Sawtooth Software

RESEARCH PAPER SERIES

Garbage Can Cluster Analysis: A Comparison of Variable Selection Techniques for Mixed Scale Data

Keith Chrzan
Sawtooth Software, Inc.

3210 N. Canyon Rd., Provo, Utah
+1 801 477 4700
www.sawtoothsoftware.com

Garbage Can Cluster Analysis:
A Comparison of Variable Selection Techniques for Mixed Scale Data

Keith Chrzan, Sawtooth Software
November 2022

Introduction

Many segmentation studies suffer from having too many basis variables. Frequently clients lack a strategy for pruning the many variables in their segmentation data sets (and they don't know that they need one). Faced with this garbage can approach, we struggle with too many variables, hampered by the curse of dimensionality and the need for larger samples than we can usually get. Moreover, such garbage cans frequently include masking variables which add noise that obscures cluster structure and damages segmentation results.

At our most recent conference, Joseph White and I found that when the garbage can contains only metric variables, some automatic variable selection methods exist that work quite well (Chrzan and White 2022). At the time we were unaware of methods for handling data sets with mixes of categorical and metric variables, but Joe Retzer pointed us toward one method and our subsequent research discovered two others. We now have three methods for variable selection of mixed scale data to choose from.

Using data sets generated with mixtures of metric and categorical variables, with both noisy and informative variables, we examine the relative success of these methods to identify known segment membership. It turns out that one of these methods performs much better than the others, so that we have a viable method for culling a garbage can of variables into something useful for segmentation.

Background

Why variable selection?

Segmentation is sample size intensive. By one measure (Formann 1984), we should have a sample size at least equal to 5×2^d, where d is the number of basis variables: for a study with 100 basis variables, for example, this suggests an impossible 6.34×10^{30} respondents. That estimate may be way to high, but even a more conservative recent source (Dolnicar *et al* 2018) suggests at least 100 observations per basis variable or 10,000 respondents in our case of 100 basis variables. Because such large sample sizes will usually exceed research budgets, we have good reason to want to reduce the number of basis variables.

We have a better reason still: the "curse of dimensionality" damages segmentation studies with large numbers of basis variables. The clearest statement of how the curse of dimensionality affects cluster analysis comes from Yiu (2019):

> When we have too many features, observations become harder to cluster — believe it or not, **too many dimensions causes every observation in your dataset to appear equidistant from all the others.** And because clustering uses a distance measure such as Euclidean distance to quantify the similarity between observations, this is a big problem. **If the distances are all approximately equal, then all the observations appear equally alike (as well as equally different), and no meaningful clusters can be formed.**

In other words, having too many basis variables works against the objectives of a segmentation study: all else being equal, it inclines the segments to be less differentiated the more basis variables we use.

For both of these reasons, analysts need a strategy to reduce the number of variables we put into our segmentation studies.

Variable selection methods for mixed scale data

Two "implicit" variable selection techniques use different methods to produce dissimilarity matrices that can serve as inputs for clustering. Because the clustering algorithm takes in the dissimilarity matrix and not the raw variables, the total number of variables no longer matters; these methods implicitly weight the variables so that influential variables contribute more to the dissimilarities than do noisy or masking variables.

Random Forests with PAM

Brieman and Cutler (2003) recommend using an unsupervised version of Brieman's random forests (RF), available in R as the randomForest package (Liaw and Wiener 2002). The method is unsupervised in that no variable in the initial data set supervises the trees. Instead, the analysis proceeds by making a new data set with the same number of variables and cases, with each individual variable independently having the same distribution as in the original data set (but with all the relationships between variables broken because of the independence). Next the algorithm combines the two data sets, adding a label of 1 to each record in the original data set and of 2 to each record in the new data set. This new variable becomes the supervising variable for an RF analysis that produces two outputs:

- a set of importances that identify the extent to which each variable influences the forest and
- a dissimilarity matrix based on the proportion of times any pair of cases ends up in the same leaf across all the trees in the forest

Now the analyst can choose to use (a) the dissimilarity matrix as an input to a partitioning around medoids (PAM) clustering or (b) the importances to cull the list of potential basis variables to a reasonable number and then proceed to cluster the data using a clustering algorithm of choice. The empirical test below features both the RF dissimilarity matrix/PAM pairing and culling of importances followed by model-based clustering to find segments.

