INTERNATIONAL JOURNAL OF STATISTICS AND MEDICAL INFORMATICS

Vol. 15 No.1/2020, pp.: 1-5

CART and CHAID ANALYSIS

Editor, International Journal of Statistics and Medical Informatics

Classification and Regression Trees (CART) and Chi Square Automatic Interaction Detector (CHAID) works on principles of decision tree analysis. Classification and Regression (CART) classifies the data based on the categorical outcome variable (Classification) and also uses continuous outcome variable for regression problem. Chi Square Automatic Interaction Detector (CHAID) is similar to CART which uses classifies the data into multiple class labels not only binary classification. In CHAID both dependent variable and independent variables will be categorical. This paper provides an overview and CART and CHAID methods using open source R software with hypothetical data set.

Keywords: Decision Tree, Classification, Regression, CART, CHAID,

Introduction

Classification and Regression Tree (CART) [1, 2, 3] is binary decision tree algorithm which uses recursive partitioning to divide the data into subset or binary classes. Mean Square Error is used for splitting in the regression trees and impurity index such as Gini is used for classification problems. Purity is measured by the homogeneity of the members in the partitioned class and perfect partition means all the members of the classes share same property and no overlapping of members of two classes found. CART uses recursive partitioning where in the root node is split into sub nodes and this process is repeated till there is no possibility of splitting or further splitting will not help in explain more variability. The final sub node at the tree is called leaf. Pruning is used to reduce the over fitting of the trees.

Chi Square Automatic Interaction Deduction [4,5] (CHAID) uses the Chi Square test to decide the split and association between dependent variable and independent variables. CHAID uses recursive partition algorithm which maximizes the chi square statistics with respect to the cross tabulation or interaction between dependent variable and independent variables. The partition process involves splitting the data set into sub sets with respect to dependent variable categories. During the recursive partitioning if p values are greater than the critical values then two pairs are merged and this is repeated till no significant pairs is found. If the p values are significant then it causes the splitting of the data set into subsets which needs to be highest among variables. Since it involves multiple computations of p values Bonferrroni correction is used to control the type I error.

Vol. 15 No.1/2020, pp.: 1-5

We use R software [6] and its GUI R Studio [7] to build CART and CHAID model using hypothetical datasets. We will be building classification tree here. We will use xlsx[8], rpart[9] and rpart.plot[10] to plot the CART tree as below

```
library(xlsx)
library(rpart)
library(rpart.plot)
```

We will bring the test data set into r studio environment and the following variable is included in the data set

- age
- gender
- diabetic
- hypertension
- smoking
- alcohol
- Cholesterol
- Survival

Here Survival status is the dependent categorical variable. Here head (data1) statement will help us to display the first 5 rows of the dataset. We will split the data set into training and test data set

```
data1<-read.xlsx('survivalstroke.xlsx',sheetIndex = 1)
head(data1)
train<-data1[1:100,]
test<-data1[101:121,]
```

We will use the rpart function to build the CART model

```
cartmodel<-
rpart(survival~$age+gender+diabetic+hypertension+smoking+alcohol+Cholesterol
, method="class", data=train)
```

