www.ijecs.in
International Journal Of Engineering And Computer Science ISSN:2319-7242
Volume 4 Issue 3 March 2015, Page No. 10856-10858

Analysing Uncertain Data by Building Decision Tree

Megha Pimpalkar[1], Garima Singh[2],

[1] *Student,* Department of Computer Technology, WCEM, Nagpur, INDIA
megha.pimpalkar@gmail.com
[2] Assistance Professor, Department of Computer Technology, WCEM, Nagpur, INDIA
garima11makhija21@gmail.com

Abstract: ***In data processing, Classification of objects supported their options into pre-defined classes could be a wide studied downside with rigorous applications in fraud detection, computer science ways and lots of alternative fields. Among the assorted classification algorithms out there in literature the choice tree is one in every of the foremost sensible and effective ways and uses inductive learning. during this paper we tend to reviewed numerous call treefor same dataset and that we ar primarily performing on the ID3 algorithmic rule.***

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

1. Introduction

Data mining is an automatic discovery method of nontrivial, antecedently unknown and potentiallyuseful patterns embedded in databases. Researchhas shown that, knowledge doubles each 3 years.Thus data processing has become a crucial tool totransform these knowledge into data. The datasetsin data processing applications ar typically giant and sonew classification techniques are developedand ar being developed to subsume uncountable objects having maybe dozens or perhaps many attributes. therefore classifying these datasetsbecomes a crucial downside in data processing.Classification is that the downside of automaticallyassigning associate degree object to at least one of many pre-definedcategories supported the attributes of the thing.Classification is additionally called supervisedlearning [1]. In classification a given set of knowledgerecords is split into coaching and take a look at data sets.The coaching knowledge set is employed to make theclassification model, whereas the take a look at knowledge records areused in confirmatory the model. The model is thenused to classify and predict new set {of knowledge|of knowledge| of information} recordsdifferent from each the coaching and take a look at data sets.Some of the ordinarily used classification algorithms ar neural networks, logisticregression and call trees etc. Among these call tree algorithms ar most commonlyused. call tree provides a modelling technique that's straightforward for humans to comprehendand is simplifies the classification method.This paper makes an attempt to supply a close structure of call tree exploitation ID3 algorithmic program. It additionally provides concepts the way to generate completely different call tree by dynamical threshold price of entropy. during this paper we have a tendency to mentioned completely different knowledge set to get call tree [3]. It isorganized as follows: Section a pair of provides anoverview on call tree algorithms and completely different perform to create a amendment in call tree construction and its implementationpatterns. Section three provides experimental analysisand completely different call treefor same knowledge set. Section four provides a outline and conclusions.

1.1. DESCRIPTION OF TECHNICALTERMS/ NOTATIONS USED

Entropy may be a live of the amount of random ways in which within which a system could also be organized. For a data set S containing n records the informationentropy is outlined asEntropy(S) =- ∑ PIlog2...PI.(HerePIisthe proportion of S happiness to category I.)

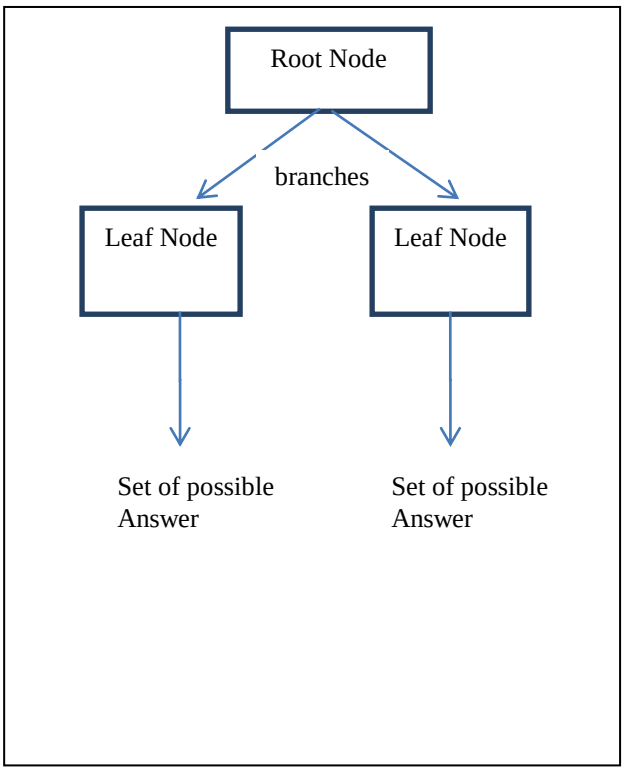
Gain or the expected data gain is thechange in data entropy from a previous state toa state that takes some data. , theinformation gain of example set S on attribute A isdefined as

Gain(S,A) = Entropy(S)-∑_(vϵValue(A))▒|S_v |/|S| Entropy(Sv) wherever Value(S)is the set of all attainable valuesof attribute A, (Sv)is the set of S for whichattribute A has price v,|S_v |is the amount of components inS.

Gini index for an information set S is outlined asgini(S)=1-∑▒P_1^2 and for a 2-split⟦gini⟧_split (s)=n_1/n gini(S_1)+n_1/n gini(S_2)andso on for a k-split.

