Implementation of Job Classification for Accounting Field using Decision Tree Algorithm

Khin Lay Nwe Oo, Yuzana
University of Computer Studies, Yangon, Myanmar
khinlaynweoo88@gmail.com , yuzana.yzn@gmail.com

Abstract

Data mining is seen as an increasingly important tool by modern business to transform data into an informational advantage. Data mining could also be described as trying to create a simplified model of the complex world described in the database. Data mining is a way of dealing with large amounts of information, and it is helpful for finding useful information faster than any human. Decision tree is mainly used for classification purposes. Decision tree is a classifier in the form of a tree structure. Rules can be easily extracted from the decision tree. The main task performed in this system is using inductive methods to the given values of attributes of an unknown object to determine appropriate classification according to decision tree rules. This paper examines the decision tree learning algorithm for classifying job related accounting field. This paper implements the decision tree using ID3 and gives advice to users about the types of job in accounting field. This system uses 900 training data set and 300 testing data set. This paper calculates the system accuracy by using Hold_Out Method and provides 84.75% after reviewing 300 testing data set.

Keywords: Job Classification, Decision Tree Induction, ID3 algorithm.

1. Introduction

Decision trees classify instances by sorting them down the tree from the root to some leaf node, which provides the classification of the instance [1]. Each node in the tree specifies a test of some attribute of the instance and each branch descending from that node corresponds to one of the possible values for this attribute. Decision tree are commonly used for gaining information for the purpose of decision-making. Decision tree starts with a root node on which it is for users to take actions. From this node, users split each node recursively according to decision tree learning algorithm. The final result is a decision tree in which each branch represents decision and outcomes. The input data of ID3 is known as sets of "training" or "learning" data instances, which will be used by the algorithm to generate the decision tree. The ID3 algorithm includes Classification Models, also called Decision Trees, from data. Each record has the same structure, consisting of a number of attribute /value pairs. Classification rules represent the classification knowledge as IF-THEN rules and are easier to understand for human users. A number of algorithms for induction decision trees have been proposed over the year ID3, C4.5, CART.

This paper examines the decision tree learning algorithm ID3. Test attributes are selected on the basis of Entropy and Information gain measure. Such a measure is referred to as an attribute selection measure or a measure of the goodness of split. Session 2 will be discussed the related work, data mining concept will be discussed in session 3, session 4 describes decision tree and ID3 algorithm, design and implementation of the system will be presented in session 5, session 6 will be described the conclusions of the system.

2. Related Work

Minos Garofalakis proposed that classification is an important problem in data mining. A number of popular classifiers construct decision trees to generate class models. The constructed trees are complex with hundreds of nodes and thus difficult to comprehend. Therefore, there is a problem of constructing simple decision trees with few nodes that are easy for humans to interpret. By permitting users to specify constraints on tree size or accuracy, and then building the best tree that satisfies the constraints, and graunteed that the final tree is both easy to understand and has good accuracy. 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. A decision tree induction is a systematic action plan for suggesting suitable classification to users. The implementation of this strategy is called a decision tree system. Decision trees have been applied to many different attributes Classification systems are necessarily more sophisticated because they must compare unlike objects. Some apply standard tools of information search, such as simple filters. More sophisticated

approaches have also been developed. Raskutti et al. (1997) introduced a recommender based on ID3 algorithm describes in [7]. Building classifiers that minimize testing costs has received much attention in the field of decision tree induction. However the problem is fundamentally different from the forensic classification problem. Several cost-sensitive algorithms have been proposed that building decision trees using non-incremental methods, such as a genetic algorithm proposed in [3]. Elkan shows a method to misclassification costs given classification probability estimates. Their systems compare pruning algorithms to minimize classification costs. As their methods act independently of the decision tree growing process, they can be incorporated with their algorithms in [5]. Integrating machine learning with program understanding is an active area of these systems. Systems that analyze root cause errors and systems that find bugs using decision tree inductions may both benefit from decision tree learning to decrease the pruning of the trees explained in [6].

