ISSN (Print) : 2319-5940
ISSN (Online) : 2278-1021

International Journal of Advanced Research in Computer and Communication Engineering
Vol. 2, Issue 9, September 2013

Data Mining Techniques and Applications to Agricultural Yield Data

D Ramesh[1] , B Vishnu Vardhan[2]

Associate Professor of CSE, JNTUH College of Engineering , Karimnagar Dist., Andhra Pradesh, India[1]

Professor of CSE, JNTUH College of Engineering, Karimnagar Dist., Andhra Pradesh, India[2]

Abstract : Data Mining is emerging research field in Agriculture crop yield analysis. In this paper our focus is on the applications of Data Mining techniques in agricultural field. Different Data Mining techniques are in use, such as K-Means, K-Nearest Neighbor(KNN), Artificial Neural Networks(ANN) and Support Vector Machines(SVM) for very recent applications of Data Mining techniques in agriculture field. In this paper consider the problem of predicting yield production. Yield prediction is a very important agricultural problem that remains to be solved based on the available data. The problem of yield prediction can be solved by employing Data Mining techniques. This work aims at finding suitable data models that achieve a high accuracy and a high generality in terms of yield prediction capabilities. For this purpose, different types of Data Mining techniques were evaluated on different data sets.

Keywords : Data Mining, K-Means, K-Nearest Neighbor, Artificial Neural Networks, Support Vector Machines, Yield Prediction.

I. Introduction

Data Mining is the process of extracting useful and important information from large sets of data. Data Mining in agriculture field is a relatively novel research field. In this paper describe an overview of Data Mining techniques applied to agricultural and their applications to agricultural related areas. Yield prediction is a very important agricultural problem. Any farmer is interested in knowing how much yield he is about to expect. In the past, yield prediction was performed by considering farmer's experience on particular field and crop. Consider that data are available for some time back to the past, where the corresponding yield predictions have been recorded. In any of Data Mining procedures the training data is to be collected from some time back to the past and the gathered data is used in terms of training which has to be exploited to learn how to classify future yield predictions.

Different techniques were proposed for mining data over the years. A detailed and elaborated 10 Data Mining techniques were discussed by the researchers [1] . In this paper present some of the most used general Data Mining techniques in the field of agriculture. The Artificial Neural Networks do not appear among the afore mentioned Data Mining techniques [1], they were considered in this paper because there are few applications of this technique in agriculture. By using Multilayer Perceptron model of Neural Networks the researchers [2] trained to predict wheat yield by considering sensor input and fertilizers as parameters. MLPs was used successfully in previous work by the researchers Georg Ruβ et al. [3,4,5]. Two different neural networks are considered in [6], one network with a Multilayered Perceptron, another one with a Radial Basis Function, as well as a Support Vector Regression and a Decision Regression Tree. A comparison of these four techniques [6] showed that the Support Vector Regression technique is the most suitable for this kind of problem. Recently spatial Auto Correlation has improved [7] the quality of the prediction.

II. Data Mining Techniques

Data Mining techniques are mainly divided in two groups, classification and clustering techniques [8]. 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, Neural Networks [3,4,5] and Support Vector Machines [9], these two classification techniques learn from training set how to classify unknown samples.

Another classification technique, K- Nearest Neighbor [10], 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. The basic assumption in the K-Nearest Neighbor algorithm is that similar samples should have similar classification. The parameter K shows the number of similar known samples used for assigning a classification to an unknown sample. The K-Nearest Neighbor uses the information in the training set, but it does not extract any rule for classifying the other.

In the event a 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. One of the most used clustering technique is the K-Means algorithm [11]. Given a set of data with unknown classification, the aim is to find a partition of the set in which similar data are grouped in the same cluster. The parameter K plays an important role as it specifies the number of clusters in which the data must be partitioned. The idea behind the K-Means algorithm is, given a certain partition of the data in K clusters, the centers of the clusters can be computed as the means of all samples belonging to a clusters. The center of the cluster can be considered as the representative of the cluster, because the center is quite close to all samples in the cluster, and therefore it is similar to all of them. There are some disadvantages in using K-Means method. One of the disadvantages could be the choice of the parameter K. Another issue that needs attention is the computational cost of the algorithm. There are other Data Mining techniques statistical based techniques, such as Principle Component Analysis(PCA) , Regression Model and Biclustering Techniques [12,13] have some applications in agriculture or agricultural - related fields.

