

Using CHAID for classification problems

Dr Ray Hoare, HRS Ltd, ray@hrs.co.nz

A paper presented at the New Zealand Statistical Association 2004 conference, Wellington.

ABSTRACT

The author believes that CHAID analysis could be much more widely used than it is, as a tool for exploring commercial and scientific data. Two examples will be presented, one of which is based on real-world New Zealand financial data, with the objective of showing that CHAID can be an everyday tool for dealing with non-linear or complex datasets, in order to find significant patterns.

Introduction

I talk to many people about statistical software, and often ask them what they know about CHAID. I usually get a blank stare, even among people who are well trained in statistics. I also make a point of reading what I can about CHAID, and often find it is presented as one of those mysterious components of insanely expensive data mining programs, and therefore not of interest to mere mortals.

My own experience in dealing with client's data, and the experience of some of my more adventurous clients, is that CHAID, in one or other of its many forms, is a great way to sift certain kinds of data to find out where interesting relationships are buried, especially when the relationships are more complex than the linear or at least monotonic ones usually sought.

I will present in this paper two datasets that show very different kinds of application of the CHAID method. This printed paper shows the overview of what will be shown in a live investigation at a conference.

What is CHAID?

The acronym CHAID stands for *Chi*-squared Automatic Interaction Detector. Although it can be used for regression problems, in this paper I will only build classification trees. The "*Chi*-squared" part of the name arises because the technique essentially involves automatically constructing many cross-tabs, and working out statistical significance of the proportions. The most significant relationships are used to control the structure of a tree diagram.

Because the goal of classification trees is to predict or explain responses on a categorical dependent variable, the technique has much in common with the techniques used in the more traditional methods of Discriminant Analysis, Cluster Analysis, Nonparametric Statistics, and Nonlinear Estimation. The flexibility of classification trees make them a very attractive analysis option, but this is not to say

that their use is recommended to the exclusion of more traditional methods. Indeed, when the typically more stringent theoretical and distributional assumptions of more traditional methods are met, the traditional methods may be preferable. But as an exploratory technique, or as a technique of last resort when traditional methods fail, classification trees are, in the opinion of many researchers, unsurpassed.

Classification trees are widely used in applied fields as diverse as medicine (diagnosis), computer science (data structures), botany (classification), and psychology (decision theory). Classification trees readily lend themselves to being displayed graphically, helping to make them easier to interpret than they would be if only a strict numerical interpretation were possible.

Example 1. Correctly classifying all cases in a dataset

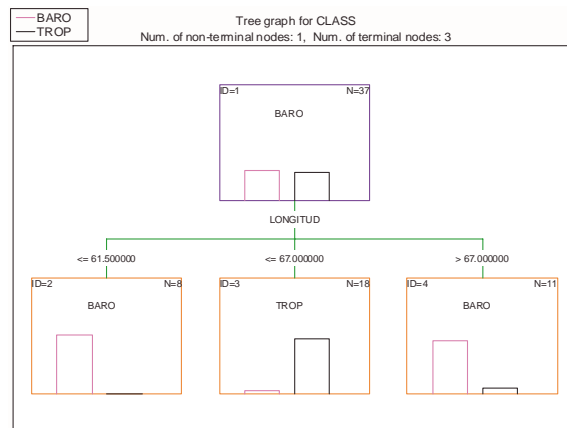
Lets take the artificial data in barotrop.sta, part of which is shown below.

	LONGITUD	LATITUDE	CLASS
1	61.50	17.00	BARO
2	61.50	19.00	BARO
3	62.00	14.00	BARO
4	63.00	15.00	TROP
5	63.50	19.00	TROP
6	64.00	12.00	TROP

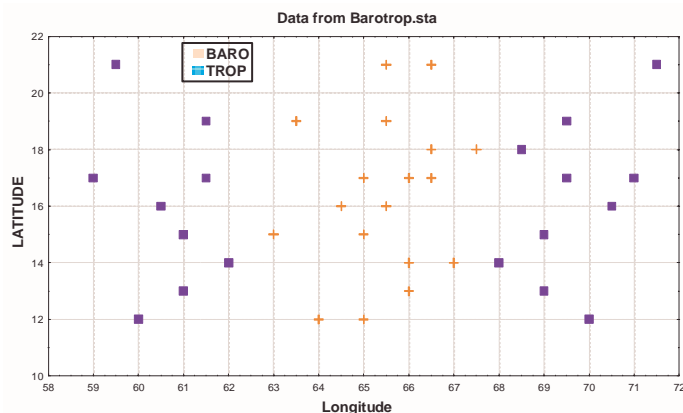
There are two kinds of hurricane, “BARO” and “TROPO”. We want to predict which kind of hurricane it will be from the longitude and latitude.

