Research Article Impact Factor: 0.621 ISSN: 2319-507X
Pravin Chunarkar, IJPRET, 2014; Volume 2 (9): 769-774 IJPRET

INTERNATIONAL JOURNAL OF PURE AND APPLIED RESEARCH IN ENGINEERING AND TECHNOLOGY

A PATH FOR HORIZING YOUR INNOVATIVE WORK

CLASSIFICATION OF BRAIN TUMORS BY USING DECISION TREE

PRAVIN N. CHUNARKAR[1], JAIKUMAR M. PATIL[2]

1. Post-graduating student of Shri Sant Gajanan Maharaj College of engineering, Shegaon, India.

2. Faculty member of department of computer science and engineering, Shri Sant Gajanan Maharaj College of engineering, Shegaon, India.

Accepted Date: 27/02/2014 ; Published Date: 01/05/2014

Abstract: Decision Trees can be used as classifiers and its approach is becoming popular in recent days. Data mining have efficiently dealt with the use of decision tree for growing available data. The purpose of this work is to present an updated survey of current methods for constructing decision tree for classifying brain tumors. The CART and Random algorithm are some examples of single decision tree classifiers which can solve the problem of classification of cancer, showing strengths and weaknesses of the proposed methodologies when compared to other popular classification methods. GINI index is specially used for calculating the weight of the node and can be further use to take decision about where the classification leads. When some selected attributes of brain tumors are used of the classification, the use of decision tree classifiers classifies it into main categories.

Keywords: Bioinformatics, Cancer classification, C4.5, Cart algorithm, Gini index, Gain ratio, Information gain.

PAPER-QR CODE

Corresponding Author: MR. PRAVIN N. CHUNARKAR

Access Online On:

www.ijpret.com

How to Cite This Article:

Pravin Chunarkar, IJPRET, 2014; Volume 2 (9): 769-774

INTRODUCTION

One of the fundamental tasks in data mining is classification. Statistics, machine learning, neural networks and expert systems also studies classification. A set of training records stands as the input for classification, where each record has several attributes. Attributes with continuous domains are referred to as numerical while those with discrete domains are referred to as Categorical. There is one distinguished attribute called the class label. In general, given a database of records, each with a class label, a classifier generates a concise meaningful description for each class in terms of the attributes.

Often the medical decision maker will be faced with a sequential decision problem involving decisions that lead to different outcomes depending on chance. If the decision process involves many sequential decisions, then the decision problem becomes difficult to visualize and to implement. Decision trees are indispensable graphical tools in such settings. They allow for intuitive understanding of the problem and can aid in decision making.

A decision tree is a graphical model describing decisions and their possible outcomes. Decision trees consist of three types of nodes

a) Decision node: Often represented by squares showing decisions that can be made. Lines emanating from a square show all distinct options available at a node.

b) Chance node: Often represented by circles showing chance outcomes. Chance outcomes are events that can occur but are outside the ability of the decision maker to control.

c) Terminal node: Often represented by triangles or by lines having no further decision nodes or chance nodes. Terminal nodes depict the final outcomes of the decision making process.

2. ALGORITHMIC FRAMEWORK FOR DECISION TREES

Decision tree inducers are algorithms that automatically construct a decision tree from a given dataset. Typically the goal is to find the optimal decision tree by minimizing the generalization error. However, other target functions can be also defined, for instance, minimizing the number of nodes or minimizing the average depth.

a) Univariate Splitting Criteria

Univariate means that an internal node is split according to the value of a single attribute. In most of the cases, the discrete splitting functions are univariate. Consequently, the inducer searches for the best attribute upon which to split. Following is the one of the various univariate criteria's.

```
Tree Growing (S, A, y)
Where:
S-Training
A-Input Feature set
y- Target feature
Create a new tree T with a single root node.
If One of the Stopping Criteria is fulfilled THEN
Mark the root node in T as a leaf with the most common value oy y in S as a label.
Else
Find a discrete function f(A) of the input attributes values such that splitting S according to f(A)'s outcomes (v1,.....,vn) gains the best splitting metric.
If
Best splitting metric> threshold THEN
Label t with f(A)
        For each outcome vi of f(A)
Set $Subtree_i$= Tree growing(S,A,y)
Connect the root node of tT to subtree(i) with an edge that is labelled as vi
END FOR
ELSE
Mark the root node in T as a leaf with the most common value of y in S as a label.
END IF
END IF
RETURN T
TreePrunning(S,T,y)
Where:
S- Training set
y- Target feature
```

```
T- The tree to be pruned
DO
    Select a node t in T such that pruning it maximally improve some evaluation criteria
    IF t!=0 THEN T=pruned(T,t)
UNTIL t=0
RETURN T
```

Algorithm 1- Top-Down algorithmic framework for decision tree induction

b) GINI Index

Gini index is an impurity-based criterion that measures the divergences between the probability distributions of the target attribute's values. The Gini index has been used in various works and it is defined as: $\mathrm{GINI}(t)=1-\sum_i[p(i/t)]^2$

Let's take three examples to calculate GINI index with six records which are distributed in two classes.

