2000

Behavioral Market Segmentation of Binary Guest Survey Data with Bagged Clustering

Sara Dolnicar
University of Wollongong, s.dolnicar@uq.edu.au

Friedrich Leisch
Vienna University of Technology, Austria

Behavioral Market Segmentation of Binary Guest Survey Data with Bagged Clustering

Abstract

Binary survey data from the Austrian National Guest Survey conducted in the summer season of 1997 were used to identify behavioural market segments on the basis of vacation activity information. Bagged clustering overcomes a number of difficulties typically encountered when partitioning large binary data sets: The partitions have greater structural stability over repetitions of the algorithm and the question of the "correct" number of clusters is less important because of the hierarchical step of the cluster analysis. Finally, the bootstrap part of the algorithm provides means for assessing and visualizing segment stability for each input variable.

Disciplines

Business | Social and Behavioral Sciences

Behavioral Market Segmentation of Binary Guest Survey Data with Bagged Clustering

Sara Dolničar[1] and Friedrich Leisch[2]

[1] Department of Tourism and Leisure Studies,
Vienna University of Economics and Business Administration,
A-1090 Wien, Austria
Sara.Dolnicar@wu-wien.ac.at
[2] Department of Statistics and Probability Theory,
Vienna University of Technology, A-1040 Wien, Austria
Friedrich.Leisch@ci.tuwien.ac.at

Abstract. Binary survey data from the Austrian National Guest Survey conducted in the summer season of 1997 were used to identify behavioral market segments on the basis of vacation activity information. Bagged clustering overcomes a number of difficulties typically encountered when partitioning large binary data sets: The partitions have greater structural stability over repetitions of the algorithm and the question of the "correct" number of clusters is less important because of the hierarchical step of the cluster analysis. Finally, the bootstrap part of the algorithm provides means for assessing and visualizing segment stability for each input variable.

1 Introduction

The importance of binary data in social sciences is growing due to manifold reasons. Yes-no questions are simpler and faster to answer for respondents. Not only does this fact increase the chances of the respondents finishing a questionnaire and answering it in a concentrated, spontaneous and motivated manner, binary question format also allows the designer of the questionnaire to pose more questions, as the single answer is less tiring. This is especially important for studies, where attitudes towards a multitude of objects are questioned, thus dramatically increasing the number of answers expected from the respondents as it is typically the case with guest surveys within the field of tourism.

These developments lead to an increasing number of medium to large empirical binary data sets available for data analysis. Turning to the field of market segmentation, empirical binary survey data sets exclude a number of clustering techniques viable for analysis due to their size which seems to be too large for hierarchical and too small for parametric approaches. Most parametric approaches require very large amounts of data in relation to the number of variables, growing exponentially. For the use of latent class analysis, Formann [4] recommends a sample size of 5×2^k, a very strict requirement, especially when item batteries of

20 items are not unusual, as it is the case in market segmentation, be it with demographic, socioeconomic, behavioral or psycho-graphic variables. Unless these huge data sets are available, exploratory clustering techniques will broadly be applied to analyze the heterogeneity underlying the population sample.

Among the exploratory approaches, the hierarchical clustering techniques require the data sets to be rather small, as all pairwise distances need to be computed in every single step of the analysis. This leaves us with partitioning approaches like learning vector quantization (LVQ) within the family of cluster analytic techniques. However, partitioning cluster methods typically give less insight into the structure of the data, as the number of clusters has to be specified a-priori and solutions for different number of clusters can often not be easily compared. Myers & Tauber [9] state in their classic book on market structure analysis that hierarchical clustering better shows how individuals combine in terms of similarities and partitioning methods produce more homogeneous groups.

2 The bagged clustering approach

The central idea of bagged clustering [7, 8] is to stabilize partitioning methods like K-means or LVQ by repeatedly running the cluster algorithm and combining the results. Bagging [1], which stands for bootstrap aggregating, has been shown a very succesfull method for enhancing regression and classification algorithms. Bagged clustering applies the main idea of combining several predictors trained on bootstrap sets in the cluster analysis framework. K-means is an unstable method in the sense that in many runs one will not find the global optimum of the error function but a local optimum only. Both initializations and small changes in the training set can have big influence on the actual local minimum where the algorithm converges.

By repeatedly training on new data sets one gets different solutions which should on average be independent from training set influence and random initializations. We can obtain a collection of training sets by sampling from the empirical distribution of the original data, i.e., by bootstrapping. We then run any partitioning cluster algorithm—called the *base cluster method* below—on each of these training sets.

