An Empirical Comparison of Tree-Based Learning Algorithms: An Egyptian Rice Diseases Classification Case Study

Mohammed E. El-Telbany
Computers and Systems Department
Electronics Research Institute Cairo,
Egypt

Mahmoud Warda
Computers Department
National Research Center
Cairo, Egypt

***Abstract*—Applications of learning algorithms in knowledge discovery are promising and relevant area of research. The classification algorithms of data mining have been successfully applied in the recent years to predict Egyptian rice diseases. Various classification algorithms can be applied on such data to devise methods that can predict the occurrence of diseases. However, the accuracy of such techniques differ according to the learning and classification rule used. Identifying the best classification algorithm among all available is a challenging task. In this study, a comprehensive comparative analysis of a tree-based different classification algorithms and their performance has been evaluated by using Egyptian rice diseases data set. The experimental results demonstrate that the performance of each classifier and the results indicate that the decision tree gave the best results.**

***Keywords*—Data Mining, Classification, Decision Trees, Bayesian Network, Random Forest, Rice Diseases.**

I. Introduction

Processing the huge data and retrieving meaningful information from it is a difficult task. Data mining is a wonderful tool for handling this task. The major components of the architecture for a typical data mining system are shown in Fig 1. The term Data Mining, also known as Knowledge Discovery in Databases (KDD) refers to the non trivial extraction of implicit, previously unknown and potentially useful information from data in databases [1]. They are several different data mining techniques such as *clustering*, *association*, *anomaly detection* and *classification* [2]. The classification process has been identified as an important problem in the emerging field of data mining as they try to find meaningful ways to interpret data sets. The goal of classification is to correctly predict the value of a designated discrete class variable, given a vector of predictors or attributes by produces a mapping from the input space to the space of target attributes [3]. There are various classification techniques each technique has its pros and cons. Recently, Fernandez-Delgado *et al.* [4] evaluate 179 classifiers arising from 17 families (e.g. statistics, symbolic artificial intelligence and data mining, connectionist approaches, and others are ensembles). The classifiers show strong variations in their results among data sets, the average accuracy might be of limited significance if a reduced collection of data sets is used [4]. For example, the largest merit of neural networks (NN) methods is that they are general: they can deal

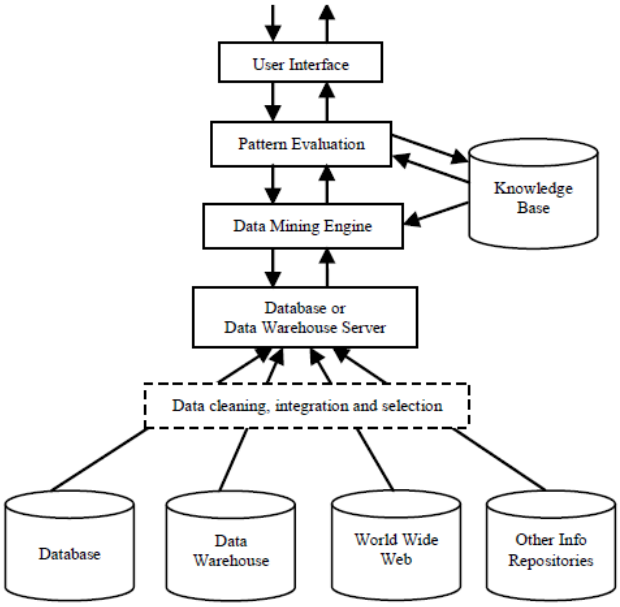

Fig. 1: Architecture of a Typical Data Mining System [1]

with problems with high dimensions and even with complex distributions of objects in the n-dimensional parameter space. However, the relative importance of potential input variables, long training process, and interpretative difficulties have often been criticized. Although the *support vector machine* (SVM) has a high performance in classification problems [5], the rules obtained by SVM algorithm are hard to understand directly and costly in computation. Due to the above-mentioned drawbacks of NN and SVM, the purpose of this paper, is to explore the performance of classification using various decision tree approaches which have the following advantages as follows [6]:

1) Decision trees are easy to interpret and understand;
2) Decision trees can be converted to a set of $if-then$ rules; and
3) Decision trees don't need priori assumptions about the nature of data, it is a *distribution-free*.