III. Applications

There are several applications of Data Mining techniques in the field of agriculture. Some of the data mining techniques are related to weather conditions and forecasts. For example, the K-Means algorithm is used to perform forecast of the pollution in the atmosphere [14], the K Nearest Neighbor(KNN) is applied for simulating daily precipitations and other weather variables [15], and different possible changes of the weather scenarios are analyzed using SVMs [16].

Data Mining techniques are applied to study sound recognition problems. For instance, Fagerlund S [17] uses SVMs to classify the sound of birds and other different sounds. Holmgren et al. [18] uses a K-Nearest Neighbor approach to evaluate forest inventories and to estimate forest variables for analyzing satellite imagery. Das KC et al. [19] uses ANNs to classify eggs as fertility and Patel VC et al. [20] uses Computer Vision to recognize cracks in eggs. Du C-J et al .[21] uses SVMs to classify pizza sauce spread and Karimi Y et al. [22] uses SVMs for detecting weed and nitrogen stress in corn.

Data Mining techniques are often used to study soil characteristics. As an example, the K-Means approach is used for classifying soils in combination with GPS-based technologies [23]. Meyer GE et al. [24] uses a K-Means approach to classify soils and plants and Camps Valls et al. [25] uses SVMs to classify crops. Apples are checked using different approaches before sending them to the market. Leemans V et al. [26] uses a K-Means approach to analyze color images of fruits as they run on conveyor belts. Shahin MA et al. [27] uses X-ray images of apples to monitor the presence of water cores, and a neural network is trained for discriminating between good and bad apples. A Mucherino et al. [28] apply a supervised biclustering technique to a dataset of wine fermentations with the aim of selecting and discovering the features that are responsible for the problematic fermentations and also exploit the selected features for predicting the quality of new fermentations. Taste sensors are used to obtain data from the fermentation process to be classified using ANNs [29]. Similarly, sensors are used to smell milk, that is classified using SVMs [30].

IV. Overview of Data

The data available in this paper is obtained for the years from 1965 to 2009 in East Godavari district of Andhra Pradesh in India. The data is taken in four input variables. They are Year, Rainfall, Area of Sowing and Production. Year attribute specifies the year in which the data available in Hectares. Rainfall attribute specifies the Rainfall in East Godavari in the specified year in Centimeters. Area of Sowing attribute specifies the total area sowed in East Godavari district in the specified year. Production attribute specifies the production of crop in East Godavari district in the specified year in Tons. The preliminary data collection is carried out for all the districts of Andhra Pradesh in India. Each area in this collection is identified by the respective longitude and latitude of the region. In this paper the evaluation is considered for only one district i.e. East Godavari. The information gathering process is done with three government units like Indian Meteorological Department, Statistical Institution and Agricultural department. Instead of restricting with few regions and few samples of data, it is aimed at applying the Expectation Maximization approach on all the regions of Andhra Pradesh in India. In this paper the estimation of the crop yield is analyzed with respect to four parameters namely Year, Rainfall, Area of Sowing and Production.

V. Result Analysis

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 i.e Rainfall and independent variables are called predictors i.e Year, Area of sowing, Production .

By adopting K-Mean clustering approach four clusters are formed by considering Rainfall as the key parameter. Table I represents four clusters which is formed by the K-Means clustering. Though the total number of years Rainfall is divided in to four clusters, ranging from 1 to 5 cm based on the Rainfall of that region. The four clusters to obtained using K-Means algorithm along with minimum, mean and maximum values are presented in the Table II.

In this process the mean Rainfall is ranging from 1.93 to 4.47. Highest number of years were observed in a clusters 2 and 3 whose yearly mean average is being 2.58 and 3.45 i.e. 21 years are in the mean average of 2.58 and another 18 years are observed in the mean average of 3.45. Only 3 years were mapped to the cluster 1 and 2 years were mapped to the cluster 4 whose mean average is 1.93 and 4.47.

The comparison of yield prediction based on Rainfall between MLR Technique and K-Means algorithm is presented in the Table III. The estimation of average production using MLR Technique is given as 98 % and using K-Means algorithm is given as 96% accuracy with respect to four parameters when it compare to the actual average production.

