ISSN: 2321-7782 (Online)
Volume 2, Issue 9, September 2014
International Journal of Advance Research in Computer Science and Management Studies
Research Article / Survey Paper / Case Study
Available online at: www.ijarcsms.com

A survey on Data Mining Techniques for Crop Yield Prediction

Ramesh A. Medar[1]
Dept. of Computer Science & Engineering
Gogte Institute of Technology
Belgaum, Karnataka, India

Vijay. S. Rajpurohit[2]
Dept. of Computer Science & Engineering
Gogte Institute of Technology
Belgaum, Karnataka, India

Abstract: ***This paper presents the various crop yield prediction methods using data mining techniques. Agricultural system is very complex since it deals with large data situation which comes from a number of factors. Crop yield prediction has been a topic of interest for producers, consultants, and agricultural related organizations. In this paper our focus is on the applications of data mining techniques in agricultural field. Different Data Mining techniques 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. Data mining technology has received a great progress with the rapid development of computer science, artificial intelligence. Data Mining is an emerging research field in agriculture crop yield analysis. Data Mining is the process of identifying the hidden patterns from large amount of data. 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.***

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

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. 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. In any of Data Mining procedures the training data is to be collected from historical data and the gathered data is used in terms of training which has to be exploited to learn how to classify future yield predictions[1].

The vision of meeting world's food demands for the increasing population throughout the world is becoming more important in these recent years. Crop models and decision tools are increasingly used in agricultural field to improve production efficiency. The combination of advanced technology and agriculture to improve the production of crop yield is becoming more interesting recently. Due to the rapid development of new higher technology, crop models and predictive tools might be expected to become a crucial element of agriculture.

Crop yield prediction has been a topic of interest for producers, consultants, and agricultural related organizations. Crop yield is a unified bio-socio-system comprised of complex interaction among the soil, the air, the water, and the crops grown in it, where a comprehensive model is required which are possible only through classical engineering expertise. As defined by the Food and Agriculture of the United Nations, crop forecasting is the art of predicting crop yields and production before the harvest actually takes place, typically a couple of months in advance. Crop forecasting philosophy is based on various kinds of data collected from different sources: meteorological data, agro-meteorological (phenology, yield), soil (water holding capacity), remotely sensed, agricultural statistics. Based on meteorological and agronomic data, several indices are derived

which are deemed to be relevant variables in determining crop yield, for instance crop water satisfaction, surplus and excess moisture, average soil moisture, etc[11].

There are two distinct crop models – simulation and regression (Malaylay, et.al., 2010). A simulation model characterizes the mathematical relationships intrinsic to the data set from previous experiments. This method can generate results under various conditions assuming extensive information used to develop and test the model. However, in agricultural data, information is rather sparse and incomplete. Because of this limitation, the regression approach is the common approach for predicting yield across large area. Furthermore, the most investigated statistical crop-yield-weather models are multivariate regression models[11].

In past decades, IT has become more & more part of our everyday lives. With IT improvements, efficiency can be made in almost any part of industry and services, now a days this is especially true for agriculture. A farmer now a days harvest not only crops but also growing amounts of data. These data are precise & small in scale. However, collecting large amounts of data often is both a blessing and a curse. There is a lot of data available containing information about certain asset. Here soil and yield properties, which should be used to the farmers advantage. This is a common problem for which the term data mining has been coined. Data mining techniques aim at finding those patterns or information in the data that are both valuable and interesting to the farmer. A common specific problem that occurs is yield prediction[3].

II. LITERATURE SURVEY

The paper titled "Generalized software tools for crop area estimation and yield forecast", Roberto Benedetti and others describes the procedure that leads to the estimates of the variables of interest, such as land use and crop yield and other sampling standard deviation, is rather tedious and complex, till to make necessary for statistian to have a stable and generalized computational system available. This paper focus on the use of this system in different steps of the survey: sample design, data editing and estimation. The information produced is however, available for one user only, the manager of the survey.

