Classification of Paddy Types by Decision Tree Induction

Myat Min Soe, Khin Sandar
Computer University, Pathein , Myanmar
myatmin.ucsy@gmail.com, drkhinsandar@gmail.com

Abstract

This system is to give the information for the knowledge worker can be decision for the best paddy type. Correct classification of measurements may in fact be the most critical part of the diagnostic process. In this paper, we implemented as a method of decision support for paddy types classification using Decision Tree Induction. Decision Support System is needed to help the farmer who selects himself the most suitable paddy type. This paper intends to develop a Decision Support System for paddy type selection to grow according to the factors of their own lands. Therefore computer based information system which stored knowledge of the cultivation experts and agricultural records of the previous year is needed to get the highest income. Classifier accuracy is also estimated to get the better decision support system with the minimum error rate by using decision tree induction.

Keywords
Decision Support System (DSS), Knowledge Base, Expert System, Decision Tree Induction and Classification.

1. Introduction

While classification is a well-studied problem, only recently has there been focus on algorithm that can handle large databases. The intuition is that by classifying large datasets, we will be improving the accuracy of the classification model. Many classification and prediction methods are imposed by researchers in machine learning, expert systems, statistics and neurobiology. Some other approaches of classifications are neural networks, statistical models like linear / quadratic discriminants, K-nearest neighbor classifiers, decision trees and genetic models. Among these model, decision trees are particularly studied for data mining. Decision Trees can be constructed relatively fast compared to other methods. Another advantage is that decision tree models are simple and easy to understand. Moreover, tree can be easily converted into statements that can be used to access databases efficiently. Finally, decision tree classifiers obtain similar and sometimes better accuracy when composed with other classification methods.

In classification, we are given a set of example records, called training set, where each record consists of several fields or attributes. Attributes are either continuously, coming from an ordered domain. One of the attributes, called the classifying attribute, indicates the class to which each example belongs. The objective of classification is to build a model of the classifying attribute based upon the other attributes. Applications of classification arise in diverse fields, such as retail target marketing, customer retention, fraud detection and diagnostic processes.

Classification is an important data mining problem and can be described as follows. The input data, also called the training set, consists of multiple examples. Each example is tagged with a special class label.

A decision tree is a class discriminator that recursively partitions the training set until each partition consists entirely or dominantly of examples from one class. Each non-leaf node of the tree contains a split point which is a test on one or more attributes and determine how the data is partitioned.

Decision Tree is one of the most popular methods used for inductive inference. They are robust for

noisy data and cable of learning disjunctive expressions. A decision tree is a k-ary tree where each of the internal nodes specifies a test on some attributes from the input feature set used to represent the data. Each branch descending from a node corresponding to one of the features specified at that node.

Rice is the daily staple food for Myanmar people. According to the historical records, there had been rice cultivation in Myanmar since the ancient days. Because of the facts that suit to the ecological conditions of Myanmar, rice has been grown traditionally and continuously and is the staple food of the whole nation while surpluses are being exported as a source of foreign exchange, rice was classified as the ”National Priority Crop” in 1994[8]. The major aims for farmer are to get the best yield per acres (baskets). Therefore it is important to select the better rice strain.

This article has been motivated by classifying problem in paddy type selection diagnosis. Some recently application domain includes different kinds of paddy and learning from the database factors about the paddy. These learning models are expected to be accurate and are further to be intelligible to expert in the field.

In this paper we describe an approach for developing knowledge based on Decision Support System. DSS is an approach for solving problems based on the solutions of similar cases. Decision Tree Induction is used to classify the paddy types and it need the previously train data. These train data are got from the cultivation experts, agricultural records, books and previous experiences of the system.

2. Related Work

Many approaches have been proposed to construct decision tree classifiers in general and particularly to improve scalability of classification. Most well-known algorithms like CART [1], ID3 [3], C4.5 [4] assume the data to be in memory and therefore are able to work only with relative small data sets efficiently. In the database research scalability is addressed by developing algorithms based on special data structures, e.g. SPRINT [6], SLIQ [6] or optimistic tree construction [2]. In [9] the Rain Forest framework is described that introduces an AVC-group data structure providing sufficient statistics for determining the split and algorithms for constructing this structure. [4] describe a similar data structure called CC table and a middleware based on a scheduler ensuring optimized scans and staging.

