Empowering Agriculture Data Mining with FCA implied Artificial Immune Recognition System

Ms.K.Mythili
Head and Associate Professor,
Department of Computer Technology,
Hindusthan College of Arts and Science,
Coimbatore, India

Abstract —Agriculture is the essential overlay for the country growth. Deploying the data mining techniques in agriculture datasets reveal essential inferences to go further. This paper proposes Formal Concept Analysis (FCA) machine learning integrated with the artificial immune recognition system to reveal seasonal crop growth district wise in Tamilnadu. Researchers proposed many methods for Agriculture Data Mining using MLT. This paper proposed FCA based AIRS for agriculture data mining. Result significantly improves the performance. The algorithm is implemented using CONEXP and simulations carried out in WEKA. Results seems to be promising.

Keywords— Agriculture, MLT, FCA, AIRS, Data Mining

I. Introduction

Researchers proposed many methods for Agriculture Data Mining using MLT. This paper proposed FCA based AIRS for agriculture data mining. Result significantly improves the performance. The algorithm is implemented using CONEXP and simulations carried out in WEKA. Results seems to be promising when compared with the existing methods for the problem.

The problem considered in this paper is seasonal crop cultivation classification district wise in Tamilnadu. The agricultural crop year in India is from July to June. The Indian cropping season is classified into two main seasons-(i) Kharif and (ii) Rabi based on the monsoon. The kharif cropping season is from July –October during the south-west monsoon and the Rabi cropping season is from October-March (winter). The crops grown between March and June are summer crops. Pakistan and Bangladesh are two other countries that are using the term 'kharif' and 'rabi' to describe about their cropping patterns. The terms 'kharif' and 'rabi' originate from Arabic language where Kharif means autumn and Rabi means spring.

The kharif crops include rice, maize, sorghum, pearl millet/bajra, finger millet/ragi (cereals), arhar (pulses), soyabean, groundnut (oilseeds), cotton etc. The rabi crops include wheat, barley, oats (cereals), chickpea/gram (pulses), linseed, mustard (oilseeds) etc.

In an agricultural year (July-June), the Directorate of Economics & Statistics (DES), Department of Agriculture & Cooperation, Ministry of Agriculture releases four Advance Estimates followed by Final Estimates of production of major agricultural crops of the country. First Advance Estimates, released in September when Kharif sowing is generally over, cover only Kharif crops. Second Advance Estimates are released in February next year when rabi sowing is also over. These estimates covering Kharif as well as rabi crops take into account firmed up figures on kharif area coverage along with available data on crop cutting experiments for yield assessment of Kharif crops and tentative figures on area coverage of rabi crops. Third Advance Estimates incorporating revised data on area coverage for rabi crops and better yield estimates of Kharif crops are released in April-May.

Fourth Advance Estimates are released in July-August and by this time fully firmed up data on area as well as yield of Kharif crops and rabi crops are expected to be available with the States. As such, Fourth Advance Estimates are considered to be almost as good as Final Estimates released in next February along with Second Advance Estimates for the subsequent agricultural year. In order to allow sufficient time to States to take into account even the delayed information while finalizing area and yield estimates of various crops, the Final Estimates are released about seven months after the Fourth Advance Estimates and no revision in the State level data is accepted after release of Final Estimates by DES.

II. Literature Review

Machine learning deals with the erection and study of systems that learns from data. To maximize the crop yield, selection of the appropriate crop that will be sown plays a vital role. It depends on various factors like the type of soil and its composition, climate, geography of the region, crop yield, market prices etc. Machine learning provides many effective algorithms which can identify the input and output relationship in crop selection and yield prediction.

Techniques like Artificial neural networks, K-nearest neighbors and Decision Trees have carved a niche for themselves in the context of crop selection which is based on various factors. Crop selection based on the effect of natural calamities like famines has been done by Washington et al [1] based on machine learning. The use of artificial neural networks to choose the crops based on soil and climate has been shown by researchers [2]. A plant nutrient management system has been proposed based on machine learning methods to meet the needs of soil and maintain its fertility levels and hence improve the crop yield [3]. A crop selection method called CSM has been proposed [4] which helps in crop selection based on its yield prediction and other factors.

