Mining the Classification Rules for Egyptian Rice Diseases

Mohammed El-Telbany[1], Mahmoud Warda[2], and Mahmoud El-Borahy[3]
[1]Computers and System Department, Electronics Research Institute, Egypt
[2]National Research Center, Egypt
[3]Mathematical Department, Alexandria University, Egypt

Abstract: *Applications of learning algorithms in knowledge discovery are promising and relevant area of research. It is offering new possibilities and benefits in real-world applications, helping us understand better mechanisms of our own methods of knowledge acquisition. Decision trees is one of learning algorithms which posses certain advantages that make it suitable for discovering the classification rule for data mining applications. This paper, intended to discover classification rules for the Egyptian rice diseases using the C4.5 decision trees algorithm. Experiments presenting a preliminary result to demonstrate the capability of C4.5 mine accurate classification rules suitable for diagnosis the disease.*

Keywords: *Data mining, classification, decision trees, neural networks*, *expert systems*.

Received April 5, 2005; accepted June 26, 2005

1. Introduction

The knowledge sector of modern economies has grown extremely rapidly, and the value of knowledge is now reckoned to be a major economic force. The Egyptian Ministry of Agriculture (MOA) has decided to investigate the usage of expert systems technology to respond to this need. The Central Laboratory for Agricultural Expert Systems (CLAES) has been established in 1991 to conduct research in the area of expert systems in agriculture. The transfer of experts from consultants and scientists to agriculturists, extension workers and farmers represent a bottleneck for the development of agriculture on the national level. The experts tend to execute the reasoning process with a series of rules. The rules are abstracted from basic principles and cases they have experienced in their fields. The knowledge acquisition process in expert system design is the most valuable asset in output accuracy. However, much of this asset is either hidden in databases as information that has not yet been tested out and made explicit, or locked up in individual principals and employees. An emerging field: Knowledge Discovery in Databases (KDD), extends the scope of knowledge engineering research to extracting knowledge from data records collected for routine use. The stepwise process in KDD includes as shown in Figure 1: Defining goal(s); data collecting, cleaning, and reduction; data analyzing and hypothesis selecting; data mining; interpreting mined pattern(s); validating and acting upon discovered knowledge. Among them, data mining is the key step that focuses on applying some specific machine learning algorithms that can discover previously unknown regularities and trends in databases and also helps people to explicate and codify their knowledge and expertise. It therefore has great potential to contribute to the economy and sector and decision support process in Egypt in many different ways.

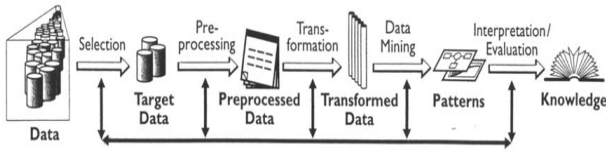

Figure 1. KDD process chain (adopted from [4]).

Decision trees, one of the machine learning algorithms, are powerful and popular tools for classification and prediction. They can be used for discovering *consistent*, *accurate*, *comprehensible* and *predictive* classification rules. It is a greedy search techniques provided consistent, relatively accurate and understandable rules by using the training data set to construct a decision tree or collection of rules, which discriminates examples. The ID3 and C4.5 algorithms are the popular machine learning methods that produce a decision tree from examples [8, 9]. They use an *information theoretic heuristic* to determine which attribute should be tested at each node, looking at all members of the training set which reach that node and selecting the attribute that most reduces the entropy of the positive/negative decision. By the nature of the algorithm, correct performance is guaranteed on the training set. The decision trees are attractive due to the fact that, in contrast to other machine learning techniques such as neural networks, they represent

rules. Rules can readily be expressed so that humans can understand them or even directly used in a database access language like SQL so that records falling into a particular category may be retrieved.

Applying the KDD principles in agriculture fields requires several practical difficulties to be evident:

1. One may want to include cases that could cover all possible situations in the problem domain before starting the decision tree buildup. However, the decision about which cases should be in the training set is always debatable.
2. There is no guarantee that the cases collected noisy free. There is no easy way to judge which case is "polluted", except by manual inspection by experts. When dealing with large amounts of data, manual inspection is not practical. Therefore, a decision tree generated from the contaminated data set would not truly reflect the domain knowledge.