Since decision trees have the described advantages, they have proven to be effective tools in classification of Egyptian rice

disease problems [7]. Specially, the transfer of experts from consultants and scientists to agriculturists, extends workers and farmers represent a bottleneck for the development of agriculture on the national. This information can be used as part of the farmers decision-making process to help to improve crop production. The aim of this paper is to evaluate the tree-based classifiers to select the classifier which more probably achieves the best performance for the Egyptian rice diseases which cause losses that estimated by 15% from the yield, malformation of the leaves or dwarfing of the plants. Discovering and controlling of diseases are the main aims and have a large effect for increasing density of Fadden and increasing gain for farmer then increasing the national income. Actually, the original contribution of this research paper is to measure and compare the performances of tree-based classification algorithms for Egyptian rice diseases. In particular, we have focused on the Bayesian network, random forest algorithms, comparing its performances with a decision tree using a variety of performance metrics. In this paper, four classification algorithms are investigated and presented for their performance. Section II, presents the related previous work. The proposed used classification algorithms are explained in section III. In section IV, our problem is formally described. Section V, describes data set used in this paper. In section VI an experimental results described for investigated types of classification algorithms including their performance measures. Finally, the conclusions are explained in section VII.

II. Related Work

The objectives of applying data mining techniques in agriculture is to increase of productivity and food quality at reduced losses by accurate diagnosis and timely solution of the field problem. Using data mining classification algorithms, it become possible to discover the classification rules for diseases in rice crop [7], [8]. The image processing and pattern recognition techniques are used in developing an automated system for classifying diseases of infected rice plants [9]. They extracted features from the infected regions of the rice plant images by using a system that classifies different types of rice disease using self-organizing map (SOM) neural network. Feature selection stage was done using rough set theory to reduce the complexity of classifier and to minimize the loss of information where a rule base classifier has been generated to classify the different disease and provide superior result compare to traditional classifiers [9]. Also, SVM is used to disease identification in the rice crop from extracted features based on shape and texture, where a three disease leaf blight, sheath blight and rice blast are classified [10]. In another work, the brown spot in rice crop is identified using K-Means method for segmentation and NN for classification of disease [11]. The NN is used to identify the three rice diseases namely (i) Bacterial leaf blight, (ii) Brown spot, and (iii) Rice blast [12]. The fuzzy entropy and probabilistic neural network are used to identify and classifying the rice plant diseases. Developed a mobile application based on android operating system and features of the diseases were extracted using fuzzy entropy [13].

III. Classification Algorithms

A total of four classification algorithms have been used in this comparative study. The classifiers in Weka have been categorized into different groups such as Bayes and Tree based classifiers, etc. A good mix of algorithms have been chosen from these groups that include decision tree, Naive Bayes net, random trees and random forest. The following sections briefly explain about each of these algorithms.

A. Decision Tree

The first algorithm used for comparison is a decision tree, which generates a classification-decision tree for the given data-set by recursive partitioning of data [14]. The algorithm considers all the possible tests that can split the data set and selects a test that gives the best information gain. For each discrete attribute, one test with outcomes as many as the number of distinct values of the attribute is considered. For each continuous attribute, binary tests involving every distinct values of the attribute are considered. In order to gather the entropy gain of all these binary tests efficiently, the training data set belonging to the node in consideration is sorted for the values of the continuous attribute and the entropy gains of the binary cut based on each distinct values are calculated in one scan of the sorted data. This process is repeated for each continuous attributes [15], [16]. In particular entropy, for an attribute is defined as in equation 1.

$$H(X) = -\sum_{j}^{m} p_j log_2(p_j) \tag{1}$$

Where p_j is defined as $P(X = V_j)$, the probability that X takes value V_j, and m is the number of different values that X admits. Due to their recursive partitioning strategy, decision trees tend to construct a complex structure of many internal nodes. This will often lead to over fitting. Therefore, the decision tree algorithm exhibits meta-parameters that allow the user to influence when to stop tree growing or how to prune a fully-grown tree.

