Developing Decision Making System to Classify Pests of Paddy Infesting

Su Mon Aung, Tin Tin Htwe
Computer University, Pathein, Myanmar
sumonaung.wka @gmail.com, htwe14@gmail.com

Abstract

Decision tree learning algorithms have been successfully used in knowledge discovery. These algorithms use induction in order to provide an appropriate classification of objects in terms of their attributes, inferring decision tree rules. There are many problems in cultivated lands because pests destroy paddy. Hence, we develop decision making system to classify the kinds of pest using Iterative Dichotomiser 3 (ID3). ID3's output is easy to read by computer officers without having previous knowledge about classification techniques. The main task performed in this system is using ID3 decision tree induction methods to the given values of attributes of an unknown object to determine appropriate classification according to decision tree rules. So, this system solves the problems in their cultivated lands about pests without helping of agriculture experts.

Keywords: Data Mining, Classification, Decision Tree Learning, Iterative Dichotomiser3 (ID3)

1. Introduction

The area of data mining in many applications is growing rapidly because of strong need for analyzing the vast amount of database stored data related with these applications. Data mining [1] are primarily database_oriented, designed for the efficient handling of huge amounts of data that are typically multidimensional and possibly of various complex types. Data mining are referred to as Knowledge Discovery Database (KDD) and data mining is the process of using tools such as classification, association rule mining, clustering, and etc [2]. This system classifies the kinds of pest using decision tree induction which is one method of classification of data mining techniques. Decision tree induction is one of the most popular algorithms in the classification. Decision trees are widely used in patern_recognition, machine learning, and data mining applications [3]. This paper examines decision tree learning algorithm ID3 and implement by using Java Programming language. By using this system, it makes to increase paddy yield for farmers and reduces the time that calling agriculture experts. The remaining of the paper is organized as follows: Section 2 includes Related Work in Many Applications; Section 3 is Classification and Prediction. Section 4 includes Decision Tree Algorithm. Section 5 is Explanation of the system and the last section is conclusion and further extension.

2. Related Work

This section deals with data mining & related researches and focuses on current research work on data mining. Madigan EA at al [4] employed the data mining approach CART (Classification And Regression Tree) to determine the drivers of home healthcare service outcomes (discharge destination and length of stay and examine the applicability of induction through data mining to home healthcare data. D.V.Chandra Shekar and V.Sesha Srinivas [1] researched that Myopia was again the most common refractive error. This study was to classify the sample data using the decision tree conducted using ID3 algorithm. "Wei Peng, University of New South Wales, School of Computer Science & Engineering, Sydney, NSW 2032, Australia" is used ID3 method to decide play tennis or not [5]. He used two classes: "Positive" and "Negative", four attributes, "Outlook", "Temperature", "Humidity", and "Windy". Victor H.Garcia, Raul Monroy, and Maricela Quintana presented web attack detection using ID3 algorithm [6]. They studied Intrusion Detection System (IDS) using ID3 algorithm based on the type of detection: (1) SQL Injection; (2) Cross Site Scripting (XSS); (3) Code Injection and (4) Directory Traversal. Pabitra Mitra, Sushimita Mitra, Member , IEEE, and Sankar k.Pal, fellow, IEEE [8] applied data mining approach soft computing methods to classify the stages of Cervical Cancer and compared accuracy that evaluated from using soft computing methods such as rough set theory , and ID3 algorithm. This system classifies the order of pests by using ID3 classifier and gives kinds of pest that are related as pest information.

3. Classification and Prediction

Classification and Prediction is an important technique in data mining. Classification is the process of finding a set of models that describes and distinguishes data classes or concepts, for the purpose of being able to use the model to predict the class of objects whose class label is unknown. The derived model is based on the analysis of a set of training data. Classification can be used for predicting the class label of data objects. When the predicted values are numerical data and is often referred to as prediction.

