software review

Deeper into the Trees

A new hybrid CHAID application analyzes multiple dependent variables.

By Ken Deal

Tree-splitting algorithms were initially developed by John Sonquist and James Morgan (1964) and rendered into THAID by Morgan and R.C. Messenger (1973) of the Institute for Social Research at the University of Michigan. G.V. Kass transformed this into CHAID (Chi-square Automatic Interaction Detection) in 1980. Leo Breiman and Jerome Friedman independently worked on the innovation that became CART (Classification and Regression Trees) in the early 1980s. Since then, those algorithms and others have been developed into viable tree-splitting computational applications by SPSS, Statistical Innovations (SI), Salford Systems, SAS, Insightful Corp., and others.

The introduction of SI-CHAID 4.0 is much more than just another entrant to this field; it is a creative approach to handling the problem of effectively segmenting databases when there are multiple dependent variables, possibly of mixed types, and a large number of candidate predictor variables. In addition, SI-CHAID 4.0 (SIC4), along with its sister product Latent GOLD Choice 4.0 (LGC4), function in concert to perform this task on choice-based conjoint data sets.

To the best of my knowledge, all other tree-splitting algorithms focus on finding segments that help to predict one categorical criterion variable. However, there are many situations where it's possible to identify multiple candidate criterion variables. In those cases, it's necessary in conventional tree-splitting applications to generate a segmentation for each criterion variable separately. Then, of course, the segmentations are likely not to be congruent and the problem of choosing among those segmentations presents itself.

SI-CHAID 4.0 has been developed to provide exploratory segmentation trees that are predictive of multiple correlated dependent variables. These multiple criterion variables are probabilities of class membership obtained from LGC4, with the latent classes from LGC4 being proxies for the several dependent variables used in the analysis. It's also possible to use the algorithm for segmenting data sets containing only one dependent variable, and it is in that use where SIC4 becomes directly comparable to other existing tree-splitting applications.

LGC4, a latent class analysis application of Statistical Innovations (SI), is integrated with SPSS and SI-CHAID 4.0 to provide a nearly seamless way to identify latent classes based on multiple criterion variables and then to identify segments based on variables that can be used to score the related customer database. Of special value is the facility for LGC4 to work directly with Sawtooth Software CBC (choice-based conjoint) files in *.cho format to generate the latent classes that are moved into SIC4 and used to produce the segments. (Please note that LGC4 contains several substantial features not available in the earlier version that was reviewed in the Winter 2003 edition of *Marketing Research*.) Exhibit 1 illustrates the analysis process depending on the type of input data. SIC4 comprises SIC4 Define, which configures the analysis mechanism, and SIC4 Explore, which generates the tree and allows for interactive exploration of the tree.

Data input. SI has been successful at making data input very easy and output very instructive and attractive in all of its applications, and this is true of SI-CHAID as well. LGC4 and SIC4 both take SPSS files as direct input, and ASCII files also can be used. When using Latent GOLD Choice prior to SIC4 to identify latent classes, the resulting latent class file can then be immediately opened by the SIC4 design program. If LGC4 is used to estimate latent classes based on a Sawtooth Software choice-based conjoint *.chd file, there is just a little more work needed before the file can be worked on by the SIC4 Define program. If more flexibility is needed for handling a variety of data formats, the excellent file translator DBMS/COPY is available for a small premium.

SI-CHAID 4.0 has been developed to provide exploratory segmentation trees that are predictive of multiple correlated dependent variables.

Analysis. The analysis stage is very easy to specify. LGC4 can be directed to produce a file automatically for use in SIC4. If the data is in an ASCII file or SPSS file and contains just one dependent variable, a new project is opened in the SIC4 Define application. The dependent variable(s) is entered into one window, predictors into another, the case identification variable and any sample weighting variable into the last. There are several additional tabs that provide control over the algorithm, tree development, and nature of the variables. A Technical tab provides

Reprinted with permission from *Marketing Research*, Summer 2005, published by the American Marketing Association.