B. Random Decision Tree

The second chosen algorithm for the comparison is the *random decision tree* presented by Fan *et al.* in [17]. The Random decision tree is an ensemble learning algorithm that generates many individual learners. It employs a bagging idea to produce a random set of data for constructing a decision tree. In the standard tree each node is split using the best split among all variables. The choice is bind on the type of the attribute, in particular if the feature can assume values in a finite set of options it cannot be chosen again in the sub tree rooted on it. However, if the feature is a continuous one, then a random threshold is chosen to split the decision and it can be chosen again several times in the same sub tree accordingly with the ancestor's decision. To enhance the accuracy of the method, since the random choice may leads to different results, multiple trees are trained in order to approximate the true mean. Considering k as the number of features of the dataset and N as the number of trees, then the confidence probability to have is:

$$1 - (1 - \frac{1}{k})^N \tag{2}$$

Considering the k features of the dataset, and the i classifying attributes, the most diversity among trees is with depth of

$$\frac{k}{2} \tag{3}$$

since the maximum value of the combination is

$$\binom{i}{k} \tag{4}$$

Once the structure is ready the training may take place, in particular each tuple of the dataset train all the trees generated in order to read only one time the data. Each node counts how many numbers of examples go through it. At the end of the training the leaves contain the probability distribution of each class, in particular for the tree *i*, considering *n[y]* the number of instances of class y at the node reached by x, is:

$$P_i(y|x) = \frac{n[y]}{\sum_y n[y]} \tag{5}$$

The classification phase retrieves the probability distribution from each tree and average on the number of trees generated in the model:

$$P(y|x) = \frac{1}{N}\sum_{i=1}^{N} P_i(y|x) \tag{6}$$

C. Bayesian Network

Bayesian Networks encode conditional interdependence relationships through the position and direction of edges in a directed acyclic graph. The relationship between a node and its parent is quantified during network training. This classifier learns from training data the *conditional probability* of each attribute X_i given the class label C. Classification is then done by applying Bayes rule to compute the probability of C given the particular instances of $X_1 \ldots X_n$ and then predicting the class with the highest posterior probability. The goal of classification is to correctly predict the value of a designated discrete class variable given a vector of predictors or attributes [2]. In particular, the naive Bayes classifier is a Bayesian network where the class has no parents and each attribute has the class as its sole parent. Bayesian and neural network seem to be identical in their inner working. Their difference exist in the construction. Nodes in a neural network don't usually have clearly defined relationship and hidden node are more "discovered" than determined, whereas the relationships between nodes in Bayesian network are due to their conditional dependencies [18], [2].

D. Random Forest

The random forest classifier, described by Ho [19], [1], works by creating a bunch of decision trees randomly. Each single tree is created in a randomly selected subspace of the feature space. Trees in different subspaces complement each other's classifications. Actually, random forest is an ensemble classifier which consists of many decision tree and gives class as outputs i.e., the mode of the class's output by individual trees. Prediction is made by aggregating (majority vote for classification or averaging for regression) the predictions of the ensemble. Random Forests gives many classification trees without pruning [20]. The success of an ensemble strategy depends on two factors, the strength (accuracy) of individual base models and the diversity among them.

TABLE I: Possible value for each attribute from the Egyptian rice database

Attribute	Possible Values
Variety	gizal71,gizal77, gizal78 sakhal0l, sakhal02, sakhal03 , sakhal04
Age	Real values
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
Colour	gray, olive, brown, brownish, whitish, yellow, green, orange, greenish black, white, pale, blackish, blac k
Temperature	Real values
Disease	Blight, brown spot, false smut, white tipe, stem rot

IV. Problem Definition

The main aim of this work is to produce a comparison among different inductive learning the *optimal model* for a target function $t = F(x)$, given a training set of size n, $(x_1; t_1), ..., (x_n; t_n)$, an inductive learner produces a model $y = f(x)$ to approximate the true function $F(x)$. Usually, there exists x such that $y \neq t$. In order to compare performance, a loss function is introduced $L(t, y)$. Given the loss function $L(t, y)$, that measures the discrepancy between our function's class and reality, where t is the true class and y is the predicted class, an optimal model is one that minimizes the average loss $L(t, y)$ for all examples. The optimal decision y^* for x is the class that minimizes the expected loss $E_t(L(t, y^*)$ for a given example x when x is sampled repeatedly.

