Data Mining Predictive Technique: CART

Ms. Sarmishtha Das[1], Prof.G.P.Saradhi Varma[2],
Mr. G. Ramesh Naidu[3], Mr. A.V.D.N. Murthy[4]

[1]Reader, VIPS College, Visakhapatnam

[2]HOD,InformationTechnology Dept,S.R.K.R.Engg.College,BHIMAVARAM

[3]M.Tech (Ph.d) HOD, Kaushik Engg. College, Visakhapatnam

[4]Lecturer, VIPS College, Visakhapatnam

Introduction:

Data Mining is an emerging technology for the automatic extraction of patterns, associations for large data sets. It is one of the key technologies which enable business to select, filter, and correlate data automatically. Classification and Prediction are two forms of data analysis that can be used to categorize applications and in prediction. Here I discuss one predictive technique i.e. **Classification And Regression Tress (CART).**

Features of CART:

i. It is a tree structured statistical analysis.
ii. It is easy to use.
iii. It uses an intuitive, windows-based interface, making it accessible to both learning and statistical research.
iv. It provides stable performance and reliable results as it is based on machine learning and statistical research.
v. It is a proven statistical methodology and digs deep into date using advanced techniques. It uses a **backward pruning technique** that identifies the tree structure of the data.

vi. CART's binary decision trees are more sparing with data and detect more structure and very little data is left for learning – whereas other decision tree approaches are based on multi way splits that fragment the data rapidly and make it difficult to delete the rules which require broad range of data to discover.

vii. Testing and Selection of the optimal CART algorithm. This test ensures that the pattern found will hold up when applied to new data.

viii. It intelligently handles the missing values by substituting surrogate splitters which are backup rules that contains information that is similar to that found in primary splitter.

ix. It includes **built-in cross validation** for smaller databases which uses every record to alternatively train and test resulting in a highly reliable determination of optimal model complexity while using all data for learning.

x. For larger databases CART flexibly select test data chosen from separate test base or selects test record randomly.

xi. CART users can specify a higher penalty for misclassifying certain data and the software will steer the tree away from that type of error.

xii. When CART cannot guarantee a correct classification it ensures that the error occurred is less expensive.

Methodology:

CART uses **binary recursive partitioning**. It is binary because the parent node is always split exactly into two child nodes and is recursive because the process can be repeated by treating each child node as a parent.

The CART analysis consists of the following:

i. Determining the rules of splitting each node into a tree.

ii. Deciding when a tree is complete.

iii. Assigning each terminal node to a class or predicted value.

Example: A splitting rule is yes-no question. CART imposes on a variable in the data set such as income <= 10,000 or is experience <=25.

CART looks at all possible splits for all variables in the data set and then chooses the best using goodness of split criteria.

CART then repeats the process on each child node until the splitting is impossible or stopped for other reasons.

Once the tree is complete, CART assigns classes to each terminal node. This is usually done using the largest percentage of cases in the node.

Therefore CART first creates maximal trees, growing until it is not possible to grow any further and then prunes away branches, creating a smaller and more efficient tree. CART does not stop in the middle of process though it is an optimal point as it might miss the important branches.

Disadvantages:

i. The predicted value of the response is discontinuous, which means that sometimes a small change in the value of a prediction variable could lead to a large change in the predicted value of the response.
ii. The decision tree model is **coarse-grained** in the sense that a model with n-nodes can only predict n-different probabilities, which can be an issue, particularly if the tree produces only a small number of nodes.
iii. It's weakness at capturing strong linear structure. A very large tree can be produced in an attempt to represent very simple relationships. The algorithm recognizes the structure but cannot represent it effectively.

Remedy:

LOGIT (Logistic Regression)and Neural Nets can compensate for CART weakness.

- We **should not run** LOGIT, Neural Nets or any model in CART terminal nodes.
- We **should run** LOGIT, Neural Nets or any model in root node.
- Hybrid of **CART-LOGIT** works effectively:

(I) General Applications:

i. It can be applied in various areas including **telecommunications, banking, financial services, insurances, healthcare, manufacturing, retail sales and education.**
ii. It is an important statistical analysis tool that can be used to segment databases and predict risk factor.

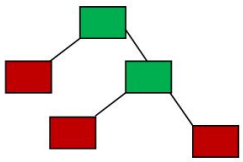
iii. CART's recursive partitioning abilities gives a statistical method for generating marketing models in an easy to understand decision tree format. So its application in various field includes **market segmentation, customer profiling,** market segment profitability, campaign targeting, credit card fraud detection, quality control, clinical and biomedical research.

iv. CART is a robust program that can support a diverse set of application ranging from **food security analysis** to pattern recognition and **remote sensing** problem. It is a tool for analyzing small, complex data sets which are inter related and which can be used in characterizing geographic areas instead of taking many expensive sample for oil well drilling.

(II) A New Approach of Cart Usage:

We require the hybrid of CART and Neural Network for effective result.

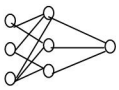
CART



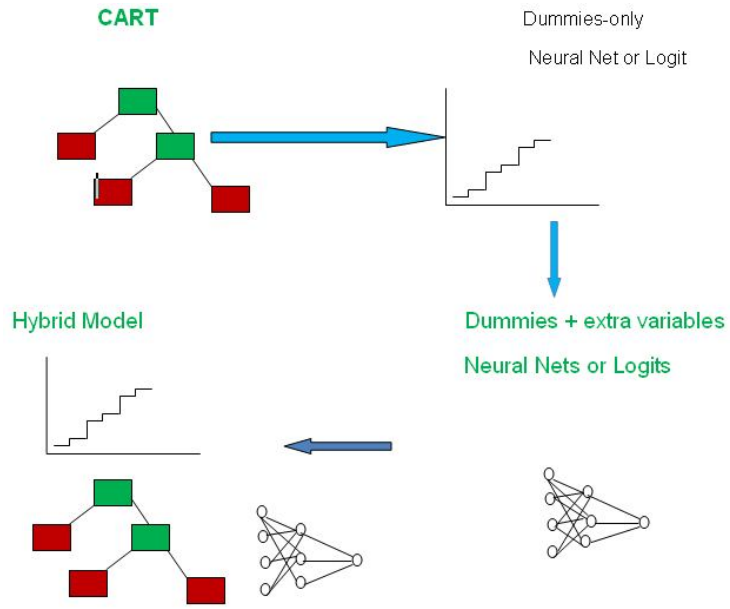
- Fast
- Robust
- Easy to use

Neural Nets

- Slow
- Prefer clean data
- Need expert guidance.
- Can always start the CART quickly.



- Knowledge discovered with CART can be used to improve Neural Nets.
- CART model applies to all records even with many missing data fields.
- Hybrid allows for easy updates.



Conclusion:

CART uses a combination of exhaustive searches and computer intensive testing techniques to identify useful tree structures of data.

It can be applied to any data set and can proceed till little or no guidance from the user.

It can be used to analyze the data in the situations where there is no idea on the data set.

Since CART does not suffer from statistical deficiencies it will be very accurate on new data.

CART is competitive when compared with best parametric models. If the list of potential predictor is large, CART can be first used to extract the most important variables.

References:

1. Breiman, L., J. Friedman, and R. Olshen, and C.Stone(1994), Classification and Regression Trees, Pascific Grove: Wadsworth.
2. Steinberg, D. and Colla, P.L., (1995), CART: Tree-Structured Nonparmetric Data Analysis, San Diego, CA: Salford Systems.
3. www.Springer.com