berlin, Heidelberg, Springer, 2009, pp. 24–37.

[15] V. Ramesh and K. Ramar, "Classification of agricultural land soils: A data mining approach," *Agricultural Journal*, pp. 82-86, 2011.

B. Vishnu Vardhan received Doctorate in CSE in 2008 from JNTU Hyderabad and published 21 research papers in National / International Journals / Conferences. He has vast academic experience in Teaching and presently working as Professor of CSE and Head of the Department of IT, JNTUH College of Engineering, Karimnagar Dist., Andhra Pradesh, India, a constituent college of JNTU Hyderabad.

D. Ramesh was graduated from ANU, Guntur, Post Graduate from JNTU Hyderabad, pursuing Ph.D from JNTU Kakinada and having 14 years of experience in Teaching. Presently working as Associate Professor of CSE in the Department of IT, JNTUH College of Engineering, Karimnagar Dist., Andhra Pradesh, India, a constituent college of JNTU Hyderabad.

O Subhash Chander Goud was graduated from Nizam College, Post Graduated from Osmania University and working with 4 years of experience in Teaching as an Assistant Professor(C) in Nizam College, Hyderabad, Andhra Pradesh, India, a constituent College of Osmania University, Hyderabad.