Machine learning methods have been used in the recent years for crop disease prediction and these efforts have been proved worthwhile. They revealed higher accuracy compared to the traditional statistical methods like regression analysis. These methods deal well with noisy and multi-faceted data [5][6][7]. Early crop disease detection and classification has been done using Support Vector Machines [8]. There are several factors like soil quality, crop rotation cycle, seed quality etc which can lead to poor health and diseases in crops. Machine learning algorithms effectively take into consideration all the possible factors, historic data as well as satellite/sensor data of fields to provide valuable disease classifiers. Disease detection using images of crop leaves has been implemented using pattern recognition branch of machine learning. It works by obtaining patterns from input data and separating them into classes of diseases [9].

From the review of the existing literature it is clear that every work tries to substantiate a small success in terms of the performance. This paper proposed FCA implied AIRS for Seasonal crop classification for agriculture domain. Result significantly improves the performance. The algorithm is implemented using CONEXP and WEKA. Results seems to be promising when compared with the existing methods for the problem.

III. Technical perspectives of working methodology

In artificial intelligence, artificial immune systems (AIS) are a class of computationally intelligent, rule-based machine learning systems inspired by the principles and processes of the vertebrate immune system. The algorithms are typically modeled after the immune system's characteristics of learning and memory for use in problem-solving.

The common techniques are inspired by specific immunological theories that explain the function and behavior of the mammalian adaptive immune system.

Clonal Selection Algorithm: A class of algorithms inspired by the clonal selection theory of acquired immunity that explains how B and T lymphocytes improve their response to antigens over time called affinity maturation. These algorithms focus on the Darwinian attributes of the theory where selection is inspired by the affinity of antigen-antibody interactions, reproduction is inspired by cell division, and variation is inspired by somatic hypermutation. Clonal selection algorithms are most commonly applied to optimization and pattern recognition domains, some of which resemble parallel hill climbing and the genetic algorithm without the recombination operator.[8]

Negative Selection Algorithm: Inspired by the positive and negative selection processes that occur during the maturation of T cells in the thymus called T cell tolerance. Negative selection refers to the identification and deletion (apoptosis) of self-reacting cells, that is T cells that may select for and attack self tissues. This class of algorithms are typically used for classification and pattern recognition problem domains where the problem space is modeled in the complement of available knowledge. For example, in the case of an anomaly detection domain the algorithm prepares a set of exemplar pattern detectors trained on normal (non-anomalous) patterns that model and detect unseen or anomalous patterns.[10]

Immune Network Algorithms: Algorithms inspired by the idiotypic network theory proposed by Niels Kaj Jerne that describes the regulation of the immune system by anti-idiotypic antibodies (antibodies that select for other antibodies). This class of algorithms focus on the network graph structures involved where antibodies (or antibody producing cells) represent the nodes and the training algorithm involves growing or pruning edges between the nodes based on affinity (similarity in the problems representation space). Immune network algorithms have been used in clustering, data visualization, control, and optimization domains, and share properties with artificial neural networks.[11]

Dendritic Cell Algorithms: The Dendritic Cell Algorithm (DCA) is an example of an immune inspired algorithm developed using a multi-scale approach. This algorithm is based on an abstract model of dendritic cells (DCs). The DCA is abstracted and implemented through a process of examining and modeling various aspects of DC function, from the molecular networks present within the cell to the behaviour exhibited by a population of cells as a whole. Within the DCA information is granulated at different layers, achieved through multi-scale processing.[11]

Considerations:

affinityThresholdScalar -- Affinity threshold scalar (ATS). Used with the system calculated affinity threshold to determine whether or not a candidate memory cell can replace the previous best matching memory cell. This occurs if the affinity between the candidate and the best match cell is < (AT * ATS).

arbInitialPoolSize -- Initial ARB cell pool size. Specifies the number of randomly selected training data instances used to seed the ARB cell pool. This paramter must be in the range [0, num training instances].

clonalRate -- Clonal rate. Used to determine the number of mutated clones to create of an ARB during the ARB refinement stage. Calculated as (stimulation * clonal rate).

debug -- If set to true, classifier may output additional info to the console.

hypermutationRate -- Hypermutation rate. Used with the clonal rate to determine the number of clones a best matching memory cell can create to then seed the ARB pool with. This is calculated as (stimulation * clonal rate * hypermutation rate).