Exhibit 1 Generating segmentation trees

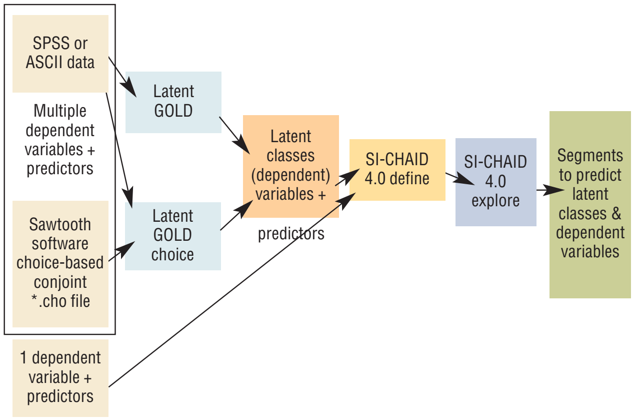
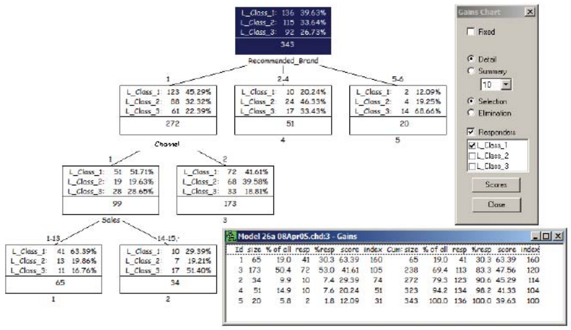

SPSS or ASCII data
Multiple dependent variables + predictors
Sawtooth software choice-based conjoint *.cho file
1 dependent variable + predictors
Latent GOLD
Latent GOLD choice
Latent classes (dependent) variables + predictors
SI-CHAID 4.0 define
SI-CHAID 4.0 explore
Segments to predict latent classes & dependent variables

Exhibit 2 Choice-based conjoint study



☐ Fixed
● Detail
○ Summary
10
● Selection
○ Elimination
☑ Responders
☑ L_Class_1
☐ L_Class_2
☐ L_Class_3

Model 26a 08Apr05.chd:3 - Gains

```
Id size % of all resp %resp  score index Cum: size % of all resp %resp score index
 1   65     19.0   41  30.3  63.39   160        65    19.0   41  30.3 63.39   160
 3  173     50.4   72  53.0  41.61   105       238    69.4  113  83.3 47.56   120
 2   34      9.9   10   7.4  29.39    74       272    79.3  123  90.6 45.29   114
 4   51     14.9   10   7.6  20.24    51       323    94.2  134  98.2 41.33   104
 5   20      5.8    2   1.8  12.09    31       343   100.0  136 100.0 39.63   100
```

control over the statistical calculations, such as whether log likelihood Chi-square or Pearson Chi-square is used, whether the Bonferroni adjustment is applied, and control over the amount of information provided in the report logs.

After defining the tree-splitting process, the Explore button is selected and the SIC4 Explore application is opened with the root node, the first level split or the full tree diagrammed, as requested. There is a great deal of tree control provided through a "Tree Node Display" window under the View menu. This control is so extensive that the tree can be reduced to just a set of branches or a set of boxes with no numbers whatsoever. The tree can be expanded to provide a large amount of information. In addition, control over the spacing between nodes vertically and horizontally is provided. This control helps the analyst to comfortably view large trees while also being able to expand the view when needed. These alterations are all done interactively so no time is wasted by changing settings and then rerunning the tree until it looks acceptable. The only control that seems not to be available is color formatting of the output tree.

Exhibit 2 shows a fully annotated small tree. A choice-based conjoint survey was conducted using Sawtooth Software SSI Web 5.0 to discover business customers' preferences for various brands of a product including additional attributes of price, construction material, brand of a component part, and warranty. Other brand and product preference information, sales, and firmographics were collected.

The data was analyzed initially using the Sawtooth Software Market Research Tools (SMRT) application. The Sawtooth Software *.chd file was processed by SMRT to include several covariates. That file was then taken into Latent GOLD Choice 4.0, which identified that three latent classes best captured the basic structure of the choice relationships. Sawtooth Software Latent Class Analysis was also used to identify the optimal number of latent classes and that analysis very closely matched the structure provided by LGC4. The set of three posterior probabilities of class membership was merged into the SPSS file of survey data and taken into SIC4 Define. After defining the analysis, SIC4 Explore produced the segmentation tree seen in Exhibit 2.

The one piece of information not provided on SIC4 tree diagrams directly is the statistical calculation for the splitting of each parent node. For example, Answer Trees from SPSS automatically provide that information, as do other tree-splitting applications. In SI-CHAID 4.0, the statistical information is obtained by selecting a node of interest and then asking for a "New Table" from the Windows menu. The table provided is a basic cross-tab with either the likelihood ratio Chi-square statistic or the Pearson Chi-square that was selected. The cross-tab can be configured as row percents, column percents, total percents, or frequencies.

In addition, the quality of the split at any node can be investigated by expanding the table request to include all other possible statistically significant splits of the selected parent node or, in the extreme, all possible splits of that node. For some, this might seem like overkill. However, if you are very diligent in investigating the data, this feature can save the work of recasting the tree to explore other options. If a more preferred split is found from the cross-tabs, that split can be selected for the parent node being investigated and calculated for that node immediately rather than having to recalculate the complete tree. This information regarding the significance of splits from any node can be obtained also by requesting Select under the Tree menu. That action produces a table listing significant and other splits from that node along with their significance levels, categories, and groups.

