International Journal of Current Engineering and Technology
E-ISSN 2277 – 4106, P-ISSN 2347 - 5161

Available at http://inpressco.com/category/ijcet

Research Article

Analysing Uncertain Data by Building Decision Tree using ID3 Algorithm

Bhagyashree R. Patle[Ȧ*] and P. M. Chawan[Ȧ]

[Ȧ]Department of Computer Technology, VJTI, Mumbai, INDIA

Accepted 11 July 2014, Available online 01 Aug 2014, **Vol.4, No.4 (Aug 2014)**

Abstract

In Data mining, Classification of objects based on their features into pre-defined categories is a widely studied problem with rigorous applications in fraud detection, artificial intelligence methods and many other fields. Among the various classification algorithms available in literature the decision tree is one of the most practical and effective methods and uses inductive learning. In this paper we reviewed various decision tree for same dataset and we are mainly working on the ID3 algorithm.

Keywords: *Data mining, Decision Tree, Uncertain Data, Entropy*

1. Introduction

Data mining is an automated discovery process of nontrivial, previously unknown and potentially useful patterns embedded in databases. Research has shown that, data doubles every three years. Thus data mining has become an important tool to transform these data into information. The datasets in data mining applications are often large and so new classification techniques have been developed and are being developed to deal with millions of objects having perhaps dozens or even hundreds of attributes. Hence classifying these data sets becomes an important problem in data mining. Classification is the problem of automatically assigning an object to one of several pre-defined categories based on the attributes of the object. Classification is also known as supervised learning. In classification a given set of data records is divided into training and test data sets. The training data set is used to build the classification model, while the test data records are used in validating the model. The model is then used to classify and predict new set of data records different from both the training and test data sets. Some of the commonly used classification algorithms are neural networks, logistic regression and decision trees etc. Among these decision tree algorithms are most commonly used. Decision tree provides a modelling technique that is easy for humans to comprehend and is simplifies the classification process. This paper attempts to provide a detailed structure of decision tree using ID3 algorithm. It also gives ideas how to generate different decision tree by changing threshold value of entropy. In this paper we mentioned different data set to generate decision tree. It is organized as follows: Section 2 provides an overview on decision tree algorithms and different function to make a change in decision tree construction and its implementation patterns. Section 3 provides experimental analysis and different decision tree for same data set. Section 4 provides a summary and conclusions.

2. Description of Technical terms/Notations Used

Entropy is a measure of the number of random ways in which a system may be arranged. For a data set S containing n records the information entropy is defined as Entropy(S)=- $\sum P_I \log_2 P_I$.(Here P_I is the proportion of S belonging to class I.)

Gain or the expected information gain is the change in information entropy from a prior state to a state that takes some information, the information gain of example set S on attribute A is defined as

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

where Value(S)is the set of all possible values of attribute A, (S_v)is the subset of S for which attribute A has value v,$|S_v|$is the number of elements inS.

Gini index for a data set S is defined asgini(S)=$1 - \sum P_1^2$ and for a 2-split$gini_{split}(s) = \frac{n_1}{n}\ gini(S_1) + \frac{n_1}{n}\ gini(S_2)$andso on for a k-split.

Hunts method algorithm for decision tree construction trains the data set with recursive partition using depth first greedy technique, till all the record data sets belong to the class label.

3. Decision Tree Algorithm

Decision tree algorithm is a data mining induction techniques that recursively partitions a data setof records using depth-first greedy approach (Hunts et al, 1966) or breadth-first approach (Shaferet al, 1996) until all the data items belong to a particular class. A decision tree structure is made of root, internal and leaf nodes. The tree structure

*Corresponding author **Bhagyashree R. Patle** is a student and **P. M. Chawan** is working as Associate Professor

is used in classifying unknown data records. At each internal node of the tree, a decision of best split is made using impurity measures (Quinlan, 1993). The tree leaves is made up of the class labels which the data items have been group.

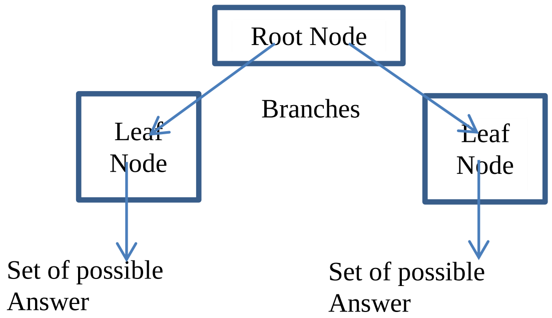

Fig1: Format of Decision Tree

Decision tree classification technique is performed in two phases: tree building and tree pruning. Tree building is done in top-down manner. It is during this phase that the tree is recursively partitioned till all the data items belong to the same class label (Hunts et al, 1966). It is very tasking and computationally intensive as the training data set is traversed repeatedly. Tree pruning is done is a bottom-up fashion. It is used to improve the prediction and classification accuracy of the algorithm by minimizing over-fitting (noise or much detail in the training data set) (Mehta et al, 1996). Over-fitting in decision tree algorithm results in misclassification error. Tree pruning is less tasking compared to the tree growth phase as the training data set is scanned only once. In this study we will review Decision tree algorithms implemented in a serial pattern, identify the algorithms commonly used and compare their classification accuracy and execution time by experimental analysis.

