Developing Decision Tree Using ID3 for Paddy Classification

May Latt Cho, Moe Sanda Htun
Computer University, Loikaw
maylattcho@googlemail.com, moesdhtun@gmail.com

Abstract

Data Mining is the task of discovering interesting pattern from large amounts of data where the data can be stored in database, data warehouse. Data classification is the process of building a model from available data called the training data set and classifying objects according to their attributes. Decision tree algorithms have been used for classification in a wide range of application domains. The aim of this paper is to study about decision tree algorithm. This system is intended to develop the type of Myanmar's paddy data by using decision tree induction classification algorithm, Depending upon the data tuples of paddy dataset, the system can classify the type of paddy data whether it is good or bad quality and quantity.

Keyword: Data mining, classification, decision tree Induction.

1. Introduction

Paddy is our staple food. There is no person in Myanmar who does not eat the paddy. And then Myanmar is an agricultural country. This proposed system support for farmers and other researcher. This proposed system can be studied who interested in the quality and quantity of the paddy.

Classification is a well-studied important problem. It has many applications. The insurance industry, tax and credit card fraud detection and medical diagnosis and other application can use the classification method [1].

Data mining refers to extraction or "mining" knowledge from large amounts of data, also called "Knowledge mining". There are many other terms carrying a similar or slightly different meaning to data mining, such as knowledge mining from database, knowledge extraction, data/pattern analysis, data archeology and data dredging. Data mining is all multidisciplinary field, drawing work from areas including database technology, artificial intelligence, machine learning, neural networks, statistics, pattern recognition, knowledge-based systems, knowledge acquisition, information retrieval, high-performance computing and data visualization. Data mining present the material database perspective. That is, issues relate to the feasibility, usefulness, efficiency and scalability of techniques for the discovery of patterns hidden in large database [2].

2. Classification

Classification is the process of finding a set of models (or functions) that describe the model and also distinguish data classes or concepts, for the purpose of being able to use the model to predict the class of object whose class label is unknown. The derive model is based on the analysis of a set of training data.

The derived model may be represented in various forms, such as classification (IF-THEN) rules, decision trees, mathematical formulae, or neural networks. Classification can be used for predicting the class label of data objects; user may wish to predict some mission or unavailable data values.

A model is built describing a predetermined set of data classes. The model is constructed by analyzing data tuples described by attributes. Each tuple is assumed to belong to a predefined class, as determined by one of the attributes, called the class label attribute in the context of classification, data tuples are also referred to as samples, examples or objects [3].

3. Decision Tree Induction

A decision tree is a flow chart like tree structure, where each internal node denotes a test on an attribute, each branch represents an outcome of the test, and lead nodes represent classes or class distribution. The top-most node in a tree is the root node. Internal nodes are denoted by rectangles, and leaf nodes are denoted by ovals. In order to classify an unknown sample, the attribute values of the sample are tested against the decision tree. A path is traced from the root to a lead node that holds the class prediction for the sample. Decision tree can easily be converted to classification rules. The attribute with the

highest information gain (or greatest entropy reduction) is chosen as the test attribute for the current node.

Let S be a set consisting of s data samples. The class label attribute has m distinct values defining in distinct classes, Ci (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

$$I(s1, s2, \ldots, sm) = -\Sigma P_i \log_2(P_i) \quad (1)$$

where p_i is the probability that an arbitrary sample belongs to class C_i and is estimated by s_i/s. A log function to the base 2 is used since the information is encoded in bits.

Let attributes A above v distinct values, $\{a_1, a_2, \ldots, a_v\}$. Attribute A can be used to partition S into v subsets, $\{S_1, S_2, \ldots S_v\}$, where S_j contains those samples in S that have value 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,

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

The term $\frac{s_{1j} + \ldots + s_{mj}}{s}$ acts as the weight of the jth subset and is the number of samples in the subset (i.e., having value a_j of A) divided by the total number of samples in S. The smaller the entropy value, the greater the purity of the subset partitions. Note that for a given subset S_j,

$$I(s_{1j}, s_{2j}, \ldots, s_{mj}) = -\sum_{i=1}^{m} P_{ij} \log_2(p_{ij})$$

where $P_{ij} = \frac{S_{ij}}{|S_j|}$ and is the probability that a sample in S_j belongs to class C_i. The encoding information that world be gained by branching on A is

$$\text{Gain}(A) = I(s_1, s_2, \ldots, s_m) - E(A) \quad (3)$$

In others words, Gain (A) is the expected reduction in entropy caused by knowing the value of attribute A.