TABLE I. CLUSTERS WITH YEARLY RAINFALL

Cluster	Yearly Rainfall in cms
1	1.89,1.93,1.97
2	2.12,2.15,2.17,2.37,2.38,2.52,2.52,2.53,2.53, 2.55,2.55,2.57,2.68,2.69,2.77,2.78,2.79,2.81, 2.84,2.93,2.99
3	3.04,3.18,3.21,3.23,3.26,3.26,3.28,3.36,3.4, 3.47,3.55,3.55,3.56,3.62,3.65,3.83,3.85,3.97
4	4.47,4.47

TABLE II. MIN, MEAN AND MAX RAINFALL IN EACH CLUSTER

Cluster	MIN	MEAN	MAX
1	1.89	1.93	1.97
2	2.12	2.58	2.99
3	3.04	3.45	3.97
4	4.47	4.47	4.47

TABLE III. MLR TECHNIQUE VS K-MEANS ALGORITHM

Cluster	Actual Data		MLR Technique	K-Means Algorithm
	Average Area of Sowing	Average Production	Average Production	Average Production
1	208971	426671	412506	464495
2	276816	502582	512596	447685
3	232978	473213	469635	419095
4	238542	463056	444828	456596

By this one can conclude that the years belong to the clusters have witnessed abnormal climatic conditions. Inspite of the regions which are being affected the production is almost double to that of the Area of Sowing in this region. This phenomena may not be similar to all the districts. It depends on the region which is being affected by the natural calamity. In general if the flood hits to costal belt of Andhra Pradesh, this region (East Godavari) is one of the most effected one. As this comparison goes some districts may not come in to such type of clusters and this clustering years changing per cluster per region. In this process, given the Rainfall in a specific year the system is in a position to predict the average yield production by considering the cluster in which the estimated Rainfall belongs to. The cluster identification is carried out based on the input average year Rainfall. This value is mapped with all the clusters by identifying the similar cluster then it predicts the production and approximates the Area of Sowing too.

VI. CONCLUSIONS

In this paper certain Data Mining techniques were adopted in order to estimate crop yield analysis with existing data. The applications that use the K-Means approach , utilize only the basic algorithm, while many other improvements are available. Some Data Mining techniques have not yet been applied to agricultural problems. As an example, Biclustering techniques may be employed for discovering important information from agricultural-related sets of data. The K-Means algorithm is able to partition the samples in clusters, but no considerations are made on the compounds that are responsible for this partition. Biclustering can provides this kind of information.

The future work aimed at the analysis of the entire set of data and will be devoted to suitable strategies for improving the efficiency of the proposed algorithm.

REFERENCES

[1] Wu X, Kumar V, Quilan JR, Ghosh J, Yang Q, Motoda H, McLanchlan GJ, Ng A, Liu B, Yu PS, Zhou Z-H, Steinbach M, Hand DJ, Steinberg D, Top 10 algorithms in data mining. Knowl Inf Syst 14 : 1-37, 2008.

[2] P. Wagner and M. Schneider. Economic benefits of neural network-generated site-specific decision rules for nitrogen fertilization. In J. V. Stafford, editor, Proceedings of the 6th European Conference on Precision Agriculture, 775–782, 2007.

[3] Georg Ruß, Rudolf Kruse, Martin Schneider, and Peter Wagner. Estimation of neural network parameters for wheat yield prediction. In Max Bramer, editor, Artificial Intelligence in Theory and Practice II, volume 276 of IFIP International Federation for Information Processing, 109–118. Springer, July 2008.

[4] Georg Ruß, Rudolf Kruse, Martin Schneider, and Peter Wagner. Optimizing wheat yield prediction using different topologies of neural networks. In Jos´e Luis Verdegay, Manuel Ojeda-Aciego, and Luis Magdalena, editors, Proceedings of IPMU-08, 576–582. University of M´alaga, June 2008.

[5] Georg Ruß, Rudolf Kruse, Peter Wagner, and Martin Schneider. Data Mining with neural networks for wheat yield prediction. In Petra Perner, editor, Advances in Data Mining (Proc ICDM 2008), 47–56, Berlin, Heidelberg, July 2008. Springer Verlag.

[6] G. Ruß, Data Mining of Agricultural Yield Data: A Comparison of Regression Models, Conference Proceedings, Advances in Data Mining – Applications and Theoretical Aspects, P. Perner (Ed.), Lecture Notes in Artificial Intelligence 6171, Berlin, Heidelberg, 24–37, Springer, 2009.

