Vol 3, Issue 5, May 2023 E-ISSN: 2582-9734
International Journal of Engineering, Science, Technology and Innovation (IJESTI)

PREDICTION OF CROPS METHODOLOGY USING DATA MINING TECHNIQUES

Kunal Bhushan Ranga

Assistant Professor, Dept. of Computer Applications, Engineering College Bikaner

Email: kunalranga@gmail.com

ABSTRACT

Agricultural system is very difficult since it deals with large data condition which comes from a number of issues. Crop harvest prediction has been a matter of awareness for producers, consultants, and agricultural associated establishments. In this research an effort has been made to review the research studies on application of data mining techniques in the field of agriculture. Some of the techniques, like 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 farming is a comparatively new methodology for forecasting / predicting of agricultural crop/animal management. This research explores the applications of data mining techniques in the field of agriculture and allied sciences. This research grants the numerous crop yield prediction methods 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)

1. Introduction

In the cultivation sector, data mining can help government to growth yield benefit largely to support resolution creation, consistent and timely information on crop area, crop production and land use is of great importance to planners and policy makers for well-organized agricultural development and for pleasing decisions on obtaining, storage, public distribution, export, import and many other connected concerns to participate in the trade of crop design. Data mining can be defined as the process of selecting, exploring and modeling large amounts of data to uncover previously unknown patterns. In the agriculture sector, data mining can assistance agriculturalists to gain profit and country development. For example, by applying data mining techniques, government can fully exploit data about agriculturalists' buying configurations and behavior – as well as gaining a greater understanding of their land to protect them managing invertebrate pests and vertebrate pests, diseases, improve underwriting and enhance risk on crop cultivation. This research discusses how agriculturalists can benefit by using modern data mining methodologies and thereby reduce costs, increase profits, acquire new agriculturalists, retain current agriculturalists and cultivate new crops. Data mining procedure frequently can recover upon traditional statistical methods to solving business solutions. For example, linear regression may be used to solve a problem because insurance industry regulators require easily interpretable models and model parameters. Data mining often can improve existing models by finding additional, important variables, recognizing collaboration terms and noticing nonlinear relationships.

2. Literature Survey

From the research article “Data mining of agricultural yield Data: A comparison of regression models” George RuB 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 advantage. This research deals with appropriate regression techniques on selected agriculture data. “Classification of agricultural land soils: A data mining approach” In this research V. Ramesh and K. Ram explains comparison of different classifiers and the outcome of this research could improve the management and systems of soil uses throughout a large field that include agriculture, horticulture, environmental and land use management. D. R. Mehata and others are worked on “Rainfall variability analysis and its impact on crop productivity”

In this case study they collected the weekly rainfall data and number of rainy days recorded at the main Dry farming research station from 1958 to 1996 (39 yrs). 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. From “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 statistical to have a stable and generalized computational system available. The SAS is also often the ideal instrument to face with these needs, because it permits the handling of data effectively and provides all necessary functions to manage easily surveys with thousands of micro data. This research 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.

3. 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 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 techniques 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 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 Blustering 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 procedure; this is very rare and has many uses in a host of corrections. The network proposal the additional improvement of existence able to create a 'training' segment, where example inputs are presented and the networks learn to extract the relevant information form these patterns. With this, the network can simplify results and lead to logical and other unforeseen conclusions through the model [11].

4. Methodology

4.1 Data Mining

Data Mining is the process of discovering previously unknown and potentially increasing pattern in large datasets. The extracted evidence is used for demonstrating as a model for prediction or classification. Datasets which are collected from Kolhapur district appear to be significantly more complex than the dataset traditionally used in the machine learning. Data mining is mainly categorized as descriptive and extrapolative data mining. But in the agricultural area predictive data mining is mainly used. There are two main techniques namely classification and clustering. [5] Some of the following procedures are used for getting the solution from collected data.