The algorithm computes the information gain of each attribute. The attribute with the highest information gain is chosen as the test attribute for the given set 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]

3.1. General Decision Tree Construction Template

In decision tree classification, the training data set is recursively partitioned until the records in the sub-partitions are entirely or mostly from the same classes [1].

The following is a general template for all decision tree classification algorithms:

1. Partition (Dataset S)
2. If (all records in S are of the same class) then return;
3. Compute the splits for each attribute;
4. Choose the best split to partition S into S_1 and S_2;
5. Partition (S_1);
6. Partition (S_2);

3.2. Decision Tree Works

In classification techniques, the decision tree classifier is widely used. The tree has three types of nodes.

A node that has no incoming edges and zero or more outgoing edges is called a root node. Leaf nodes, each of which has exactly one incoming edge and no outgoing edges, also called terminal node.

In a decision tree, each leaf node is assigned a class label. The non-terminal nodes contain root and other internal nodes; include test attribute conditions to separate records that have different characteristics. Classifying a test record is straightforward once a decision tree has been generated. The test conditions is started from the root node, and applies to the record and follow the appropriate branch based on the outcome of the test. This will lead to another internal node, for which a new condition is applied, or to a leaf node. This class label associated with a leaf node is then assigned to the record [4].

4. Attribute Information

In this training data set, there are ten attributes and two classes. The following table described name of attributes and description of these attributes.

Table 1. Name and Description of attribute.

Attribute Name	Description
Variety Name	AyeYarMin, LoneThweHmway, SinAKaYee, YeSin 1, …….
Origin	Malaysia, Thailand, Myanmar,…..
Variety Type	EaeMaHta, MeiDone, Selection, MeiDone
Life Day	2.5, 3.5, 4, 4.5, 5
Plant Height	2.0-2.5, 2.5-3.0, 3.0-3.5, 3.5-4.0, 4.0-4.5, 4.5-5.0
Grains Per Panicle	65, 85, 93, 95, 101,104, 105,…..
Milling Recovery	35-44, 45-54, 55-59, 60-64
Amylose	17-21, 21-25, 25-29, 29-32
Grain Appearances	Clear, Not Clear, White Belly, White Midst
Yield	25-40, 40-50, 50-70, 70-100, 100-110

5. Extracting Classification Rule from Tree

Represent knowledge in the form of IF-THEN rules. One rule is created for each path from root to a leaf. Each (attribute, value) pair along a path forms a conjunction. Leaf node holds the class prediction [3]. Rules are easier for user to understand. In this system, fifty-two training data sets have been used to get that rules. Amylose is the highest information gain.

Rules for quality of paddy data by using Decision Tree Induction.

RULES (1) – IF Amylose 17-21, Grain Appearance= Clear THEN Decision=”Good”
RULES (2) – IF Amylose 17-21, Grain Appearance= White Midst THEN Decision=”Bad”
RULES (3)-IF Amylose 21-25, Origin=America THEN Decision=”Good”
RULES (4)-IF Amylose 21-25, Origin=Bengalardish THEN Decision=”Good”
RULES (5) – IF Amylose 21-25, Origin=India THEN Decision=”Good”
RULES (6) – IF Amylose 21-25, Origin=Malaysia THEN Decision=”Bad”
RULES (7) – IF Amylose 21-25, Origin=Myanmar, Plant Height= 3.5-4.0 THEN Decision=”Good”
RULES (8) – IF Amylose 21-25, Origin=Myanmar, Plant Height=4.0-4.5 THEN Decision=”Good”
RULES (9) – IF Amylose 21-25, Origin=Myanmar, Plant Height=4.5-5.0 THEN Decision=”Good”
RULES (10) – IF Amylose 21-25, Origin=Myanmar, Plant Height=5.0-5.5 THEN Decision=”Good”
RULES (11) – IF Amylose 21-25, Origin=Myanmar, Plant Height=5.5-6.0, Yield= 40-50 THEN Decision=”Good”
RULES (12) – IF Amylose 21-25, Origin=Myanmar, Plant Height=5.5-6.0, Yield=50-70 THEN Decision=”Bad’
RULES (13) – IF Amylose 21-25, Origin=Myanmar, Plant Height=5.8-6.0 THEN Decision=”Good”
RULES (14) – IF Amylose 21-25, Origin=Phillipine, THEN Decision=”Good”
RULES (15) – IF Amylose 21-25, Origin=Thailand, THEN Decision=”Good”
RULES (16) – IF Amylose=25-29, Yield=100-110 THEN Decision=”Good”
RULES (17) – IF Amylose=25-29, Yield=40-50 THEN Decision=”Good”
RULES (18) – IF Amylose=25-29, Yield=50-70 THEN Decision=”Bad’
RULES (19) – IF Amylose=25-29, Yield=70-100 THEN Decision=”Bad”
RULES (20) – IF Amylose=29-32 THEN Decision=”Bad”