Bagged clustering explores the independent solutions from several runs of the base method using hierarchical clustering. Hence, it can also be seen as an evaluation of the base method by means of the bootstrap. This allows the researcher to identify structurally stable (regions of) centers which are found repeatedly.

The algorithm works as follows:

1. Construct B bootstrap training samples $\mathcal{X}_N^1, \ldots, \mathcal{X}_N^B$ by drawing with replacement from the original sample $\mathcal{X}_N$.
2. Run the base cluster method (K-means, LVQ, . . .) on each set, resulting in $B \times K$ centers $c_{11}, c_{12}, \ldots, c_{1K}, c_{21}, \ldots, c_{BK}$ where K is the number of centers used in the base method and c_{ij} is the j-th center found using $\mathcal{X}_N^i$.

3. Combine all centers into a new data set $\mathcal{C}^B = \mathcal{C}^B(K) = \{c_{11}, \ldots, c_{BK}\}$.
4. Run a hierarchical cluster algorithm on $\mathcal{C}^B$ (or $\mathcal{C}^B_{\text{prune}}$), resulting in the usual dendrogram.
5. Let $c(x) \in \mathcal{C}^B$ denote the center closest to x. A partition of the original data can now be obtained by cutting the dendrogram at a certain level, resulting in a partition $\mathcal{C}^B_1, \ldots, \mathcal{C}^B_m$, $1 \le m \le BK$, of set $\mathcal{C}^B$. Each point $x \in \mathcal{X}_N$ is now assigned to the cluster containing $c(x)$.

The algorithm has been shown to compare favorably to several standard clustering methods on binary and metric benchmark data sets [7]; please see [8] for a detailed analysis and experiments using artificial data with known structure, as space constraints do not allow us to include the results in this paper. Especially the exploratory nature of the approach is attractive for practitioners [3].

3   Behavioral segmentation of tourist survey

Our application consists of the segmentation of tourist surveys for marketing purposes. A data set including 5365 respondents and 12 variables was used. The respondents were tourists spending their vacation in the rural area of Austria during the summer season of 1997, city tourists were excluded from the study. These visitors were questioned in the course of the Austrian National Guest Survey. The vacation activities used for behavioral segmentation purposes were:

Activity	Agreement (%)
cycling	30.21
swimming	62.65
going to a spa	14.61
hiking	75.62
going for walks	93.25
organized excursions	21.62
excursions	77.04
relaxing	80.17
shopping	71.50
sightseeing	78.02
museums	45.09
using health facilities	13.61

The task is to find market segments having homogeneous preferences in some of the activities. In addition to the variables that were used as segmentation base, a number of demographic, socioeconomic, behavioral and psycho-graphic background variables is available in the extensive guest survey data set: age, daily expenditures per person, monthly disposable income, length of stay, intention to revisit Austria, intention to recommend Austria, number of prior vacations in Austria, etc. These variables were not used as input in the cluster analysis, as only homogeneous groups with respect to vacation activities were of interest. The background variables are only used to describe the market segments in more detail.
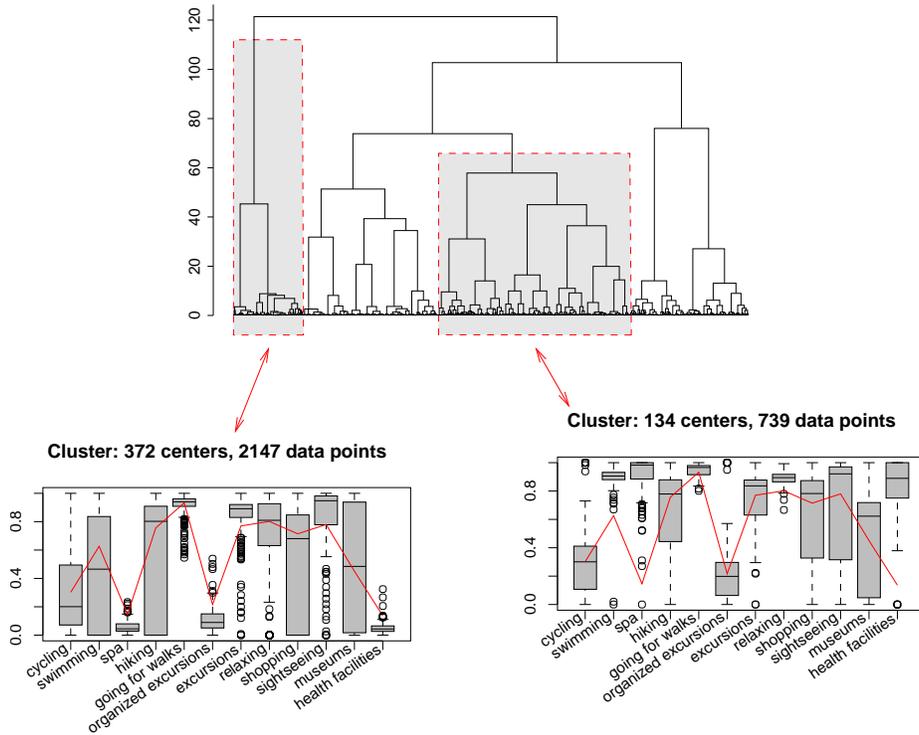
Fig. 1. Bagged clustering dendrogram together with boxplots for two selected clusters