V. Data Set Description

Rice is the worlds most common staple food for more than half of mankind. Because of its important, rice is considered a strategic resource in Egypt has been assigned as a high priority topic in its Agricultural Strategic Plans. 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 Gizal77 needs 125 day and Sakhal04 needs 135 day. Climate for the Egyptian rice is that daily temperature maximum = $30 - 35^o$, and minimum = $18 - 22^o$; humidity = 55%-65%; wind speed = $1 - 2m$. 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 $(2.4t/fed.)$ during the lowest period $1984 - 1986$ to $(3.95t/fed.)$ in 2002 [21]. 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. Each case in the data set is described by seven attributes. We have a total of 206 samples and the attribute and possible values are listed in Table I.

VI. Experimental Evaluation

To gauge and investigate the performance on the selected classification methods or tree-based learning algorithms, many experiments are implemented within the WEKA framework

[22]. The Weka is an open source data mining workbench software which is used for simulation of practical measurements. It contains tools for data preprocessing, classification, regression, clustering, association rule and visualization. It does not only supports data mining algorithms, but also data preparation and meta-learners like bagging and boosting [22]. In order to test the efficiency of tree-based classification algorithms, training and test sets are used. Usually disjoint, subsets, the training set to build a classification tree(s) and the test set so as to check and validate the trained model. Also, cross-validation process applied where same sized disjoint sets are created so as to train the model fold wise. n-fold cross-validation, (usually $n = 10$) is used to divide the data into equally sized k subsets/ folds. In such case the model is trained using $(k-1)$ folds and the k^{th} fold is used as a test set. The whole process is repeated n times in an attempt to use all the folds for testing thus allowing the whole of the data to be used for both training and testing. In our data, ten cross-validation bootstraps, each with 138 (66%) training cases and $68(34\%)$ testing cases, were used for the performance evaluation. The simulation results are partitioned into two parts for easier analysis and evaluation. On the first part, correctly and incorrectly classified instances will be partitioned in percentage value and subsequently Kappa statistics, mean absolute error and root mean squared error will be in numeric value only. We also show the relative absolute error and root relative squared error in percentage for references and evaluation. The results of the simulation are shown in Tables II and III. Table II mainly summarizes the result based on accuracy and time taken for each simulation. Meanwhile, Table III shows the result based on error during the simulation.

TABLE II: Evaluation results of different classification algorithms

Alg.	Correctly %	Incorrectly %	time (sec.)	Kappa statistics
Decision Trees	97.57	2.42	0.01	0.97
Random Trees	94.66	5.33	0.07	0.92
Bayes Net	93.68	6.31	0.06	0.93
Random Forest	95.63	4.36	0.07	0.94

TABLE III: The errors of different classification algorithm

Alg.	Mean Abs. Error	Root Mean Squ. Error	Relative Abs. Error(%)	Root Relative Squ. Error(%)
Decision Tree	0.04	0.12	12.8	30.7
Random Trees	0.06	0.133	19.61	33.7
Bayes Net	0.129	0.199	41.31	50.4
Random Forest	0.036	0.124	11.44	31.4

Figure 2 shows the evaluation of different classification algorithms which are summarized in Table III. From the confusion matrix to analyse the performance criterion for the classifiers in disease detection accuracy, precision, recall and Mathews correlation coefficient (MCC) have been computed for the dataset as shown in Table IV. MCC is a special case of the linear correlation coefficient, and therefore also scales between +1 (perfect correlation) and -1 (anti correlation), with 0 indicating randomness. Accuracy, precision (specificity), recall (sensitivity) and MCC are calculated using the equations (7), (8), (9) and (10) respectively, where T_p is the number of true positives, T_n is the number of true negatives, F_p is the number of false positives and F_n is the number of false negatives.

