International Journal of Emerging Trends & Technology in Computer Science (IJETTCS)
Web Site: www.ijettcs.org Email: editor@ijettcs.org, editorijettcs@gmail.com
Volume 1, Issue 2, July – August 2012 ISSN 2278-6856

Data Mining: An effective tool for yield estimation in the agricultural sector

Raorane A.A.[1], Kulkarni R.V.[2]

[1] Department of computer science, Vivekanand College,
Tarabai park Kolhapur INDIA.

[2]Head of the Department, Chh. Shahu Institute of business Education and Research Centre
Kolhapur 416006 INDIA

Abstract: ***Agriculture is a business with risk. Crop production depends on climatic, geographical, biological, political and economic factors. Because of these factors there are some risks, which can be quantified when applied appropriate mathematical or statistical methodologies. Actually accurate information about the nature of historical yield of crop is important modeling input, which are helpful to farmers & Government organization for decision making process in establishing proper policies. The advances in computing and information storage hove provided vast a most of data. The challenge has been to extract knowledge from this raw data; this has lead to new methods and techniques such as data mining that can bridge the knowledge of the data to the crop yield estimation. This research aimed to assess these new data mining techniques and apply them to the various variables consisting in the database to establish if meaningful relationships can be found.***

Keywords- Yield estimation, Data mining, regression analysis, crop cutting experiments

1. INTRODUCTION

Indian agriculture is known for its diversity which is mainly result of variation in resource and climate, to topography and historical, institutional and socio economic factors. Policies followed in the country and nature of technology that became available over time has reinforced some of the variations resulting from natural factors. As a consequence, production performance of agriculture sector has followed on uneven path and large gaps have development in productivity between different geographic locations across the country.

Agriculture as a business is unique crop production is dependent on many climatic, geographical, biological political and economic factors that are mostly independent of one another. This multiple factor introduces risk. The efficient management of these risks is imperative for the successful agricultural and consistent output of food.

The Agricultural yield is primarily depends on weather conditions, diseases and pests, planning of harvest operation. Effective management of these factors is necessary to estimate the probability of such unfavorable situation & to minimize the consequences. Accurate and reliable information about historical crop yield is thus vital for decisions relating to agricultural risk management.

Historical crop yield information is also 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.[1],[2]

2. APPLICATION

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 a day this is especially true for agriculture. A farmer now a day harvests 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 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.

Estimation of productivity of different crops is one of the important activities undertake by the government departments in order to monitor the progress of the sector & provide insurance to the sector. Revenue, agriculture & Economics & statistics departments are jointly involved in the estimation process. Researcher & many other agencies use the data so generated by the Government departments. But these are usually available only in an aggregate form & maximum of taluka level satellite images of crop slate are being used increasingly to estimate the area but productivity data have to come from crop cutting experiment.

Article 243-9 of constitution of India requires the panchayat Raj institutions to be the decision making bodies in various aspects of agricultural sector and especially the implementation of the schemes. Crop Insurance is one of the important schemes of the agricultural sector. The debate in implementation of this scheme indicated requirement of the yield estimates of lower than the talaka level and especially of panchayat level. [3]

3. LITERATURE REVIEW

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 paper deals with appropriate regression techniques on selected agriculture data.

“Classification of agricultural land soils: A data mining approach” In this research paper V. Ramesh and K. Ramr explains comparison of different classifiers and the outcome of this research could improve the management and systems of soil uses throughout a large fields 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 statistian 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 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.

4. SAMPLE DESIGN

Researcher uses data which is proposed by directorate of economics & statistics of India. State Governments Statistical & agricultural department as well as soil department.

Generally the government employee called as talathis, is collecting the required data for the department. In each village he use to select plot and the respective crops randomly, means the department is collecting the required information for yield estimation from each and every village.

For this research study researcher has selected following crops in Kolhapur district in Maharashtra state in India. He selected these crops because maximum of the farmers are cultivating these crops though out the district as cash crops, which are as follows:

- Rice
- Ground nut
- Soybean
- Sugarcane

Figure 1: Kolhapur shown wrt India Talukas in Kolhapur

Following table shows the statistics of total number of talukas & villages coming under Kolhapur districts along with crop area in respective taluka.