Classification and prediction are two forms of data analysis that can be used to extract models describing important data classes or to predict future data trends. Whereas classification predicts categorical labels, prediction models continuous_valued function. The use of prediction to predict class label is classification. Basic techniques for data classification are decision tree induction, Bayesian classification and Bayesian belief networks and Neural network. Decision tree induction method is used in this system because the outputs of decision tree induction are easily understood by humans.

4. Decision Tree Induction

In machine learning, a decision tree is a predictive model; that is, a mapping observation about an item to conclusions about the item's target attribute. The machine learning technique for including a decision tree from inducing a decision tree learning or decision trees. Decision tree algorithms [3, 9, 10] represent a widely used family of machine learning algorithms for building pattern classifier from labeled training data. Decision trees are powerful and popular tools for classification and prediction. Decision tree induction is the best algorithm to generate rules because of human understood than others.

Decision trees [11] classify instances by sorting them down the tree from the root to some leaf node which provides the classification of the instances. Decision tree is a classifier in the form of a tree structure, where each node is either: a leaf node indicates the value of the target attribute of instances and a decision node specifies some test to be carried out on a single attribute_value, with one branch and sub tree for each positive outcome of the test.

The key requirements to do mining with decision tree are:

(1) Attribute : Value description;
(2) Predefined Classes : category to which instances are assigned;
(3) Discrete Classes : A case does not belongs to a particular class;
(4) Sufficient Data: Usually hundred and thousands of training sets;

4.1. Iterative Dichotomiser 3 (ID3) Method

ID3 is a simple inductive, non_ incremental classification algorithm. Using a top down, greedy search through a fixed set of instances, it builds a decision tree, which is then applied for classifying samples. Each sample has several attributes and belongs to a class. Each node of the decision tree is a decision node, which each leaf node corresponds to a class name.

When deciding attribute to make decision node is the best, ID3 uses a measure called information gain. Information gain referred to as Attribute Selection Measure.ID3 operates only on examples described by the same attributes. Attributes must take values from a fixed, finite set. ID3 is not tolerant to noisy or missing attributes and classes must be sharply defined.

4.1.1. Top Down Induction of Decision trees: ID3

(1) Compute information gain of all attributes
(2) A← the best decision attribute for node
(3) Assign A as decision attribute for node
(4) For each value of A create new descendant
(5) Sort training samples to leaf node according to the attribute value of the branch
(6) If all training samples are perfectly classified (same value of target attribute), stop
(7) else iterate over new leaf nodes

4.1.2. Attribute Selection Measure

ID3 uses attribute selection measure to select which attribute at each node in the tree. Firstly, the entropy of total dataset is evaluated. If the target attribute takes on c different value, then the entropy S relative to the c values is defined as

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

Where S = a set of consisting data samples,

S_i = number of consisting of S in class c_i,

p_i = the probability of S belonging to class i

Let attribute A have v distant values $\{a_1, a_2, \ldots\ldots\ldots\ldots, a_v\}$. Attribute A can be used to partition S into v subset, $\{s_1, s_2, \ldots\ldots\ldots\ldots, s_m\}$ where s_j contained those samples in S that value of a_j of A. The entropy based on the partitioning into subsets of A is given by

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

The information gain of attribute A is defined as

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

Then, the algorithm computes information gain of each attribute. The attribute with the attribute with the highest information gain is chosen as the test attribute for the given set. A node is created and labeled with the test attribute. Below node, branches are created for each value of the test attribute and samples are partitioned until leaf node is reached

5. System Design

In Figure 3, the required data are collected from the agriculture experts store in dataset. And then, two third of dataset is used for training data and remaining one third is used for testing data. Training data is applied ID3 algorithm to generate decision tree for this system. Testing data are used to evaluate accuracy of this system. New user who used this system chooses data of all attributes firstly. The selected data is compared decision tree that appeared from applying ID3 algorithm and produces result (pest order) to the user. The user retrieve kinds of pest in pest information form as shown in Figure 2.

