EasyChair Preprint

№ 2458

A Study on the use of Agriculture Data Mining

Vyoma Srivastava, Vishnu Sharma and Dr.Abhay Srivastava

January 25, 2020

A Study on the use of Agri Data Mining

Review Article

Vyoma Srivastava
Dept. of Computer Science
Jaipur National University
Jaipur, India
vyoma.phdscholar@jnujaipur.ac.in

Dr. Vishnu Sharma
Dept. of Computer Science
Jaipur National University
Jaipur, India

Dr. Abhay Kumar Srivastava
Department in Business School
Amity Business School
Noida, India

ABSTRACT

In this paper, we study various review papers on use of data mining in the field of agriculture. Researches have used various data mining techniques, machine learning methods to real life agricultural datasets to very positive conclusions. Most of the papers concluded the results from application of data mining much more accurate compared to even experts. There researchers have used techniques like ID3 decision tree, Optimization algorithms, Bayesian classification, WEKA, Clustering techniques, MBA algorithms and many others. One if the biggest challenges faced by the researchers is the dataset itself. The dataset available in the field of agriculture is unclean. The datasets come with lot of missing values, duplicate entries and many other wide differences requiring multiple efforts in cleaning of data itself though many researches used this challenge as an opportunity as well to use data mining techniques to arrive to a usable dataset.

INTRODUCTION

Search is a very integral part of "Research". While Data Mining and Agriculture become the classic combo of the one of the oldest profession i.e. Agriculture and one of the best and latest

analytics i.e. Data Mining, any study can never be complete till we go through the wonderful done by researchers across the world. Various researchers have used Data Mining techniques, Machine learning techniques to make some great decision making tools which can help in delivering the best results for this sector. Agriculture is the backbone of any country and for India, it is the very basis of our Economy. Using Data Mining techniques for Agriculture can help in making tools for finding the exact relation between the parameters which effect the output and also in making predictive models which can help in making some good decisions for a better output. In this paper, we study various review papers on use of data mining in the field of agriculture. Most of the papers concluded the results from application of data mining much more accurate compared to even experts. There researchers have used techniques like ID3 decision tree, Optimization algorithms, Bayesian classification, WEKA, Clustering techniques, MBA algorithms and many others. One if the biggest challenges faced by the researchers is the dataset itself. The dataset available in the field of agriculture is unclean. The datasets come with lot of missing values, duplicate entries and many other wide differences requiring multiple efforts in cleaning of data itself though many researches used this challenge as an opportunity as well to use data mining techniques to arrive to a usable dataset.

KEYWORDS:

ID3 Decision Tree, Agri Data Mining, WEKA, MBA, Bayesian.

CHALLENGES ABOUT AGRI DATA:

Over years of experience, hit and trial, man-kind has developed very good understanding of the Agriculture and how to get the best out of it. With the knowledge of parameters effecting the output, the complexity has grown multi-fold now. It is now the time when the combination of various parameters are being studied and that's where Data Mining becomes most handy. Though, there are few serious roadblock that Agri Data suffers inherently:

- The Data Size itself: History and Historic Data for Agriculture is huge in size and same applies to various parameters which effect it. This also generates the pertinent need of using Data warehousing[6] techniques for this data.

- Agricultural data, unlike various other datasets, is non-numeric[4] mostly. It is more qualitative and not quantitative. Non-numericity of the data or parameters make the application of Data Mining techniques very complex.
- Cleaning[4] of the data e.g. duplication, missing values is a cumbersome process and can lead to gross errors in bringing up the right models, tools, predictions.

To overcome these challenges, this paper cover studies of various other global researches on Datamining, Agriculture and specially in their merging zone.