Hunts technique rule for call treeconstruction trains the info set with recursivepartition exploitation depth 1st greedy technique, until allthe record information sets belong to the category label [1].

2. DECISION TREE ALGORITHM



Decision tree formula may be a data processing induction techniques that recursively partitions a knowledge setof records exploitation depth-first greedy approach (Hunts et al, 1966) or breadth-first approach (Shaferet al, 1996) till all the info things belong to a selected category. a call tree structure is created of root, internal and leaf nodes. The tree structure is employed in classifying unknown information records. At every internal node of the tree, a call of best split is created exploitation impurity measures (Quinlan, 1993). The tree leaves is created from the category labels that the info things are cluster [2].

Decision tree classification technique is performed in 2 phases: tree building and tree pruning. Tree building is finished in top-down manner. it's throughout this section that the tree is recursively partitioned off until all the info things belong to constant category label (Hunts et al, 1966). it's terribly tasking and computationally intensive because the coaching information set is traversed repeatedly. Tree pruning is finished may be a bottom-up fashion. it's accustomed improve the prediction and classification accuracy of the formula by minimizing over-fitting (noise or a lot of detail within the coaching information set) (Mehta et al, 1996). Over-fitting in call tree formula ends up in misclassification error []. Tree pruning is a smaller amount tasking compared to the tree growth section because the coaching information set is scanned one time. during this study we'll review call tree algorithms enforced in an exceedingly serial pattern, establish the algorithms unremarkably used and compare their classification accuracy and execution time by experimental analysis..

2.1. Uncertain Data

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

2.1.1. Faulty Measurement

In several instrument there's error of twenty-two in measure values. for instance whenever we tend to square measure measure temperature through measuring system there's error of zero.2o C. whenever we tend to get a rather completely different reading. to search out out precise temperature we tend to take scores of reading and averaging them. Such variety of error makes the info unsure.

2.1.2. Repeating Process

If we have a tendency to take a survey of student learning at school, if we have a tendency to raise them what percentage hours they're studying? we have a tendency to get totally different answer from every student. we are able to conclude that the actual age bracket of student finding out what percentage hours. This repetition of method for each student offers the unsure information.[3]

3. ID3 ALGORITHM

Step 1: If all instances in C area unit positive, then produce affirmative node and halt. If all instances in C area unit negative, produce a NO node and halt. Otherwise choose a feature, F with values v1... vn and make a choice node.

Step 2: Partition the coaching instances in C into subsets C1, C2, ...,Cn in keeping with the values of V.

Step 3: apply the algorithmic program recursively to every of the sets Ci. Note, the trainer (the expert) decides that feature to pick out [4].

ID3 improves on construct Learning System by adding a feature choice heuristic. ID3 searches through the attributes of the coaching instances and extracts the attribute that best separates the given examples. If the attribute absolutely classifies the coaching sets then ID3 stops; otherwise it recursively operates on the n (where n = variety of doable values of Associate in Nursing attribute) divided subsets to induce their "best" attribute. The algorithmic program uses a greedy search, that is, it picks the most effective attribute and ne'er appearance back to rethink earlier selections.

4. CONCLUSION

Thus we've studied the the way to build a unique call tree on the premise of data gain and entropy. In our experiment we tend to like highest data Gain for choosing the attribute to separate. And calculate the entropy to search out whether or not node are going to be dividing more or not. If the entropy is one then that may be leaf node. during this experiment we tend to use stack as a knowledge structure to stay record on every node for more split.

REFERENCE

[1] Jiawei Han, MichelineKamber, "Data Mining Concepts And Technique", 2nd Edition

[2] Margaret H. Dunham,"Data Mining-Introductory And Advanced Topocs" Pearson Education,SixtyhImnpression ,2009.

[3] Smith Tsang, Ben Kao, Kevin Y. Yip, Wai-ShingHo, AndSau Dan Lee "Decision Trees For Uncertain Data" IEEE TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING, VOL. 23, NO. 1, JANUARY 2011

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

[5] Charu C. Aggarwal, Philip S. Yu "A Survey Of Uncertain Data Algorithms And Applications" IEEE TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING, VOL. 21, NO. 5, MAY 2009

[6] M. Suresh Krishna Reddy1, R. Jayasree "EXTENDING DECISION TREE CLASIFIERS FOR UNCERTAIN DATA" INTERNATIONAL JOURNAL OF ENGINEERING SCIENCE & ADVANCED TECHNOLOGY ISSN: 2250–3676 Volume-2, Issue-4, 1030 – 1034

[7] Miss Pragati Pandey , Miss PrateekshaPandey,Mrs. MriduSahu, " 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, May 2012

[8] Chunquan Liang, Yang Zhang "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.

[9] SwapnilAndhariya, KhushaliMistry, Prof.SahistaMachchhar, Prof.Dhruv Dave "Prodtu: A Novel Probabilistic Approach To Classify Uncertain Data Usingdecision Tree Induction" International Journal Of Engineering Research & Technology (IJERT) ISSN: 2278-0181Vol. 2 Issue 6, June – 2013