ISSN No. 0976-5697

Volume 3, No. 7, Nov-Dec 2012

International Journal of Advanced Research in Computer Science

RESEARCH PAPER

Available Online at www.ijarcs.info

Impact of Datamining Techniques in Forecasting Plant Disease

R.Rajalakshmi*
Research Scholar, Department of Computer Science,
St. Peter's University, Chennai, Tamil Nadu, India.
rajielangor@yahoo.co.in

Dr.K.Thangadurai
Assistant Professor ,
Department of Computer Science,
Government Arts College (Autonomous),
Karur, Tamil Nadu, India.
ktramprasad04@yahoo.com

M.Uma
Research Scholar,
Dravidian Univeristy,
Kuppam, Andra Pradesh, India

Dr.M.Punithavalli
Director of M.C.A,
Sri Ramakrishna Engineering College,
Coimbatore, Tamilnadu, India.

Abstract: In this article a challenge has been made to analysis the explore studies on significance of data mining techniques in the field of agriculture. Couple of the techniques, such as decision algorithms ID3, the CHAID algorithm, C4.5, and Cluster analysis applied in the field of agriculture was presented. Data mining in application of agriculture is a relatively new approach for forecasting / predicting of fungal diseases of agricultural crop. This article explores the applications of data mining techniques in the field of farming and similar sciences.

Keywords: ID3 algorithm, CHAID algorithm, C4.5, Cluster analysis.

I. INTRODUCTION

Data mining methods for an early detection of plant diseases are vital for precision crop protection. Since the plant diseases causes major hazards in the agricultural locale, it leads to a great loses in the vegetation in various places due to some climatic variations. Agriculture and correlated activities constitute the only major component of India's gross domestic product, contributing nearly 25% of the total and nearly 60% of Indian population depends on this profession. Due to vagaries of climate factors the agricultural productivities in India are continuously decreasing over a decade[1]. Data mining tasks predict future trends and behaviors, allowing farmers to make proactive, knowledge-driven decisions. This information can be used as part of the farmer's decision-making process to help to forecast crop disease . The main contribution of this paper is a procedure for the early detection and differentiation of fungal disease based on decision algorithms.

II. AN OVERVIEW OF FUNGAL DISEASES

It is the major disease limiting fruit production in all countries where mangoes are grown, especially where high humidity prevails during the cropping season. The post-harvest phase is the most damaging and economically significant phase of the disease worldwide. It directly affects the marketable fruit rendering it worthless. This phase is directly linked to the field phase where initial infection usually starts on young twigs and leaves and spreads to the flowers, causing blossom blight and destroying the inflorescences and even preventing fruit set. Fungus infects almost all parts including floral panicles, twigs, leaves, and fruits of mature and immature trees. Next to this, stem-end rot is considered a major problem limiting the storage and shelf life of various fruits. This disease occurs only on ripe fruits where rotting usually begins at the stem end[2][3].

III. TECHNIQUES INVOLVED

Data mining is the extraction of hidden predictive information from large databases, is a powerful new technology. Data mining tools predict future trends and behaviors, allowing one to make proactive, knowledge-driven decisions.

Data mining Techniques includes different parameters namely:

a) Association
b) Sequence analysis
c) Classification
d) Clustering
e) Prediction

A. Classification:

Rule discovery in classification is an important data mining task since it generates a set of symbolic rules that describe each class or category in a natural way. In the context of classification, data tuples are also referred to as samples, examples, or objects. Data classification is a two-step process. In the first step, a model is built describing a predetermined set of data classes or concepts. The model is constructed by analyzing database tuples described by attributes. Each tuple is assumed to belong to a predefined class, as determined by one of the attributes, called the class label attribute.

a. Decision algorithm:

Decision tree is a popular classification model and is further used to predict the hidden patterns of data.

They are capable of handling both numerical as well as categorical attributes.