STUDIES ON AGRI DATA MINING:

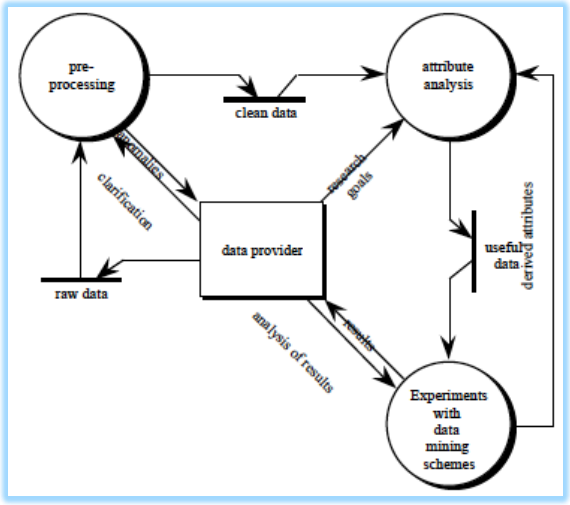
In the research done by Prof. M. S. Prasad, Babu N. V. Ramana Murty, S. V. N. L. Narayana in their paper on Expert System using AI and Machine Learning algorithms for Tomato Crop[1], extensive research has been to make a rule based expert system. This web based system uses ID3 Decision Tree & Optimization algorithms through its Jave front-end, SQL back-end to decide the disease and it's control measures, based on the user's inputs. User has to provide inputs to this expert system, which, based on these techniques, help the user with symptons, Prevention for the Tomato crop along with other details.

In a similar research in the Animal Husbandry sector, researchers Robert J. McQueen, Stephen R. Garner, Craig G. Nevill-Manning, Ian H. Witten, in their paper, named "Applying Machine Learning to Agricultural Data"[2], found that using WEKA along with machine learning gave a 97% accuracy in the farmer's strategy of culling less productive cows. This was a huge improvement as compared to experts driven rules which gave a 72% accuracy. The interactive tool developed in the study was just not able to derive new attributes using the combination of existing attributes but was also able to help inter-record the calculations like the change rate of time-series data.

With the advent of more and more satellite imaging and GIS data getting available, researchers have also moved to new line of research w.r.t. agriculture like in research done by Chi-Chung LAU, Kuo-Hsin Hsiao[3]. In this study, the paper studied the paddy distribution using multitemporal images, cadaster GIS, run through Bayesian posteriori probability classifier. Bayesian is a Soft classifier, which uses PTF (Probability to Feature) instead of DTF (Distance

to feature). Bayesian decision method with the GIS data, presented a very high level of accuracy, based on a simple procedure of computation.

Agricultural data is just not huge, it also needs lots of cleansing and a huge task of using non-numeric, qualitative data, including image data to be used for deriving the targets. Sally Jo Cunningham and Geoffrey Holmes, in their research[4], worked substantially on data cleaning like removing outliers, detecting error values, covering missing value to get data in ARFF format. This task was very difficult with over 60 image based attributes of total 68 attributes of mushrooms. Finally the attributes were brought down to derived or most directly effecting 6 to 8 attributes. While the paper also studied MBA, Market Basket Analysis theorem, but ultimately worked using Clustering Techniques and WEKA



Picture 1: Process Model for Machine Learning Application (data flow Diagram)

```
weka.classifiers.ZeroR
weka.classifiers.OneR
weka.classifiers.NaiveBayes
weka.classifiers.DecisionTable
weka.classifiers.Ibk
weka.classifiers.j48.J48
weka.classifiers.j48.PART
weka.classifiers.SMO
weka.classifiers.LinearRegression
weka.classifiers.m5.M5Prime
weka.classifiers.LWR
weka.classifiers.DecisionStump
```

Picture 2: The basic learning schemes of WEKA

Agriculture or crops are not untouched by the usage of Pesticides. A research, "Effect of Pesticides on Human Life through Visual Data Mining"[5], used Chernoff faces for applying clustering techniques on the agricultural data. Using cartoon like human faces to represent multivariate data is the bases of this method. The researches suggest that COF Clustering tool is useful just not in case of agriculture, but in case of any numeric data. In this study, 18 paramters were used to define the facial features like eyes, eyebrows, mouth etc.

In another research covering, "Geospatial Data Mining Techniques: Knowledge Discovery in Agricultural"[6], researcher Shital Hitesh Bhojani, worked on Geographical data mining, which included Clustering and Classification for Spatial segmentation, dependency, trend detection etc. The study covered different kind of query languages but mostly SQL, while used commands like DDL, DML for editing relational databased. OLAP was also studied to cover multi-dimensional analysis. It is generally used for conducting special analysis on the high-volume databases. The study also touches upon the Data Warehousing with respect to Agricultural data.

