Implementation of Classification System for Dry Zone Plantation Based on ID3 Algorithm

Thazin Hlaing
Computer University, Mandalay
thazintz@gmail.com

Abstract

Classification is data analysis process that can be used to extract models describing important data classes or predict future data. Classification of large data sets is an important data mining problem. Decision tree, mainly used for classification purpose, is a classifier method in the form of a tree structure. Decision tree algorithms are a method for approximating discrete-valued target functions, in which the learned function is represented by a decision tree. The data set of plantation system for Dry Zone is used to provide the aim of the system. The aim of this paper is to present how to construct decision trees (ID3), show its rules and evaluate the performance of model and to test the unknown data set. An accuracy method, hold-out, is allowed to use in this system for the validity of rules.

Keywords: *ID3, Classification, Data mining, Accuracy*

1. Introduction

Classification is a form of data analysis that can be used to extract models describing important data class or to predict future data trends. Data classification is a two-steps process. In the first step, a model is built, describing a predetermined set of data classes. In the second step, the model is used for classification. If the accuracy of the model is considered acceptable, the model can be used to classify future data tuples for which the class label is not known. Classification rules represent the classification knowledge as IF-THEN rules and are easier to understand for users.

This paper examines accuracy of decision tree learning algorithm ID3 and implements this algorithm by using C# programming language. The main task performed in this system is using decision tree induction methods to the given values of attributes and to determine the unknown object according to decision tree rules. Test attributes are selected on the basis of a heuristic or statistical measure. Such a measure is referred to as an attribute selection measure or a measure of the goodness of split.

The remainder of the paper is structured as follows. Section 2 introduces the related work of classifier methods. Section 3 explains the theoretical framework. The proposed system framework is discussed in section 4. The design of system is also described in Section 5 and Section 6 describes the experimental result of system. Finally, this paper is concluded with further extension of some methods.

2. Related work

The decision tree is one of the most popular classification algorithms in Data Mining and Machine Learning [3]. A number of algorithms for induction decision trees have been proposed over the year ID3, C4.5, SPRINT, PUBLIC and BOAT [4]. Decision trees are especially attractive for a data mining environment for three reasons. First, due to their intuitive representation, ones are easy to assimilate by humans [1]. Second, they can be constructed relatively fast compared to other methods. Last, the accuracy of decision tree classifier is comparable or superior to other models. The classification is an important problem in data mining was proposed in [2]. A number of popular classifiers construct decision trees to generate class models.

Frequently, the constructed trees are complex with hundreds of nodes and thus difficult to comprehend, a fact that calls into question and often cited benefit that decision trees are easy to interpret. The Decision tree learning algorithm has been successfully used in expert systems in capturing knowledge was proposed [5]. The main task performed in these systems is using inductive methods to the given values of attributes of an unknown object to determine appropriate classification according to decision tree rules. One

uses the decision tree learning algorithm ID3 and implements basic ID3 in which one deals with the target function that has discrete output values. One also extends the domain of ID3 to real-valued output, such as numeric data and discrete outcome rather than simply Boolean value.

3. Theoretical framework

3.1. Decision tree induction

Decision trees are powerful and popular tools for classification and prediction. The main task performed in this method is using inductive methods to the given values of attributes of an unknown object to determine appropriate classification according to decision tree rules. There are two steps in decision tree induction. They are model construction and model usage.

A decision tree is classified instances by sorting them down the tree from the root to some leaf node, which provides the classification of the instance.

ID3 is a typical decision-tree algorithm. It introduces information entropy as the splitting attribute's choosing measure. It trains a tree from root to leaf, a top-down sequence. Each path from that form is a decision rule. The ID3 algorithm is to search the attribute with maximum information gain, and to use the attribute as the splitting attribute.

The basic structure of ID3 is iterative. A subset of training set is chosen at random and a decision tree formed from it; this tree correctly classified all objects. All other objects in the training set are then classified using the tree.

3.2. ID3 algorithm

Generate_Decision_Tree ()
Create a node N
If samples are all of the same class, C then
Return N as a leaf node labeled with the class C
If attributes-list is empty then
Return N as a leaf node labeled with the most common class in sample
Select test-attribute, the attribute among attribute list with highest gain
Label node N with test-attribute
For each known value a_i of test-attribute
Grow a branch from node N for the condition test-attribute= a_i
Let s_i be the set of samples in samples for which test-attribute=ai
If s_i is empty then
Attach a leaf labeled with the most common class in samples
Else attach the node returned by
Generate_Decision_Tree ()

3.3. Attribute selection in decision tree classifier based on information

Entropy is the expected information based on the partitioning into subsets by an attribute. The smaller the entropy value, the greater is the purity of the subset partitions. The information gain measure is used to select the test attribute at each node in the tree.