Uncertain Data

Uncertain data arises in many applications due faulty measurements, repeating process or missing values.

A. Faulty Measurement

In many instrument there is error of 2% in measuring values. For example whenever we are measuring body temperature through thermometer there is error of 0.2^{o} C. every time we get a slightly different reading. To find out exact temperature we take lots of reading and averaging them. Such type of error makes the data uncertain.

B. Repeating Process

If we take a survey of student learning in school, if we ask them how many hours they are studying? We get different answer from each student. We can conclude that the particular age group of student studying how many hours. This repetition of process for every student gives the uncertain data.

3. ID3 Algorithm

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 v_1... v_n 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.
Note, the trainer (the expert) decides which feature to select. ID3 improves on Concept Learning System by adding a feature selection heuristic. ID3 searches through the attributes of the training instances and extracts the attribute that best separates the given examples. If the attribute perfectly classifies the training sets then ID3 stops; otherwise it recursively operates on the n (where n = number of possible values of an attribute) partitioned subsets to get their best attribute. The algorithm uses a greedy search, that is, it picks the best attribute and never looks back to reconsider earlier choices.

4. Experimental Results

In this section, we present the experimental results of the proposed decision tree technique. We studied that; we can draw number of decision tree for same database on the basis of different attribute datasets.

Based on the ID3algorithm implemented on java applet, we implemented the DTUas described in Section 5. The experiments are executed on a PC with an Intel Dual-core processor and 2.0 GB main memory. A collection contains 5 real-world datasets. Wetried to draw all type of decision tree, likely split first node, last node or randomly select node. The 5 datasets are namely Iris, Tennis, German, Titanic and Zoo contains numerical as well as symbolic attributes.

The experiment gives simple as well as Detail structure of Decision tree. In detailed view we use 25 different colours to differentiate various attribute.

Data Set: Play Tennis

Table 1: Data Set: Play Tennis

Outlook	Temp	Humidity	Wind	Play
Sunny	Hot	High	False	No
Sunny	Hot	High	True	No
Overcast	Hot	High	Weak	Yes
Rain	Mild	High	Weak	Yes
Rain	Cool	Normal	False	Yes
Rain	Cool	Normal	True	No
Overcast	Cool	Normal	True	Yes
Sunny	Mild	High	False	No
Sunny	Cool	Normal	False	Yes
Rain	Mild	Normal	False	Yes
Sunny	Mild	Normal	True	Yes
Overcast	Mild	High	True	Yes
Overcast	Hot	Normal	False	Yes
Rain	Mild	High	True	No

