ISSN: 2277-9655
Impact Factor: 1.852

IJESRT

INTERNATIONAL JOURNAL OF ENGINEERING SCIENCES & RESEARCH TECHNOLOGY

An Efficient Approach of Decision Tree for Classifying Brain Tumors

Pravin N. Chunarkar
Student of SSGMCE, Shegaon, India
chunarkar.pravin@gmail.com

Abstract

Decision Trees are considered to be one of the most popular approaches for representing classifiers. Statistics, machine learning, pattern recognition, and Data Mining have dealt with the issue of growing a decision tree from available data. The purpose of this work is to present an updated survey of current methods for constructing decision tree for classifying brain tumours. The main focus is on solving the cancer classification problem using single decision tree classifiers (CART and Random algorithm) showing strengths and weaknesses of the proposed methodologies when compared to other popular classification methods. This paper presents a literature review of articles related to the use of decision tree classifiers which classifies brain tumours into main categories.

Keywords: Bioinformatics, cancer classification, CART algorithm, Decision Tree, Gain ration, GINI index, Information gain

Introduction

A decision tree is a powerful method for classification and prediction and for facilitating decision making in sequential decision problems. This entry considers three types of decision trees in some detail. The first is an algorithm for a recommended course of action based on a sequence of information nodes; the second is classification and regression trees; and the third is survival trees.

Decision Tree

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.

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.

Induction of an optimal decision tree from a given data is considered to be a hard task. It has been shown that finding a minimal decision tree consistent with the training set is NP–hard. Moreover, it has been shown that constructing a minimal binary tree with respect to the expected number of tests required for classifying an unseen instance is NP–complete

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 criterias.

```
TreeGrowing(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 Stoppng 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
```

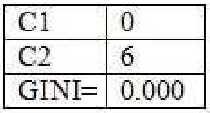
Algorithm 1- Top-Down algorithmic framework for decision tree induction

b) GINI Index

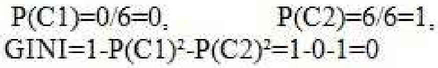
Gini index is an impurity-based criterion that measures the divergences between the probability distributions of the target attribute's values. The Gini index has been used in various works and it is defined as:

$$GINI(t)=1-\sum_i[p(i/t)]^2$$

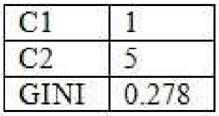
Let's take an example to calculate GINI index with six records which are distributed in two classes.



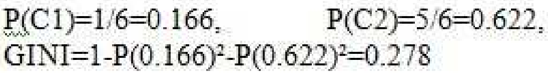
C1	0
C2	6
GINI=	0.000



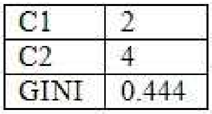
$P(C1)=0/6=0$, $P(C2)=6/6=1$, $GINI=1-P(C1)^2-P(C2)^2=1-0-1=0$



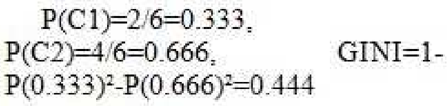
C1	1
C2	5
GINI	0.278



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



C1	2
C2	4
GINI	0.444

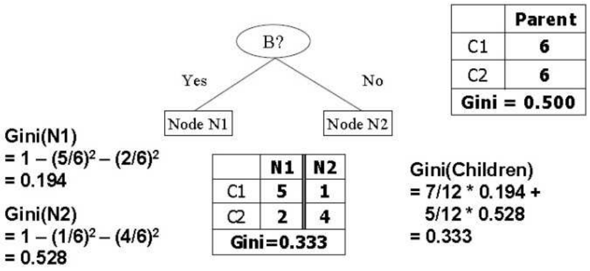


$P(C1)=2/6=0.333$, $P(C2)=4/6=0.666$, $GINI=1-P(0.333)^2-P(0.666)^2=0.444$

c) Splitting based on GINI index

$GINIsplit=\sum i=1$ to k $[n_i/n\ GINI(i)]$

n_i=number of records at child i,
n= nmber of records at node p.



	Parent
C1	6
C2	6
Gini = 0.500	

$Gini(N1) = 1 - (5/6)^2 - (2/6)^2 = 0.194$

$Gini(N2) = 1 - (1/6)^2 - (4/6)^2 = 0.528$

	N1	N2
C1	5	1
C2	2	4
Gini=0.333		

$Gini(Children) = 7/12 * 0.194 + 5/12 * 0.528 = 0.333$

d) Entropy

The average amount of information needed to classify an object is given by entropy. It measures homogeneity of node. It turns out maximum when records are equally distributed in classes and minimum when all records belong to one class, implying most information.

$$Entropy(t)=-\sum_i p(i/t)\log_2 p(j/t)$$

C1	0
C2	6

$P(C1)=0/6=0$, $P(C2)=6/6=1$
Entropy=-0log0-1log1=-0-0=0

C1	1
C2	5

$P(C1)=1/6$, $P(C2)=5/6$
Entropy=$-(1/6)\log_2(1/6)-$

$(5/6)\log_2(5/6)=0.65$

C1	2
C2	4

$P(C1)=2/6$, $P(C2)=4/6$ Entropy=-$(2/6)\log_2(2/6)-(4/6)\log_2(4/6)=0.92$

e) Splitting criteria based on Classification of errors

It gives maximum value when records are equally distributed among all classes, implying least information and minimum when all records belong to oe class, implying most information.

Error(t)=1-maxP(i/t)

C1	1
C2	5

P(C1)=0/6=0, P(C2)=6/6=1 Error=1-max(0,1)=0

C1	0
C2	6

P(C1)=1/6, P(C2)=5/6 Error=1-max(1/6,5/6)=1-(5/6)=1/6

C1	2
C2	4

P(C1)=2/6, P(C2)=4/6 Error=1-max(2/6,4/6)=1-(4/6)=1/3

Decision Tree Induction

Decision tree induction is a very popular and practical approach for pattern classification. Decision tree induction is the learning of decision trees from class-labelled training tuples. A decision tree is a flow chart like tree structure, where each internal node denotes a test on an attribute, each branch represents an outcome of the test, and each leaf node holds a class label.

The decision tree classifier has two phases:

i) Growth phase or Build phase.

ii) Pruning phase.

The tree is built in the first phase by recursively splitting the training set based on local optimal criteria until all or most of the records belonging to each of the partitions bearing the same class label. The tree may overfit the data.

The pruning phase handles the problem of over fitting the data in the decision tree. The prune phase generalizes the tree by removing the noise and outliers. The accuracy of the classification increases in the pruning phase. Pruning phase accesses only the fully grown tree. The growth phase requires multiple passes over the training data. The time needed for pruning the decision tree is very less compared to build the decision tree.

Cart

CART stands for Classification And Regression Trees introduced by Breiman. It is also based on Hunt's algorithm. CART handles both categorical and continuous attributes to build a decision tree. It handles missing values. 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.

An Example of an Approach for Implementing decision tree for classifying brain tumours

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.

ISSN: 2277-9655
Impact Factor: 1.852

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 tumours 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.

*[7] H. Ren, ―Clinical diagnosis of chest pain R. Abraham, S. van den Bergh, and P. Nair, A new approach to galaxy morphology, I: Analysis of the Sloan digital sky survey early data release, Astrophysical Journal **588** (2003) 218–229. doi:10.1086/37391*