Let *S* be a set consisting of s data samples. Suppose the class label attribute has *m* distinct values defining m distinct classes: C_i (for i = 1,.., m). Let s_i be the number of samples of *S* in class C_i. The expected information needed to classify a given sample is given by:

$$I(s_1, S_2, \ldots, S_m) = \sum_{i=1}^{m} p_i \log_2 (P_i) \qquad (3.1)$$

Where P_i is the probability that an arbitrary belongs to class C_i and is estimated is estimated by s_i/s. Note that a log function to the base 2 is used since the information is encoded in bits.

Let attribute A have v distinct values, $\{a_1, a_2, .., a_v\}$. Attribute A can be used to partition S into v subsets, $\{S_1, S_2, .., S_v\}$, where S_j contain those samples in S that have a_j of A. If A were selected as the test attribute (i.e., the best attribute for splitting), then these subsets would correspond to the branches grown from the node containing the set S. Let s_{ij} be the number of samples of class C_i in a subset S_j. The entropy, or expected information based on the partitioning into subsets by A, is given by

$$E(A) = \sum_{j=1}^{v} \frac{S_{ij} + \ldots + S_{mj}}{S} I\left(S_{1j} + \ldots + S_{mj}\right) \qquad (3.2)$$

The term $\frac{S_{ij} + \ldots + S_{mj}}{S}$ acts the weight of the j^{th} subset and is the number of samples in the subset divided by the total number of samples in S. Note that for a given subset S_j,

$$I(s_{1j}, S_{2j}, \ldots, S_{mj}) = -\sum_{i=1}^{m} p_{ij} \log_{2j} (P_{ij}) \qquad (3.3)$$

Where $p_{ij} = s_{ij} / | s_j |$ and is the probability that a sample in Sj belongs to class C_i.

The encoding information that would be gained by branching on A is

$$Gain(A) = S(s_1, s_2, \ldots s_m) - E(A). \qquad (3.4)$$

Gain (A) is the expected reduction in entropy caused by knowing the value of attribute A. The attribute with the highest information gain is chosen as the test attribute for given set of S. A node is created and labeled with the attribute, branches are created for each value of the attribute, and the samples are partitioned accordingly.

3.4. Tree pruning

When a decision tree is built, many of the branches will reflect anomalies in the training data due to noise or outliers. Tree pruning methods are used to solve this problem that is over fitting the data. Such methods typically use statistical measures to remove the least reliable branches, generally resulting in faster classification and an improvement in the ability of the tree to correctly classify independent test data.

There are two common approaches to tree pruning. In the prepruning approach, a tree is pruned by halting its construction early. Upon halting, the node becomes a leaf. The leaf may hold the most frequent class among the subset samples or the probability distribution of those samples.

The second approach, postpruning, removes branches from "fully grown" tree. A tree node is pruned by removing its branches. In this system, postpruning method is used when the tree is fully grown.

3.5. Classifier accuracy

Estimating classifier accuracy is important in that it allows one to evaluate how accurately a given classifier will label feature data, that is, the data on which the classifier has not been trained. An accuracy method, Hold-out, is used to evaluate the method's accuracy.

$$\text{Accuracy} = sensitivity\frac{\text{pos}}{(\text{pos}+\text{neg})} + specifity\frac{neg}{(pos+neg)} \quad (3.5)$$

Where, pos is the number of positive samples, neg is the number of negative samples, and sensitivity can be calculate,

$$sensitivity = \frac{t\_pos}{pos}, \quad (3.6)$$

specifity can be calculate,

$$specifity = \frac{t\_neg}{neg} \quad (3.7)$$

and the precision can be calculate as follows

$$precision = \frac{t\_pos}{(t\_pos + f\_pos)} \quad (3.8)$$

Where t_pos is the number of true positives, t_neg is the number of true negatives and f_pos is the number of false positives.