As a purely exploratory vehicle, the tree provides a way to identify the rank order of significance of the covariates and merging categories that are not significantly different. The merging process is somewhat controllable by specifying whether any categories can be combined or just adjacent categories. When a node is split by a covariate whose categories are merged, a cross-tab can be produced to show the relationship and to unmerge the categories and recalculate the Chi-square statistic. SI-CHAID 4.0 is billed as an exploratory application and, to enhance this process, multiple trees can be open simultaneously and displayed on the split screen.

The value of the segmentation to marketing strategies can be investigated through calculation of gains, profits, or costs and this information is provided in a gains table within SI-CHAID 4.0. The gains table is obtained from the Windows menu and can be adapted to provide information on any level of the tree and in a variety of formats. A gains table is included in Exhibit 2.

The value of the segmentation to marketing strategies can be investigated through calculation of gains, profits, or costs and this information is provided in a gains table within SI-CHAID 4.0.

The final key deliverable from data mining is the resource for scoring the data file and identifying segment membership for all cases. SIC4 produces if-then-else code in either SPSS or C format when the "New Source" option is selected under the Windows menu. This syntax code can be pasted into the syntax window of SPSS to assign each respondent to one of the several segments produced by SIC4. At that point, tabular and other output can identify how accurately the segments predict each of the several dependent variables on which the latent classes were calculated and how well the latent classes are predicted. Jay Magidson and Jeroen Vermunt (2005) stated, "Since the attributes are now included as additional dependent variables (the latent classes are proxy for these dependent variables), we might expect that the resulting segments might predict any single dependent variable less well than CHAID based on only that dependent variable."

If the analysis is based on a choice-based conjoint study produced using Sawtooth Software, the SIC4 segments could be merged back into SMRT for additional analysis. For example, simulations could be conducted by each segment identified by SIC4. Alternatively, individual coefficients could be calculated using hierarchical Bayes analysis, merged into the study SPSS files, and analyzed by segment. Another option is to bring each of the original dependent variables into the final tree and assess the ability of the tree to estimate those dependent variables.

Output. The analysis produces a rich range of graphs and tables. These include the tree diagram; a gains chart; cross-tabs for each specified node that include the selected Chi-square statistics; a table "select tree" that lists all splits from a node along with the statistics; the syntax code in SPSS or C for scoring the data file; the option to produce the tree in a more compact diagram; and a log of the output process.

Gains tables are produced to direct the researcher to those subgroups that promise the greatest returns for their relative marketing efforts, which is often related to subgroup size. Naturally, one would like to get the greatest benefit from the least amount of work. The gains table in Exhibit 2 shows several columns: the size of the segment node, the percentage of all respondents in that segment, the number of responders (i.e., the number of respondents in latent class 1 who are classed in segment 1), the percentage of responders out of the total number of responders (i.e., 41 out of 136), the average score of the dependent variable, and the index, which is the score for a specific segment relative to the average score for the total sample in the study. Cumulative information is also provided. The gains table can be changed to reflect one segment only by either clicking on that segment's terminal node in the tree or on the corresponding row in the gains table.

Summary. SI-CHAID 4.0 provides a level of sophistication for gaining insight into segmentation of complex structures containing multiple correlated dependent variables using simple-to-visualize trees. This application makes a valuable contribution to data analysis and provides an important step to the segmentation process that has not existed previously in accessible and easy-to-use commercial software. Statistical Innovations has produced another application that stands with Latent GOLD, Latent GOLD Choice, and GOLDMineR as being statistically sound, very easy to use, and visually rewarding. As with the other SI products, SI-CHAID 4.0 initially appears to be a very simple and basic application. This is due to the SI devotion to simplicity in the basic application with almost hidden flexibility that provides outstanding analysis and visual output. In my opinion, SI continues to achieve the highest levels of elegance in its analytics and its output. The manuals and tutorials are of professional quality and very informative.

SI-CHAID 4.0 is available from Statistical Innovations in Belmont, Mass. and is priced at $995 for the commercial license and $695 for an academic license. Please visit www.statisticalinnovations.com for additional information about SI-CHAID 4.0 and other SI products. ●

Ken Deal is chairman of strategic market leadership and health services management at the DeGroote School of Business, McMaster University in Hamilton, Ontario, and president of marketPOWER research inc. in Burlington, Ontario. He may be reached at kendeal@marketpowerresearch.com.