Soil and its properties are not different world, when it comes to Agriculture. Kris Verheyen, Dries Adriaens, Martin Hermy, Seppe Deckers[7], attempt to make a numeric soil classification system. The study uses non-heirarchial clustering method to apply partitioning method along with Wiki's criterion Ž L. The researches used GENSTAT 5r4.1 Ž1997. tool for clustering and then used SPSS 8.0 Ž1998 to calculate canonical discriminant functions for every horizon. They also used 'fuzzy k-means with extragrades' algorithmfor continuous classification.

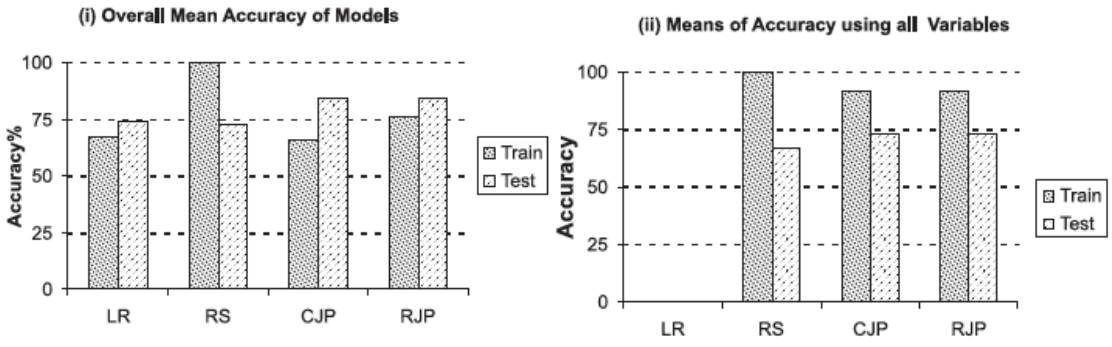
Bayesian Networks are one of the most commonly used Data Mining technique, when it comes to the field of agriculture. Yungang Zhu, Dayou Liu, Guifen Chen, Haiyang Jia, Helong Yu, in 2013[8], used Bayesian Network to workout the linkage between the crop diseases and symptons. The study used the concept of Markov blanket in Bayseian network and used it to develop incremental learning algorithms. Diagnosis model developed using the Bayesian network developed in a self-sustaining manner to adapt to dynamic changes in the environment.

In another study to do automatic detection of disease on Mango by S. B. Ullagaddi, Dr. S.Vishwanadha Raju[9], the paper studied over 21 researches done on various finding, detecting, recognizing methods of diseases on crops. Research also surveyed comprehensively various image processing techniques. The challenge was to find the most cost effective , robust and most accurate technique.

With the advent of Digital imaging and the devices and the technology becoming more and more within the reach, usage of of it in Agriculture and it's decision making, is becoming more and more common. George. E. Meyer, Joao Camargo Neto, David D. Jones, Timothy W. Hindman[10], used customized MATLAB version 6.1 script to do image processing using fuzzy logic tools. Study also covers use of Fuzzy C-Means function (FCM) & Gustafson–Kessel (GK) algorithm. The research developed clustering methods and unsupervised fuzzy color index to identify green plants from soil and residue from the Digital images.

While the agricultural industry suffered with the overuse of Pesticides, researchers, Ahsan Abdullah, Stephen Brobst, Ijaz Pervaiz, studied Dynamics of Pesticide Abuse through data Mining[11]. They use Clustering by Recursive Noise Removal technique, commonly called as RNR algorithm. This algorithm is able to do clustering of evidence spread across the data, which a Human mind can not comprehend. The study uses RNR algorithm to use the data of pest scouting, pesticide usage and metereological data to comprehend how excessive overuse of pesticides results in loss of yield.

Prediction of Crop diseases using Data Mining and Machine Learning has been of great interest for the researchers. In one such study, Rajni Jain, Sonajharia Minz and Ramasubramanian V[12], compared newer techniques like Rough Set based Decision tree (RDT), ID3 algorithm, CJP (Java of C4.5) against the traditional Logistic Regression (LR) methods). The combination of CJP and RJP (variant of RDT) were found most suitable in predicting PWM (P owdery Mildew of Mango) disease much better than LR or RS methods.