3. Data Mining

Data mining is the process of extracting patterns from data. The database which the data mining system tries to extract knowledge from, is called the training set. By examining the data in this database, the system tries to create general rules and descriptions of the patterns and relations in the database. The goal is to gain knowledge which is valid not only in the specific database considered, but also for other similar data. A deductive system reasons about data by using a pre-defined set of rules. These rules limit how the information given to the deductive system may be used to draw conclusions and infer information. A correct deductive system is inferring information which is a logical consequence of the database contents.

4. Decision tree Induction

Decision tree is a classifier in the form of a tree structure, where each node is either leaf node or decision node. A leaf node indicates the value of the target attribute (class) of examples. A decision node specifies some test to be carried out on a single attribute-value, with one branch and sub-tree for each possible outcome of the test.

From this node, users split each node recursively according to decision tree learning algorithm. Among methods, decision tree learning is attractive for 3 reasons:

1. Decision tree is a good generalization for unobserved instance, only if the instances are described in terms of features that are correlated with the target concept. [1]
2. The methods are efficient in computation that is proportional to the number of observed training instances.
3. The resulting decision tree provides a representation of the concept those appeals to human because it renders the classification process self-evident. [4]

A decision tree is a tree in which each branch node represents a choice between a number of alternatives, and each leaf node represents a decision.

4.1. ID3 Algorithm

ID3 is a simple decision learning algorithm. ID3 constructs decision tree by employing a top-down, greedy search through the given sets of training data to test each attribute at every node. It uses statistical property call information gain to select which attribute to test at each node in the tree. Information gain measures how well a given attribute separates the training examples according to their target classification [2]. Figure 1 describes the ID3 algorithm for decision tree.

ID3 Algorithm for Decision Tree

ID3 (Examples, Target_Attribute, Attributes)

-Create a root note for the tree
-If all examples are positive Return the single-note tree Root, with label = -
If all examples are negative Return the single-note tree Root, with label = +
-If number of predicting attributes is empty, then Return the single node tree Root, with label = most common value of the target attribute in the examples
-Otherwise Begin

- A ← The Attribute that best classifies examples
- Decision Tree attribute for Root ← A
- For each positive value, vi, of A,
  - Add a new tree branch below Root, corresponding to the test A = vi
  - Let Examples(vi), be the subset of examples that have the value vi of A
  - If Examples(vi) is empty
    - Then below this new branch add a leaf node with label = most common target value in the examples
    - Else below this new branch add the subtree ID3 (Examples(vi), Target_Attribute, Attributes{A})

-End
-Return Root

Figure 1. ID3 Algorithm

4.2. Data Description

The sample data used by ID3 has certain requirements, which are: Attribute-value description - the same attributes must describe each example and have a fixed number of values. Predefined classes - an example's attributes must already be defined, that is, they are not learned by ID3. Discrete classes - classes must be sharply delineated. Continuous classes broken up into vague categories such as a metal being "hard, quite hard, flexible, soft, quite soft" are suspect. Sufficient examples - since inductive generalization is used (i.e. not provable) there must be enough test cases to distinguish valid patterns from chance occurrences.

4.3. Attribute Selection Measure

In information theory, entropy is a measure of the uncertainty about a source of messages. The more uncertain a receiver is about a source of messages, the more information that receiver will need in order to know what message has been sent. The one with the highest information (information being the most useful for classification) is selected. In order to define gain, Entropy concept is very important. A measure used from Information Theory in the ID3 algorithm and many others used in decision tree construction is that of Entropy. Informally, the entropy of a dataset can be considered to be how disordered it is. It has been shown that entropy is related to information, in the sense that the higher the entropy, or uncertainty, of some data, then the more information is required in order to completely describe that data. Entropy measures the amount of information in an attribute.