Table No. 1

Taluka (Village)	Gross cropped Area (In Arcs			
	Khaiff		Rabi	
	Food crops	Non Food crops	Food crops	Non Food crops
Ajara(22)	45688	28444	359	--
Bawada (11)	21704	5431	123	--
Bhudargad(37)	47191	12640	676	--
Gadhinglaj(42)	62946	39757	1309	20
Hatkanangale(17)	72599	47662	384	--
Kagal(17)	62546	58985	3588	6
Karveer(67)	74553	36204	5309	--
Panhala (43)	48326	29060	4021	--
Radhanagari(34)	48233	20531	1946	--
Shahuwadi(45)	48441	20967	7602	--
Shirol(41)	63897	41710	5292	1231
Total	595026	341391	37466	1257

Table No. 2 : Distribution of cropped Area in Kolhapur District (in Arcs)

Taluka	Gross cropped area	total Food crops	Grand Total
Ajara	45688	28444	74832
Bawada	21704	5431	27135
Bhudargad	47191	12640	59831
Gadhinglaj	62946	39777	102723
Hatkanangale	72599	47662	126261
Kagal	62546	58991	121537
Karveer	74553	36204	110757
Panhala	48328	29060	77388
Radhanagari	48233	20531	68764
Shahuwadi	48441	20967	69408
Shirol	63897	42941	106838
Total	595626	342648	937674

Table No. 3 : A crops under different crops in olhapur Districts

Crop	Ajara	Bawada	Bhudargad	Gadhinglaj
Cereals	41745	26662	41533	50316
Pulse	2488	112	1481	5761
Sugarcane	857	926	3925	2980
Oil seeds	5644	2	4314	16911
Foodgrain	2872	4896	7496	15183

Crop	Hatkanangale	Kagal	Karveer	Panhala
Cereals	51569	49383	54400	38829
Pulse	8792	5834	5712	2796
Sugarcane	6725	4114	12687	5457
Oil seeds	25148	18737	10861	7211
Foodgrain	11828	33647	24033	21536

Crop	Radhanagari	Shahuwadi	Shirol	Total
Cereal	39874	43144	4482	476311
Pulses	1735	2937	1466	51694
Sugarcane	6212	2122	2356	48361
Oilseeds	2741	5808	1773	115086
Food grain	16637	15126	4993	177247

Source: Kolhapur Gazetteer

The data is collected from the district level or state level Directorate of Economics & statistics considered a reputed government of organization within India. This organization prepare yield estimation by conducting crop cutting experiment (CCES) taken under scientifically designed general crop estimation surveys (GCES). The crop cutting plots of a specified size and shape in a selected field, on the principle of random sampling, threshing the produce and recording of the produce harvested for determining the percentage of recovery of the economic or marketable form of produce.

The GCES are done by caring out stratified multi- stage random sampling design with Tehsil / Taluka as strata, revenue villages within a stratum as first stage unit of sampling, survey number or field within each selected village as sampling unit at the second stage an experimental plot of specified shapes and size as the ultimate unit of sampling. The government statistical department used scientific methodology for a riving of the estimation.

Identification of a suitable statistical technique is necessary to analyze the data and arrive at conclusions. Understanding of previous methodologies followed by other researcher and the merits and demerits of these different techniques helps in identification of the appropriate methodology.

For this survey stratified sampling is used also procedure of multivariate allocation, whose development require generalization of the classical formulas of calculation of optimal size.

The stratified random sampling selection without replacement of the units is make through the use of the well known technique of the permanent random numbers in which for every unit I of the frame (Information about their geographical location and other information that can be used for sampling as well as producing estimation of certain basic characteristics as sample aggregations and tabulation)of N dimension is associated independently by

the other, meaning pseudo random (pseudo because it is generated by a computer)by a rectangular variable.[4]

5. MATERIAL AND METHODS

5.1 Data Mining

Data Mining is the process of discovering previously unknown and potentially increasing pattern in large datasets. The mined information is used for representing 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 predictive 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 techniques are used for getting the solution from collected data.

5.2 Artificial Neural Network

Artificial Neural Network is a new technique used in flood forecast. The advantage of ANN system over the other system is it can model the rainfall also it predicts the pest attack incidence for one week in advance. Data mining tools are beginning to show value in analyzing massive data sets from complicated systems and providing high-quality information (White and Frank, 2000). 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. 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]

5.3 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 aims to learn a general rule from the observed 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 constructed. [9][14]

5.4 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 technique explicitly deals with uncertainty of data and relationships, and can include both qualitative and quantitative variable. It facilitates effective communication with stakeholders, while promoting a focus on key variables and relationships of the system, rather than being bogged down in details.[10][11]

5.5 Support Vector Machine

SVM is able to classify data samples in two disjoint clusters. 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. An SVM is used in model building which is a representation of the examples as points in space, mapped so that the examples of the separate categories are divided by a clear gap that is as wide as possible. New examples are then mapped into that same space and predicted to belong to a category based on which side of the gap they fall.[12],[13]

6. RESULTS ON DISCUSSION

Several Data mining techniques used in agriculture study area. We are discussed the few techniques here. Also one technique called K means method is used to forward the pollution in atmosphere. Different changes of weather are analyzed using SVM. K means approach is used to classify the soil and plants. Wine fermentation process monitored using Data mining techniques.