Whereas the Rain Forest framework does not address SQL databases, the middleware is implemented on a commercial DBMS. Our COMPUTENODESTATICTICS is derived directly from this both approaches. Other approaches consider approximation techniques for scaling up the classification, e.g. sampling [1] and discrimination, as well as permitting the user to specify constraints on tree size [7]. Particularly, approximations techniques could be supported by the database systems very well and thus could lead to further primitives.

There has been much work in the field of failure diagnosis, though most previous work explicitly models causal or dependence interactions between the various components of the system. Our approach, in comparison, makes only implicit use of the underlying structure during the node merging phase. It is also necessary to stress that, though we have made references to “cause-finding,” we do not attempt to infer any causal relationships between any of the components and the outcome.

There has been much work in causal network modeling [3], and also on inferring causal relationships from observational data [3]. However, that is not the approach taken in this paper. There are many commercial management systems that aid failure diagnosis. These systems typically either employ expert systems with human-generated rules or rely on the use of dependency models [7].

However, these systems do not consider how the required dependency models are obtained. More recent research has focused on automatically generating dependency models based on dynamic observations.

The dependence graph models different layers within each host and linkage pattern between hosts. Each layer is associated with multiple possible failure modes. After observing certain symptoms in the system, belief propagation algorithms are run on the graph, and the posterior beliefs are examined to pick out the most likely causes of the symptoms.

3. Proposed System

This system intends to develop the Decision Support System for analyzing paddy types classification and helping in diagnosing it. The system uses Decision Tree Induction Classifier to classify the paddy types. It must maintain the Knowledge Base by collecting the knowledge of the cultivation experts and previous experiences.

It also contains knowledge that has been organized and analyzed to make it understandable and applicable to problem solving or decision making. The system will prepare the training data using expert knowledge and experiences stored.

It also intends to advise the farmer for choosing the most suitable paddy type in order to grow in their firms. If a user inputs his firms' factors, the system will find the highest information gained of each field and extract the result according to the rules that were drawn by the highest information gained.

In this paper we introduce inductive retrieval algorithm, a technique that determines which features do the best job in discriminating and generating a decision tree type structure to organize the cases in memory. This approach is very useful when a single feature is required as a solution and, when that case feature is dependent upon others.

In this system, training data is used to build the model by using Decision Tree Induction and testing data is used to test the accuracy of the training data.

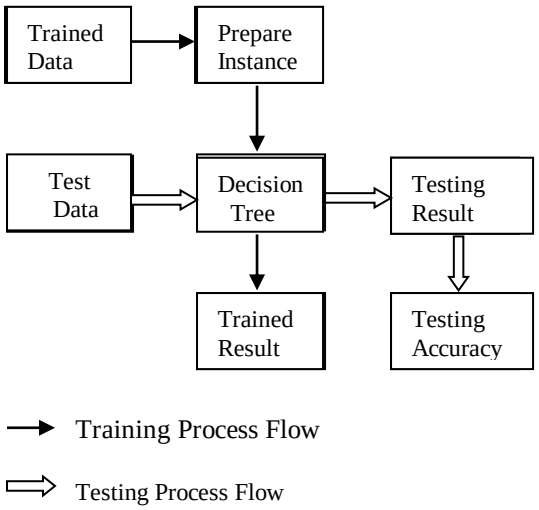

Figure 1. Process Flow of the System

3.1 Decision Tree Induction

Decision Tree is a greedy algorithm that constructs a tree in a top-down recursive divide-and-conquer manner. It is also a flow-chat like tree structure, where each internal node denotes a test on an attribute, each branch represents an outcome of the test and leaf node represents class or class distribution. We wish to predict the test sample.

ID3 and C4.5 are algorithms introduced by Quinlan for inducing Classification Models, also called Decision Trees, from data [6]. We are given a set of records. Each record has the same structure, consisting of a number of attribute/value pairs. One of these attributes represents the category of the record. The problem is to determine a decision tree that on the basis of answers to questions about the non-category attributes predict correctly the value of the category attribute. Usually the category attribute takes only the values {true, false}, or {success, failure}, or something equivalent. In any case, one of its values will mean failure.

ID3 can deal with very large data sets by performing induction on subsets or windows onto the data.

1. Select a random subset of the whole set of training instances.
2. Use the induction algorithm to form a rule to explain the current window.
3. Scan through all of the training instances looking for exceptions to the rule.