Given a collection S of c outcomes

$$Entropy(S) = S - p_i \log 2\ p_i \qquad (1)$$

In Equation (1), p_i is the proportion of S belonging to class I. Log2 is log base 2. Note that S is not an attribute but the entire sample set. First the entropy of the total dataset is calculated. The dataset is then split on the different attributes. The entropy for each branch is calculated. The resulting entropy is subtracted from the entropy before the split. The result is information Gain.

Example of a part of a decision tree is shown in Figure 2.

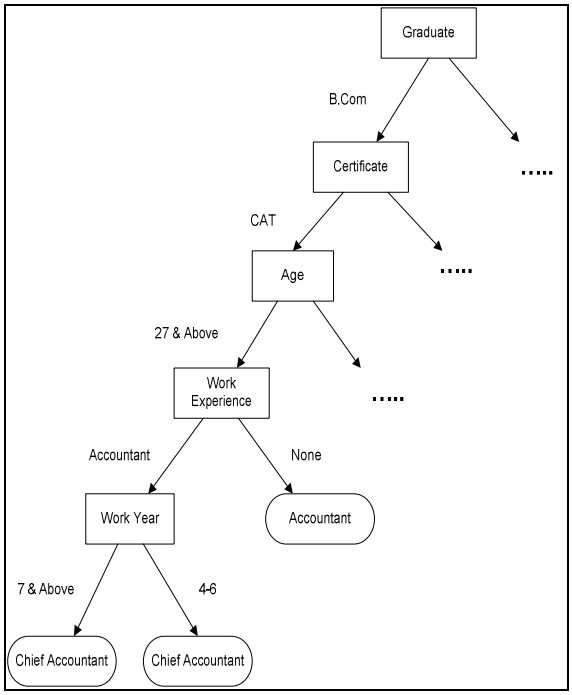

Figure 2. An Example of a part of decision tree

4.4. Extracting Classification Rules from Decision Tree

The knowledge represented in decision tree can be extracted and represented in the form of classification IF-THEN rules [3]. 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. Some examples of classification rule for job classification are following:

- If Graduate= B.Com and Certificate= CAT and Age=27&Above and Work Experience= Accountant and Work Year= 7&Above Then Job Type= Chief Accountant
- If Graduate=B.Com and Certificate=CAT and Age=27&Above and Work Experience= Accountant and Work Year= 4-6 Then Job Type=Chief Accountant
- If Graduate= B.Com and Certificate= CAT and Age=27&Above and Work Experience=None Then Job Type=Accountant

5. Implementation of the System

The process flow of the system is as shown in Figure 3. In this system there are two parts, admin and user. In admin part, the admin passes the log in. After passing the log in, the admin can insert, delete and edit the training data in training dataset and testing data in testing dataset and then check the accuracy result of this system. ID3 algorithm uses training dataset and then produces the rules and decision trees. These producing rules are stored in rule dataset. Using the testing dataset and rule dataset, the admin can calculate the system accuracy by using holdout method. In the user part, user input the job entry information and computes the final job result. There are five attributes and four classes in the system. The attributes of using this system are: Graduate, Certificate, Age, Past Work Experience and Past Work Year. The classes of this system are Accountant, Chief Accountant, Auditor and Chief Auditor. The algorithm takes user samples and produces class label as output and checking the classification results with accuracy testing.