LR: Logistic Regression (SAS) RS: Rough Set Theory (Rosetta) CJP: C4.5 Java Implementation (Weka)
RJP: Variant of RDT (Rough Set based Decision Tree) (Rosetta, Weka, C++)

Table 6. Comparison of average of test accuracy (in per cent) of various algorithms

Variables used	LR	RS	CJP	RJP
Pairwise	75	74	83	84
All	*	62	74	74

Regression methodologies become more important for our research, especially in case of Agricultural data, which is both nominal and mostly non-numeric. While researching on considering 80 atributes of a cotton variety, researchers, Ahsan Abdullah, Rizwan Bulbul, Tahir Mehmood[13], found about 60 of them as non-numeric. They proposed mapping of nominal to numeric values based on the statistical properties of the crop is a tedious process. Regression and classification are most known for doing classification data mining.

In a much more scientific study by Namait Allah Y. Osman, M. K. Sadik, A. M. A. ABD El-Haleem, H. M. Eid and H. M. Salem[14], use Cropsyst Simulation model to predict Wheat crop growth w.r.t. water and nitrogen levels. The study concluded that after proper verification, crop modelling can help extrapolate on other crop prediction studies with the changing climate. Weather generator ClimGin v 4 and GSP techniques provide a great helping hand as well.

In another detailed study about SVMs (Support Vector Machines), researchers G. Camps-Valls, L. G´omez-Chova, J. Calpe-Maravilla, E. Soria-Olivas, J. D. Mart´ın-Guerrero, and J. Moreno[15], compared it with other machine learning method i.e. Co-Active Neural Fuzzy Inference System (CANFIS), Radial Basis Functions (RBF), multilayer perceptrons (MLP) etc. SVM was found pretty good while analyzing data in two classes only. The study uses multi classification schemes for all methods to avoid problem of "false positives". For, hyperspectral data classification, the study, proposed the use of kernel methods.

Data Warehousing can not remain untouched when it comes to Data Mining in Agricultural sector. Ahsan Abdullah, Stephen Brobst, Muhammad Umer, M. Farooq Khan[16], studied the process of setting up data warehouse with respect to Agri data. They also use OLAP tool over the data warehouse to get decision support. Data warehouse provides reliable and secured structure for storing huge amount of data over multiple categories, multiple years.

SUMMARY

Agriculture as a sector is very data intensive. A lot of data is generated with a good historic database. What effects the agriculture or factors effecting the agriculture itself have dependency as being one of the oldest profession, the sector has got good understanding of the factors which can effect the output. Most of the data is in the public domain and that's where Data Mining techniques come very handy to help the mankind to get the best results. Researches have done some amazing work in getting the best methods either in case of pesticide usage or soil information or animal culling or bettering the yield itself. The huge amount of data can help us devise ways and means to increase our agricultural productivity and efficiency. Better, scientific advises on crop based on forecasted demand. All this can inturn help directly on the forming of farmer policies, government policies and can bring back the agricultural economy of India.

Just to make a quick snapshot of our findings of these researches, we hereby present a tabulated format of the study of various researches that we have conducted in this paper. It is a total of 16 researches that we have covered in this paper:

Area of Application	Major Contributions	Tools Used	Year
Web Based Tomato expert Information System	Based on information from different species the expert system decides the disease and displays its control measure of disease	ID3 Decision Tree Algorith Optimization Algorithm	2010
Applying Machine Learning for culling less productive cows	The computer-generated rules outperformed the expert-derived rules. They gave the correct disease top ranking just over 97% of the time, compared to just under 72% for the expert derived rules	WEKA	1994
Bayesian Classification for Rice Paddy distributions	Interpreting paddy distributions using multitemporal imageries together with cadastre GIS by Bayesian posteriori probability classifier	Bayesian Classification	2002
Induce a classification system capable of sorting mushrooms into quality grades	The average accuracy of the models was compared favorably with that of the human inspectors and the level of agreement with the human experts was, on average, acceptable	WEKA	2000
Effect of Pesticides on Humans	Icon based technique which uses features in cartoon-like human faces, each representing variables in order to depict multivariate data	Chernoff faces COF Clustering Tool	2010
Geospatial Data Mining Techniques	Application of computational characteristic to the needs of agriculture data, as they are uncertain and fundamentally seasonal so use of data mining techniques be helpful in some aspect of agriculture	Knowledge Discovery from Databases OLAP - Online Analytical Processing	2013