Tree clustering

The second implicit method, tree clustering, also uses a large number of trees to compute a dissimilarity matrix. As implemented in R's treeClust package (Buttrey and Whitaker 2015), the method takes every variable in the data file in turn and uses it as the supervising variable for construction of a classification or a regression tree, predicted by all the others. Again, a dissimilarity matrix arises based on how often each pair of cases appears together in the leaves of the various trees. The algorithm then uses this dissimilarity matrix to support either PAM or k-means cluster generation. To guarantee a thorough search for good solutions, the k-means algorithm in the empirical study below uses 10,000 random initial sets of starting seeds.

Unfortunately, neither of these implicit approaches prevent redundant variables from influencing segments, a potential disadvantage relative to the automatic variable selection described below.

Automatic variable selection

Marbac and Sedki (2017, 2020) provide an automatic variable selection method for model-based clustering. Implemented as the VarSelLCM package in R, the method relies on an information criterion

the authors call Maximum-Integrated Complete Data Likelihood to do the variable selection. Usefully, the criterion doesn't require parameter estimation of any model-based clustering to do the model selection, so the optimal model can be identified in a relatively short time (less than a minute in the case of the analyses below). Also useful is the very clear direction the method provides about which variables are most discriminating: like the RF method above, analysts can use VarSelLCM to identify a small set of discriminating variables and then choose to perform any sort of clustering they like on that small set.

The comparisons below feature

- VarSelLCM, the automatic variable selection approach
- Two uses of the unsupervised RF:
    - the RF/PAM combination
    - using the RF importances to cull the variable list (dropping variables with importances - MeanDecreaseAccuracy - of less than 2.0) prior to model-based clustering
- Two flavors of treeClust
    - PAM and
    - k-means.

Empirical Test

Data

To test these methods, we generated a total of 120 data sets using the clusterGeneration package in R (Qiu and Joe 2020). The package improves upon the widely-used cluster generation approach described by Milligan (1985).

Each data set contains respondents known to belong to 4 clusters of random size ranging from n=50 to n=200 and with a "close" cluster structure featuring a small separation between clusters (a separation index of 0.01). The first 90 data sets had 18 variables, half of which we recoded to be categorical variables with 2-6 categories each. Of the 18 variables,

- 30 data sets have 9 informative and 9 "noisy" (masking) variables
- 30 data sets containing 6 informative and 12 noisy variables
- 30 data sets have 3 informative and 15 noisy variables

The final 30 data sets mimic a more extreme case and include 6 informative and 94 noisy variables

Previously Chrzan and White (2021) showed how difficult it is to identify the correct number of clusters. We assume that the analyst has succeeded in that difficult task so we can focus on the matter of variable selection.