In this paper, an expert guided decision tree construction strategy using C4.5 decision tree algorithm is proposed for automatic rule discovery to help farmers in organizing their reasoning processes from the evidence of collected cases. There are many advantages for using decision trees in classification. The *first benefit* of trees over many other classifiers such as neural network is: *Interpretability*. It is a straightforward matter to render the information in such a tree as logical expressions. Such interpretability has two manifestations. *First*, we can easily interpret the decision for any particular test patterns as the conjunction of decisions a long the path to its corresponding leaf node. *Second*, we can occasionally get clear interpretations of the categories themselves, by creating logical descriptions using conjunctions and disjunction. *Another benefit* of trees is that they lead to rapid classification. *The third benefit*, we note that trees provide a natural way to incorporate prior knowledge from human experts. There are other benefits of decision trees; decision trees are perfect tools for making corporate or financial decisions where a lot of complex information has to be considered. Decision trees also help in forming a balanced, accurate picture of the risks and rewards that can result from a particular decision.

The evaluation of C4.5 algorithm performance over the Egyptian rice crop diseases as real data is an important matter, especially; the Egyptian rice diseases cause losses that estimated by 15% from the yield, malformation of the leaves or dwarfing of the plants. Discovering and control of disease is the main aim and have a large effect for increasing density of faddan and increasing gain for farmer then increasing the national income. Actually, the original contribution of this research paper is to provide the usage of decision tree for Egyptian rice diseases classification. The paper is organized as follows. Section 2 describes the related work. Section 3 examines the data used to assess the Egyptian rice crop diseases. Section 4 represents the methodology and the classification results. Section 5 concludes the paper.

2. Related Work

Agriculture is sometimes referenced as a weak theory domain, in which a large part of the reasoning knowledge is vague and described differently by various experts. Though the entry points in reviewing a case among different experts could not be the same, the conclusion should be similar. The precedence factors related to outcome from different expert's viewpoint also varies. So, the Central Laboratory for Agricultural Expert Systems (CLAES) has been established in 1991 to conduct research in the area of expert systems in Egyptian agriculture. Four expert systems for cucumber [11], tomato [3], orange [13], and lime [10] have been built using the developed methodology and one expert system for wheat [7], have been built using the Generic Task (GT) methodology. In effect, each expert system consists of a set of subsystems covering different areas of crop management namely: Variety selection, planting, irrigation, fertilization, pest control and others. However, the machine learning techniques that are used frequently to address problems in different domains do not use the agriculture.

3. Data Domain

Rice crop is one of the major cereal crops in Egypt, due to its importance as the main food and for exporting. The rice cultivation area in Egypt is approximately (1.529 million feddans) in 2002, which is about 17% of Egypt's total cultivated area. Successful Egyptian rice production requires for growing a summer season (May to August) of 120 to 150 days according to the type of varieties as Giza177 needs 125 day and Sakha104 needs 135 day. Climate for the Egyptian rice is that daily temperature maximum = 30-35°, and minimum = 18-22°; humidity = 55%-65%; wind speed = 1-2 m. Egypt increase productivity through a well organized rice research program, which was established in the early eighties. In the last decade, intensive efforts have been devoted to improve rice production. Consequently, the national average yields of rice increased by 65% i. e., from (2.4 t/fed.) during the lowest period 1984-1986 to (3.95 t/fed.) in 2002 [1, 2, 5, 12]. Many affecting diseases infect the Egyptian rice crop; some diseases are considered more important than others. In this study, we focus into the most important diseases, which are five; blight, brown spot, false smut, white tip nematode and stem rot sequence. We have a total of 206 sample, somewhat arbitrarily took the 138 data points for training, and the reset points for validation and test.