Soil classification using morphological soil profile descriptions	The aptness of semi-quantitative morphological soil profile descriptions for numerical soil classification is explored	Wilk' s criterion Ž L GENSTAT 5r4.1 Ž1997 SPSS 8.0 Ž1998	2001
Representing the relationships among the symptoms and crop diseases	Tool can be used to diagnose many other diseases by extending the values of the ''Diseases'' variable and the ''Symptoms'' variables to certain other disease situations, especially for diseases whose diagnosis process is prone to be affected by temporal changes in a dynamic environment	Bayesian networks	2013
Automatic detection and diagnose of mango pathologies	21 researches / data mining techniques study done for Identification, Detection, Recognition methods	21 Data Mining Techniques researched	2016
Intensified fuzzy clusters for classifying plant, soil, and residue	Unsupervised fuzzy color index and clustering methods were developed and employed for identifying green plants from soil and residue	Gustafson–Kessel (GK) algorithm MATLAB Fuzzy clustering techniques	2004
Dynamics of Pesticide Abuse through Data Mining	Unsupervised clustering of pest scouting, pesticide usage and meteorological data to dig out the answers for a complex scenario where pesticide usage is increasing with a simultaneous decrease of yield.	Recursive Noise Removal (RNR) Clustering	2004
Forewarning Crop Diseases	Recommendation of CJP (C4.5 Java Implementation (Weka)) and RJP (Variant of Rough Set based Decision Tree) for prediction in Powdery Mildew of Mango (PWM) Disease as it performs better than LR and RS in terms of performance parameters	Logistic Regression Rough Set Theory (Rosetta) C4.5 Java Implementation (Weka) RJP Rosetta, Weka, C++	2009
Mapping non-numeric or nominal of Agricultural data by Data Mining Spectral Properties of Leaves	Mechanism of performing the mapping from nominal to numeric values (actually ranking) based on the transmittance as well as the statistical properties of the plant	Linear Regression Curve Classification	2005
Predicting Wheat growth under different water and nitrogen regimes	CropSyst model was able to track the aboveground biomass, grain yield, ET crop and N uptake progress throughout the season when compared with observed data from the filed experiments	CropSyst model	2009
Crop Classification Using Hyperspectral Data	Proposed the use of kernel methods for both hyperspectral data classification. SVM have revealed very efficient in different situations when a preprocessing stage is not possible	Multilayer Perceptrons (MLP) Radial Basis Functions (RBF) Co-Active Neural Fuzzy Inference Systems (CANFIS) Support Vector Machines (SVM)	1970
Agri Data Warehousing	A data warehouse provides a flexible yet efficient and reliable storage structure for vast amount of data while OLAP techniques provide mechanisms for ad hoc and in depth analysis of this data	RDMS OLAP	2004

REFERENCES

[1] Prof. M.S. Prasad Babu , N.V .Ramana Murty and S.V.N.L. Narayana ”A Web Based Tomato Crop Expert Information System Based On Artificial Intelligence Issn: 0975-9646 And Machine Learning Algorithms” International Journal of Computer Science and Information Technologies, Vol. 1 (1), 2010, 6-15..

[2] McQueen, R.J., Garner, S.R., Nevill-Manning, C.G. & Witten, I.H. (1994). “Applying machine learning to agricultural data”. (Working paper 94/13). Hamilton, New Zealand: University of Waikato, Department of Computer Science.

[3] Lau, Chi-Chung and Kuo-Hsin Hsiao. “BAYESIAN CLASSIFICATION FOR RICE PADDY INTERPRETATION.” (2002)..

[4] Cunningham, Sally & Holmes, Geoffrey. (2000). Developing Innovative Applications in Agriculture Using Data Mining.

[5] Imtiaz, Rabia & Khiyal, Malik & Khalil, Shahid & Khan, Aihab & Abdullah, Ahsan. (2010). Effect of Pesticides on Human Life through Visual Data Mining.. Journal of Theoretical and Applied Information Technology. 17. 104-109.

[6] Shital Hitesh Bhojani Geospatial Data Mining Techniques: Knowledge Discovery in Agricultural Indian Journal of Applied Research, Vol.III, Issue.I January 2013.

[7] Verheyen, Kris & Adriaens, Dries & Hermy, Martin & Deckers, Seppe. (2001). High-resolution continuous soil classification using morphological soil profile descriptions. Geoderma. 101. 31-48. 10.1016/S0016-7061(00)00088-4.