CHAID and other kinds of tree algorithm can classify the data so that if you know the latitude and longitude you will be able to say what kind of hurricane it is likely to be.

The data shown was analysed with Interactive CHAID in *STATISTICA*, and the tree graph shown here shows the algorithm has classified the data so that nearly all storms are correctly predicted, based entirely on the longitude, even though the latitude was supplied to the program.



The graph of longitude against latitude, with the hurricane type shown as different markers, shows why the classification is so good. In fact, it shows that the classification should be better! The imperfect classification arises because CHAID uses cross-tabs of categorical variables. When you have a continuous variable, such as



longitude, it is automatically broken up into sets of ranges. In this case, the sets do not divide at the value which would lead to perfect classification.

Not many real examples are as clear as this one. However, many cases occur where the object is similar – come up with a rule that leads to as good a classification as possible. Examples of this occur when you want to do a direct marketing campaign, and you need to hit as many good prospects as possible while minimising the wasted mailings.

Example 2: Locating “interesting” subsets in a dataset

Data used

I have been inviting customers or prospective customers to supply me with real data to explore, so that I can find out for myself how much of the manufacturer’s enthusiastic ravings to believe.

In most practical cases, I have found that the data does not lend itself to simplistic use such as the one above. Instead, it can be used to provide useful insights on subgroups of your data.

I will illustrate this with a dataset whose origin I am not at liberty to disclose, but which is local, and real apart from some descriptive labels I invented for clarity.

The table shows the descriptions of the variables available for the analysis. We want to find those factors that indicate whether someone is likely to default on a payment. Most of the data is from credit agency, but some is from the application form used to sign up for the service.

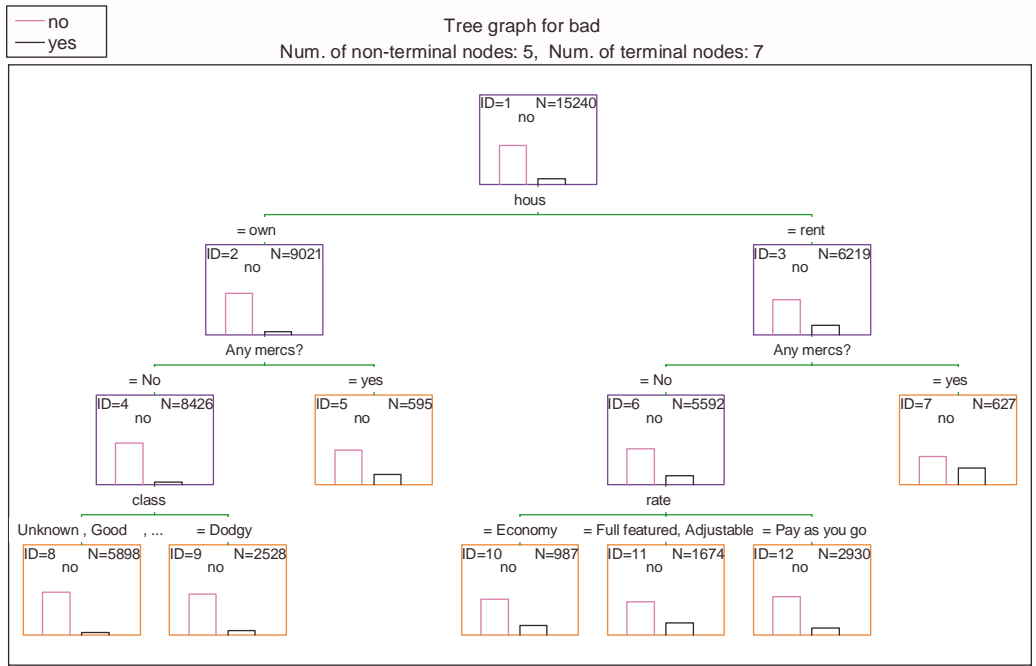
client defaulted on payment
client discontinued the service early
Age Group
number of defaults
number of mercantiles
number of judgements
number of bankruptcies
number of defaults last 180 days
number of mercantiles last 180 days
number of judgements last 180 days
Number of bankruptcies last 180 days
number of defaults last 365 days
number of mercantiles last 365 days
number of judgements last 365 days
Number of bankruptcies last 365 days
company's classification of the applicant
housing status
the category of the service applied for
occupation status
time with employer
time at address
number of enquiries at credit agency
first enquiry at credit agency for this applicant
application is the result of a promotion
switching from another supplier
month of application - don't use records where this is missing
Any mercs?
Any Judgements?
Any Mercs this year?
Any Judgements this year?

