Mining the Classification Rules: The Egyptian Rice Diseases as Case Study

MOHAMMED E. EL-TELBANY[1*]
[1]Computer Engineering Department
Faculty of Engineering
Amman University
JORDAN

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, and expert systems.

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 represents 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 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 includes: 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.

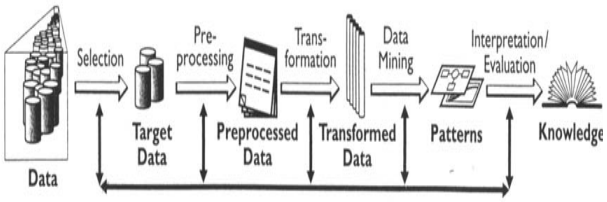

Fig. 1: KDD processes chain (adopted from [4]).

Decision trees one of the machine learning algorithms are powerful and popular tools for classification and prediction. It 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 [7-8]. 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

*Assist. Prof. Computers and Systems Dept., Electronics Research Institute, El-Tahrir St. Dokki, EGYPT.

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 any 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 organize their reasoning processes from the evidence of collected cases. 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 estimate 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 organized as follows:. Section 2 examines the data used to assess the Egyptian rice crop diseases. Section 3 represents the methodology and the classification results. Section 4 describes the related work Section 5 concludes the paper.

2 DATA DOMAIN

Rice Crop is one of the major cereal crops in Egypt 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 Aug.) 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 that Temperature is daily maximum = 30-35°, and minimum = 18-22°; Humidity (55%-65%); Wind speed (1-2 m). Egypt must increase productivity through a well-organized rice research program 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][11]. 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 206 sample, somewhat arbitrarily took the 138 data points for training, and the reset points for validation and test.

3 DISCOVERING THE CLASSIFICATION RULES FOR EGYPTIAN RICE DISEASES

3.1 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* [8].

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

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 [8]

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

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 comes 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.

3.2 Preliminary Results

The C4.5 decision tree learning algorithms is tested for discovering a classification rule for predicting the diseases of rice crop using WEKA which a library of Machine Learning Algorithms in Java [13]. 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. 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 and values in the training data which degrades the performance of neural networks.

Table 1. The 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. The Classification Accuracy: A Comparison.

	NN algorithm		C4.5 algorithm	
	Training	Test	Training	Test
Accuracy	97.18%	96.4%	98.52%	97.10%

4 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. For that, 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. They developed four expert systems, for cucumber [9], tomato [3], orange [12], and lime [10], have been built using the developed methodology and one expert system, for wheat [6], 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 be used frequently to address problems in different domains is not used the agriculture.

5 Conclusion and Discussion

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. 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.

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., 2003, “Studies on Control of Rice Blight Disease,” Master Thesis, Faculty of Agriculture, Zagazig University.

[2] Central Laboratory for Agricultural Expert Systems, 2002, “Rice Expert System Software,” Ver. 1.2.

[3] El-Shishtawy, T., Wahab, A., El-Dessouki, El Azhary, S., 1995, “From Dependence Networks to KADS: Implementation Issues", 2nd IFAC/IFIP/EnrAgEng workshop on Artificial Intelligence in Agriculture. The Netherlands.

[4] Fayyad, U., G. Piatetsky-Shapiro, and P. Smyth, 1996, “From Data Mining to knowledge Discovery in Databases,” AI Magazine, pp.37-54.

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

[6] Kamel, A.; Schroeder, K.; Sticklen, J.; Rafea, A.; Salah, A.; Schulthess, U.; Ward, R.; Ritchie, J., 1995, “An Integrated Wheat Crop Management System Based on Generic Task Knowledge Based Systems and CERES Numerical Simulation,” AI Applications, 9(1).

[7] Quinlan, R., 1993, *C4.5: Programs for machine learning*. Morgan Kaufmann.

[8] Quinlan, R., 1986, “Induction of decision trees,” Machine Learning 1 (1), pp. 81–106.

[9] Rafea,A., El-Azhari, S., Ibrahim, I., Edrees, S., ahmoud, M., 1995, “Experience with the Development and Deployment of Expert Systems in Agriculture,” Proceedings of IAAI-95 conference, Montreal-Canada.

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

[11] Sakha Research Center, 2002,”The Results of Rice Program for Rice research and Development,” Ministry of Agriculture, Egypt.

[12] Salah, A.,Hassan,H., Tawfik, K., Ibrahim,I., Farahat, H., 1993, “CITEX: An Expert System for Citrus Crop Management", (ESADW-93), MOALR, Cairo - Egypt.

[13] Holmes, G.; Donkin, A.; Witten, I.H., 1994, “WEKA: a machine learning workbench,” In: Proceedings Second Australia and New Zealand Conference on Intelligent Information Systems, Brisbane, Australia 1994, 357-361