1) C1=0, C2=6, C1+C2=6

$P(C1)=0/6=0$, $P(C2)=6/6=1$, $GINI=1-P(C1)^2-P(C2)^2=1-0-1=0$

2) C1=1, C2=5, C1+C2=6

$P(C1)=1/6=0.166$, $P(C2)=5/6=0.622$, $GINI=1-P(0.166)^2-P(0.622)^2=0.278$

3) C1=2, C2=4, C1+C2=6

$P(C1)=2/6=0.333$, $P(C2)=4/6=0.666$, $GINI=1-P(0.333)^2-P(0.666)^2=0.444$

3. CART

The acronym CART stands for Classification And Regression Trees. Both categorical and continuous attributes to build a decision tree can efficiently handle by CART. It handles missing values also. CART uses Gini Index as an attribute selection measure to build a decision tree. CART uses cost complexity pruning to remove the unreliable branches from the decision tree to improve the accuracy.

It is characterized by the fact that it constructs binary trees, namely each internal node has exactly two outgoing edges. The splits are selected using the twoing criteria and the obtained tree is pruned by cost–complexity Pruning. When provided, CART can consider misclassification costs in the tree induction. It also enables users to provide prior probability distribution. An important feature of CART is its ability to generate regression trees. Regression trees are trees where their leaves predict a real number and not a class. In case of regression, CART looks for splits that minimize the prediction squared error (the least–squared deviation). The prediction in each leaf is based on the weighted mean for node.

4. C4.5

This algorithm is an extension to ID3 developed by Quinlan Ross. It is also based on Hunt's algorithm.C4.5 handles both categorical and continuous attributes to build a decision tree. In order to handle continuous attributes, C4.5 splits the attribute values into two partitions based on the selected threshold such that all the values above the threshold as one child and the remaining as another child. It also handles missing attribute values. C4.5 uses Gain Ratio as an attribute selection measure to build a decision tree. It removes the biasness of information gain when there are many outcome values of an attribute.

Very firstly, calculate the gain ratio of each attribute. The root node will be the attribute whose gain ratio is maximum. C4.5 uses pessimistic pruning to remove unnecessary branches in the decision tree to improve the accuracy of classification.

5. An example of an approach for implementing decision tree for classifying brain tumors

The cancer classified by taking nine different variables collected from biopsy. The first split of the tree (at the root node) is taken on the basis of variable "unsize," which measures uniformity of cell size. All patients having values less than 2.5 for this variable are assigned to the left node (the left daughter node); otherwise they are assigned to the right node (right daughter node). The left and right daughter nodes are then split on the variable "unshape" for the right daughter node and on the variable "nuclei" for the left daughter node), and patients are assigned to subgroups defined by these splits. These nodes are then split, and the process is repeated recursively in a procedure called recursive partitioning. When the tree construction is completed, terminal nodes are assigned class labels by majority voting (the class label with the largest frequency). Each patient in a given terminal node is assigned the predicted class label for that terminal node.

5. CONCLUSION

Data Mining is gaining its popularity in almost all applications of real world. Decision trees are so popular because they produce human readable classification rules and easy to interpret than other classification methods. Frequently used decision tree classifiers are studied and the experiments are conducted to find the best classifier for Medical Diagnosis. Using this approach, brains tumors could be efficiently classify into their types. It is also observed that CART performs well for classification on medical data sets of increased size.

REFERENCES

1. DECISION TREES, Lior Rokach, Department of Industrial Engineering, Tel-Aviv University, liorr@eng.tau.ac.il, Oded Maimon, Department of Industrial Engineering, Tel-Aviv University, maimon@eng.tau.ac.il

2. Decision Tree Classifiers in Bioinformatics, Inese Polaka, Riga Technical University, Igor Tom, United Institute of Informatics Problems of the National Academy of Sciences of Belarus, Arkady Borisov, Riga Technical University

3. An extensive comparison of recent classification tools applied to microarray data, J. W. Lee, J. B. Lee, M. Park, S. H. Song, Computational Statistics & Data Analysis, Vol. 48, Issue 4, pp. 869-885, Apr. 2005.

4. U.M. Fayyad, G. Piatetsky-Shapiro and P. Smyth, From Data Mining to knowledge Discovery in Databases , AI Magazine, vol 17, pp. 37-54, 1996.

5. Antonia Vlahou, John O. Schorge, Betsy W.Gregory and Robert L. Coleman, Diagnosis of Ovarian Cancer Using Decision Tree Classification of Mass Spectral Data , Journal of Biomedicine and Biotechnology • 2003:5 (2003) 308–314.

6. Kuowj, Chang RF,Chen DR and Lee CC, Data Mining with decision trees for diagnosis of breast tumor in medical ultrasonic images ,March 2001.