Analysis of CART algorithm on large dataset

Deepanjali Sharma

Integrated M.Sc Computer Science

2013IMSCS005

IXth semester

INTRODUCTION

CART:- Classification trees are designed for dependent variables that take a finite number of unordered values, with prediction error measured in terms of misclassification cost. Regression trees are for dependent variables that take continuous or ordered discrete values, with prediction error typically measured by the squared difference between the observed and predicted values.

Goal-

To classify an instance based on a set of predictors and also analyse the performance of CART algorithm on large dataset.

Dataset-

All the analysis has been performed on wave500k dataset. It is a well known dataset described in the CART book(Breiman et al., 1984) . Baisc Statistics of the dataset can be seen in **Table 1**

Total Instances	65536
No. of features	21
Class labels	A , B , C

Table1:- Basic Statistics of Dataset

Method-

CART analysis uses recursive partitioning to create a tree where each node represents a cell of the partition. The dataset which is used for the analysis has 21 attributes and one class column with three different classes. Here we have encoporated a decicison tree classifier model based on *CART* algorithm which is trained with 21 attributes as input to the model and class label is taken as output for the model. We have built the classifier models with different criterion, splitter and other parameters to evaluate the performance of model.

Parameter Tuning

Parameter Tuning of *CART* is done i.e, parameters are given different value and corresponding accuracy and time complexity is evaluated for each value. Eventually the parameters with value which gives the best precision compared to the other values is selected as the result for that *CART* algorithm. Parameter Tuning for CART can be clearly visualised from **Table 2**

Criterion	Splitter	max_Depth	min_samples_split	min_samples_leaf	Precision	Time
gini	best	4	4	1	74.71%	8.44
gini	random	8	6	3	79.65%	8.66
entropy	best	4	4	1	74.07%	8.003
entropy	random	8	6	3	77.13%	7.68

Table 2:- Parameter Tuning for CART

Results

CART algorithm gives the highest precision when splitter is selected as random and purity criterion is selected as gini also the precision of the model increases when depth, split, and no. of leaf is increased with negligible increase in time complexity whereas when the splitter is chosen as best and split,depth and no. of leaf is decreased the precision of the algorithm also falls for the dataset . Also gini is proved to be the best criterion for splitting compared to entropy which can be clearly seen in **Table 2.** The tree obtained from the rules after considering all the parameters can be seen in **Figure 1.**

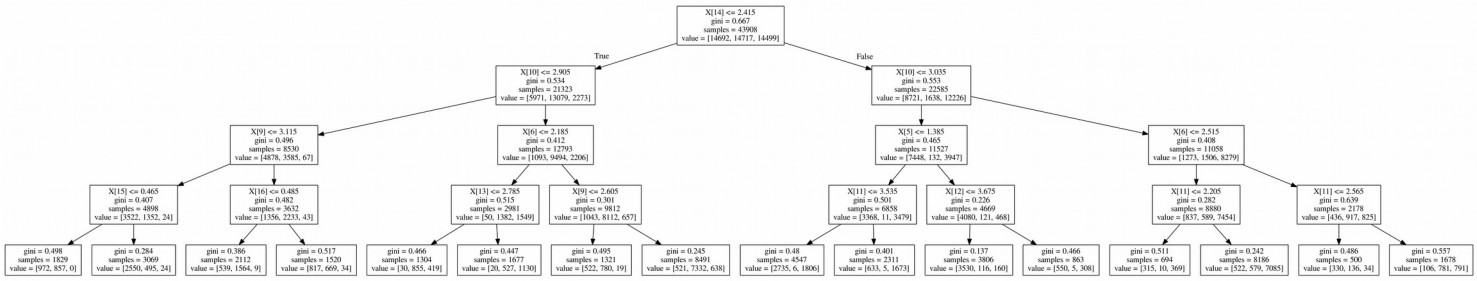

Figure 1:- Decision Tree for the dataset based on CART

Conclusion & Discussion

CART classification seems to be the efficient algorithm even for the large dataset as it gives quite satisfactory performance with minimal time complexity. Also the criterion for measuring the purity should be GINI as it gives better accuracy compare to the entropy measure. The accuracy can be further improved if, a smart measure for selecting the features is encoporated .