$$accuracy = \frac{T_p + T_n}{T_p + T_n + F_p + F_n} \tag{7}$$

$$specificity = \frac{T_p}{T_p + F_p} \tag{8}$$

$$sensitivity = \frac{T_p}{T_p + F_n} \tag{9}$$

$$MCC = \frac{T_p * T_n - F_p * F_n}{\sqrt{(T_p + F_n)(T_p + F_p)(T_n + F_n)(T_n + F_p)}} \tag{10}$$

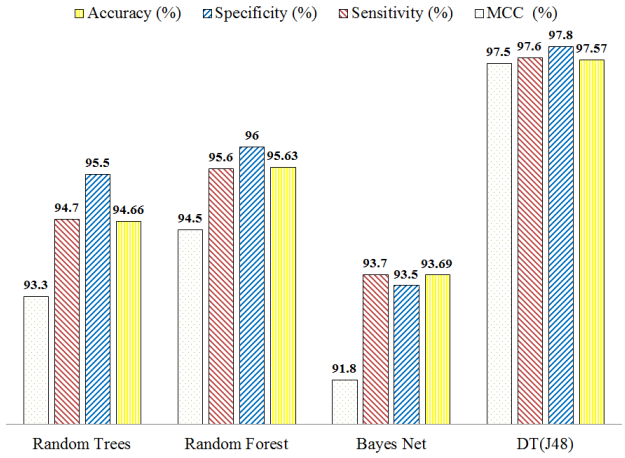

Fig. 2: The Roote Mean Square (RMSE) of each algorithm

TABLE IV: Accuracy, Specificity, Sensitivity and MCC of different classification algorithm

Alg.	Accuracy (%)	Specificity (%)	Sensitivity (%)	MCC
Decision Tree	97.57	97.8	97.6	0.95
Random Tree	93.69	93.5	93.7	0.92
Bayes Net	95.63	96.0	95.6	0.95
Random Forest	94.66	95.5	94.7	0.068

VII. Conclusions and Future Work

Data mining in agriculture is a very interesting research topic and can be used in many applications such as yields prediction, disease detection, optimizing the pesticide usage and so on. There are many algorithms that have been presented for classification in diagnosing the Egyptian rice diseases data set. However, we have choose four algorithms the J48 decision tree, Bayes net, random trees and random forest that belongs to the Tree-based category which are easy to interpret and understand. we conduct many experiments to evaluate the four classifiers for Egyptian rice diseases. The above analysis shows that for the J48 decision tree achieves highest sensitivity, specificity and accuracy and lowest RMS error, than Bayes net, random trees and random forest. J4.8 gave the best results due to the pruning process which simplify the tree and remove

unrelevant branches. Moreover, the random forest superior over random trees due to boosting process [23], [24].

Lastly, it should be mentioned that the predictive accuracy is the probability that a model correctly classifies an independent observation not used for model construction. A tree that involves irrelevant variables is not only more cumbersome to interpret but also potentially misleading. selecting an informative features and removing irrelevant/redundant features drastically reduces the running time of a learning algorithm and yields a more general classifier [25], [26]. So, in future works we intend to apply relevant methods for *feature selection* in classification to improve our results as a preprocessing stage before the classification process.

ACKNOWLEDGEMENT

The authors would like to thank the anonymous reviewers for their revision and help to enhancement the paper writing. Also, we are indebted to Central Laboratory for Agricultural Expert Systems staff for providing us with their experiences and data set. And above all, God for His continuous guidance.

REFERENCES

[1] J. Han, M. Kamber and J.Pei, Data mining: concepts and techniques, Elsevier Inc., 3^{rd} edition, 2012.

[2] Y. Nong, Data mining: theories, algorithms, and examples, CRC Press, 2014.