7. Conclusion

It is observed that efficient technique can be developed and analyzed using the appropriate data, the data which is collected from Kolhapur district to solve complex agricultural problems using Data mining techniques.

Recommendation

There can be more advanced techniques developed in agriculture area. After studying more techniques some of the algorithms, statistical methods will give good results in agricultural growth.

REFERENCES

[1] Data mining Techniques for Predicting Crop Productivity – A review article 1S.Veenadhari, 2Dr. Bharat Misra, 3Dr. CD Singh IJCST Vol. 2, Iss ue 1, March 2011

[2] Chapman P. Gleason LARGE AREA YIELD ESTIMATION/ FORECASTING USING PLANT PROCESS MODELS By Chapman P. Gleason For Presentation at the 1982 Winter Meeting AMERICAN SOCIETY OF AGRICULTURAL ENGINEERS Palmer House, Chicago, Illinois December r 14-17, 1982

[3] R S Deshpande AN ANALYSIS OF THE RESULTS OF CROP CUTTING EXPERIMENTS Agricultural Development and Rural Transformation Unit Institute for Social and Economic Change February 2003

[4] Georg Ruß Data Mining of Agricultural Yield Data: A Comparison of Regression Models, ICDM'09,. Leipzig, Germany, July 2009

[5] David B. Lobell, J. Ivan Ortiz-Monasterio, Gregory P. Asner, Rosamond L. Naylor, and Walter P. Falcon. Combining field surveys, remote sensing, and regression trees to understand yield variations in an irrigated wheat landscape. Agronomy Journal, 97:241–249, 2005.

[6] Chengquan Huang, Limin Yang, Bruce Wylie, and Collin Homer. A strategy for estimating tree canopy density using landsat 7 etm+ and high resolution images over large areas. In Proceedings of the Third International Conference on Geospatial Information in Agriculture and Forestry, 2001.

[7] 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), pages 47–56, Berlin, Heidelberg, July 2008. Springer Verlag.

[8] 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, pages 576–582. University of M´alaga, June 2008.

[9] 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, pages 109–118. Springer, July 2008.

[10] J. R. Quinlan. Induction of decision trees. Machine Learning, 1(1):81–106, March 1986.

[11] Applying Naive Bayes Data Mining Technique for Classification of Agricultural Land Soils P.Bhargavi, Dr.S.Jyothi, IJCSNS International Journal of Computer Science and Network Security, VOL.9 No.8, August 2009 117

[12] Bernhard E. Boser, IsabelleM. Guyon, and Vladimir N. Vapnik. A training algorithm for optimal margin classifiers. In Proceedings of the 5th Annual ACM Workshop on Computational Learning Theory, pages 144–152. ACM Press, 1992

[13] Ronan Collobert, Samy Bengio, and C. Williamson. Svmtorch: Support vector machines for large-scale regression problems. Journal of Machine Learning Research, 1:143–160, 2001.

[14] Iv´an Mej´ıa-Guevara and ´Angel Kuri-Morales. Evolutionary feature and parameter selection in support vector regression. In Lecture Notes in Computer Science, volume 4827, pages 399–408. Springer, Berlin, Heidelberg, 2007.

[15] Georg Ruß Data Mining of Agricultural Yield Data: A Comparison of Regression Models, ICDM'09,. Leipzig, Germany, July 2009

[16] L. Breiman, J. Friedman, R. Olshen, and C. Stone. Classification and Regression Trees. Wadsworth and Brooks, Monterey, CA, 1984.

[17] Gazetteer of Kolhapur District (2001)

[18] Giudici Paulo. Applied Data mining :Statistical Methods for business and industry, -ISBN 9812-53-178-5 (2003)

Abhijit A. Raorane received M.C.M. and M.C.A. degree in computer from Shivaji University, Kolhapur, Maharashtra in 1993 and 1998 respectively. He also completed his research work as a part of his M.phil. degree, naming "A quantitative approach for data mining and its application in a selected business organizations" in 2009. He is now pursuing his P. Hd. He is now working as Head of the department of computer in Vivekanand college Kolhapur.