In the Decision Support System for selecting a suitable paddy type, the system needs to be trained before it is in practically used. The followings are attributes, attribute values.

```
@ attribute field type {'flooded field', 'deep-water field ','saline water field', 'irrigated'}
@ attribute soil type {'alluvial', 'meadow and meadow alluvial ', 'gluey and gluey swampy ',' saline swampy and meadow gluey '}
@ attribute rainfall {'<20', '20-30','30-40','>40'}
@ attribute temperature {'<20', '20-30', '>30'}
@ attribute water gain {'less',' air', 'excellent'}
@ attribute germ resistance {'less', 'fair', 'excellent'}
@ attribute utilization of fertilizer {'high', 'low',' medium'}
@ attribute cultivation method {'move', 'scatter seeds'}
@ attribute harrow type {'machine', 'animals'}
@ attribute multiple crop {'yes', 'no'}
@ attribute result {A, B, C, D}
```

Figure 2. Attributes Used in the System

In this system, we used ten attributes and one class label. The critical step in the decision tree is the selection of the best test attributes. The information gain measure is used to select the test attribute at each node in the tree.

First, another related term called entropy needs to be introduce. In general, E is measure of the purity in an arbitrary collection of example S.

Let S be a set consisting of a s data samples. Suppose the class label attributes has four distinct values defining four distinct classes g(namely {A,B,C,D}), therefore there are four distinct classes c1='A', c2='B', c3='C' and c4='D', Let si be the number of samples of S in class Ci. The expected information need to classify a given sample is given by

$$I(s_1,s_2,\ldots,s_m) = -\sum_{i=1}^{m} p_i \log_2 (p_i) \qquad (1)$$

Where pi means probability that an arbitrary samples belongs to class Ci and is estimated by si/s.

Let attribute A have {field type, soil type, …, multiple crop, etc.}. Attribute A can be used to partition S into subsets, {S1, S2,…, Sm}, where Sj contains those samples in S that have value aj of A.

The entropy based on the partitioning into subsets by A, is given by

$$E(A) = \sum_{j=1}^{v} \frac{s_{1j}+\ldots+s_{mj}}{s} I(s_{1j},\ldots,s_{mj}) \qquad (2)$$

The smaller the entropy value, the greater the purity of the subsets of partitions. Note that for a given subset Sj,

$$I(s_{1j},s_{2j},\ldots,s_{mj}) = -\sum_{i=1}^{m} p_{ij} \log 2 (p_{ij}) \qquad (3)$$

Hence, the gain in information from such a partitioning would be

$$Gain(A) = I(s_1,\ldots,s_m) - E(A) \qquad (4)$$

Gain (A) is expected reduction in entropy caused by knowing the attribute A.

In this way we can compute the information gain of each attributes. The attributes with highest information gain is chosen as the test attributes for the given set S. Therefore we select highest information gain among the attributes, it is selected as the test attribute. This attribute become a test attributes at the root node of the decision tree. A node is created and labeled with the attribute, branches created for each value of the attributes, and samples are partitioned accordingly. This process is repeated for each subset belong in a single class, The partitioning procedure attempts to find the most informative attribute in order to create the shortest tree possible. At each node in the decision tree, the criterion is evaluated for all the attributes which are relevant and the one is picked which yields the highest increase of the information gain measure.

4. Implementation of the System

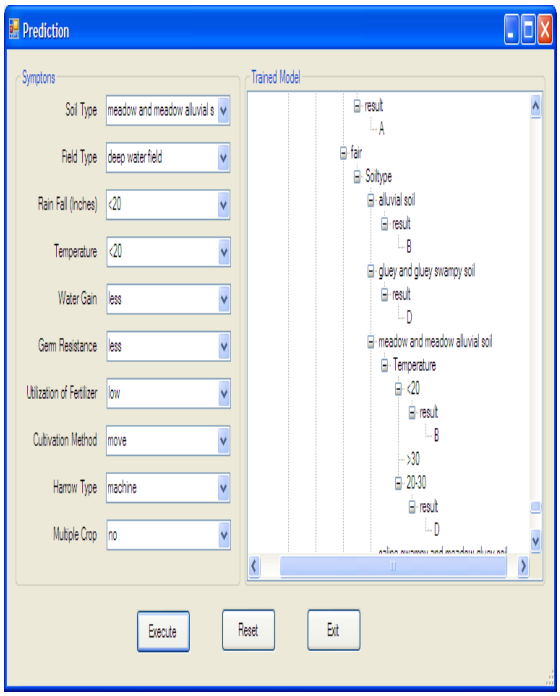

Figure 3. Prediction the result

Figure 3 shows the prediction section of the system. When we fill the information needed by the system, we can see the result of the paddy type in the following.