"Risk in Agriculture: A study of crop yield distribution and crop insurance" by Narsi Reddy Gayam in his research study examines the assumption of normality of crop yields using data collected from INDIA involving sugarcane and Soybean. The null hypothesis (Crop yield are normally distributed) was tested using the Lilliefore method combined with intensive qualitative analysis of the data. Result show that in all cases considered in this thesis, crop yield are not normally distributed[10].

The paper titled "APPLYING DATA MINING TECHNIQUES IN THE FIELD OF AGRICULTURE AND ALLIED SCIENCES" an attempt has been made to review the research studies on application of data mining techniques in the field of agriculture. Some of the techniques, such as ID3 algorithms, the k-means, the k nearest neighbour, artificial neural networks and support vector machines applied in the field of agriculture were presented. Data mining in application in agriculture is a relatively new approach for forecasting / predicting of agricultural crop/animal management[6].

F.P. Lansigan, et.al (2010) used climate change scenarios to simulate corn performance under the anticipated seasonal climate change conditions and used a weather generator (simmeto) to produce sequences of daily weather data. A multiple regression was developed to predict corn yield using climatic variables, such as temperature, humidity, rainfall, wind speed, and solar radiation. The results of the study showed that climate variability significantly affects crop yields[7].

The paper titled "Data mining Techniques for Predicting Crop Productivity – A review article" an attempt has been made to review the research studies on application of data mining techniques in the field of agriculture. Some of the techniques, such as ID3 algorithms, the k-means, the k nearest neighbor, artificial neural networks and support vector machines applied in the field of agriculture were presented. Data mining in application in agriculture is a relatively new approach for forecasting / predicting of agricultural crop/animal management. This article explores the applications of data mining techniques in the field of agriculture and allied sciences. Historical crop yield information is important for supply chain operation of companies engaged in industries that use agricultural produce as raw material. Livestock, food, animal feed, chemical, poultry, fertilizer pesticides,

seed, paper and many other industries use agricultural products as intergradient in their production processes. An accurate estimate of crop size and risk helps these companies in planning supply chain decision like production scheduling. Business such as seed, fertilizer, agrochemical and agricultural machinery industries plan production and marketing activities based on crop production estimates[5].

III. DATA MINING TECHNIQUES

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, Neural Networks and Support Vector Machines, these two classification techniques learn from training set how to classify unknown samples[1]. 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. 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[1].

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 . 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.

Artificial neural network (ANN) is based on the human brain's biological neural processes. ANN learns to recognize the patterns or relationships in the data by observing a large number of input and output examples. Once the neural network has been trained, it can predict by detecting similar patterns in future data. They include the ability to learn and generalize from examples to produce meaningful solutions to problems even when input data contain errors or are incomplete, and to adapt solutions over time to compensate for changing circumstances and to process information rapidly. Furthermore, a system may be *nonlinear* and *multivariate*, and the variables involved may have complex interrelationships. ANNs are capable of adapting their complexity, and their accuracy increases as more and more input data are made available to them. They are capable of extracting the relationship between the input and output of a process without the any knowledge of the underlying principles. The recent increased interest and use of neural models stems primarily from its nonlinear models that can be trained to map past and future values of the input-output relationship. This adds analytical value, since it can extract relationships between governing the data that was not obvious using other analytical tools. ANNs are also used because of its capability to recognize pattern and the speed of its techniques to accurately solve complex processes in many applications. They help to characterize relationships via a *nonlinear, nonparametric inference technique*, this is very rare and has many uses in a host of disciplines. The network offer the added advantage of being able to establish a 'training' phase, where example inputs are presented and the networks learn to extract the relevant information form these patterns. With this, the network can generalize results and lead to logical and other unforeseen conclusions through the model[11].

IV. 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, the K Nearest Neighbor(KNN) is applied for simulating daily precipitations and other weather variables , and different possible changes of the weather scenarios are analyzed using SVMs. 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. Meyer GE et al. uses a K-Means approach to classify soils and plants and Camps Valls et al. uses SVMs to classify crops[1].