There are about 15,000 cases in use for the analysis.

Variable screening

Very often when you have large numbers of variables, some of them are irrelevant to the task in hand. In this case we will see if we can ignore some of the variables, by applying a screening algorithm. This will tell us which variables have a p-value less than 0.01. In this case, 24 out of 28 variables are significant. The most significant, according to the screening, are whether they own their own house, the number of mercantile disputes, the age of the applicant, and the occupation status. Bankruptcies do not have any importance.

I have chosen to put all the significant variables into a CHAID model, to see how well the model works.



The tree down to 3 levels is shown above. The height of each bar shows the proportion of “yes” and “no” answers in the Bad variable. No node has a preponderance of “Yes” answers in a cell.

This means that when you prepare the confusion matrix, or classification matrix, as shown here, no cases are classified “Yes”.

	no	yes
no	13214.00	2024.000
yes		

This may imply that the model is of no use, but really it means that it is of no use if the task is to classify a new respondent as one likely to default.

There are other uses for this sort of tree, of course.

The first split on the tree tells us that the best predictor of being a bad customer is whether the person owns or rents. Bad customers are significantly more common among renters. Within the renters, those with a “Merc” are more likely to default. This sounds reasonable.

We need to get more quantitative about the good and bad nodes. The tree is great for getting an overall impression, but *STATISTICA* allows you to save the data in a form that enables you to process it further.

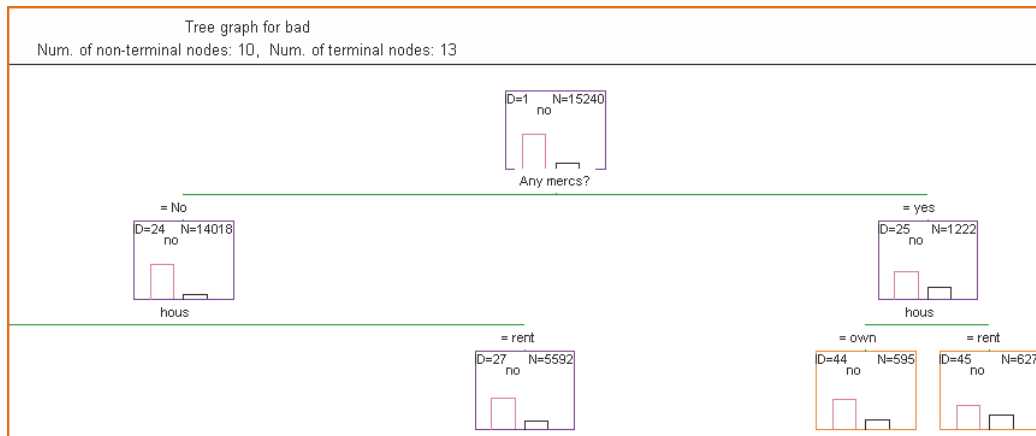
I have created a table of the numbers in each class in the terminal nodes, and then calculated the proportions of bad clients in each node.

Node	Class No	Class Yes	Proportion =v2/(v1+v2)
7	393	234	37%
11	1222	452	27%
5	456	139	23%
10	782	205	21%
12	2477	453	15%
9	2292	236	9%
8	5592	306	5%

We can see that 37% of the people in node 7 are bad clients. These are the people who rent their house and have “MerCs” against them. Node 11 has slightly better performance, but there are more of them. These are people on two specific payment plans.

Node 5 shows that people who own their own home but have “MerCs” are bad bets, too.

Since the “MerCs” show up so often, maybe we could explore whether the split on this is highly significant.



STATISTICA allows you to see the p-value for the split at any node, and in this case several variables are about equally probable. (Very low p-values are characteristic of large data sets, and are not always an advantage.) In the above graph I have forced the tree to be split on “MerCs”, and you can see that those with “MerCs” are poorer bets. Within those, then house ownership is still important.

Conclusion

Although the real world dataset did not contain relationships that enabled us to define groups that were most probably bad customers, the factors that pointed to this quality of a customer could be revealed by looking at the details within the table. Most of the measured data was relevant, except the previous occurrence of bankruptcy, and the investigation could be worth the effort of manually overriding the automatic selection of splits.