Decision Tree

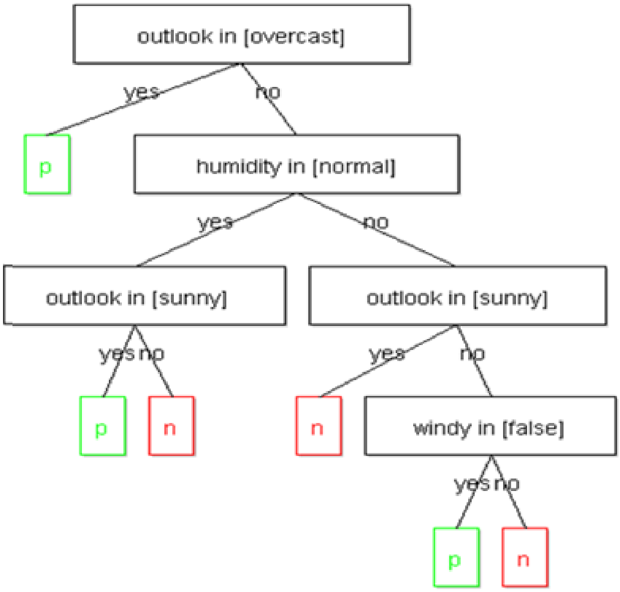

Fig. 2: Decision tree for data set

Entropy Calculation For above Graph

Table 2: Entropy Calculation For above Graph

node0	LS : n:4 p:7 total:11 - ENTROPY : 0.9456603046006401	Outlook-values:[overcast]-IG:0.31869694 Humidity-values:[high]-IG:0.0720566251 Temp-values:[mild]-IG:0.00343048854 Windy-values:[false]-IG:0.00343048854
Node1	LS : n:0 p:4 total:4 - ENTROPY : 0.0	Entropy low enough (threshold=0.0)a leaf
Node2	LS : n:4 p:3 total:7 - ENTROPY : 0.9852281360342511	Humidity-values:[normal]- IG: 0.12808527889139398 Temp-values:[hot]-IG: 0.12808527889139365 Outlook-values:[rain]-IG: 0.020244207153755633 Windy-values:[true]-IG: 0.020244207153755633
Node3	LS : n:1 p:2 total:3 - ENTROPY : 0.9182958340544896	Outlook-values:[sunny]-IG: 0.9182958340544896 Temp-values:[sunny]-IG: 0.2586291673878229 Windy-values:[false]-IG: 0.2586291673878229
Node4	LS : n:3 p:1 total:4 - ENTROPY : 0.8112781244591328	Outlook-values:[sunny]-IG: 0.31127812445913283 Temp-values:[hot]-IG 0.1225562481826566 windy-values:[false]-IG 0.1225562481826566
Node5	LS : n:0 p:2 total:2 - ENTROPY : 0.0	Entropy low enough(threshold=0.0)a leaf
Node6	LS : n:1 p:0 total:1 - ENTROPY : 0.0	Entropy low enough(threshold=0.0)a leaf
Node7	LS : n:2 p:0 total:2 - ENTROPY : 0.0	Entropy low enough(threshold=0.0)a leaf
Node8	LS : n:1 p:1 total:2 - ENTROPY : 1.0	windy-values:[false]-IG 1.0
Node9	LS : n:0 p:1 total:1 - ENTROPY : 0.0	Entropy low enough(threshold=0.0)a leaf
Node10	LS : n:1 p:0 total:1 - ENTROPY : 0.0	Entropy low enough(threshold=0.0)a leaf

The above table gives the calculation at each node of the decision tree. At node 0 means root node of the decision tree shows the following result: LS: n:4 p:7 total:11 it means that entry having overcast outlook is present or not. N denotes not present and P denotes present. Entropy of the node 0 is: 0.9456603046006401. Now we will calculate the information gain of the remaining attribute for selection of next split so that we get better result. For best selection we choose highest information gain. The node 0 gives the following Information gain values:

Outlook-values:[overcast]-IG: 0.31869694
Humidity-values:[high]-IG: 0.0720566251
Temp-values:[mild]-IG: 0.00343048854
Windy-values:[false]-IG: 0.00343048854

Here outlook having highest Information Gain so we will select overcast outlook for the split. In this way reaming all the nodes are splits. If the entropy is zero it means that node will not further divide. It means this is a leaf node.

Conclusion

Thus we have studied the how to make a different decision tree on the basis of Information gain and entropy. In our experiment we prefer highest Information Gain for selecting the attribute to split. And calculate the entropy to find whether node will be dividing further or not. If the entropy is 1 then that will be leaf node. In this experiment we use stack as a data structure to keep record on each node for further split.

References

Jiawei Han, MichelineKamber, Data Mining Concepts and Technique, 2nd Edition

Margaret H. Dunham(2009), Data Mining-Introductory And Advanced Topocs *Pearson Education, Sixtyh Imnpression* .

Smith Tsang, Ben Kao, Kevin Y. Yip, Wai-ShingHo(2011), AndSau Dan Lee Decision Trees For Uncertain Data *IEEE Transactions On Knowledge And Data Engineering, Vol. 23, No. 1, January*.

Varshachoudhary, Pranita Jain (2013)Classification: A Decision Tree For Uncertain Data Using CDF, *International Journal Of Engineering Research And Applications (IJERA)* ISSN: 2248-9622 Vol. 3, Issue 1, Pp.1501-1506

Charu C. Aggarwal, Philip S. Yu (2009) A Survey Of Uncertain Data Algorithms And Applications *IEEE Transactions On Knowledge And Data Engineering*, Vol. 21, No. 5.

M. Suresh Krishna Reddy1, R. Jayasree(2011) Extending Decision Tree Clasifiers For Uncertain Data *International Journal Of Engineering Science & Advanced Technology* ISSN: 2250–3676 Volume-2, Issue-4, 1030 – 1034

Pragati Pandey , Miss Prateeksha Pandey, Mrs. MriduSahu(2012), Mining Uncertain Data Using Classification Feature Decision Trees ISSN: 2277 – 9043 *International Journal Of Advanced Research In Computer Science And Electronics Engineering* Volume 1, Issue 3

Chunquan Liang, Yang Zhang (2010)Decision Tree For Dynamic And Uncertain Data Streams *JMLR: Workshop And Conference Proceedings 13: 209-224 2nd Asian Conference On Machine Learning (ACML2010)*, Tokyo, Japan, Nov. 8{10, 2010.

Swapnil Andhariya, Khushali Mistry, Prof. Sahista Machchhar, Prof. Dhruv Dave (2013) Prodtu: A Novel Probabilistic Approach To Classify Uncertain Data Using decision Tree Induction *International Journal of Engineering Research & Technology* (IJERT) ISSN: 2278-0181Vol. 2 Issue 6.