knn -- k-Nearest Neighbour. Specifies the number of best match memory cells used during the classification stage to majority vote hte classification of unseen data instances.

memInitialPoolSize -- Initial memory cell pool size. Specifies the number of randomly selected training data instances used to seed the memory cell pool. This paramter must be in the range [0, num training instances].

mutationRate -- Mutation rate of cloned ARBs. Used to determine the degree of mutation a cloned ARB is subjected to. Must be in the range of [0,1].

numInstancesAffinityThreshold -- Total training instances to calculate affinity threshold (AT). Specifies the number of trainign data instances used to calculate the affinity threshold (AT) which is the mean affinity between data instances. A value of -1 indicates to use the entire training dataset.

seed -- Random number seed. The seed used in for random number generator.

stimulationValue -- Stimulation threshold. Used to determine when to stop refining the pool of ARBs for an antigen. This occurs when the mean normalised ARB stimulation value is >= the stimulation threshold. Must be in the range of [0,1].

totalResources -- Total allocatable resources. Specifies the maximum number of resources (B-cells) that can be allocated to ARBs in the ARB pool. Those ARBs with the weakest stimulation are removed from the pool until the total allocated resources is less than the maximum allowable resources.

IV. Experimental Results

The problem is implemented using CONEXP and WEKA under 32 bit Vista operating system. Experiments are conducted on a laptop with Intel(R) CoreTM 2 Duo 2.00 GHz CPU, and 3 GB of RAM. The values of parameters of the proposed algorithm are selected based on some preliminary trials. The selected parameters gave the best results concerning both the solution quality and the computational time needed to reach this solution.

