RESEARCH SCHOLARS EVALUATION BASED ON GUIDES VIEW USING ID3

Sathiyaraj.R[1], Sujatha.V[2]

[1, 2]*Assistant Professor, Department of CSE, MITS, SVCET,* ***sathiyarajr@mits.ac.in, suji4068@gmail.com***

Abstract

Research Scholars finds many problems in their Research and Development activities for the completion of their research work in universities. This paper gives a proficient way for analyzing the performance of Research Scholar based on guides and experts feedback. A dataset is formed using this information. The outcome class attribute will be in view of guides about the scholars. We apply decision tree algorithm ID3 on this dataset to construct the decision tree. Then the scholars can enter the testing data that has comprised with attribute values to get the view of guides for that testing dataset. Guidelines to the scholar can be provided by considering this constructed tree to improve their outcomes.

---***---

1. INTRODUCTION

Data mining an interdisciplinary subfield of computer science, is the computational process of discovering patterns in large data sets involving methods at the intersection of artificial intelligence, machine learning, statistics, and database systems. The overall goal of the data mining process is to extract information from a data set and transform it into an understandable structure for further use. The actual data mining task is the automatic or semi-automatic analysis of large quantities of data to extract previously unknown interesting patterns such as groups of data records (cluster analysis), unusual records (anomaly detection) and dependencies (association rule mining). This usually involves using database techniques such as spatial indices. These patterns can then be seen as a kind of summary of the input data, and may be used in further analysis or, for example, in machine learning and analytics. Data Mining can be used to solve many real time problems. Decision tree is an efficient method that can be used in classification of data. A decision tree is a decision support tool that uses a tree-like graph or model of decisions and their possible consequences, including chance event outcomes, resource costs, and utility. In this paper, we use decision tree algorithm ID3 for analyzing feedback given by guides. The training dataset consists of attributes such as Research proposal, Qualification, Experience, Way of Problem solving, Knowledge level, Interaction with guide, Journals published, Implementation of algorithm, Relating with real-life applications, Assessment, Subject knowledge, Punctual and Nature. The outcomes in the training dataset are specified with values like Excellent, Good, Poor and Average. The ID3 Algorithm can be applied on this training dataset to form a decision tree with view of guide as a leaf node. Whenever any research scholars provide testing data consisting of attribute values to the formed tree. Also, we can suggest the possible area where he/she has scope for improvement. This will help the scholar for self-evaluation and improvement where they lag.

The Next section describes about the decision tree algorithm and also defines entropy and gain ratio which are necessary concepts for constructing decision tree using ID3 and the next section by describing the problem statement and how we can analyze the dataset and evaluate the problem by using ID3 algorithm; finally, the conclusions and future works are outlined.

2. ID 3 ALGORITHM

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. 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 a possible scenario of decision and its outcome.

Decision tree learning is a method for approximating discrete-valued target functions, in which the learned function is represented by a decision tree.

ID3 is a simple decision learning algorithm developed by J. Ross Quinlan (1986) at the University of Sydney. ID3 is based off the Concept Learning System (CLS) algorithm. The basic CLS algorithm over a set of training instances C:

Step 1: If all instances in C are positive, then create YES node and halt.

If all instances in C are negative, create a NO node and halt.

Otherwise select a feature, F with values v1, ..., vn and create a decision node.

Step 2: Partition the training instances in C into subsets C1, C2, ..., Cn according to the values of V.

Step 3: apply the algorithm recursively to each of the sets Ci.

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. The algorithm uses a greedy search, that is, it picks the best attribute and never looks back to reconsider earlier choices.