For example, user want to know the name of Ye Baw Sain’s quality is good or bad. Its origin is Myanmar, Amylose is between 21 and 25, and Plant height is 5.0 to 5.5 feet. And then, checking this testing data with rule. Result is equal to rule (10). So, Ye Baw Sain is good quality for eat.

6. Implementation

This system can create the paddy dataset with attributes values containing user defined values. The user can choice quantity or quality for classifies the paddy data. The paddy dataset contain ten attributes and two classes. This attributes include variety name, origin, variety

type, life day, plant height, grain per panicle, milling recovery, amylose, grain appearances and yield of the paddy data set and two classes are quality and quantity. Then, the system compute the information gain of each attribute by using equation1 and compute the entropy or expected information of each attribute by using equation2. By using equation 3, the highest information gain among the attribute is selected and created as root node. Finally, the system can generate the decision tree by using decision tree induction algorithm.

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 [5].The decision tree converts the understandable rules for the user. After generating the rules, the system can classify the type paddy data which quality and quantity whether good or bad.

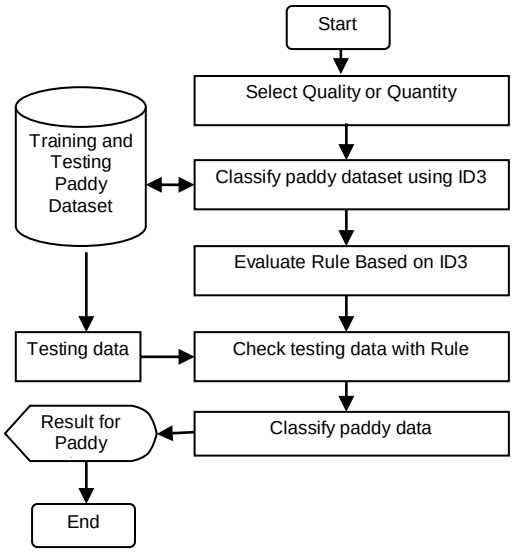

Figure 1: System Flow Diagram

Firstly, user has to choose the quantity or quality of paddy data. Training and testing paddy data are already existed in the database. Secondly, these paddy dataset are classified by ID3 algorithm. And then, calculated rules based on ID3 algorithm. The user will check with testing data and rules. Finally, results of paddy data will come out.

7. Conclusion

This proposed system is intended to know about how to select and plant available on the region and weather knowing Genetical characters and Agronomical Characters of paddy, by using decision tree induction algorithm .The knowledge represented in decision trees can be extracted and represented in the form of classification IF-THEN rules.

The aim of this paper is to study the concepts of classification under data mining and to know how to generate the decision rules by using decision tree induction algorithm. This paper presents how to classify the type of paddy data are whether good or bad quality and quantity. The user can classify the type of paddy data from the data set in a short time. So this system provides for the researcher who wants to learn the type of paddy.

REFERENCES

[1] A.seime, “Web Mining; Application and Techniques”, State University of New York College at Brockport, USA.

[2] H.Lu.R.Setino, and H.Liu, “Neurorule: A connectionist approach to data mining” VLDB, Swizerland, 1995.

[3] J. Ham and M. Kamber, "**Data Mining Concepts and Techniques**', Morgan Kaufmann, 2001.

[4] L.D.Radet, “Principles of Data Mining and Knowledge Discovery”, ISBN 3540425349, Oct 1, 2001 by Springer.

[5]http://decisiontrees.net