The upper part of Figure 1 depicts the dendrogram resulting from a bagged clustering analysis. Learning vector quantization (e.g., [10]) was used as base method with $K = 20$ centers in each run on $B = 50$ training sets. The resulting 1000 centers were then hierarchically clustered using Euclidean distance and Ward's linkage method (e.g., [6]). We also tried other parameter combinations, but results were very similar, as the algorithm is not very sensitive to B and K once these are large enough.

Bagged clustering has been implemented using the R package for statistical computing (`http://www.R-project.org`), a free implementation of the S language; R functions for bagged clustering can be obtained from the authors upon request. The software allows interactive exploration of the dendrogram: by clicking on a subtree of the dendrogram one gets (in another window) a box-whisker plot of the centers in the corresponding cluster $\mathcal{C}_i^B$. Figure 1 shows as example boxplots corresponding to two such subtrees.

The boxes range from the 25% quantile to the 75% quantile, the line in the middle represents the median, the whiskers and circles depict outliers. See the documentation of any modern statistics package for more details on box-whisker plots. The horizontal polygon depicts the overall sample mean such that one can

easily compare which variables are so-called marker variables of the segment, i.e., are different in the segment than in the overall population and can be repeatedly found having similar values such that the corresponding boxes are small.

The market segments corresponding to the two boxplots in Figure 1 can be described as follows:

– **Individual sightseers (left plot):** This large segment (40 percent of the tourists questioned) have a clear focus when visiting Austria: They want to hop from sight to sight. Therefore both the items sightseeing and excursions are strongly and commonly agreed upon in this group. Neither sports nor shopping are of central importance, although some members do spend some of their leisure time undertaking those activities. Well reflecting the individualist character of this group is the heterogeneity of this segment concerning a number of activities, as e.g. swimming, hiking, shopping or visiting museums.

– **Health oriented holiday-makers (right plot):** This niche segment represents a very stable and distinct interest group. Clearly, these tourists spend their vacation swimming and relaxing in spas and health facilities. Also, they all seem to enjoy going for a walk (after the pool is closed?). As far as the remaining activities are concerned, homogeneity decreases as indicated by the large dispersion of mean values.

The information which variables "define" a segment (small boxes) and with respect to which variables a segment is heterogenous (large boxes) is unique to the bagged cluster approach. A small box indicates that the corresponding cluster center was stably found over all repetitions of the base method. Bagged clustering bootstraps the base cluster method, hence the sizes of the boxes visualize the dispersion of the segment mean for each input variable and indicate how "correlated" an input variable is with the segment. This information is not available if the data set is partitioned only once. Note that we have only $\{0, 1\}$-valued data, hence data dispersion in a segment can also not be used.

The analysis of the background variables shows that the sightseeing tourists are rather young (median 48 years) very fond of Austria, intend to revisit the country to a high extent and spend an average amount of money per day (52 Euro). The health-oriented tourists are moderately older (median 53 years), have similar intent to revisit the country, however spend significantly more money per day (68 Euro).

4 Stability analysis

We have also compared the stability of standard K-means and LVQ with bagged versions thereof. K-Means and LVQ were independently repeated 100 times using $K = 3$ to 10 clusters. Runs where the algorithms converged in local minima (SSE more than 10% larger than best solution found) were discarded. Then 100 bagged solutions were computed using $K = 20$ for the base method and $B = 50$ training sets. The resulting dendrograms were cut into 3 to 10 clusters.