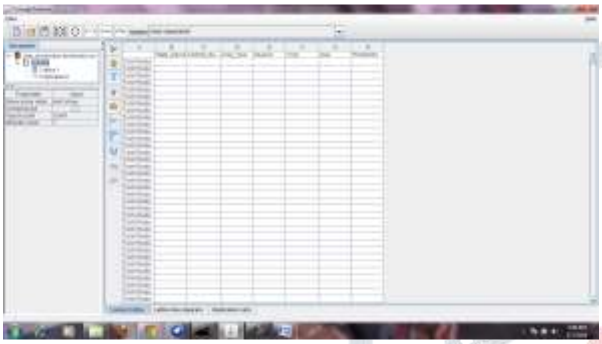


Figure 1 FCA for the Agriculture Seasonal Crop Prediction Preview

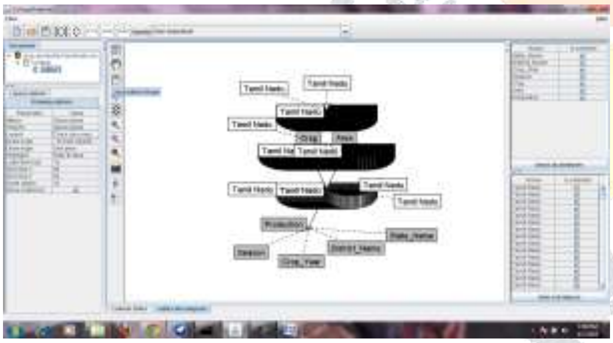


Figure 2 Concept Lattice

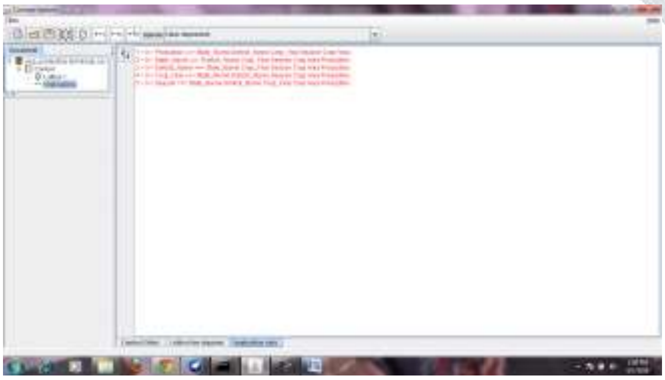


Figure 3 Implication Rules of the Problem

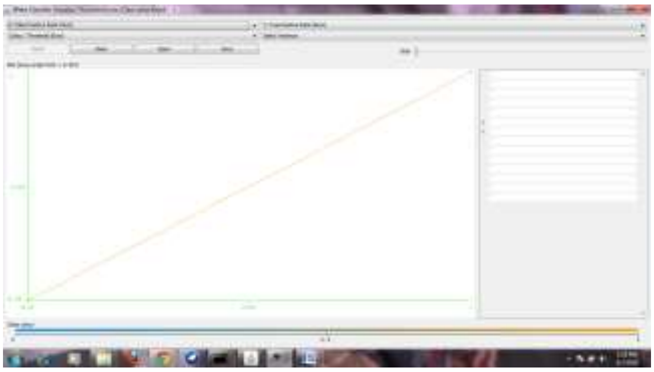


Figure 4 ROC of the FCA implied AIRS

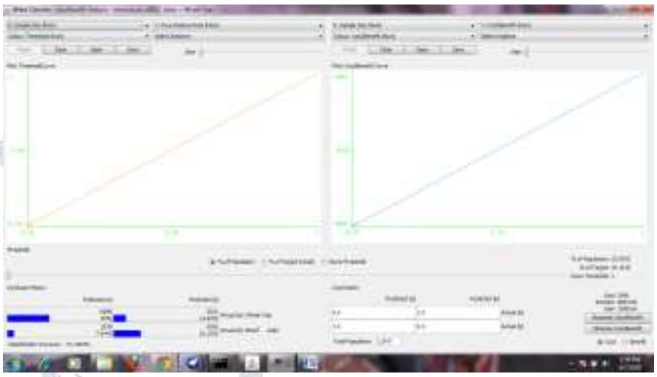


Figure 5 Cost-Benefit Analysis

Figure 6 Area vs. crop prediction Plot

V. Findings and Discussion

The development environment that is used was CONEXP with WEKA. Initially the data was fed into the FCA for the concept analysis and based on the findings of important features the dataset is preprocessed and again subject to the AIRS classifier. results show that this proposed approach performs well.

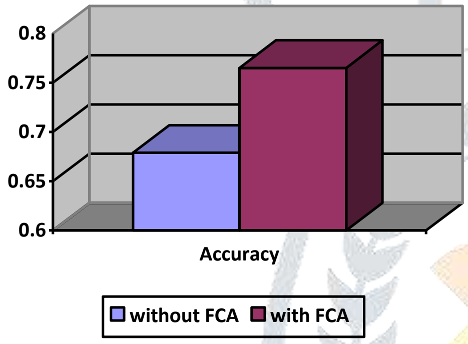
Time taken to build model: 25.85 seconds		
=== Stratified cross-validation ===		
=== Summary ===		
Correctly Classified Instances	10015	73.9278 %
Incorrectly Classified Instances	3532	26.0722 %
Kappa statistic	0.5047	
Mean absolute error	0.1738	
Root mean squared error	0.4169	
Relative absolute error	50.5743 %	
Root relative squared error	100.5763 %	
Total Number of Instances	13547	

=== Detailed Accuracy By Class ===

TP Rate	FP Rate	Precision	Recall	F-Measure	ROC Area	Class
0.789	0.255	0.626	0.789	0.698	0.767	Kharif
0.751	0.198	0.85	0.751	0.797	0.777	Whole Year
0.253	0.017	0.442	0.253	0.322	0.618	Rabi
0.739	0.209	0.751	0.739	0.739	0.765	Weighted Avg.

=== Confusion Matrix ===

a	b	c	<-- classified as
3747	944	60	a = Kharif
1864	6096	157	b = Whole Year
376	131	172	c = Rabi



VI. Conclusion

Agriculture is the essential overlay for the country growth. Deploying the data mining techniques in agriculture datasets reveal essential inferences to go further. This paper proposes Formal Concept Analysis (FCA) machine learning integrated with the artificial immune recognition system to reveal seasonal crop growth district wise in Tamilnadu. Researchers proposed many methods for Agriculture Data Mining using MLT. This paper proposed FCA based AIRS for agriculture data mining. Result significantly improves the performance. The algorithm is implemented using CONEXP and simulations carried out in WEKA. Results seems to be promising.