[3] M. Zaki and W. Meira, Data mining and analysis: foundations and algorithms, Cambridge University Press, 2014.

[4] M. Fernandez-Delgado, E. Cernadas, S. Barro and D. Amorim, "Do we need hundreds of classifiers to solve real world classification problems?," in Machine Learning Research, 15, pp. 3133-3181, 2014.

[5] C. Bishop, Pattern recognition and machine learning, Springer New York, 2006.

[6] Y. Zhao and Y. Zhang, *Comparison of decision tree methods for finding active objects*, arXiv:0708.4274v1, 2007.

[7] M. El-Telbany, M. Warda and M. El-Borahy, "Mining the classification rules for Egyptian rice diseases," in International Arab Journal of Information Technology (IAJIT), Jordan, Vol. 3, No. 4, 2006.

[8] A. Nithya, V. Sundaram, "Classification rules for Indian Rice diseases," in International Journal of Computer Science (IJCSI), Vol. 8, Issue 1, 2011.

[9] S. Phadikar, J. Sil and A. Das, "Rice diseases classification using feature selection and rule generation techniques," in Comput. Electron. Agric., 90, pp. 7685, 2013.

[10] Q. Yao, Z. Guan, Y. Zhou, J. Tang, Y. Hu and B. Yang, "Application of support vector machine for detecting rice diseases using shape and color texture features," in International Conference on Engineering Computation, pp.79-83, 2009.

[11] D. Al-Bashish, M. Braik, S. Bani-Ahmad, "A Framework for Detection and Classification of Plant Leaf and Stem Diseases," in International Conference on Signal and Image Processing, pp. 113-118, 2010.

[12] J. Orillo, J. Cruz, L. Agapito, P. Satimbre and I.Valenzuela, "Identification of diseases in rice plant (oryza sativa) using back propagation Artificial Neural Network." in 7^{th} International Conference on Humanoid, Nanotechnology, Information Technology, Communication and Control, Environment and Management (HNICEM), 2014.

[13] K. Majid, Y. Herdiyeni and A.Rauf, "I-PEDIA: Mobile application for paddy disease identification using fuzzy entropy and probabilistic neural network," in International Conference on Advanced Computer Science and Information Systems (ICACSIS), 2013.

[14] T. Hastie, R. Tibshirani and J. Friedman, The elements of statistical learning: data mining, inference and prediction, the Mathematical Intelligence, 27(2): pp. 83-85, 2005.

[15] J. Quinlan, "Induction of decision trees," in Machine Learning, 1(1), pp. 81-106, 1986.

[16] T. Mitchell, Machine Learning, McGraw Hill, 1997.

[17] W. Fan, H. Wang, P. Yu, and S. Ma, "Is random model better? On its accuracy and efficiency," in 3^{rd} IEEE International Conference on Data Mining, 2003.

[18] N. Friedman, D. Geiger and M. Goldszmidt, "Bayesian network classifiers," in Machine Learning, 29, pp. 131-163, 1997.

[19] K. Ho , "Random decision forests," in IEEE Proceedings of the 3^{d} International Conference on Document Analysis and Recognition, pp. 278-282, 2005.

[20] L. Breiman, Random Forests, Machine learning, Springer, pp. 5-32, 2001.

[21] Sakha Research Center, "The results of rice program for rice research and development," Laboratory for Agricultural Expert Systems, Ministry of Agriculture, Egypt, 2002.

[22] M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann and I. Witten,", The WEKA data mining software an update," in ACM SIGKDD Explorations Newsletter, 2009.

[23] L. Rokach and O. Maimon, Data mining with decision trees: theory and applications, World Scientific Publishing, 2^{nd} ed. 2015.

[24] T. G. Dietterich, "An experimental comparison of three methods for constructing ensembles of decision trees: Bagging, Boosting, and Randomization," in Machine Learning, pp. 1-22, 1999.

[25] C. Aggarwal, Data mining: the textbook, Springer, 2015.

[26] S. Garca, J. Luengo and F. Herrera, Data preprocessing in data mining, Springer, 2014.