5.1. Explanation of the System

In this system, there are two major parts in this system. First is classification and prediction function. Second is to support not only the user and farmers but also the people which are interesting about agriculture with pest information. In the decision support system, the system needs to be trained before it is practically used. Therefore, ten attributes and one class label are used in this system. In this system, all attributes are nominal (categorical) and there are 4 classes (Coleoptera, Hemiptera, lepidotera, Thysanoptera). All attributes and their values description are shown in Table 1.

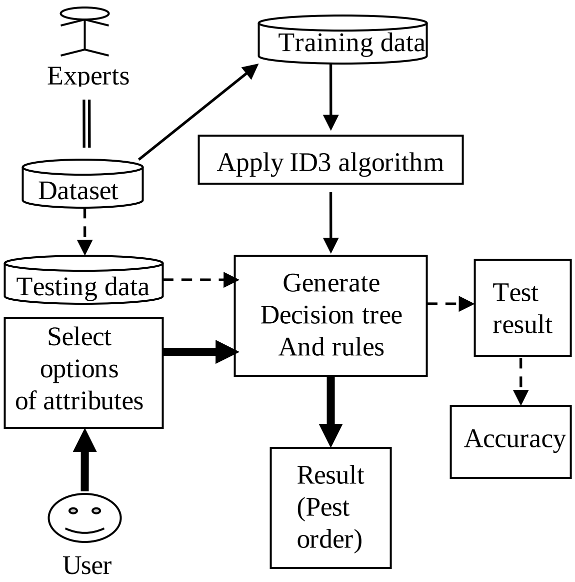


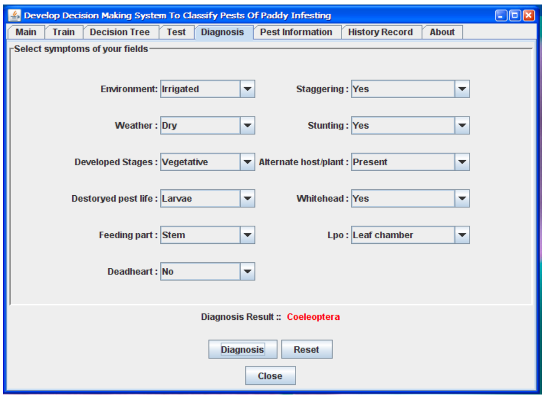
Figure 3. System Design

Table 1. Attributes and Their Values

Attribute Name	Values Description
Environment	Rainfed, Irrigated, Aquatic, Wetland, Dryland, Upland, Non_floaded
Weather	Dry, Rainy, Winter
Developed Stages	Seedling,Vegetative, Milking,Tillering, Reproductive,Stem elongation, Panicle initiation
Destroyed pest life	Nymph,Adult,Larvae
Feeding part	Leave, Seed, Stem, Root, Plant sap
Deadheart	Yes, No
Stunting	Yes, No
Staggering	Yes, No
Alternate host/plant	Present, Absent
Whitehead	Yes, No
Location of pupa occurred	Ground Causeway, Stubble, Grassy Area, Tassel, Folded leaf, Rolled leaf, leaf chamber, Outer of swarmed leaf

6. Experimental Result

The user can choose any data of attributes from Diagnosis Form. After selecting data of each Figure 1. If the user want to retrieve kinds of pest information that are related with the pest order, the user is going to Pest Information Form and read information by clicking pest order radio button as shown in Figure 2.