4. Discovering the Classification Rules for Egyptian Rice Diseases

4.1. Constructing Decision Trees

Most algorithms that have been developed for learning decision trees are variations on a core algorithm that employs a top down, greedy search through the space of possible decision trees. Decision tree programs construct a decision tree from a set of training cases. The central focus of the decision tree growing algorithm is selecting which attribute to test at each node in the tree. For the selection of the attribute with the most inhomogeneous class distribution the algorithm uses the concept of *entropy* [9].

$$Entropy(S) = \sum_{i=1}^{c} - p_i \log_2 p_i \qquad (1)$$

Where *pi* is the proportion of *S* belonging to class *i*. However, a good quantitative measure of the worth of an attribute is a statistical property called *information gain* that measures how well a given attribute separates the training examples according to their target classification. The information gain, *Gain* (*S, A*) of an attribute *A*, relative to a collection of examples S, is defined as [9]:

$$Gain(S, A) = Entropy(S) - \sum_{v \in V(A)} Entropy \frac{|S_v|}{S}(S_v) \qquad (2)$$

Where *V* (*A*) is the set of all possible values for attribute *A*, and *Sv* is the subset of *S* for which attribute A has value *v* (i. e., *Sv* = {s S | *A* (*s*) = *v*}). The decision tree construction strategy partitioned all the learning cases come from the experts along the decision paths in the tree and allows farmer to enter a new set of an ordered list of attributes, which they believed to be significant in discriminating and diagnosis, the rice disease.

4.2. Preliminary Results

The C4.5 decision tree learning algorithms are tested for discovering a classification rule for predicting the diseases of rice crop using WEKA which is a library of machine learning algorithms in Java [6]. The parameters for those classifiers were chosen to be the default one used by WEKA. Each case in the data set is described by seven attributes. The attribute and possible values are listed in Table 1. Ten *cross-validation* bootstraps, each with 138 (66%) training cases and 68 (34%) testing cases, were used for the performance evaluation. In order to compare the performance of C4.5, which is an orthogonal classifier, we also conducted experiments using Neural Networks (NN) results. In this paper, a three-layer, fully connected feed-forward network as shown in Figure 2 is used. In this type of network, the first layer is composed of the input variables, the second layer is composed of hidden nodes, and the last layer is composed of the output nodes. The neural network architecture is 52-33-5. The network is trained using back propagation algorithm with learning rate is 0.3, and momentum is 0.2 for 500 iterations.

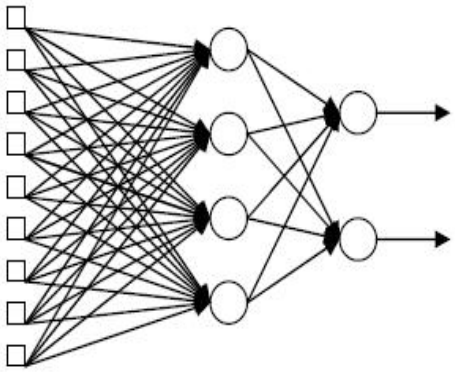

Figure 2. Network configuration.

The mean accuracy of results of training and test data is presented in Table 2. It can be shown that the decision tree is perform better than multi-layer neural network which is non-linear classifier. This results due to the huge numbers of attributes' values in the training data which degrades the performance of neural networks and the capability of decision tree in reducing *overfitting error* using pruning techniques.

Table 1. Possible value for each attribute from the rice database.

Attribute	Possible value
Variety	Giza171, giza177, giza178, sakha101, sakha102, sakha103, sakha104
Age	Real value
Part	Leaves, leaves spot, nodes, panicles, grains, plant, flag leaves, leaf sheath, stem
Appearance	Spots, elongate, spindle, empty, circular, oval, fungal, spore balls, twisted, wrinkled, dray, short, few branches, barren, small, deformed, seam, few stems, stunted, stones, rot, empty seeding
Color	Gray, olive, brown, brownish, whitish, yellow, green, orange, greenish black, white, pale, blackish, black
Temperature	Real values
Disease	Blight, brown spot, false smut, white tipe, stem rot

Table 2. Classification accuracy: A comparison.

	NN algorithm		C4.5 algorithm	
	Training	Test	Training	Test
Accuracy	97.18%	96.4%	98.55%	97.25%

Two of the resulting rule set that are discovered by decision tree is listed in Table 3.

Table 3. Some discovered classification rules.