4. Proposed system framework

4.1. Decision tree implements from dry zone data set

In this Dry Zone Plantation system, there are 12 classes and twelve attributes. The consists of twelve attributes are stony, composed, land, sandy clay, rain, zone, sandy lome, rain, cinnamon, sandy clay, sunlight and red brown and twelve classes are Kok-ko, Kab-wi, Kyatsu, Tha-naung, Thit-si, Thi, Tha-dut, Htin-shoe, Thit-cho, Tamar, Than and Ngu. So number of classes (m=12). Let class C_1 correspond to class Than, C_2 correspond to class Kyatsu, C_3 correspond to class Htin-shoe, C_4 correspond to class Kokko, C_5 correspond to class Tamar and so on. S_i means that total count of each class from training data set and S is total count of training data set. To compute all classes of expected information of each attributes and the first uses in Equation (3.1). Next, it needs to compute the entropy of each attribute when class label is not the same class. Let's start with the attribute Stony; it can need to look at the distribution of 12 class samples for each value of Stony. And then, compute the expected information of Stony attribute for each of these distributions by using Equation (3.2). The expected information is achieved by classifying a given sample if the samples are partitioned according to Stony attribute value. To compute information gain of Stony attribute, it can use Equation (3.4) of information Gain. The rest of attributes are also computed like this.

Sunlight attribute have the highest information gain among the attributes, it is selected as the test attribute. A node is created and labeled with Sunlight and branches are grown for each of the attribute's value. The sample are then partitioned accordingly each of the Sunlight attribute's value, 80F-110F, 60F-80F and 46F-60F. A sub class, Dry Zone, is classified into derived classes according to Magwe, Mandalay. And then, the next sub class, Cinnamon, is also too classified into derived classes according to between Yes or No conditions. And then, the next sub classes are continued with recursive each partition. Figure 1 display some part of tree of the dry zone data set,

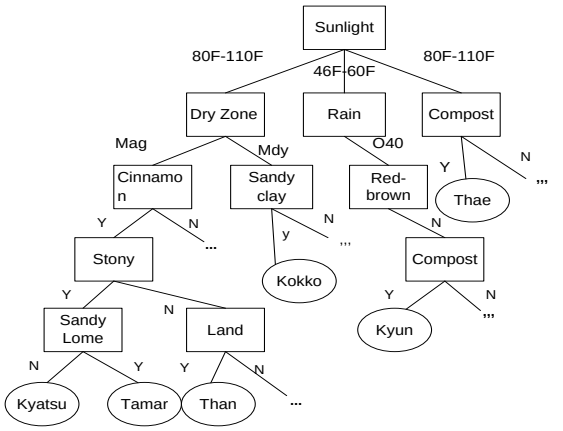

Figure1. An example of decision tree for dry zone plantation

4.2. Extracting classification rules from decision tree

The knowledge represented in decision trees can be extracted and represented in the form of classification IF-THEN rules. One rule is created for each path from the root to a leaf node. Each attribute value pair along a given path forms a conjunction in the rule antecedent ("IF" part). The leaf node holds the class prediction, forming the rule consequent ("THEN" part). The IF-THEN rules may be easier for humans to understand, particularly if the given tree is very large. This system produces 28 rules for sample dry zone data set. Some parts of classification rule for Dry Zone Plantation is as follows:

IF Sunlight = "80F to 110F" AND Dry Zone = "Sag" AND Cinnamon = "Y" AND Stony = "Y" AND Red Brown = "N" THEN "Kokko"
ELSE IF Sunlight = "80F to 110F" AND Dry Zone = "Mag" AND Cinnamon = "Y" AND Stony = "Y" AND RedBrown= "Y" THEN "Kyetsu"
ELSE IF Sunlight = "60F to 80F" AND Dry Zone = "Sag" Compost = "Y" AND Cinnamon = "Y" AND Land = "N" AND SandyClay = "N" THEN "Pyinkado"
ELSE IF Sunlight = "60F to 80F" AND Dry Zone = "Sag" AND Compost = "Y" AND Cinnamon = "Y" AND Stony = "Y" AND Rain = "Between 20 and 40" THEN "Thi"
ELSE IF Sunlight = "60F to 80F" AND Dry Zone = "Mag" AND RedBrown = "N" AND Stony = "Y" AND Land = "Y" AND Compost = "Y" THEN "Sha"
ELSE IF Sunlight = "46F to 60F" AND Dry Zone = "Mdy" AND Stony = "N" AND RedBrown = "N" AND Cinnamon = "Y" AND SandyClay = "N" THEN "Dahat"

4.3. Classifier accuracy

The size of data set used in this system is 299 of Dry Zone data sets. This system is also provides 98.99% accuracy after 99 testing datasets is implied with the rules generated by training data sets, 196 items.