[8] Yungang Zhu, Dayou Liu, Guifen Chen, Haiyang Jia, Helong Yu, "Mathematical modeling for active and dynamic diagnosis of crop diseases based on Bayesian networks and incremental learning," Mathematical and Computer Modelling, Volume 58, Issues 3–4, August 2013, Pages 514-523. [Online]. Available: https://doi.org/10.1016/j.mcm.2011.10.072 [Accessed 18th Jan 2020].

[9] S. B. Ullagaddi, Dr. S.Vishwanadha Raju, “A review of techniques for Automatic detection and diagnose of mango pathologies”, International Journal of Engineering and Computer Science ISSN: 2319-7242 Volume 5 Issue 5 May 2016.

[10] Meyer, George & Camargo Neto, Joao & Jones, David & Hindman, Timothy. (2004). Intensified fuzzy clusters for classifying plant, soil, and residue regions of interest from color images. Computers and Electronics in Agriculture. 42. 161-180. 10.1016/j.compag.2003.08.002.

[11] Abdullah, Ahsan & Brobst, Stephen & Pervaiz, Ijaz. (2004). Learning Dynamics of Pesticide Abuse through Data Mining.. 151-156.

[12] Rajni Jain, Sonajharia Minz, Ramasubramanian V., "Machine learning for forewarning crop diseases", Journal of the Indian Society of Agricultural Statistics 2009 Vol.63 No.1 pp.97-107 ref.18. [Online]. Available: https://www.cabdirect.org/cabdirect/abstract/20103190359. [Accessed 18th Jan 2020].

[13] Ahsan Abdullah, Rizwan Bulbul, Tahir Mehmood, Mapping Nominal Values to Numbers by Data Mining Spectral Properties of Leaves, 3rd International Symposium on Intelligent Information Technology in Agriculture, October 14-16, 2005, Beijing, China.

[14] Osman, Namait & Sadik, M. & El-Haleem, A. & Eid, H. & Salem, Haythum. (2009). Using cropsyst simulation model to predict wheat crop growth under different water and nitrogen regimes in a middle Egypt type of environment (Giza region). J. Biol. Chem. Environ. Sci. 4.

[15] Camps-Valls, Gustau & Gómez-Chova, Luis & Calpe, Javier & Olivas, Emilio & Martín-Guerrero, José & Moreno, Jose. (1970). Support Vector Machines for Crop Classification Using Hyperspectral Data. Lecture Notes in Computer Science - LNCS. 2652. 134-141. 10.1007/978-3-540-44871-6_16.

[16] Abdullah, Ahsan & Brobst, Stephen & Umer, Muhammad & Khan, Muhammad. (2004). The Case for an Agri Data Warehouse: Enabling Analytical Exploration of Integrated Agricultural Data.. 139-144.

ABOUT AUTHOR

Vyoma Srivastava is an engineering professional with an extensive analytics and IT experience of over 6 years. With the rich experience of premier Organizations like Quality Council of India, NEISBUD and many others, Vyoma has gathered a very practical approach and experience towards the application of Data Mining Techniques in the real world, thus helping her research from the perspective of a very thorough researcher. Vyoma has completed her Bachelor of Technology from Vishveshwarya Institute of Technology, Gautam Buddha Technical University, India and has done her Master of Engineering from Indraprastha Engineering College, Mahamaya Technical University, India. She excelled in both BTech and MTech in 1st Division. Vyoma is currently pursuing her PhD. in Data Mining from Computer Sciences Department of Jaipur National University.

Dr. Vishnu Sharma, Ph.D is an Assistant Professor in School of Computer & Systems Sciences at Jaipur National University. HeHe has published 7 peer reviewed research articles, 2 book chapters. Prior to this, Dr. Sharma has held position as Head of Department at Computer Systems Sciences at Mewar University, Chittorgarh. He has served various institutions and having around 17 Years vast teaching and research experience in educational in University.

Dr. Abhay Kumar Srivastava has over twenty years of diversified teaching experience. He has taught to engineering students as well as management students in different reputed colleges of across India. He has published 17 research papers in various national and international journals of repute. He has guided students for PhD in the area of data mining. Prior joining academics, Dr. Srivastava has served industry in the area of Supply Chain and Vendor Management.