References

[1] Washington Okori, Joseph Obua,"Machine Learning ClassificationTechnique for Famine Prediction". Proceedings of the WorldCongress on Engineering 2011 Vol II WCE 2011, July 6 - 8, London, U.K, 2011.

[2] Miss.Snehal, S.Dahikar, Dr.Sandeep V.Rode, "Agricultural Crop Yield Prediction Using Artificial Neural Network Approach". International Journal of Innovative Research in Electrical, Electronic, Instrumentation and Control Engineering, Vol. 2, Issue 1, January 2014.

[3] Shivnath Ghosh, Santanu Koley, "Machine Learning for Soil Fertility and Plant Nutrient Management using Back Propagation Neural Networks". International Journal on Recent and Innovation Trends in Computing and Communication Volume: 2 Issue: 2, 292 297 ISSN: 2321-8169

[4] Kumar, Rakesh, M.p. Singh, Prabhat Kumar, and J.p. Singh. "Crop Selection Method to Maximize Crop Yield Rate Using Machine Learning Technique." 2015 International Conference on Smart Technologies and Management for Computing, Communication, Controls, Energy and Materials (ICSTM) (2015). Web

[5] Gutierrez D. D. (2015). Machine Learning and Data Science: An Introduction to Statistical Learning Methods with R. Basking Ridge, NJ: Technics Publications.

[6] Mehra, Lucky K. et al. "Predicting Pre-Planting Risk of Stagonospora Nodorum Blotch in Winter Wheat Using Machine Learning Models." Frontiers in Plant Science 7 (2016): 390. PMC. Web. 16 Apr. 2016.

[7] Development of a disease risk prediction model for downy mildew (Peronospora sparsa) in boysenberry. Kim KS, Beresford RM, Walter M Phytopathology. 2014 Jan; 104(1):50-6.

[8] Rumpf, T., A.-K. Mahlein, U. Steiner, E.-C. Oerke, H.-W. Dehne, and L. Plümer. "Early Detection and Classification of Plant Diseases with Support Vector Machines Based on Hyperspectral Reflectance." Computers and Electronics in Agriculture 74.1 (2010): 91-99. Web.

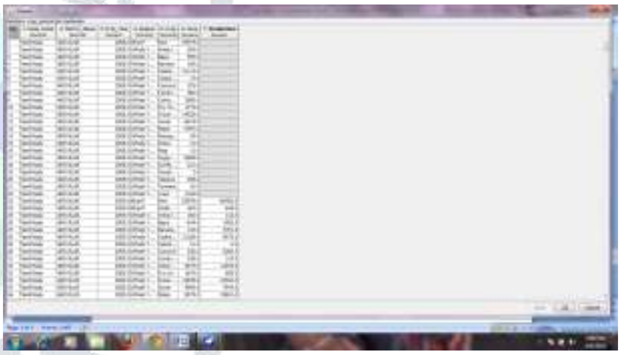
[9] M.P.Raj, P,R,Swaminarayan, J.R.Saini, D.K.Parmar "Application Of Pattern Recognition Algorithm In Agriculture : A Review" Int.J.Advanced Networking and Application, vol:6, issue:5, 2015

[10] Forrest, S.; Perelson, A.S.; Allen, L.; Cherukuri, R. (1994). "Self-nonself discrimination in a computer" (PDF). Proceedings of the 1994 IEEE Symposium on Research in Security and Privacy. Los Alamitos, CA. pp. 202–212.

[11] Jump up^ Timmis, J.; Neal, M.; Hunt, J. (2000). "An artificial immune system for data analysis". BioSystems. 55 (1): 143–150.doi:10.1016/S0303-2647(99)00092-1. PMID 10745118.

[12] Jump up^ Greensmith, J.; Aickelin, U. (2009). "Artificial Dendritic Cells: Multi-faceted Perspectives" (PDF). Human-Centric Information Processing Through Granular Modelling: 375–395.