4.2. Artificial Neural Network

Artificial Neural Network is ainnovativemethod used in flood forecast. The benefit of ANN system over the other system is it can model the rainfall also it forecasts the pest occurrence incidence for one week in advance. Data mining implements are commencement to demonstration value in examining enormous data sets from problematical systems and providing high-quality information (White and Frank, 2000). An artificial neural network (ANN) is an striking unusual 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. In the past few years, several yield simulation models have been built. Ambuel et al. (1994) used a fuzzy logic expert system to predict corn yields with promising results. The functional relationship using the fuzzy logic expert system was expressed linguistically instead of mathematically. The authors suggested the use of a neural network to predict within-field yields. [6][7][8]

4.3 Decision tree

Decision tree is one of the classification algorithms which can be used in Data mining. Application of data mining techniques on drought related for drought risk management shows the success on advanced Geospatial Decision Support System (GDSS). Learning decision tree is paradigm of inductive learning. A model is built from data or observations according to some criteria. The model purposes to acquireaall-purpose rule from the experiential instances. Decision trees can therefore accomplish two different tasks depending on whether the target attribute is discrete or continuous. In the forest case a classification tree would result where as in the second cases regression tree would be fabricated. [9][14]

4.4. Bayesian network

Bayesian network is a powerful tool for dealing uncertainties and widely used in agriculture datasets. Bayesian network is a graphical model which encodes probabilistic relationship among variable of interest when it is used with statistical technique, the graphical model has several advantages for data analysis. This performance obviously deals with uncertainty of data and relationships, and can include both qualitative and quantitative variable. It facilitates effective communication with stakeholders, while endorsing a focus on key variables and relationships of the system, slightly than being bogged down in details. [10][11]

4.5 Support Vector Machine

SVM is able to classify data samples in two disjoint collections. SVM are a set of related supervised learning technique used for classification and regression. i.e. the SVM can build a model that predicts whether a new example falls into classification or the other. A support vector machine is a concept is statistics and computer science for a set of related supervised learning methods that examine data and distinguish decorations 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 nonprobability binary linear classifier. An SVM is used in model building which is a representation of the examples as points in space, mapped so that the instances of the separate classes are divided by a clear gap that is as wide as possible. New illustrations are then mapped into that same space and projected to belong to a category based on which side of the gap they fall. [12], [13]

5. Conclusion

Data Mining is an emerging investigation field in agriculture crop yield exploration. Data Mining is the process of distinguishing the unknown configurations from large amount of data. Yield prediction is a very important agricultural problematic that leftovers to be solved based on the obtainable data. The tricky of yield prediction can be solved by commissioning data mining techniques.With the improvement of data mining technologies, principally those without any evidences or humans subjective, data mining can be applied in many areas. In this research some data mining techniques were implemented in order to approximation crop yield examination with prevailing data and their use in data mining. This research grants new exploration prospects for the application of modern classification methodologies to the problem of yield prediction. There are anincreasing number of submissions of data mining methods in farming and a growing quantity of data that are currently accessible from many assets.

