Editorial: Some musings on cluster analysis

When market researchers are asked what they do, many of them reply that they do market segmentation at least some of the time. It is natural for the market researcher to assume that their market is not homogeneous, and therefore they desire to identify segments of customers to be able to target them in distinct ways. The method of choice for finding segments is cluster analysis. In this editorial, I look at cluster analysis from a number of points of view.

SLICING THE MELON

A market researcher once likened cluster analysis to slicing a melon. That is, the melon can be sliced this way, or it can be sliced that way, or it can be sliced this way and that way.

To make this concrete, suppose the annual household income for a sample of individuals is known. The individuals can be clustered into two clusters, which would turn out to be low income and high income. Or, the individuals can be clustered into three clusters, which would turn out to be low, medium and high income. While the descriptions of the clusters can be interesting, it is much more interesting to relate the clusters to behaviour, such as a purchase.

ARE CLUSTERS NATURAL GROUPINGS?

Are clusters natural distinct groupings? In some of the sciences, botany for example, clusters are natural and

correspond to species or hybrids. In marketing, it often seems that clusters are 'convenient fictions'. That is, the market researcher might try all cluster solutions from K equals two to ten, on the thought that two are too few and ten are too many. Here, 'too many' means that ten target markets are too many to administer. Through inspection of results, the market researcher arrives at a provisional value for K . While K in the range of two to ten is normative, customer relationship management articles have indicated that at one casino market researchers work with 90 segments. Are these 90 segments natural groupings, or are they convenient fictions sorted by, for example, propensity to spend?

A LOOK AT K-MEANS CLUSTERING

Cluster analysis has the goal of grouping a collection of objects into subsets (or 'clusters') such that objects within each cluster are more closely related to one another than objects assigned to different clusters. Of the cluster methods available, K-means clustering is the method most used in practice among market researchers. This is because K-means clustering does not require computation and storage of the matrix of pairwise distances between objects, making it suited to files with larger numbers of records.

Intuitively, in cluster analysis, marketers attempt to maximise

between-cluster variance or, equivalently, minimise within-cluster variance. In principle, K could be specified and then an attempt could be made to minimise within-cluster variance over all possible assignments of the data points to the K clusters. Such an operation, however, is combinatorially complex and not feasible for anything beyond very small data sets. Hence, the heuristic approach of K -means clustering is used.

In general terms, K -means clustering proceeds as follows. Start by specifying K along with a guess for initial cluster centres. Then, alternate the following steps until convergence:

- for each data point, identify the closest cluster centre (in Euclidean distance) and assign the case to that cluster
- re-compute each cluster centre based on points assigned to the cluster.

The assignment of cases to clusters is done in a way that improves the within-cluster variance, but the K -means algorithm does not consider all possible assignments of points to clusters, so the solution converges only to a local optimum that can be highly suboptimal when compared to the global optimum.

I suspect that many users of K -means assume that it magically and correctly finds clusters. These users might not be aware of some of its limitations and drawbacks. Here they are:

- *ad hoc* choice of K : the number of clusters must be specified, at least provisionally. Most of the time, K is not known *a priori*. Instead, the researcher tries a range of values for K . Since within-cluster variance will almost certainly decrease as K is increased, a smaller within-cluster variance cannot be used as a selection criterion for K . Moreover, this will be

the case even when within-cluster variance is assessed on an independent holdout data set, thus cross-validation cannot be used to select K

- when K is varied, solutions are not nested
- an initial set of cluster centres must be specified. It is not clear how to choose initial centres. Some software performs a data pass and selects K ‘well-separated’ points. Another approach is to use multiple random starts. One recommendation is to first run hierarchical cluster on a subset of records to identify a tentative K along with cluster centroids that can then be used as inputs to K -means cluster. In any event, poor choice of cluster centres can lead to a bad cluster solution
- the solution is contingent on the numerical scale of the variables. That is, perform a linear transformation on one or more variables, and you will in general get different assignment of cases to clusters relative to the solution on the untransformed variables. What is more, other things equal, the variables with larger variance will tend to dominate the solution
- when squared Euclidean distance is used, the largest distances have the highest influence, and K -means can lack robustness against outliers
- in many situations, there tend to be variables of mixed scale types to work with, some continuous, some ordinal, some nominal, and some counts. K -means clustering is limited to quantitative variables having interval or ratio scaling
- K -means cluster tends to fit spherical clusters in the metrics of the variables. In effect, K -means assumes that the variables in the analysis have equal variance and zero correlation. If, in reality, variables have different

variances or different associations across clusters, then K-means cluster will fit spurious clusters.

The above points on the limitations of K-means cluster should make one think about the many existing published segmentation studies. Articles featuring segmentation should give the details of how segments were arrived at, and whether the author made any attempt to validate the clusters.

SOME EMERGING CLUSTER APPROACHES

There are some emerging cluster methods that will see more use in the near future.

- Tibshirani, Walther and Hastie have proposed a gap statistic that helps the analyst identify the 'optimal' number of clusters
- finite mixture models posit that the data are a sample from a population described by a probability density function. The probability density function is described by a parameterised model that is a mixture of component density functions, each of which describes a cluster. A mixture-of-normals model can accommodate different variances across variables and different degrees

of correlation between variables in different clusters. Viewed through the lens of finite mixture models, K-means cluster is seen to be a very restrictive mixture model. The mixture model approach provides a goodness-of-fit test for the number of clusters. One drawback of the mixture model approach is that high-dimensional data (number of variables very large) would lead to within-groups covariance matrices with many elements that might be at best imprecisely estimated

- another class of approaches is mode seeking or bump hunting. These methods take a nonparametric approach, attempting to estimate directly distinct modes of the probability density function of the data. Friedman's patient rule induction method (PRIM) approach is an example of this
- methods will arise for cluster validation.

These and other approaches will undoubtedly see wider use in the future. Let us look forward to their use in targeting in marketing.

TONY BABINEC
Editorial Board
March 2002