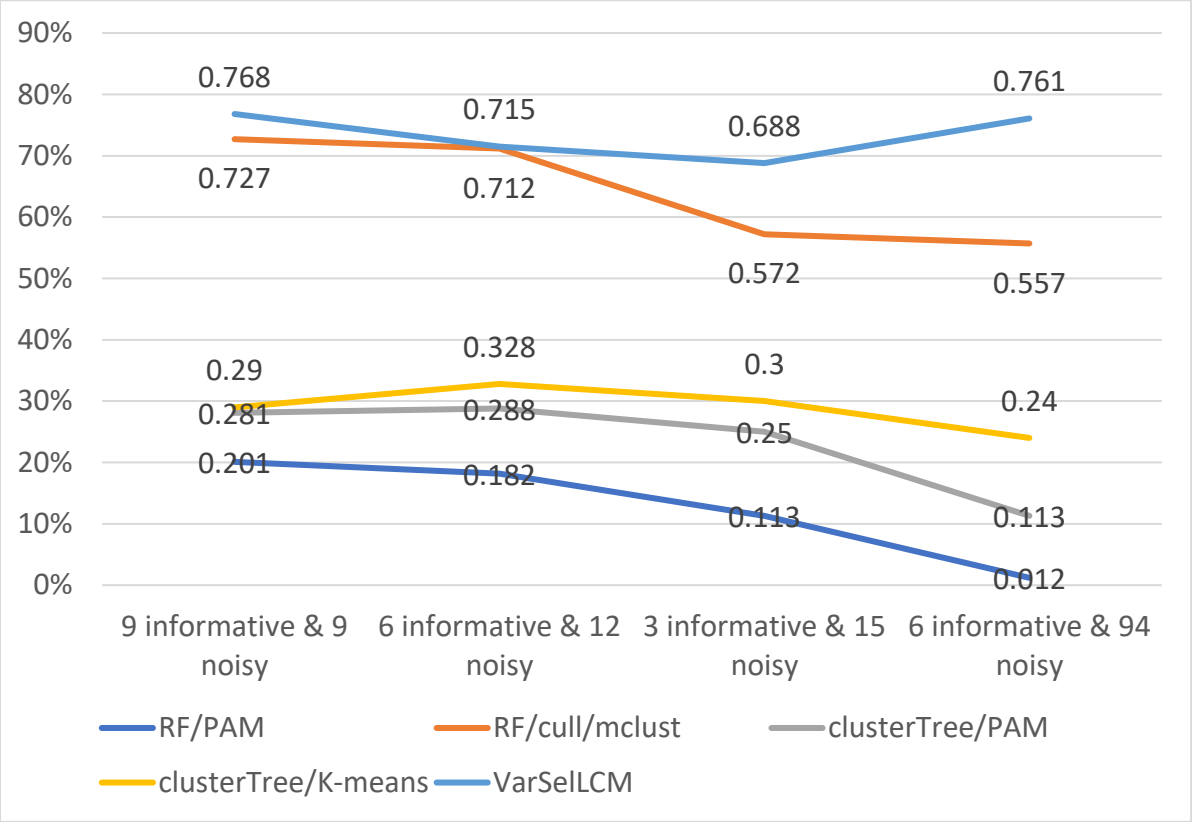
Success Criterion

Because we know segment membership in our data sets, we can use the Adjusted Rand Index (ARI) to assess the success of each variable selection approach in reproducing the known cluster structure. ARI ranges from 0 (actual and predicted segment membership are independent) to 1 (predicted membership maps perfectly to actual membership). An example of what an ARI of 0.708 (similar to some of the better results below) looks like appears as Appendix 1.

Results

Across the four experimental treatments, the automatic variable selection outperforms the four RF-based and treeClust-based methods in terms of ARI. Appendix 2 contains the tabular data to support this chart:

Exhibit 1: Mean ARI by Method and Treatment



In each case VarSelLCM has the highest level of ARI, which doesn't seem to depend on the ratio of noisy to informative variables. Using RF to cull the variable list before clustering succeeds roughly at parity with VarSelLCM when the ratio of noisy to informative variables is no more than 2:1 but beyond that it is has significantly lower fit than VarSelLCM . The two methods based on clusterTrees have much lower ARI and the RF/PAM combination has the worst ARI of all. VarSelLCM dominates the other methods for variable selection among mixed scale data.

While the combination of unsupervised RF with PAM performed significantly worse than the other methods at reproducing known cluster membership, using RF to identify informative variables performed better. Evidently, the variable importance measures that come out of unsupervised RF have more value than the variable weighting implicit in any of the dissimilarity matrices.

VarSelLCM allows, but does not require, categorical basis variables, so it can also work in the case where all variables are metric. Based on a limited number of runs, it appears to work as well as (if not a little better than) a program called clustvarsel (Scrucca and Raferty 2018), only many times faster. Chrzan and White (2022) found clustvarsel to outperform various manual variable selection methods for selection of metric variables. Given its parity performance and faster speed, analysts may want to use VarSelLCM even for garbage can segmentations based only on metric variables.

A cursory examination with a limited number of additional data sets confirms some previous findings, namely that increasing the number of clusters makes prediction harder for all the methods while increasing the separation between clusters makes it easier for all the methods.

Conclusions

The automatic variable selection method VarSelLCM dominated two methods that implicitly select influential variables for cluster analysis: the tree-based clusterTree method (paired with either PAM or K-means clustering) and unsupervised RF with PAM applied to the dissimilarity matrix. It also dominated model-based clustering based on variables culled using the Rf importances. VarSelLCM had a significantly greater average Adjusted Rand Index and it was most often the best fitting of the methods tested. That it also performs well when all basis variables are metric makes VarSelLCM a handy variable selection method solution no matter the mix of scales among our basis variables.