It can also be combined with other decision based techniques

a) Tree Induction[4][5]: For constructing a decision tree with N attributes the three parameters used are

- Data partition or database
- List of attributes (N)
- Splitting criteria for best split

b) Nodes in decision tree:

i. An internal node is a test on an attribute.
ii. A branch represents an outcome of the test, e.g., Color=red.
iii. A leaf node represents a class label or class label distribution.
iv. At each node, one attribute is chosen to split training examples into distinct classes as much as possible
v. A new case is classified by following a matching path to a leaf node.

c) **To choose a split attribute[6]:** At each node, available attributes are evaluated on the basis of separating the classes of the training examples. A Goodness function is used for this purpose. Typical goodness functions are

- information gain (ID3/C4.5)
- information gain ratio
- gini index

d) Criterian for attribute selection:

i. **Strategy:** choose an attribute that results in greatest information gain
ii. ***Information gain*** is a measure of how good an attribute is for predicting the class of the training data.
iii. Given a probability distribution, the info required to predict an event is the distribution's ***entropy***

e) Entropy:

i. Entropy provides the information-theoretic measure of the goodness of the split.
ii. For a given set of n events, and probability p = (p_1 , p_2 , …., pn),

$$\text{Entropy (p)} = - [p_1 \log(p_1)+p_2\log(p_2)+\ldots. p_n \log(p_n)]$$

iii. It is otherwise known as Info (p).

f) Gain Ratio:

Gain ratio is computed as:

$$\text{Gain (X, D)} = \text{info(D)} - \text{info(X,D)} \quad \text{--(2)}$$

i. **Where D stands for database and X stands for an event in D.**

b. Chi-square Automatic Interaction Detector (CHAID):

CHAID stands for **Chi-square Automatic Interaction Detector**[8][9]. The CHAID technique was created by Gordon V. Kass in 1980. CHAID is a technique of decision tree or regression tree, and is the best tool used to discover the relationship between variables. CHAID analysis determines how the variables best combine to explain the outcome in given dependent variables. In CHAID analysis, categorical or ordinal data is used. CHAID technique converts continuous data into ordinal data during analysis. The best use of CHAID analysis in contingency tables is to decide which variable is the maximum impotency in classification. CHAID analysis has the ability to build the non-binary classification tree as well. This is where more than two branches may go from the node. In the CHAID technique, we can visually see the relationship between the variable and the associated related factor with a tree. In most of the surveys, the answer is a categorical value instead of a continuous value. Discovering the relationship between the categorical values is a challenging job

The CHAID technique is the best tool to answer the survey research question. In CHAID analysis, we develop the decision tree or classification tree. The analysis starts with identifying the target variable or dependent variable. CHAID analysis splits the target in two or more categories that are called the initial nodes, then the, nodes are split using statistical algorithms. In CHAID analysis, there are two components: predictor variables and target variables. In CHAID analysis, predictor variables may be one, and should be ordinal, nominal or continuous in nature. Target variables should be one, and should be nominal, ordinal or continuous in nature. The CHAID technique is a better technique than regression analysis technique because in CHAID analysis, normal distribution is not required. Like Cluster Analysis

In CHAID analysis, categories of the independent variables are merged.

a) ***Merging:*** In CHAID analysis, a two way cross tabulation is formed between each dependent and independent variable and categories are merged within and across the independent variable. In CHAID analysis, Bonferroni adjusted p-value is calculated for merged crosstab.

b) ***CHAID algorithm:*** CHAID algorithm is the process of merging the categories based on their similarity in relation to their dependent variable. CHAID algorithm is a decision tree, which is constructed by splitting the subset of space into two or more nodes. This process is continued until the non- significant pair is not found[8].

a). Decision tree components in CHAID analysis:

In CHAID analysis, the following are the components of the decision tree:

i. ***Root node:*** Root node is the dependent variable or the target variable. For example, CHAID can be used if a bank wants to predict the credit card risk based upon information like age, income, number of credit cards, etc. In this example, the credit card risk dependent variable will be the root node.

ii. ***Parent's node:*** When algorithm splits the target variable into two or more categories. These categories are called parent's node or initial node. For the bank example, high, medium and low categories are the parent's nodes.

iii. ***Child node:*** Independent variable categories which come below the parent's categories in the CHAID analysis tree are called the child node.

iv. ***Terminal node:*** The last categories of the CHAID analysis tree are called the terminal node. In the CHAID analysis tree, the category that is a major influence on the dependent variable comes first and

the less important category comes last. Thus, it is called the terminal node.