2.1. Entropy

Entropy is a measure of the uncertainty in a random variable. Entropy is typically measured in bits, nats, or bans. It is a measure in the information theory, which characterizes the impurity of an arbitrary collection of examples. If the target attribute takes on c different values, then the entropy S relative to this c-wise classification. Entropy is formally defined as follows: If a data set S contains examples from m classes, then the Entropy(S) is defined as following:

$$Entropy(S) = -\sum_{j=1}^{m} P_j \log P_j$$

Where Pj is the probability of class j in S

Given a database state, D, Entropy (D) finds the amount of order in that state. When that state is split into s new states S = {D1, D2,…, Ds}, we can again look at the entropy of those states. Each step in ID3 chooses the state that orders spitting the most. A database state is completely ordered if all tuples in it are in the same class.

2.2. Information Gain

ID3 chooses the splitting attribute with the highest gain in information, where gain is defined as the difference between how much information is needed to make a correct classification before the split versus how much information is needed after the split. Certainly, the split should reduce the information needed by the largest amount. This is calculated by determining the difference between the entropies of the original dataset and the weighted sum of the entropies from each of the subdivided datasets. The entropies of the split datasets are weighted by the fraction of dataset being placed in that division. The ID3 algorithm calculates the Information Gain of a particular split by the following formula:

If attribute A is used to partition the data set S,

$$Gain(S, A) = Entropy(S) - \sum_{v \in A} \left(\frac{|S_v|}{|S|} * Entropy(S_v) \right)$$

Where, v represents any possible values of attribute A;
Sv is the subset of S for which attribute A has value v;
|Sv| is the number of elements in Sv;
|S| is the number of elements in S.

$$Entropy(S) = \sum_{i=1}^{c} -p_i \log_2 p_i$$

$$Gain(S, A) = Entropy(S) - \sum_{v \in Values(A)} \frac{|S_v|}{|S|} Entropy(S_v)$$

ID3 Algorithm for Decision Tree can be given as

ID3 (Examples, Target_Attribute, Attributes)

1. Create a root node for the tree
2. IF all examples are positive, Return the single-node tree Root, with label = +
3. If all examples are negative, Return the single-node tree Root, with label = -
4. 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
5. Otherwise Begin
   5.1 A ← The Attribute that best classifies examples
   5.2 Decision Tree attribute for Root ← A
   5.3 For each positive value, vi, of A,
      5.3.1 Add a new tree branch below Root, corresponding to the test A = vi
      5.3.2 Let Examples (vi), be the subset of examples that have the value vi for A
      5.3.3 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})
6. End
7. Return Root

The ID3 algorithm works by recursively applying the splitting procedure to each of the subsets produced until “pure” nodes are found—a pure node contains elements of only one class—or until there are no attributes left to consider.

3. DETERMINING IN-GENERAL-VIEW OF GUIDES ABOUT THE SCHOLARS

The problem that we are considering here is to determine the in-general-view of guides about the scholar. Based on the outcome value, we can suggest ways for scholars to improve. To achieve our goal, we use ID3 algorithm that is described in the previous section. First, we can have the training dataset containing following attributes:

Qualification, Experience, Way of Problem solving, Knowledge level, Interaction with guide, Journals published, Implementation of algorithm, Relating with real-life applications, Assessment, Punctual and Nature

i. Qualification (GRADUATE, POSTGRADUATE, DOCTORATE)
ii. Experience (LESS_THAN 2, 2-4, 4-8, 8-10, 10 ONWARDS)
iii. Way of Problem Solving (POOR, AVERAGE, GOOD, EXCELLENT)
iv. Knowledge level (POOR, AVERAGE, GOOD EXCELLENT)
v. Interaction with guide (POOR, AVERAGE GOOD, EXCELLENT)
vi. Journals published (LESS_THAN 1, 2-4, 5 ONWARDS)
vii. Implementation of algorithm (YES, NO)
viii. Relating with real-life applications (YES, NO)
ix. Assessment (YES, NO)
x. Subject Knowledge (POOR, AVERAGE, GOOD, EXCELLENT)
xi. Punctual (RARE, SOMETIMES, ALWAYS)
xii. Nature (COURTEOUS, RUDE, INDIFFERENT)

The outcome class is: In-general-view (POOR, GOOD, EXCELLENT). Here, we have converted the continuous attributes to the discrete/categorical attributes by considering the particular range as a class for simplicity and applicability of ID3 algorithm.