In past decades, IT has become more & more part of our everyday lives. With IT improvements in efficiency can be made in almost any part of industry and services, now days this is especially true for agriculture. A farmer now days harvest not only crops but also growing amounts of data. These data are precise & small in scale. However, collecting large amounts of data often is both a blessing and a curse. There is a lot of data available containing information about certain asset. Here soil and yield properties, which should be used to the farmers advantage. This is a common problem for which the term data mining has been coined. Data mining techniques aim at finding those patterns or information in the data that are both valuable and interesting to the farmer. A common specific problem that occurs is yield prediction. As early into the growing season as possible, a farmer interested in knowing how much yield he is about to expect. In the past, this yield prediction has actually relied on farmer's long–term experience for specific yield, crops and climatic conditions. However, this knowledge might also be available, but hidden in the small–scale. Precise data which can now days collected in seasons using a multitude of seasons. Upgrading and stabilizing the agricultural production at a faster pace is one of the basic conditions for agricultural development. Productions of any crop lead either by attention of area or improvement in productivity or both. In India, the possibility of extending the area under any crop, almost, does not exist except by restoring to increased cropping intensity or crop substitution. Moreover, area and Raorane A.A. et al, / (IJCSIT) International Journal of Computer Science and Information Technologies, Vol.4(2), 2013, 270-272 www.ijcsit.com 270 productivity of different crops are the results, and as well as the reflection of the combined effect of many factors like agro-climatic conditions resource endowment technology level, techniques adopted infrastructure, social & economic conditions many schemes have been devised to maximize the productivity of various crops in different agro-climate region, state departments, credit institution, seed/fertilizer pesticide agencies & many other partners in public & private sections are actively engaged in enhancing the productivity of different crops in different regions and under different condition. However fluctuations in crop productivity continue to dog the sector and create severe distress[3].

V. METHODOLOGY

A. Inputs:

One of the main goals of crop prediction models is to estimate agricultural production as a function of weather and soil conditions as well as crop management. Dynamic crop production model systems, as decision supporting tools, have extensively been utilized by agricultural scientists to evaluate possible agricultural consequences from inter-annual climate variability and/or climate change (Paz et al., 1998; Semenov et al., 1996, cited by Zinyengere, 2010).