All partitions of each method were compared pairwise using one compliance measure from supervised learning (Kappa index, [2]) and one compliance measure from unsupervised learning (corrected Rand index, [5]). Suppose we want to compare two partitions summarized by the contingency table $T = [t_{ij}]$ where $i, j = 1, \ldots, K$ and t_{ij} denotes the number of data points which are in cluster i in the first partition and in cluster j in the second partition. Further let $t_{i\cdot}$ and $t_{\cdot j}$ denote the total number of data points in clusters i and j, respectively:

<div align="center">

Partition 2

		1	2	$\ldots$	K	$\sum$
	1	t_{11}	t_{22}	$\ldots$	t_{1K}	$t_{1\cdot}$
Partition 1	2	t_{21}	t_{22}	$\ldots$	t_{2K}	$t_{2\cdot}$
	$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\vdots$
	K	t_{K1}	t_{K2}	$\ldots$	t_{KK}	$t_{K\cdot}$
	$\sum$	$t_{\cdot 1}$	$t_{\cdot 2}$	$\ldots$	$t_{\cdot K}$	$t_{\cdot\cdot} = N$

</div>

In order to compute the Kappa index for an unsupervised classification problem, we first have to match the clusters from the two partitions such that they have maximal agreement. We do this by permuting the columns (or rows) of matrix T such that the trace $\sum_{i=1}^{K} t_{ii}$ of T gets maximal. In the following we assume that T has maximal trace.

Then the Kappa index is defined as

$$\kappa = \frac{N^{-1} \sum_{i=1}^{K} t_{ii} - N^{-2} \sum_{i=1}^{K} t_{i\cdot} t_{\cdot i}}{1 - N^{-2} \sum_{i=1}^{K} t_{i\cdot} t_{\cdot i}}$$

which is the agreement between the two partitions corrected for agreement by chance given row and column sums.

The Rand index measures agreement for unsupervised classifications and hence is invariant with respect to permutations of the columns or rows of T. Let A denote the number of all pairs of data points which are either put into the same cluster by both partitions or put into different clusters by both partitions. Conversely, let D denote the number of all pairs of data points that are put into one cluster in one partition, but into different clusters by the other partition. Hence, the partitions disagree for all pairs D and agree for all pairs A and $A + D = \binom{N}{2}$. The original Rand index is defined as $A/\binom{N}{2}$, we use a version corrected for agreement by chance [5] which can be computed directly from T as

$$\nu = \frac{\sum_{i,j=1}^{K} \binom{t_{ij}}{2} - \sum_{i=1}^{K} \binom{t_{i\cdot}}{2} \sum_{j=1}^{K} \binom{t_{\cdot j}}{2} / \binom{N}{2}}{\frac{1}{2}\left[\sum_{i=1}^{K} \binom{t_{i\cdot}}{2} + \sum_{j=1}^{K} \binom{t_{\cdot j}}{2}\right] - \sum_{i=1}^{K} \binom{t_{i\cdot}}{2} \sum_{j=1}^{K} \binom{t_{\cdot j}}{2} / \binom{N}{2}}$$

