Some shallow introduction about CHAID decision tree

Zhang Zhiyuan 15220162202517

May 3, 2019

1 Definition

.

Chi-squared automatic interction detector, CHAID, was originally put forward by Kass. V. Goden in 1980. It's a tool to find the correlation between variables. It can be used to predict (like regression analysis, that's wht it was first called XAID), and classification. It can also be used to detect the interaction between variables.

It's core idea is: based on the given reponse variable and explanatory variable, enforce the optimal division on the training sample, and according to the significance of chi-square test, automatic judgment grouping of multivariate contingency tables is carried out. This way, by taking advantage of the adjusted singnificance test decision tree technology, it can quickly and effectively dig out the main determinant factor. It can not only process non-linear and highly correlated data, but also take missing value into consideration, thus overcome the restrictions of the traditional parametric tests.

2 Logic

2.1 Use chi-square automatic interactive detection method testing for determinant factors

.

First select the response variable to be divided, then apply cross classification, calculate the chi-squared value for the generaged two-dimensional classification tables, confirm the one with the smallest P-value. Repeat this process recursively for deeper classification, until a certain standard is reached.

2.2 Use the ROC curve to evaluate the performance of the CHAID method

.

This part is very similiar to the ROC curve we have learned in the previous lecture. This ROC curve is widely used in the assessment of medical diagnosis. The vertical and horiaontal axis are respectively the rate of false confirmnation and ture confirmnation. An area of 0.5 suggests total failure of performance. An area of 1 means perfect performance.

3 Features

.

CHAID method are generally more efficient in processing data with many variables and multi-classification or multi-range data. Compare to logistic regression, the analysis process of CHAID can demonstrate the interal influence among the variables, including variables in the subsamples.

Also, CHAID method is not very demanding regarding to the type of variables, expecially when the response variables are continious variables, escaping from the restriction of normal distribution.

Note: the application of CHAID method may generate a rather big decision tree, depending on the complexity of the data itself. So like the other decision tree method, the following pruning process is also very important.

4 Reference

.

The major source of information is:

Huang Qi, personal income analysis-based on CHAID decision tree, mathematical theory and application, 2009(4):33-37

He Fan, Shen Yi, Ye Zhong, CHAID method and its application, Chinese medical journal of precaution, 2005, 39(2):133-135

baidubaike-CHAID