If appearance = spot *and* color = olive *Then* disease = blight
If appearance = spot *and* color = brown *and* age <= 55 *Then* disease =brown-spot

5. Conclusion

Machine learning is a burgeoning new technology with a wide range of potential applications. This paper represents a first step toward redressing this imbalance by grounding machine learning techniques in important agriculture applications by explores the synergy of decision trees in discovering classification rules from the data of the rice disease as the key crop in the Egyptian. Specially, the large numbers of expert systems that have been developed for agricultural problems worldwide provide further evidence that formalizing knowledge can benefit agriculture. It seems likely that machine learning can contribute to the economy on several different fronts. The decision tree algorithm (i. e., C4.5) provides many *benefit*s of trees over many other classifiers such as neural network. The most important benefit is *interpretability*. Moreover, the C4.5 can effectively create comprehensive tree with greater predictive power and able to get a prediction error about 1.5% on data of test set. The enhancement in classification results due to the capability of decision tree in reducing *overfitting error* using pruning techniques and handling the huge numbers of attributes' values.

Acknowledgments

We are indebted to Central Laboratory for Agricultural Expert Systems staff for fruitful discussion and for providing us with their experiences.

References

[1] Ahmed E. M., "Studies on Control of Rice Blight Disease," *Master Thesis*, Faculty of Agriculture, Zagazig University, 2003.

[2] Central Laboratory for Agricultural Expert Systems, "Rice Expert System Software," Version 1.2, 2002.

[3] El-Shishtawy T., Wahab H., El-Dessouki A., and El-Azhary E., "From Dependence Networks to KADS: Implementation Issues," *in Proceedings of the 2nd Workshop on Artificial Intelligence in Agriculture (IFAC/IFIP/ EnrAgEng)*, Netherlands, 1995.

[4] Fayyad U., Piatetsky-Shapiro G., and Smyth P., "From Data Mining to Knowledge Discovery in Databases," *AI Magazine*, vol. 17, no. 3, pp. 37-54, 1996.

[5] Hesen H. A., *Plant Disease Nematode*, Faculty of Agriculture, Cairo University 2001.

[6] Holmes G., "WEKA: A Machine Learning Workbench," *in Proceedings 2nd Australia and New Zealand Conference on Intelligent Information Systems*, Brisbane, Australia, pp. 357-361, 1994.

[7] Kamel A., Schroeder K., and Sticklen J., "An Integrated Wheat Crop Management System Based on Generic Task Knowledge Based Systems and CERES Numerical Simulation," *AI Applications*, vol. 9, no. 1, 1995.

[8] Quinlan, R., *C4.5: Programs for Machine Learning,* Morgan Kaufmann, 1993.

[9] Quinlan R., "Induction of Decision Trees," *Machine Learning*, vol. 1, no. 1, pp. 81-106, 1986.

[10] Rafea M. and Rafea A., "LIMEX: An Integrated Multimedia Expert System," *in Proceedings of the International Conference on Multimedia Modeling,* Singapore, 1997.

[11] Rafea A., El-Azhari S., Ibrahim I., Edres S., and Mahmoud M., "Experience with the Development and Deployment of Expert Systems in Agriculture," *in Proceedings of IAAI-95 Conference*, Montreal-Canada, 1995.

[12] Sakha Research Center, *The Results of Rice Program for Rice Research and Development*, Ministry of Agriculture, Egypt, 2002.

[13] Salah A., Hassan H., Tawfik K., Ibrahim I., and Farahat H., "CITEX: An Expert System for Citrus Crop Management," *in Proceedings of ESADW'93*, *MOALR,* Cairo, Egypt, 1993.

Mohammed El-Telbany received his BS degree in computer engineering and science from the University of Minufia in 1991, and his MSc and PhD degree in Computer Engineering from Electronics and Communication Department, Cairo University, in 1997 and 2003, respectively. He has been a researcher at the Electronics Research Institute since 1993 till 2004. From 1998 until 1999, he has worked at the ESA of the European Space Research Institute (ESRIN), in Italy. He was working at the Faculty of Engineering, Amman Al Ahliyya University, Jordan. Currently he is an assistant professor at the Electronics Research Institute, Egypt. He has been involved in the field of autonomous mobile robots and machine leaning. His research interests include work in robotics and reinforcement learning, data mining, and swarm intelligence.

Mahmoud Warda received his BSc in science in 2000. He has been an assistance researcher at the National Research Center since 2001. His research interest includes using data mining techniques in agriculture.

Mahmoud El-Borahy received his BSc degree in mathematics from the University of Alexandria in 1961, and his PhD degree in mathematics from Moscow University, in 1968. Currently, he is a professor at the Department of Mathematics, Faculty of Science, Alexandria University, Egypt. His research interest includes deferential equations and operation researches.