Figure 2 shows the mean and standard deviation of κ and ν for $K = 3, \ldots, 10$ clusters and $100 * 99/2 = 4950$ pairwise comparisons for each number of clusters. Bagging considerably increases the mean agreement of the partitions for all number of clusters while simultaneously having a smaller variance. Hence, the procedure stabilizes the base method. It can also be seen that LVQ is more stable than K-Means on this binary data set.
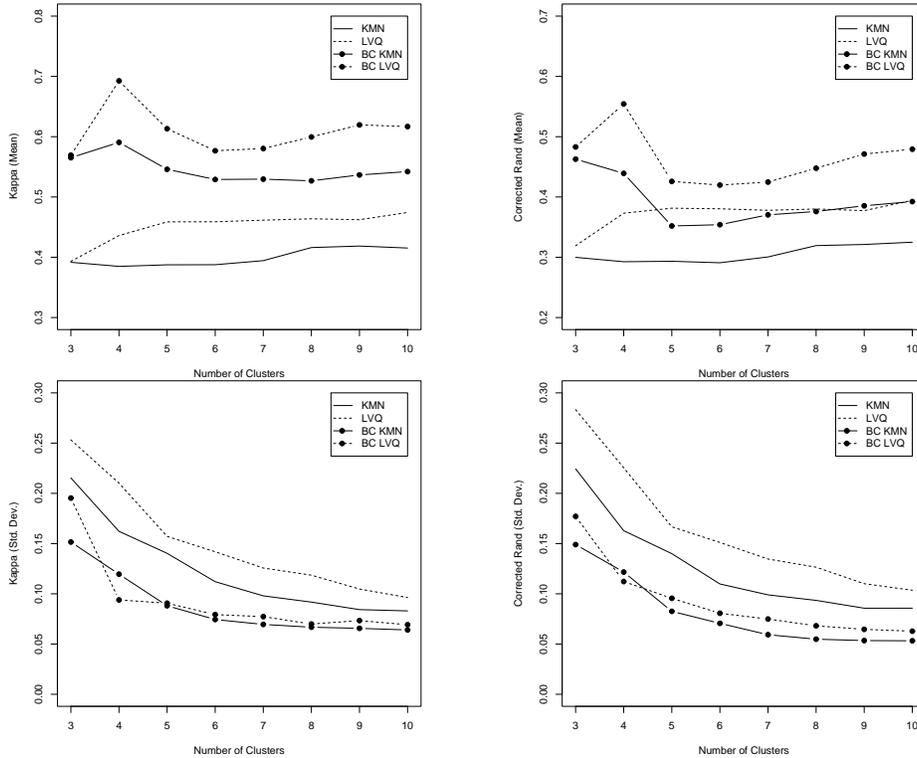
Fig. 2. Stability of clustering algorithms over 100 repetitions for 3 to 10 clusters: Mean kappa (top left), mean corrected Rand (top right), standard deviation of kappa (bottom left) and standard deviation of corrected Rand index (bottom right).

5   Summary

The bagged cluster algorithm has been applied to a binary data set from tourism marketing. Categorical data sets with very few categories are very common in the marketing sciences, yet most cluster algorithms are designed for metric input spaces, especially with Gaussian distributions. Hierarchical cluster methods allow for arbitrary distance measures (and hence arbitrary input spaces) but get quickly infeasible with increasing numbers of observations.

Bagged clustering overcomes these difficulties by combining hierarchical and partitioning methods. This allows for new exploratory data analysis techniques and cluster visualizations. Clusters can be split into sub-segments, each branch of the tree can be explored and the corresponding market segment identified and described.

By bootstrapping partitioning cluster methods we can measure the variance of the cluster centers for each input variable, which is especially important for binary data where usually only cluster centers without any variance information

are available. This leads to easy seperation of variables in which a segment is homegenous, and variables where a segment is rather heterogenous.

Finally, building complete ensembles of partitions also has a stabilizing effect on the base cluster method. The average agreement between repetitions of the algorithm is considerably increased, while the variance is reduced. The partitions found in 2 independent runs are more similar to each other, reducing the need for subjective decisions of the practitioner which solution to choose.

Our current work tries to generalize the approach to partitioning methods which are not necessarily represented by centers, e.g., fuzzy clusters. Using distance measures that operate on partitions directly (instead of representatives) these could then also be clustered using hierarchical techniques.

References

1. Leo Breiman. Bagging predictors. *Machine Learning*, 24:123–140, 1996.
2. J. Cohen. A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, 1960(20):37–46, 1960.
3. Sara Dolnicar and Friedrich Leisch. Behavioral market segmentation using the bagged clustering approach based on binary guest survey data: Exploring and visualizing unobserved heterogeneity. *Tourism Analysis*, 5(2–4):163–170, 2000.
4. Anton K. Formann. *Die Latent-Class-Analyse: Einfhrung in die Theorie und Anwendung*. Beltz, Weinheim, 1984.
5. Lawrence Hubert and Phipps Arabie. Comparing partitions. *Journal of Classification*, 2:193–218, 1985.
6. Leonard Kaufman and Peter J. Rousseeuw. *Finding Groups in Data*. John Wiley & Sons, Inc., New York, USA, 1990.
7. Friedrich Leisch. *Ensemble methods for neural clustering and classification*. PhD thesis, Institut für Statistik, Wahrscheinlichkeitstheorie und Versicherungsmathematik, Technische Universität Wien, Austria, 1998.
8. Friedrich Leisch. Bagged clustering. Working Paper 51, SFB "Adaptive Information Systems and Modeling in Economics and Management Science", Wirtschaftsuniversitt Wien, Austria, August 1999.
9. James H. Myers and Edward Tauber. *Market Structure Analysis*. American Marketing Association, Chicago, 1977.
10. Brian D. Ripley. *Pattern recognition and neural networks*. Cambridge University Press, Cambridge, UK, 1996.