Appendix – A
Sample Data



@relation 'crop_production tamilnadu'

@attribute State_Name {'Tamil Nadu'}

@attribute District_Name {ARIYALUR,COIMBATORE,CUDDALORE,DHARMAPURI,DINDIGUL,ERODE,KANCHIPURAM,KANNIYAKUMARI,KARUR,KRISHNAGIRI,MADURAI,NAGAPATTINAM,NAMAKKAL,PERAMBALUR,PUDUKKOTTAI,RAMANATHAPURAM,SALEM,SIVAGANGA,THANJAVUR,'THE NILGIRIS',THENI,THIRUVALLUR,THIRUVARUR,TIRUCHIRAPPALLI,TIRUNELVELI,TIRUPPUR,TIRUVANNAMALAI,TUTICORIN,VELLORE,VILLUPURAM,VIRUDHUNAGAR}

@attribute Crop_Year numeric

@attribute Season {'Kharif ','Whole Year ','Rabi '}

@attribute Crop {Rice,Arhar/Tur,Bajra,Banana,Cashewnut,'Castor seed','Coconut ',Coriander,Cotton(lint),'Dry chillies',Groundnut,Jowar,Maize,'Moong(Green Gram)',Onion,Ragi,Sugarcane,Sunflower,'Sweet potato',Tapioca,Turmeric,Urad,'Small millets',Sesamum,Horse-gram,Tobacco,'Black pepper',Cardamom,Gram,'Pulses total','Total foodgrain',Wheat,Sannhamp,Korra,Samai,'Guar seed','Other Cereals & Millets','Other Kharif pulses','Rapeseed &Mustard',Varagu,Arecanut,'Ash Gourd','Beans & Mutter(Vegetable)','Beet Root',Bhindi,'Bitter Gourd','Bottle Gourd',Brinjal,Cauliflower,'Citrus Fruit',Cucumber,'Drum Stick',Garlic,Grapes,'Jack Fruit',Lab-Lab,Mango,Orange,'Other Citrus Fruit','Other Fresh Fruits','Other Vegetables',Papaya,'Pome Fruit','Pome Granet',Redish,'Ribed Guard','Snak Guard',Tomato,'Water

```
Melon',Yam,Cabbage,'Pump                                         Kin','Dry
ginger',Soyabean,Potato,Carrot,Pineapple,Mesta,Apple,Peach,Pear,Plum
s,Turnip,Jute,Litchi,'Niger seed',Ber}
@attribute Area numeric
@attribute Production numeric

@data
'Tamil Nadu',ARIYALUR,2008,'Kharif     ',Rice,24574,?
'Tamil Nadu',ARIYALUR,2008,'Whole Year ',Arhar/Tur,209,?
'Tamil Nadu',ARIYALUR,2008,'Whole Year ',Bajra,565,?
'Tamil Nadu',ARIYALUR,2008,'Whole Year ',Banana,190,?
'Tamil Nadu',ARIYALUR,2008,'Whole Year ',Cashewnut,31113,?
'Tamil Nadu',ARIYALUR,2008,'Whole Year ','Castor seed',27,?
'Tamil Nadu',ARIYALUR,2008,'Whole Year ','Coconut ',335,?
'Tamil Nadu',ARIYALUR,2008,'Whole Year ',Coriander,460,?
    'Tamil Nadu',ARIYALUR,2008,'Whole Year ',Cotton(lint),3566,?
'Tamil Nadu',COIMBATORE,1999,'Kharif     ',Rice,16875,60494
'Tamil Nadu',COIMBATORE,1999,'Kharif     ',Sesamum,1100,521
'Tamil Nadu',COIMBATORE,1999,'Kharif     ',Sunflower,978,984
'Tamil Nadu',COIMBATORE,1999,'Kharif     ',Urad,5178,2682
'Tamil Nadu',COIMBATORE,1999,'Rabi       ',Gram,5030,3723
'Tamil Nadu',COIMBATORE,1999,'Rabi       ',Samai,134,73
'Tamil Nadu',COIMBATORE,1999,'Rabi       ',Wheat,9,4
'Tamil Nadu',COIMBATORE,1999,'Whole Year ',Banana,5619,183740
'Tamil Nadu',COIMBATORE,1999,'Whole Year ',Cashewnut,36,11
'Tamil Nadu',COIMBATORE,1999,'Whole Year ',Coriander,324,59
'Tamil        Nadu',COIMBATORE,1999,'Whole        Year        ','Dry
chillies',1418,1235
'Tamil        Nadu',COIMBATORE,1999,'Whole        Year        ','Guar
seed',11939,118374
  'Tamil Nadu',COIMBATORE,1999,'Whole Year ',Onion,2813,37188
```