INTERNATIONAL JOURNAL OF PROGRESSIVE RESEARCH IN ENGINEERING MANAGEMENT AND SCIENCE (IJPREMS)

www.ijprems.com
editor@ijprems.com

Vol. 04, Issue 03, March 2024, pp: 541-543

e-ISSN : 2583-1062

Impact Factor: 5.725

A REVIEW PAPER ON DECISION TREE METHODS

S. durga[1]

[1]Fatima College, Madurai, India.

ABSTRACT

Decision tree methodology is a commonly used data mining method for establishing classification systems based on multiple covariates or for developing prediction algorithms for a target variable. This method classifies a population into branch-like segments that construct an inverted tree with a root node, internal nodes, and leaf nodes. The algorithm is non-parametric and can efficiently deal with large, complicated datasets without imposing a complicated parametric structure.

When the sample size is large enough, study data can be divided into training and validation datasets. Using the training dataset to build a decision tree model and a validation dataset to decide on the appropriate tree size needed to achieve the optimal final model.

Keywords: decision tree, data mining, classification, prediction

1. INTRODUCTION

Data mining is used to extract useful information from large datasets and to display it in easy-to-interpret visualizations. First introduced in 1960's, decision trees are one of the most effective methods for data mining; they have been widely used in several disciplines because they are easy to be used, free of ambiguity, and robust even in the presence of missing values.

Both discrete and continuous variables can be used either as target variables or independent variables. More recently, decision tree methodology has become popular in medical research.

2. COMMON USAGES OF DECISION TREE MODELS

- Variable selection
- Assessing the relative importance of variables
- Handling of missing values
- Prediction
- Data manipulation

Variable Selection:

The number of variables that are routinely monitored in clinical settings has increased dramatically with the introduction of electronic data storage. Many of these variables are of marginal relevance and, thus, should probably not be included in data mining exercises.

Assessing the relative importance of variables:

Once a set of relevant variables is identified, researchers may want to know which variables play major roles. Geneally, variable importance is computed based on the reduction of model accuracy (or in the purities of nodes in the tree) when the variable is removed. In most circumstances the more records a variable have an effect on, the greater the importance of the variable.

Handling of missing values:

A common - but incorrect - method of handling missing data is to exclude cases with missing values; this is both inefficient and runs the risk of introducing bias in the analysis. Decision tree analysis can deal with missing data in two ways:

a. It can either classify missing values as a separate category that can be analyzed with the other categories or
b. Use a built decision tree model which set the variable with lots of missing value as a target variable to make prediction and replace these missing ones with the predicted value.

Prediction:

This is one of the most important usages of decision tree models. Using the tree model derived from historical data, it's easy to predict the result for future records.

3. DATA MANIPULATION

Too many categories of one categorical variable or heavily skewed continuous data are common in medical research. In these circumstances, decision tree models can help in deciding how to best collapse categorical variables into a more manageable number of categories or how to subdivide heavily skewed variables into ranges.

Basic Concepts:

The main components of a decision tree model are

- Nodes
- Branches

The most important steps in building a model are

- Splitting
- Stopping
- Pruning

Nodes:

There are three types of nodes.

A. A root node, also called a decision node, represents a choice that will result in the subdivision of all records into two or more mutually exclusive subsets.

B. Internal nodes, also called chance nodes, represent one of the possible choices available at that point in the tree structure; the top edge of the node is connected to its parent node and the bottom edge is connected to its child nodes or leaf nodes.

C. Leaf nodes, also called end nodes, represent the final result of a combination of decisions or events.

Branches:

Branches represent chance outcomes or occurrences that emanate from root nodes and internal nodes. A decision tree model is formed using a hierarchy of branches. Each path from the root node through internal nodes to a leaf node represents a classification decision rule. These decision tree pathways can also be represented as 'if-then' rules.

Splitting:

This splitting procedure continues until pre-determined homogeneity or stopping criteria are met. In most cases, not all potential input variables will be used to build the decision tree model and in some cases a specific input variable may be used multiple times at different levels of the decision tree.

Stopping:

Complexity and robustness are competing characteristics of models that need to be simultaneously considered whenever building a statistical model. The more complex a model is, the less reliable it will be when used to predict future records.

Common parameters used in stopping rules include:

(a) The minimum number of records in a leaf;

(b) The minimum number of records in a node prior to splitting; and

(c) The depth of any leaf from the root node. Stopping parameters must be selected based on the goal of the analysis and the characteristics of the dataset being used.

Pruning:

In some situations, stopping rules do not work well. An alternative way to build a decision tree model is to grow a large tree first, and then prune it to optimal size by removing nodes that provide less additional information.

There are two types of pruning,

- ✓ Pre-pruning (forward pruning)
- ✓ Post-pruning (backward pruning).

Pre-pruning uses Chi-square tests or multiple-comparison adjustment methods to prevent the generation of non-significant branches.

Post-pruning is used after generating a full decision tree to remove branches in a manner that improves the accuracy of the overall classification when applied to the validation dataset.

4. CONCLUSION

The decision tree method is a powerful statistical tool for classification, prediction, interpretation, and data manipulation that has several potential applications in medical research. Using decision tree models to describe research findings has the following advantages:

- Simplifies complex relationships between input variables and target variables by dividing original input variables into significant subgroups.
- Easy to understand and interpret.

- Non-parametric approach without distributional assumptions.
- Easy to handle missing values without needing to resort to imputation.
- Easy to handle heavy skewed data without needing to resort to data transformation.
- Robust to outliers.

5. REFERENCES

[1] Hastie T, Tibshirani R, Friedman J. The Elements of Statistical Learning. Springer; 2001. p. 269-272 [Google Scholar]

[2] Quinlan RJ. C4.5: Programs for Machine Learning. San Mateo California: Morgan Kaufmann Publishers, Inc.; 1993. [Google Scholar]

[3] Kass GV. Anexploratory technique for investigating large quantities of categorical data. Appl Stat. 1980;29: 119–127. [Google Scholar]

[4] Loh W, Shih Y. Split selection methods for classification trees. Statistica Sinica. 1997;7: 815–840. [Google Scholar]

[5] SAS Institute Inc. SAS Enterprise Miner12.1 Reference Help, Second Edition. USA: SAS Institute Inc; 2011. [Google Scholar]

[6] IBM Corporation. IBM SPSS Modeler 17 Modeling Nodes. USA: IBM Corporation; 2015. [Google Scholar]