\

References

1) Abdullah, A., Brobst, S., M.Umer M. 2004. "The case for an agri data ware house: Enabling analytical exploration of integrated agricultural data". Proc. of IASTED International Conference on Databases and Applications. Austria. Feb
2) Abdullah, A., Brobst, S, Pervaiz.I.,Umer M.,A.Nisar. 2004. "Learning dynamics of pesticide abuse through data mining". Proc. of Australian Workshop on Data Mining and Web Intelligence, New Zealand, January.
3) Abdullah, A., Bulbul.R., Tahir Mehmood. 2005. "Mapping nominal values to numbers by data mining spectral properties of leaves". Proc. of 3rd International Symposium on Intelligent Information Technology in Agriculture. Beijing, China. Oct, 2005.
4) Babu, MSP.,Ramana Murthy, NV, SVNL Narayana, 2010. "A web based tomato crop expert information system based on artificial intelligence and machine learning algorithms". Int. J. of Comp. Sci., and Information Technologies. Vol. 1(1) . pp. 6-15.
5) Basak J., Sudharshan, A., Trivedi D., M.S.Santhanam. 2004. "Weather Data Mining Using Independent Component Analysis". J. of Machine Learning Research 5: pp. 239- 253.
6) Camps-Valls G, Gomez-Chova L, Calpe-Maravilla J, SoriaOlivas E, Martin-Guerrero JD, Moreno J., 2003, "Support vector machines for crop classification using hyperspectral data". Lect Notes Comp Sci 2652: pp. 134–141
7) Chi-Chung LAU, Kuo-Hsin HSIAO, 2005. "Bayesian Classification For Rice Paddy interpretation". Research presented in Conference on data mining held at China Tapei. December, 2005
8) Cunningham S.J., G. Holmes. 2005. "Developing innovative applications in agriculture using data mining". Proc. of 3rd International Symposium on Intelligent Information Technology in Agriculture. Beijing, China. Oct, 2005.
9) Jain Rajni, Minz, S., V. Rama Subramaniam. 2009. "Machine learning for forewarning crop diseases". J. Ind. Soc. Agri. Stat. 63(1): pp. 97-107.
10) Jianlin Ji Dan, Qiu Chen, Jianping Chen, Li He Peng , 2010. "An improved decision tree algorithm and its application in maize seed breeding". Sixth Internation Conference on Natural Computation, held at Yantai, Shandon 10-12th January. pp. 117-121.

11) Jones JW, Tsuji GY, Hoogenboom G, Hunt LA, Thornton PK, Wilkens PW, Imamura DT, Bowen WT, Singh U., (1998), "Decision support system for agrotechnology transfer: DSSAT v3". In: Tsuji GY, Hoogenboom G, Thornton PK (eds) , "Understanding options for agricultural production". Kluwer Academic Publishers, Dordrecht, pp 157–177
12) Kiran Mai, C., Murali Krishna, I.V., A.Venugopal Reddy, 2006. "Data Mining of Geo-spatial Database For Agriculture Related Application". Proc. of Map India. New Delhi.
13) [13] Leonard RA, Knisel WG, Still DA., 1987, GLEAMS: groundwater-loading effects of agricultural management systems. Trans Am SocAgricEng 30(5): pp. 1403–1418
14) McQueen Robert J, Garner S.R.,Nevill-Manning C.G. , Ian H. Witten, 1995. "Applying machine learning to agricultural data".Comptuers and Electronics in Agriculture. Vol. 12: pp. 275-293.
15) Meyer GE, Neto JC, Jones DD, Hindman TW, 2004, "Intensified fuzzy clusters for classifying plant, soil, and residue regions of interest from color images". Computer Electronics Agric Vol. 42: pp. 161–180.
16) RabiaImitiaz, Malik Sikandar Hayat Khiyal, ShahidKhalil , Ahsan Abdullah, 2005, "Effect of pesticides on human life through visual data mining". Journal of Theoretical and Applied Information Technology. pp. 104-109.
17) Stockle CO, Martin SA, Campbell GS, 1994, "CropSyst, a cropping systems model: water/nitrogen budgets and crop yield". AgricSyst Vol. 46(3): pp. 335–359.
18) Veenadhari, S. 2007. "Crop productivity mapping based on decision tree and Bayesian classification". Unpublished M.Tech Thesis submitted to MakhanlalChaturvedi National University of Journalism and Communication, Bhopal.
19) Verheyen K, Adriaens D, Hermy M, Deckers S., 2001, "High-resolution continuous soil classification using morphological soil profile descriptions". Geoderma Vol. 101: pp. 31–48
20) Yue Jin Hai, Song Kai, 2010. "IBLE Algorithm in agricultural disease diagnosis". In third International Conference on Intelligent Networks and Intelligent Systems held at Shenyang, Liaoning China during November 01-November 03.