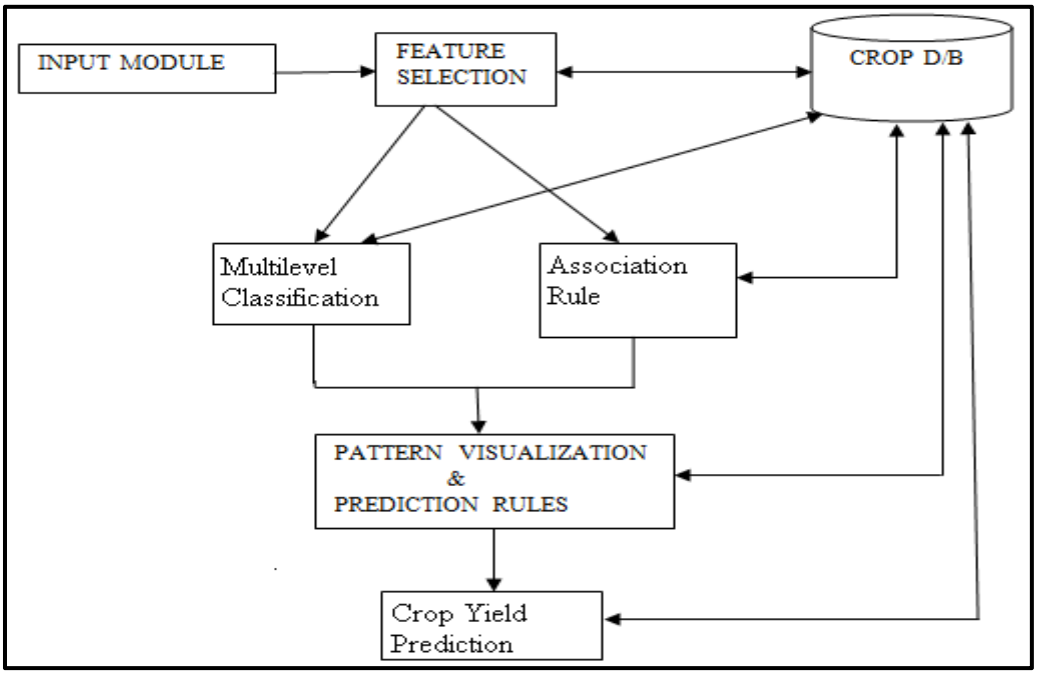

Fig. 1. Architecture of crop yield prediction

Fig. 1 shows the architecture of crop prediction which includes an input module which is responsible for taking input from farmer. In that the farmer has to provide area of land, region, economic status and city. The farmer is also responsible for interacting with predicted results. After selecting the city parameter based on altitude, longitude and latitude automatic climatic data will be reflected from crop knowledge base. The feature selection module is responsible for subset selection of attribute from crop knowledge base for robust learning. The crop knowledge base is consist of farm knowledge such as region-id, region-name, soil-type, water ph, rainfall, humidity, sunlight, land information, environmental parameter, city, pesticides information, crop knowledge such crop type, seed type. The knowledge-base also includes the samples of crop with corresponding farm knowledge, environmental parameter, and pesticides information. After subset selection of attribute, the data goes to classification and association rule for grouping similar contents. Then prediction rules will be applied to output of clustering to get results in terms of crop, pesticide and cost.

VI. CONCLUSION

With the improvement of data mining technologies, especially those without any premises or humans subjective, data mining can be applied in many areas. In this paper some data mining techniques were adopted in order to estimate crop yield analysis with existing data and their use in data mining. This paper presents new research possibilities for the application of modern classification methodologies to the problem of yield prediction. There is a growing number of applications of data mining techniques in agriculture and a growing amount of data that are currently available from many resources.

References

1. D Ramesh, B Vishnu Vardhan, “Data Mining Techniques and Applications to Agricultural Yield Data”, International Journal of Advanced Research in Computer and Communication Engineering Vol. 2, Issue 9, September 2013.
2. Mr. Abhishek B. Mankar, Mr. Mayur S. Burange, “Data Mining - An Evolutionary View of Agriculture”, International Journal of Application or Innovation in Engineering & Management (IJAIEM), Volume 3, Issue 3, March 2014.
3. Raorane A.A., Kulkarni R.V, “Review- Role of Data Mining in Agriculture”, International Journal of Computer Science and Information Technologies, Vol. 4 (2) , 2013, 270 – 272.
4. YETHIRAJ N G, “APPLYING DATA MINING TECHNIQUES IN THE FIELD OF AGRICULTURE AND ALLIED SCIENCES”, International Journal of Business Intelligents, Vol 01, Issue 02, Dec- 2012.
5. S.Veenadhari, Dr. Bharat Misra, Dr. CD Singh, “Data mining Techniques for Predicting Crop Productivity – A review article”, International Journal of Computer Science and technology, march 2011.
6. YETHIRAJ N G. “APPLYING DATA MINING TECHNIQUES IN THE FIELD OF AGRICULTURE AND ALLIED SCIENCES”, International Journal of Business Intelligents ISSN: 2278-2400, Vol 01, Issue 02, December 2012.

7. Lansigan, et.al., “Analysis of Climatic Risk and Coping Strategies in Two Major Corn Growing Areas in the Philippines” 2010.
8. Georg Rub, 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)”, pages 47–56, Berlin, Heidelberg, July 2008, Springer Verlag.
9. Georg Ruß Data Mining of Agricultural Yield Data: A Comparison of Regression Models, ICDM'09,. Leipzig, Germany, July 2009.
10. Roberto Benedetti A, Remo Catenaro A, Federica Piersimoni B, “GENERALIZED SOFTWARE TOOLS FOR CROP AREA ESTIMATES AND YIELD FORECAST ”2010.
11. Maria Rossana C. de Leon, Eugene Rex L. Jalaok, “A Prediction Model Framework for Crop Yield Prediction”, Asia Pacific Industrial Engineering and Management System, 2013.
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.