Suggestions for Future Research

VarSelLCM could also be used to estimate a range of solutions so as to identify the optimal number of clusters. As the number of clusters may be the single hardest decision that faces analysts in a segmentation study (Chrzan and White 2021), a valuable contribution by a future paper could be an empirical comparison of VarSelLCM and other methods for determining the number of clusters in a dataset.

Appendix 1 – What ARI of 0.708 Looks Like

Confusion matrix of true vs predicted segment membership:

	1	2	3	4
1	2	8	0	140
2	0	148	35	1
3	0	5	176	9
4	54	0	0	6

Appendix 2 – Mean, Range and (s.e.) or ARI by Method and Treatment

Variables	RF/PAM	RF/Cull/ mclust	clusterTree/ PAM	clusterTree/ K-means	VarSelLCM
9 informative/ 9 noisy	0.201 (0.020)	0.727 (0.024)	0.281 (0.019)	0.290 (0.019)	0.768 (0.026)
6 informative/ 12 noisy	0.182 (0.017)	.0712 (0.038)	0.288 (0.028)	0.328 (0.027)	0.715 (0.039)
3 informative/ 15 noisy	0.113 (0.010)	0.572 (0.048)	0.250 (0.021)	0.300 (0.036)	0.688 (0.041)
6 informative/ 94 noisy	0.012 (0.002)	0.557 (0.051)	0.113 (0.009)	0.240 (0.031)	0.761 (0.041)

References

- Breiman, L., and Cutler, A. (2003) "Random forests manual v4.0", technical report, UC Berkeley, available online at ftp://ftp.stat.berkeley.edu/pub/users/breiman/Using random forests v4.0.pdf.
- Buttrey, S.E. and L.R. Whitaker (2015) "treeClust: An R package for tree-based clustering dissimilarities," The R Journal, **7** (2), 227-236
- Chrzan, K. and J. White (2021) "Replication of known segment structure and membership," *Sawtooth Software Conference Proceedings*, 217-226.
- Chrzan, K. and J. White (2022) "Variable selection in segmentation," Forthcoming in the *Sawtooth Software Conference Proceedings*.
- Dolnicar, S., B. Grün, & F. Leisch (2018) *Market segmentation analysis: Understanding it, doing it, and making it useful.* Singapore: Springer.
- Formann, A. (1984) Die Latent-Class-Analyse: Einführung in die Theorie und Anwendung. Beltz, Weinheim.
- Liaw, A. and M. Wiener (2002). "Classification and Regression by randomForest." *R News*, **2**(3), 18-22. https://CRAN.R-project.org/doc/Rnews/.
- Marbac, M. and M. Sedki (2017) "Variable selection for model-based clustering using the integrated complete-data likelihood," *Statistics and Computing*, **27** (4), pp 1049–1063.
- Marbac, M. and M. Sedki (2020) "VarSelLCM: Variable selection for model-based clustering of mixed-type data set with missing values," https://cran.r-project.org/web/packages/VarSelLCM/VarSelLCM.pdf.
- Milligan G. W. (1985) "An algorithm for generating artificial test clusters," *Psychometrika*, **50:** 123–127.
- Qiu, W. & H. Joe (2020) "clusterGeneration: random cluster generation (with specified degree of separation)." https://cran.r-project.org/web/packages/clusterGeneration/clusterGeneration.pdf.
- Scrucca, L. & A.E. Raferty (2018)"clustvarsel: A package implementing variable selection for Gaussian model-based clustering in R," *Journal of Statistical Software*, **84**(1): 1-28.
- Shi, T. & S. Horvath (2006) "Unsupervised learning with random forest predictors," *Journal of Computational and Graphical Statistics*, **1**: 118-138.
- Yiu, T. (2019) "The curse of dimensionality: Why high dimensional data can be so troublesome," *Towards Data Science*, July 20, 2019, https://towardsdatascience.com/the-curse-of-dimensionality-50dc6e49aa1e. Accessed 8/13/2022