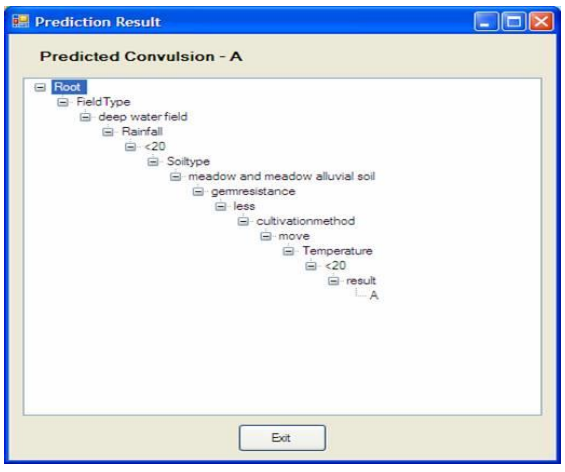

Figure 4. Prediction Result Tree

5. Classifier Accuracy

The primary metric for evaluating classifier performance is classification accuracy _ the percentage of test samples that are correctly classified. The other important matrices are classification time. The ideal goal for a decision tree classifier is to produce compact, accurate tree in a shortest classification time.

Estimating classifier accuracy is important because it determines to evaluate how accurately a given classifier is. And accuracy also help in the comparison of different classifiers.

Decision Tree generates clear description of how the method arrives at a particular classification while the Nave Bayesian Classifier was included for comparison purposes. Decision Tree has been used to discover logical patterns within dataset for many years.

Decision Tree predicts more than 98% of diagnosis process correctly. This is a further indication that the classification system is good for distinguishing diagnosis articles. In addition, Decision Tree method gives slightly better than Nave Bayesian Classification performance. The training set is similar to that on the test set , indicating that there is no over fitting on the training data. As in all diagnosis tasks, the type of misclassification made by different methods are particularly interesting.

We assume this by means of the relative performance of the classifier's in the different classes. In practice, other classifiers give less success rate than the Decision Trees Induction.

6. Conclusion

The database research community's contribution to classification and prediction for data mining have emphasized the scalability aspect, particularly with respect to Decision Tree Induction with the amount of data are being collected electronically and wide spread availability of cheap. Many researchers have stated exploring these data. Decision Tree Induction has been applied to problem such as learning to classify probability (highest information gained) by applying economical facts. The computer will make the decision instead of the user. This system supports users in classifying paddy type by using decision tree induction. The farmer can easily test himself at home by entering data in paddy type prediction. So, he can get highest yield per acres and highest income. So the farmer will know what kinds of paddy are suitable for growing in their firms by using this system.

7. Reference

[1] C. Wallace and J. Patrick. Coding Decision trees. Machine Learning, 11:7-22, 1993.

[2] J. Catlett. Megainduction: Machine Learning on Very Large Databases. PhD the- sis, University of Syney, 1991.

[3] J. Wirth and J. Catlett. Experiments on the costs and benefits of windowing in ID3. In 5th International Conference on Machine Learning, 2001.

[4] J. Ross Quinlan. C4.5: Programs for Machine Learning. Morgan Kaufman, 1992.

[5] L .Breiman et. al. Classification and regression Trees. Wadsworth , Belmont , 1984.

[6] Manish Mehta, Rakesh Agrawal, and Jorma Risanen. SLIQ: A fast scalable classifier for data mining. In Proc; of the fifth international Conference on Extending database Technology (EDBT), Avignon, France, March 1999.

[7] R. Agrawal, T. Imielinski, and A. Swami. Database mining: A performance per-spective. IEEE Trans . on Knowledge and Data Engineering, 5(6) , Dec . 1993.

[8] Rice Varieties in Myanmar, Commemoration of the Internal Year of Rice 2004. Distribution of Ministry of Agriculture & Irrigation.

[9] S. M. Weiss and C.A. Kulikowski. Computer Systems that Learn: Classification and Prediction Methods from Statistics, Neural Nets, Machine Learning, and Ex-pert Systems. Morgan Kaufman, 1991.

[10] S. M. Weiss and C.A. Kulikowski. Computer Systems that Learn: Classification and Prediction Methods from Statistics, Neural Nets, Machine Learning, and Ex-pert Systems. Morgan Kaufman, 1999.