[7] G. Ruß, A. Brenning, Data Mining in Precision Agriculture: Management of Spatial Information, Conference Proceedings, Computational Intelligence for Knowledge- Based Systems Design, E. H¨ullermeier, R. Kruse, and F. Hoffmann (Eds.), Lecture Notes in Artificial Intelligence 6178, Berlin, Heidelberg, 350–359, Springer, 2010.

[8] A. Mucherino, P. Papajorgji, P.M. Pardalos, A Survey of Data Mining Techniques Applied to Agriculture, Operational Research: An International Journal 9(2), 121–140, 2009.

[9] M. Kovacevic, B. Bajat, B. Gajic, Soil Type Classification and Estimation of Soil Properties using Support Vector Machines, Geoderma 154(3–4), 340–347, 2010.

[10] Cover TM, Hart PE, Nearest Neighbor pattern classification. IEEE Trans Info Theory 13(1) : 21-27, 1967.

[11] J. Hartigan, Clustering Algorithms, John Wiles & Sons, New York, 1975.

[12] A. Mucherino, A. Urtubia, Consistent Biclustering and Applications to Agriculture, IbaI Conference Proceedings, Proceedings of the Industrial Conference on Data Mining (ICDM10), Workshop "Data Mining in Agriculture" (DMA10), Berlin, Germany, 105-113, 2010.

[13] A. Mucherino, S. Cafieri, A New Heuristic for Feature Selection by Consistent Biclustering, arXiv e-print, arXiv:1003.3279v1, March 2010.

[14] Jorquera H, Perez R, Cipriano A, Acuna G Short term forecasting of air pollution episodes. In: Zannetti P (eds) Environmental modeling 4. WIT Press, UK, 2001.

[15] Rajagopalan B, Lall U, A K-Nearest Neighbor simulator for daily precipitation and other weather variables. Wat Res Res 35(10) : 3089–3101, 1999.

[16] Tripathi S, Srinivas VV, Nanjundiah RS Downscaling of precipitation for climate change scenarios: a Support Vector Machine approach. J Hydrol 330:621–640, 2006.

[17] Fagerlund S Bird species recognition using Support Vector Machines. EURASIP J Adv Signal Processing, Article ID 38637, p 8, 2007.

[18] Holmgren P, Thuresson T Satellite remote sensing for forestry planning: a review. Scand J For Res 13(1):90–110, 1998.

[19] Das KC, Evans MD Detecting fertility of hatching eggs using machine vision II: Neural Network classifiers. Trans ASAE 35(6):2035–2041, 1992.

[20] Patel VC, McClendon RW, Goodrum JW Crack detection in eggs using computer vision and neural networks. Artif Intell Appl 8(2):21–31, 1994.

[21] Du C-J, Sun D-W Pizza sauce spread classification using colour vision and support vector machines. J Food Eng 66:137–145, 2005.

[22] Karimi Y, Prasher SO, Patel RM, Kim SH Application of support vector machine technology for Weed and nitrogen stress detection in corn. Comput Electronics Agricult 51:99–109, 2006.

[23] Verheyen K, Adriaens D, Hermy M, Deckers S High-resolution continuous soil classification using morphological soil profile descriptions. Geoderma 101:31–48, 2001.

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

[25] Camps-Valls G, Gomez-Chova L, Calpe-Maravilla J, Soria-Olivas E, Martin-Guerrero JD, Moreno J Support Vector Machines for crop classification using hyperspectral data. Lect Notes Comp Sci 2652:134–141 , 2003.

[26] Leemans V, Destain MF A real time grading method of apples based on features extracted from defects. J Food Eng 61:83–89, 2004.

[27] Shahin MA, Tollner EW, McClendon RW Artificial intelligence classifiers for sorting apples based on watercore. J Agric Eng Res 79(3):265–274, 2001.

[28] A. Mucherino, A. Urtubia, Feature Selection for Datasets of Wine Fermentations, I3M Conference Proceedings, 10th International Conference on Modeling and Applied Simulation (MAS11), Rome, Italy, September 2011.

[29] Riul A Jr, de Sousa HC, Malmegrim RR, dos Santos DS Jr, Carvalho ACPLF, Fonseca FJ, Oliveira Jr ON, Mattoso LHC Wine classification by taste sensors made from ultra-thin films and using Neural Networks. Sens Actuators B98:77–82, 2004.

[30] Brudzewski K, Osowski S, Markiewicz T Classification of milk by means of an electronic nose and SVM neural network. Sens Actuators B98:291–298, 2004.

BIOGRAPHY

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.

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.