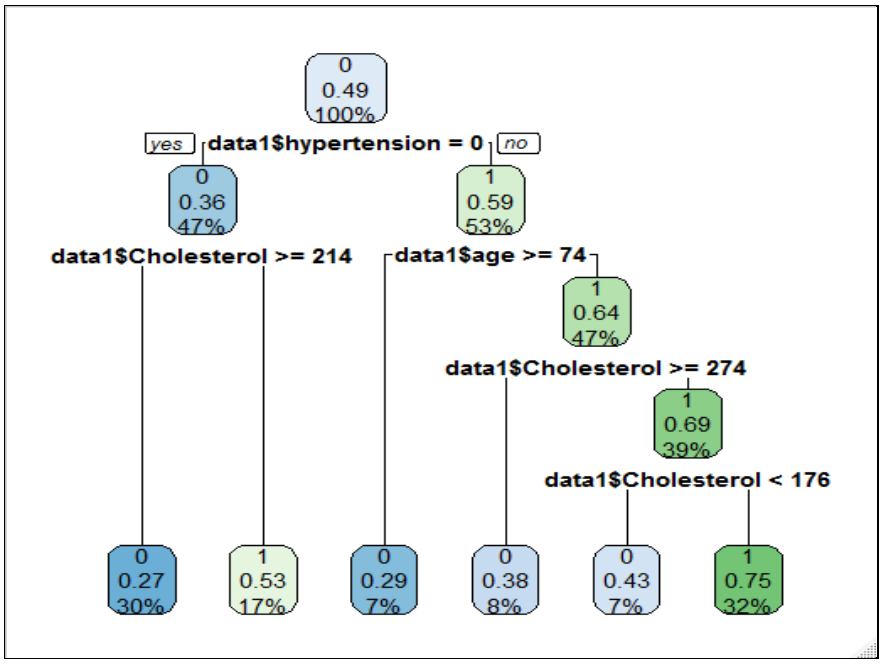
The following model summary is obtained

INTERNATIONAL JOURNAL OF STATISTICS AND MEDICAL INFORMATICS

Vol. 15 No.1/2020, pp.: 1-5

```
1) root 101 49 0 (0.5148515 0.4851485)
  2) data1$hypertension< 0.5 47 17 0 (0.6382979 0.3617021)
    4) data1$Cholesterol>=213.5 30  8 0 (0.7333333 0.2666667) *
    5) data1$Cholesterol< 213.5 17  8 1 (0.4705882 0.5294118) *
  3) data1$hypertension>=0.5 54 22 1 (0.4074074 0.5925926)
    6) data1$age>=73.5 7  2 0 (0.7142857 0.2857143) *
    7) data1$age< 73.5 47 17 1 (0.3617021 0.6382979)
     14) data1$Cholesterol>=274 8  3 0 (0.6250000 0.3750000) *
     15) data1$Cholesterol< 274 39 12 1 (0.3076923 0.6923077)
       30) data1$Cholesterol< 175.5 7  3 0 (0.5714286 0.4285714) *
       31) data1$Cholesterol>=175.5 32  8 1 (0.2500000 0.7500000) *
```

The above model is displayed as a decision tree as below



Here the model start with splitting the data set using hypertension is present of not again if it is present then it checks whether cholesterol is more than 214 and accordingly it classifies the survival status into yes or no. Similarly in the other case of hypertension which is absent it includes variable age also in the classification. Here the model ignored variables gender, smoking and alcohol in determining survival status which is due to the nature and size of the data set.

Vol. 15 No.1/2020, pp.: 1-5

We can calculate the prediction accuracy of the above model using test data

```
cartpred <- cartmodel %>% predict(test, type = "class")
table(cartpred,test$survival)
```

The following actual vs predicted table is obtained

```
cartpred 0 1
       0 6 8
       1 3 4
```

From the above table we can compute model accuracy is 52% and it might be due to sample data set and variable selection.

For CHAID analysis we need to have data set with categorical and ordered dataset. The CHAID package [11] is needed to build the CHAID model.

We will bring the test data set into r studio environment and the following variable is included in the data set

- age
- gender
- diabetic
- hypertension
- smoking
- alcohol
- type
- severity
- Survival

Here Survival status is the dependent categorical variable. Here head (data1) statement will help us to display the first 5 rows of the dataset.

```
Data2<-read.xlsx('survivalcancer.xlsx',sheetIndex = 1)
head(data2)
```

INTERNATIONAL JOURNAL OF STATISTICS AND MEDICAL INFORMATICS

Vol. 15 No.1/2020, pp.: 1-5

We need to convert the required variables in to factors

```
data2$d1 <- factor(ifelse(data2$survival==1,"Survived", "Not Survived"))
data2$h1 <- as.factor(data2$severity)
data2$h2 <- as.factor(data2$hh)
data2$h4 <- as.factor(data2$age)
data2$h3 <- as.factor(data2$type)
```