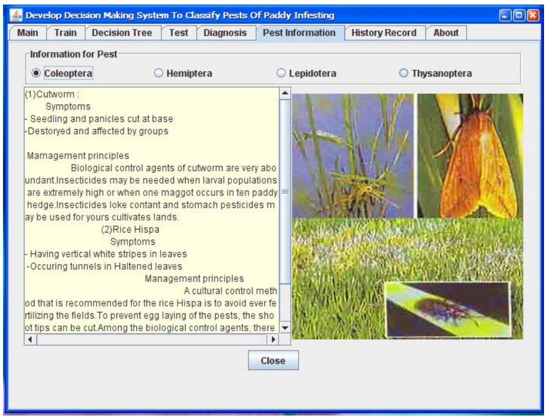


Select symptoms of your fields

Environment:	Irrigated	Staggering :	Yes
Weather :	Dry	Stunting :	Yes
Developed Stages :	Vegetative	Alternate host/plant :	Present
Destoryed pest life :	Larvae	Whitehead :	Yes
Feeding part :	Stem	Lpo :	Leaf chamber
Deadheart :	No		

Diagnosis Result :: Coeleoptera

Figure 1. Diagnosis Form



Information for Pest

● Coleoptera ○ Hemiptera ○ Lepidotera ○ Thysanoptera

```
(1)Cutworm :
      Symptoms
- Seedling and panicles cut at base
-Destoryed and affected by groups

 Mamagement principles
              Biological control agents of cutworm are very abo
undant.Insecticides may be needed when larval populations
 are extremely high or when one maggot occurs in ten paddy
 hedge.Insecticides loke contant and stomach pesticides m
ay be used for yours cultivates lands.
                        (2)Rice Hispa
                           Symptoms
- Having vertical white stripes in leaves
 -Occuring tunnels in Haltened leaves
                                      Management principles
                                                  A cultural control meth
od that is recommended for the rice Hispa is to avoid ever fe
rtilizing the fields.To prevent egg laying of the pests, the sho
ot tips can be cut.Among the biological control agents, there
```

Figure 2. Pest Information

6. Conclusion and Further Extension

The aims of this system are to learn about decision tree induction and to support the farmers at their cultivated lands. As paddy is the main staple food for ours, paddy agriculture is the main task of ours and pest detection is very important. The user can know order type accurately because decision tree predicts more than 95% of accuracy of process. Therefore, the results are generated by the system will give convenient answer to user. This system saves time consuming.

In this system, we develop decision making system to classify pests of paddy infesting. We are going to decide kinds of pests infesting in other crops instead of paddy using ID3 method.

7. References

[1] D.V.Chandra Shekar and V.Sesha srinivas, Clinical Data Mining, "An approach For Identification of Refractive Errors".

[2] Jiawei Han, Micheline Kamber, Data Mining, Concepts and Techniques.

[3] Quinlan. Induction on Decision Trees. Machine learning, 1(1):81_106, 1986.

[4] Madigan EA Curet OL _ A data mining approach in home healthcare _ outcomes and services use, 2006 Feb 24, 6:18.

[5] Wei pang, "An implementation of ID3 decision tree learning algorithm", Juhua Clean and Haiping Zhou Project of Camp 9417: Machine Learning university of New South Wales, School of Computer Science & Engineering, Sydney, NSW 2032, Australia .

[6] Victor H.Garcia, Roul Monroy, and Maricela Quintana, "Web Attack Detection Using ID3", Computer Science Department, Tecnologico de Monterrey, Campus Estado de Mexico Carretera al lago de Guadulape, km3.5, Atizapan, 52926 Mexico.

[7] Arbitral Mitra, Sushmita Mitra, Member IEEE, and Sankar k.Pal, Fellow, IEEE, "Staging of Cervical Cancer with Soft Computing".

[8] L.Breiman, J.Friedman, R.Olshen, C.Stone, Classification and Regression Trees, Wadsworth, Pacific Grove, CA, 1984.

[9] A.Buja, Y.Lee, Data Mining Criteria for Tree_Based Regression and Classification, 2000.

[10] Tom Mitchell, Mc Graw Hill, "Machine Learning", 1977.

[11] Anand Bahety, Department of Computer Science, University of Maryland, College Park, "Extension and Evaluation of ID3_ Decision Tree Algorithm".

[12] Minos.G Dongjoon.H, Rajeev R. and Kyuseok S, "Efficient Algorithm for Constructing Decision Tree with Constraints".