5. Design of system

The system data is divide two-third for training data and one- third for testing data. From training data, a classification schema is derived by decision tree induction algorithm. The system shows growing plants class from decision tree with relevant member information. There are twelve attributes and twelve classes in the system. The algorithm takes Dry Zone Plantation data sets and produces class label as output.

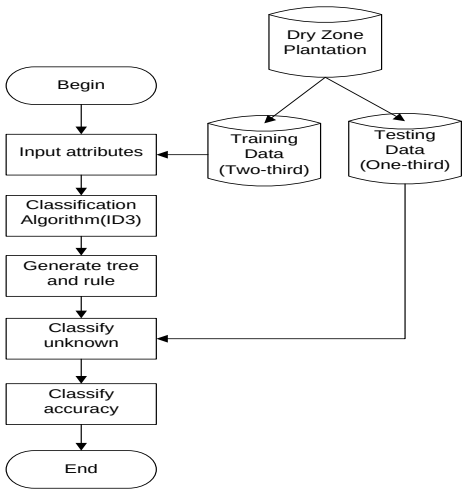

Figure 2. The overview of the system

Firstly, user entered data set is separated into training set and testing set. By using training set, ID3 is implemented to classify plantation system for dry zone. Then, the result rules and tree is generated according to iterative the process of ID3. Additionally, the unknown data included in testing data and it can also classify through the produced rules. Finally, the process of accuracy is used to classify testing data.

6. Experimental result

This system is experimented with classification algorithm, ID3 decision tree induction, using the plantation data from Dry Zone Plantation data sets. The main purpose of this system is to show the result of possible plants which are grown in the dry zone. This system uses 196 plants as training data and 99 plants as testing data for experiments. The output decision tree is a flow chart like tree structure, where each internal node denotes a test on an attribute, each branch represents on outcomes of the test, and leaf node represent classes. The top-most node in a decision tree is root node. An example part of decision tree is shown in Figure 3.

In order to classify an unknown sample, the attribute values of the sample are tested against a leaf node that holds the class prediction for that sample. If the system uses new input plants from user, the system will produce to the user from Decision Tree output that is classified decision tree classifier. Decision tree induction can be easily converted to classification rules.

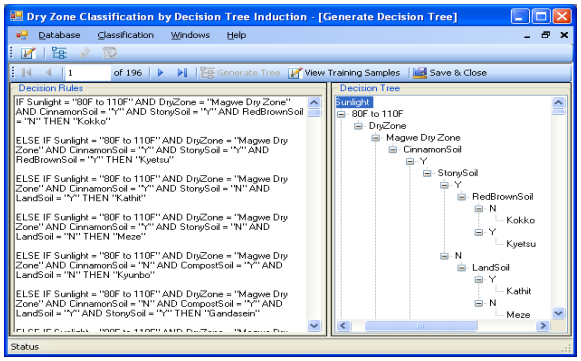

Figure 3. An example part of decision tree and rules

7. Conclusion

This paper is aimed to demonstrate that the technology for building decision trees from dry zone plantation is fairly robust. When using decision tree induction, the decision making process itself can be easily validated. This system utilizes twelve attributes in the dry zone plantation datasets and makes decision about new data or unknown data sets. This system is to predict the value of growing plants using the above attributes. In this paper, by using decision tree induction, it can get more accuracy and high performance. Finally, its gives a right decision whether or not to grow the tree on the basis of the conditions of Dry Zone plantation. As an extension, this system will be implemented with another method and k-fold accuracy method which will be used to measure the standard of classification methods.

8. References

[1]. D.J.Hand. "Construction and Assessment of Classification Rules", 1997.

[2]. G.Minos, H.Dongjoon, R.Rajeev and S.Kyuseok. *"Efficient Algorithms for Constructing Decision Trees with Constraints".*

[3]. J.R.Quinlan, "Induction of decision trees", Machine learning.

[4]. R.Rajeev and S.Kyuseok. "Public: A Decision Tree Classifier that Integrates Building and pruning". *In Proceedings of the 24th International Conference on Very Large Data Bases,* New York, USA, August 1998.

[5]. W. Peng, J. Chen and H. Zhou, "Implementation of ID3 -Decision Tree Learning Algorithm".