c. C4.5 Algorithm:

C4.5 [10] is an algorithm used to generate a decision tree developed by Ross Quinlan. C4.5 is an extension of Quinlan's earlier ID3 algorithm. The decision trees generated by C4.5 can be used for classification, and for this reason, C4.5 is often referred to as a statistical classifier.

d. Cluster Analysis:

Cluster Analysis is a class of techniques that are used to classify objects or cases into relative groups called clusters. Cluster analysis is also called classification analysis or numerical taxonomy. In cluster analysis, there is no prior information about the group or cluster membership for any of the objects[11][12].

Cluster Analysis has been used in diagonising for various purposes. Segmentation of variables in cluster analysis is used on the basis of various factors of diseases. It can be used to identify homogeneous groups of variables.

Cluster analysis involves formulating a problem, selecting a distance measure, selecting a clustering procedure, deciding the number of clusters, interpreting the profile clusters and finally, assessing the validity of clustering.

The variables on which the cluster analysis is to be done should be selected by keeping past research in mind. It should also be selected by theory, the hypotheses being tested, and the judgment of the researcher. An appropriate measure of distance or similarity should be selected, the most commonly used measure is the Euclidean distance or its square.

Clustering procedures in cluster analysis may be hierarchical, non hierarchical, or a two step procedure. A hierarchical procedure in cluster analysis is characterized by the development of a tree like structure. A hierarchical procedure can be agglomerative or divisive. Agglomerative methods in cluster analysis consist of linkage methods, variance methods and centroid methods. Linkage methods in cluster analysis are comprised of single linkage, complete linkage and average linkage.

The non-hierarchical methods in cluster analysis are frequently referred to as K means clustering. The two-step procedure can automatically determine the optimal number of clusters by comparing the values of model choice criteria across different clustering solutions. The choice of clustering procedure and the choice of distance measure are interrelated. The relative sizes of clusters in cluster analysis should be meaningful. The clusters should be interpreted in terms of cluster centroids.

a) There are certain concepts and statistics associated with cluster analysis:

i. Agglomeration schedule in cluster analysis gives information on the objects or cases being combined at each stage of the hierarchical clustering process.
ii. Cluster Centroid is the mean values of a variable for all the cases or objects in a particular cluster.
iii. A dendogram is a graphical device for displaying cluster results.
iv. Distances between cluster centers in cluster analysis indicate how separated the individual pairs of clusters are. The clusters that are widely separated are distinct and therefore desirable.
v. Similarity/distance coefficient matrix in cluster analysis is a lower triangle matrix containing pair-wise distances between objects or cases.

IV. ROLE OF DATA MINING TECHNIQUES FOR PLANT DISEASES PREDICTION

Early detection and classification of plant diseases with Support Vector Machines based on hyper spectral reflectance discrimination between[13] healthy and inoculated plants as well as among specific diseases can be achieved by a support vector machine learning with a RBF function as kernel[14].

Data were collected and analysed from various aspects on weather conditions and the severity of fungal diseases. Disease severity and weather data were analysed using data mining models. In common with multiple regression models, moisture-related variables such as rain, leaf surface wetness and variables that influence moisture availability such as radiation and wind on the day of disease severity assessment or the day before assessment were the most important weather variables in all models. An essential step in CHAID prediction model construction is selecting the relevant features for classification. The purpose of feature selection techniques helps in reduction of computation time and enhances the predictive accuracy of the model. Chi-square is the common statistical test that measures divergence from the distribution expected if one assumes the feature occurrence is actually independent of the class value.

Feature Selection via Pearson chi-square (χ^2) test is a very commonly used method and it evaluates features individually by measuring their chi-squared statistic with respect to the classes. The present investigation used data mining as a tool with CHAID classification tree as a technique to design the disease prediction model. Filtered feature selection technique was used to select the best subset of variables on the basis of the values of chi-square measure.

V. CONCLUSION

The CHAID prediction model was useful to analyze the interrelation between variables that are used to predict the outcome on the causes of the fungal diseases. The features like slightly depressed grey-black areas in the skin on ripening fruit were the strongest indicators for the diseases causes in flowering and fruit set stages. This CHAID prediction model of disease status was constructed with predictor variables like temperature, climate and humidity, whereas the earlier models in reviews were constructed with limited class predictor variables.