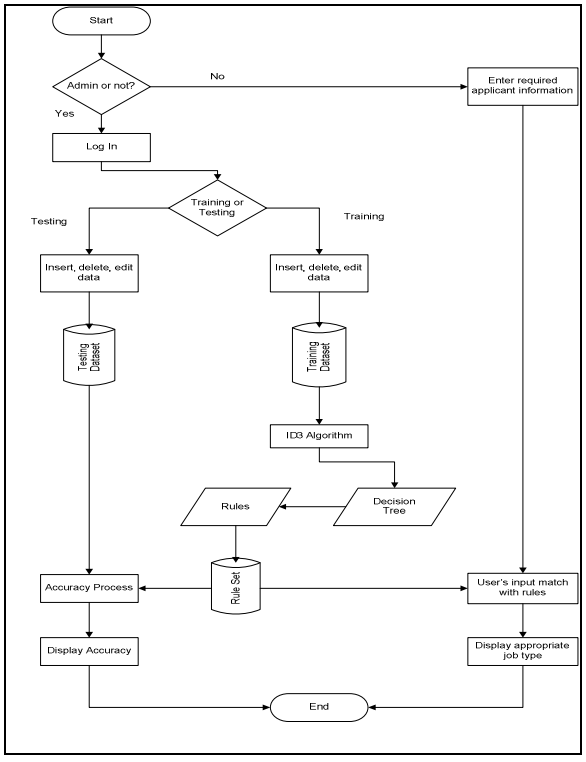

Figure 3. System Flow Diagram for this system

5.1. Classifier Accuracy

In classification problems, it is commonly assumed that all object are uniquely classification, i.e. that each training sample can belong to only one class. Classification accuracy is defined as the percentage of test examples correctly classified by the algorithm. Estimation of the true accuracy of a decision tree or rule model is one of the most important aspects of the modeling process. [7]

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. In this paper, the system accuracy will be calculated as follows:

$$Accuracy\ (\%) = \frac{No\ of\ correct\ job\ types}{No\ of\ total\ job\ types} \qquad (2)$$

In equation (2), the accuracy of the system is calculated by using correct job types (Accountant, Auditor, etc) match with rules by total job types (Accountant, Auditor) of testing data.

In the holdout method, the given data are randomly partitioned into two independent sets, a training set and testing set. Two thirds of the data are allocated to the training set, and remaining one third is allocated to the test set. The training set is used to derive the classifier, whose accuracy is estimated with the test set.

The system matched user criteria with the knowledge base that classification rule from manipulating decision tree induction algorithm and displayed user preference job type to user. If the system is used by the new user, the system has displayed user's appropriate job type to the user from decision tree output that is classified decision tree classifier. This system uses 900 user profiles as training data and 300 user profiles as testing data. The system also provides 84.75% accuracy after reviewing 300 testing data set. The system accuracy can also change as the rule changes or testing data changes.

6. Conclusion

Decision making process is applied human's intelligence in machines. When using decision tree induction, the decision making process itself can be easily validated. It intends to help the breeder and interesting person for making decision. Decision making process is applied in classifying job type. Job selection include many factors and difficult to decide accurately. This system focuses on developing architecture for giving appropriate job type by using ID3 method. It reduces time-consuming, cost, and uses easily without requiring much computer skill. In this paper, the system intends to be easily used by non-expert. The system intends for user to make a quick reference for decision making on the access job information.

References

[1] F. Berzal and Nicolas Marin, *Data Mining Concepts and Techniques*, Academic Press, USA, 2001.

[2] H.Lu.R.Setino, and H.Liu, "Neurorule: A connectionist approach to data mining", Journal of Machine Learning Research, Switzerland, 1995, pp 45-62.

[3] L.D.Radet, *Principle of Data Mining and Knowledge Discovery*, New York, USA, August 1998.

[4] H.Jerome. F. Richard, A.Olshen, and C. J.Stone, *Classifition and Regression Trees*, Chapman & Hall Academic Publisher, New York, 1984.

[5] H. Hamilton. E. Gurak, L. FindLater W. Olive, *Overview of Decision Tree*, Berlin, Germany, 1994.

[6] Kamber. J. Han, *Data Mining Concepts and Techniques*, Kluwer Academic Publishers, London GB, 1998.

[7] Qasem A. AI-Radaideh, Emad M. AI-Shawakfa, and Mustafa I. AI-Najjar, "Mining Student Data Using Decision Trees", In Proceedings of the 24th International Conference, Mumbai (Bombay), India, September 1996, pp. 131-137.