RESEARCH WHITEPAPER  June, 2009

A Framework for Market Segmentation

by Joe Retzer, Director, Marketing Sciences, Maritz Research
Market segmentation studies should ultimately inform marketing decisions. In order for this to happen, segments must be high quality (i.e., cluster members are similar to each other and clusters are distinct from other clusters) and effective in driving marketing tactics. Often, however, resultant segments do not satisfy the above requirements. This may be due to a variety of reasons, including:

- Too many variables are included in the segmentation, resulting in too much random variability from irrelevant variables that makes finding an effective cluster solution much more difficult
- The algorithm may be inherently unstable, lessening confidence in finding the “best” result
- The clustering algorithm ignores profiling requirements on tactical variables critical to translating research results into business impact
- Meaningful clusters exist which are difficult to detect with traditional clustering techniques

In the Maritz framework for market segmentation, each of these issues is addressed simultaneously using state-of-the art data mining algorithms from the fields of both unsupervised and supervised learning. This article begins by looking at each issue individually and then presents an integrated framework for segmentation.

**A Common Problem: Irrelevant variables**

Data sets used in market segmentation studies have become significantly larger and more complex in recent years. This requires clustering algorithms to adapt to maintain quality as the chances of including “noisy” irrelevant data has increased proportionally. The concern arising from the inclusion of noisy data is that meaningful segments may become undetectable because they’ve been “dispersed” by unrelated noise. In fact, in very high dimensions it is common for all of the objects/subjects in a data set to be nearly equidistant from each other, completely masking true underlying clusters.

Recent work detailed in a paper by Friedman and Meulman (2004) addresses this issue with an innovative approach referred to as COSA (Clustering Objects on Subsets of Attributes). COSA effectively identifies relevant weights for cluster basis variables allowing for the identification of underlying clusters in the presence of noisy data. Results on studies using both actual and synthetic data has shown COSA to be a very promising approach to cluster analysis when irrelevant and/or noisy variables are present in the data.

**Improving Cluster Quality using Ensembles**

Cluster Ensemble or consensus clustering analysis is a computationally intense data mining technique representing a recent advance in unsupervised learning analysis (see Strehl, A. and J. Gosh (2002)). It has been suggested as a generic approach for improving the accuracy and stability of base clustering algorithm results. Cluster Ensemble analysis begins by generating multiple cluster solutions using a collection of “base learner” algorithms (e.g., PAM (Partitioning Around Medoids), finite mixture models, k-means, etc.). It next derives a “consensus” solution that is more robust and of higher quality than any of the individual ensemble members used to create it. Cluster Ensemble solutions exhibit lower sensitivity to noise, outliers and sampling variations. Additional benefits of Cluster Ensemble analysis based on empirical studies utilizing synthetic data include:

- Cluster Ensemble analysis effectively detects and portrays meaningful clusters that cannot be identified with each individual technique.
- Cluster Ensemble analysis converges to true underlying clusters as the number of partitions increase provided the ensemble is sufficiently diverse. Facilitating Marketing Tactics using Random Forest Profiling Profiling based on a priori groupings (e.g., purchasers vs. non-purchasers) can provide information useful for informing tactical marketing decisions (e.g., direct mailing campaigns).
This is sometimes referred to as “supervised learning.” A new and highly effective approach to supervised learning is Random Forest (RF) analysis (see Breiman 2001). RF analysis is a powerful data mining technique based on the recursive partitioning algorithm CART. Empirical evidence suggests that Random Forests have much higher predictive accuracy than a single tree while not over-fitting the data. RF analysis employs bundled fits produced by an algorithm whose output is some combination of many passes through the data. RF effectively handles both categorical and continuous predictors and detects higher order interactions that need not be specified a priori. RF models are in fact unexcelled in accuracy among current predictive algorithms.

Most importantly for our market segmentation framework, RF may be used to compute proximity's between case pairs for clustering. This input may in turn be employed to facilitate marketing strategy by producing clusters which profile well on known variables of interest.

Putting it all together
The final step in developing an effective framework for market segmentation requires us to bring together the variable weighting solutions of COSA, the robustness of Cluster Ensembles and the profiling benefits of Random Forest analysis. This can be accomplished seamlessly by including both COSA and Random Forest results as members of an ensemble in order to derive a consensus solution via Cluster Ensemble analysis. Specifically, both COSA and Random Forest solutions can be specified as pair-wise distance matrices between respondents. These ensemble members, which may be large in number and possibly weighted, are used to produce a final consensus which focuses on meaningful, non-noisy variables, profiles well on marketing strategy enablers and, most importantly, exhibits high quality segment solutions. The general framework, along with various informative outputs, is illustrated in the figure below.

![Diagram](image.png)

References