Even though CHAID model handled small and unbalanced data set, it could be worked out effectively with better predictive accuracy. By applying Boosting and Bagging, which are two predominant techniques, the predictive accuracy would be further improved. Plenty of data is growing number of applications of data mining techniques in agriculture and an emergent amount of data that are currently available from lots of resources. There is several of work to be done on this up-and-coming in interesting research field. The multidisciplinary approach of integrating computational with agriculture will help in forecasting/managing agricultural crops effectively.

VI. REFERENCES

[1]. P. Revathi & R. Revathi, Knowledge Discovery in Diagnose of Crop, Diseases Using Machine Learning Techniques, International Journal of Engineering Science and Technology, ISSN: 0975-5462, Vol. 3, No. 9 September 2011,page No. 7187.

[2]. R. Pitkethley* and B. Conde,www.nt.gov.au/dpifm, Department of Primary Industry, Fisheries and Mines, Northern Territory Government, ISSN 0157-8243, 2007, Serial No. 604, Agdex No. 234/633.

[3]. S. Chakraborty, "Climate change: potential impact on plant diseases ", www.elsevier.com/locate/envpol,Environmental pollution 108 (2000), page No. 317-326.

[4]. H. A. Camdeviren, A. C. Yazici, Z. Akkus, R. Bugdayci, and M. A. Sungur, "Comparison of Logistic Regression Model and Classification Tree: An Application to Postpartum Depression Data", Expert Systems with Applications, Vol. 32, No. 4, 2007, pp. 987-994.

[5]. Introduction to data mining, author-Gupta.

[6]. I. H. Witten, and E. Frank, Data Mining – Practical Machine Learning Tools and Techniques (2nd ed.), San Francisco, CA: Morgan Kaufmann Publisher, 2005, www.cs.waikato.ac.nz/˜ml/weka/book.html.

[7]. Jiawei Han and Micheline Kamber, Classification and Prediction — Slides for Data Mining: Concepts and Techniques — Chapter 7 —, Intelligent Database Systems Research Lab, School of Computing Science, Simon Fraser University, Canada,.

[8]. M. Ramaswami and R. Bhaskaran, "A CHAID Based Performance Prediction Model in Educational Data Mining ",IJCSI International Journal of Computer Science Issues, Vol. 7, Issue 1, No. 1, January 2010 ISSN (Online): 1694-0784 ISSN (Print): 1694-0814.

[9]. J. Magidson, The Chaid approach to segmentation modeling: Chi-squared automatic interaction detection. In R. P. Bagozzi, editor, Advanced Methods of Marketing Research, pages 118-159. Blackwell Business, Cambridge Massechusetts, 1994.

[10]. J. R. Quinlan. Bagging, boosting, and c4.5. In Proc. 13th Natl. Conf. on Artificial Intelligence (AAAI'96), 725-730, Portland, OR, Aug. 1996.

[11]. S. Ganesh, "Data Mining: Should it be Included in the Statistics Curriculum?" The 6th international conference on teaching statistics (ICOTS-6), Cape Town, South Africa, 2002.

[12]. Y. Ma, B. Liu, C.K. Wong, P.S. Yu, and S.M. Lee, "Targeting the Right Students Using Data Mining", Proceedings of KDD, International Conference on Knowledge discovery and Data Mining, Boston, USA, 2000, pp. 457-464.

[13]. Juhua Luo, Wenjiang Huang, Jihua Wang And Chaoling Wei, The Crop Disease And Pest Warning And Prediction System, Computer And Computing Technologies In Agriculture II, Volume 2 IFIP Advances In Information And Communication Technology, 2009, Volume 294/2009, PN: 937-945.

[14]. Rumpf.T, Mahlein.A.K, Steiner.U, Oerke.E.C, Dehne.H.W and Plümer.L, Early Detection And Classification of Plant Diseases With Support Vector Machines Based On Hyper spectral Reflectance, Vol: 74, Issue 1, October 2010, PN: 91-99.