We need to define Ctree control function and Ctree function

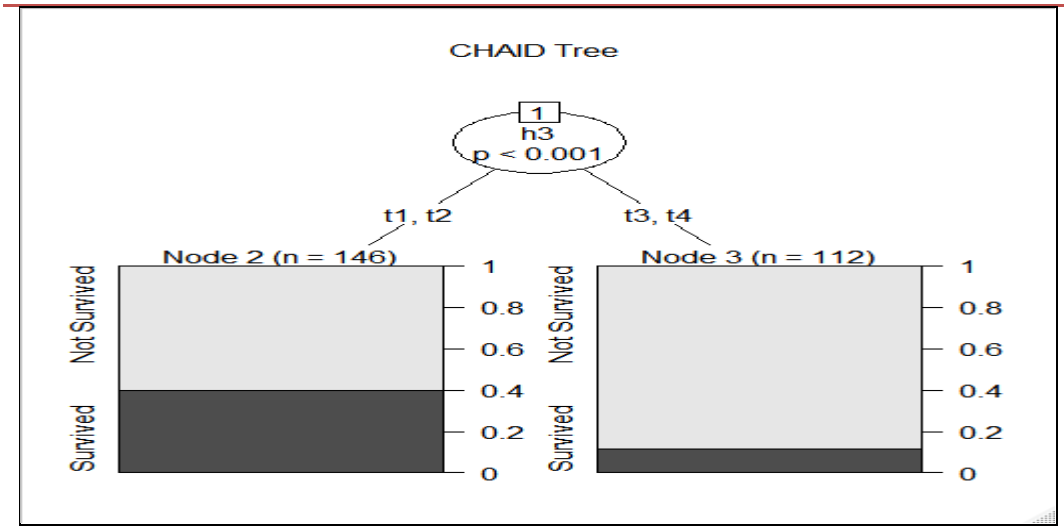
```
CHAIDctrl<- ctree_control(mincriterion = 0.95, minsplit = 100, minbucket = 100)
CHAIDTREEt <- ctree(d1~ h1 + h2 + h3 + h4,
        data=data2, control=ctrl)
```

The following model summary is obtained

```
Model formula:
d1 ~ h1 + h2 + h3 + h4

 [1] root
|   [2] h3 in t1, t2: Not Survived (n = 146, err = 39.7%)
|   [3] h3 in t3, t4: Not Survived (n = 112, err = 11.6%)

Number of inner nodes:    1
Number of terminal nodes: 2
```



The above model is displayed as a decision tree as below

Here only type variable is used to split the tree.

Vol. 15 No.1/2020, pp.: 1-5

Conclusion

The paper provided an overview of CART and CHAID tree with the help of Open Source R software and hypothetical data set. The accuracy of the model can be increased with the inclusion of more variables and samples.

Reference

1. Wilkinson, L. (1992). Tree structured data analysis: AID, CHAID and CART. Retrieved February, 1, 2008.
2. Batra, M., & Agrawal, R. (2018). Comparative analysis of decision tree algorithms. In Nature inspired computing (pp. 31-36). Springer, Singapore.
3. Shu, X. (2020). Classification and Decision Trees. In Knowledge Discovery in the Social Sciences (pp. 175-190). University of California Press.
4. Song, Y. Y., & Ying, L. U. (2015). Decision tree methods: applications for classification and prediction. Shanghai archives of psychiatry, 27(2), 130.
5. Haughton, D., & Oulabi, S. (1993). Direct marketing modeling with CART and CHAID. Journal of direct marketing, 7(3), 16-26.
6. R Core Team (2020). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/.
7. https://www.rstudio.com/products/rstudio/download/
8. https://cran.r-project.org/web/packages/xlsx/index.html
9. https://cran.r-project.org/web/packages/rpart/rpart.pdf
10. https://cran.r-project.org/web/packages/rpart.plot/rpart.plot.pdf
11. https://r-forge.r-project.org/R/?group_id=343