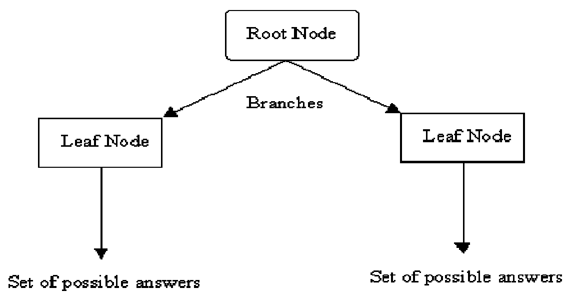

Figure 1: Decision Tree

4. DETERMINING GUIDES VIEW AND PROVIDING GUIDELINES TO SCHOLARS:

We solve the above mentioned problem using ID3 Algorithm. To solve this, a decision tree is formed by classifying the training data and then the outcome class value is determined.

The steps involved can be described as follows:

Decision Tree Construction: For each scholar registered in university, we can have collective feedback for the attributes enlisted in the problem statement. By using, ID3 algorithm, a decision tree is formed by classifying the training data and then the outcome class value is determined. The outcome class will be the leaf node of the tree and the attribute values will be the internal nodes and the arcs connecting the nodes are the decision trees made during the decision tree construction.

Determination of in-general-view about scholar: If the attribute values are provided, the decision tree formed after classification can be used to determine the outcome class, by traversing the tree using the attribute value. Scholars can provide the attribute values to the constructed tree and obtain outcome class value for self-evaluation.

Guidelines to the improvement of Scholars: Production rules can be directly obtained by traversing from root to the leves of the tree is the advantage of using decision tree. By using the production rules, we can provide the guidelines for the improvement of scholar.

For example: If a scholar gets ‘Poor’ as outcome class value due to less value for some attribute(s), we can also give the ways to get the outcome class value as ‘Excellent’, such as values for Regularity attribute should be ALWAYS instead of SOMETIMES, etc.

Thus the scholar can improve according to the guidelines. The outcome will be more accurate, when the training set is larger. If the training data set is too small, then it may not consider all the possibilities for the particular outcome and the result may not be accurate.

CONCLUSIONS AND FUTURE WORK

We conclude that ID3 Algorithm works well on classification problems. In this paper, we use decision tree algorithm to classify the dataset obtained from Guides feedback. We determine guides in-general-view about scholars and also provide guidelines to the scholars. This will be helpful for scholars to evaluate themselves and to improve accordingly. This will find its applicability in scholars’ assessment process. In future, we are trying to implement with software tools and we will assess the attribute values and calculate outcome class by getting input values from professors in universities.

REFERENCES

[1] Sonika Tiwari and Prof. Roopali Soni,' Horizontal partitioning ID3 algorithm A new approach of detecting network anomalies using decision tree', International Journal of Engineering Research & Technology (IJERT)Vol. 1 Issue 7, September – 2012.

[2] Anand Bahety,' Extension and Evaluation of ID3 – Decision Tree Algorithm' University of Maryland, College Park

[3] M. H. Dunham, DATA MINING: Introductory and Advanced Topics, Pearson Education, Sixth Impression, 2009.

[4] P. Ozer, Data Mining Algorithms for Classification, BSc Thesis, Artificial Intelligence, January 2008.

[5] A. Bahety, Extension and Evaluation of ID3 – Decision Tree Algorithm

[6] J. R. Quinlan, Simplifying decision trees, International Journal of Man-Machine Studies, 27, 221-234, 1987.

[7] P. H. Winston, Artificial Intelligence, Third Edition Addison- Wesley, 1992.