Sawtooth Software

RESEARCH PAPER SERIES

Improving K-Means Cluster Analysis: Ensemble Analysis Instead of Highest Reproducibility Replicates

Bryan Orme & Rich Johnson,
Sawtooth Software, Inc.

530 W. Fir St.
Sequim, WA 98382
(360) 681-2300
www.sawtoothsoftware.com

Improving K-Means Cluster Analysis: Ensemble Analysis Instead of Highest Reproducibility Replicates

Bryan Orme and Rich Johnson, Sawtooth Software

Introduction

Cluster analysis is a popular statistical tool for finding groups of respondents, objects, or cases that are similar to one another but different from those in other groups. In marketing, there is keen interest among managers in developing products and strategies to target segments. The challenge with cluster analysis is that it involves both art and science, and it always produces an answer whether there really are clean and separable segments or whether consumers are positioned in a continuous cloud. Complicating matters further, there are numerous cluster analysis routines, which can lead to different results.

With common methods such as k-means (convergent methods), results can depend on the starting cluster seeds. An unlucky choice of cluster seeds could lead to an uncharacteristically poor result. Other popular approaches involving hierarchical methods are sensitive to outliers and to the choice of distance/linkage criterion. It is imperative that researchers select cluster approaches likely to produce robust and reproducible solutions. It is also critical to assess whether the revealed cluster structure consistently shows more organization than could be found in random data. And, it is particularly useful if the approach can shed some light on, for example, whether a 5-group solution characterizes the data better than a 3-group solution.

Sawtooth Software developed its CCA (Convergent Cluster Analysis) package in the late 1980s. CCA used k-means clustering, but what made it stand out from other routines was that it repeated the k-means analysis from multiple, intelligently-drawn, starting points. It compared many replicates (up to 10), and selected the most reproducible (representative) replicate as the final solution. This strategy helped avoid the possibility of accepting a poor solution due to an unlucky starting seed. Reproducibility also gave an important indication of how well a particular number of groups seemed to fit the natural structure of the data, so it had secondary use as a diagnostic.

Cluster Ensemble Methods

Cluster Ensemble approaches (Strehl and Ghosh 2002, Retzer and Shan 2007) employ multiple cluster solutions as well, but rather than choose the *one* most representative solution, they develop a consensus solution based on a combination of the solutions available within the ensemble. The final solution is almost always different from all of the solutions in the ensemble. Ensemble Analysis benefits from a diverse set of cluster solutions, such as from different cluster methodologies (e.g. hierarchical, k-means, neural networks, etc.), different basis variables, and different numbers of clusters This is made possible by the fact that Ensemble Analysis does not "look at" the original data, but rather examines only the assignments of

individuals to clusters. The consensus solution combines information from those several partitionings to find one which is most representative of them all.

Ensemble analysis has been found to be robust, even when poor cluster solutions are included within the ensemble (Strehl and Gosh 2002). Importantly, it improves classification accuracy and the general quality of cluster solutions. Many authors have shown that different approaches to ensemble analysis can capture even bizarre devised patterns in synthetic data, such as doughnuts, spirals, parallel lines, concentric rings, etc.

This paper demonstrates that consensus solutions offer improvement over the previous approach offered by Sawtooth Software's CCA. We show this using a variety of synthetic data sets, where the group membership is known and the data have been perturbed by random noise.

A Direct Consensus Method Using "Clustering on Clusters"

Strehl and Ghosh (2002) discuss several approaches for developing a consensus solution, given the availability of multiple segmentation solutions within an ensemble. One method, which they call a *Meta-Clustering Algorithm*, is based on the notion of "clustering clusters."

With the Meta-Clustering Algorithm, one first develops multiple clustering solutions. These could vary in terms of:

- Method used (hierarchical, k-means under different starting points, etc.)
- Number of clusters (for example, varying from 2 to 12 groups)
- Basis variables employed
- Pre-processing options (standardization, centering)

The group assignments for multiple cluster solutions (just three in this example) could look like the following when recorded in a data file:

Caseid	Solution#1	Solution#2	Solution#3
1001	1	4	2
1002	2	2	1
1003	2	3	1
1004	1	4	2

Solutions #1 and #3 are 2-group solutions, and across the first four cases they appear to be identical (except that the labels are switched). Solution #2 is a 4-group solution.

It is very easy to modify this file to have "indicator" (dummy) coding. Strehl and Ghosh code the information for a 2-group solution (such as Solution #1) using two columns, where the first column indicates whether the respondent belongs to the first group and the second column indicates membership in the second group.

Indicator Coding for Solution #1:

1001	1	0
1002	0	1
1003	0	1
1004	1	0

All three solutions in the example above could be coded in eight total indicator columns of an "indicator matrix" as:

Indicator Coding for Solutions 1-3:

1001	1	0	0	0	0	1	0	1
1002	0	1	0	1	0	0	1	0
1003	0	1	0	0	1	0	1	0
1004	1	0	0	0	0	1	0	1

Strehl and Ghosh employ a method that involves repeatedly clustering (using a graph partitioning approach) and relabeling the clusterers, so that cluster #1 from the first solution corresponds to cluster #1 from the second solution, etc. This becomes a challenging optimization problem when many groups are included across many replicates, and with somewhat noisy datasets as would be found in practice.

We use Strehl and Gosh's first step, but have chosen to side-step the issue of relabeling altogether by simply clustering again on the indicator matrix (clustering on the cluster solutions, or "CC") without worrying about relabeling. In the example above, we simply use these eight columns as new basis variables in a secondary cluster analysis, where we are looking for a final k-group solution (and the indicator variables could represent cluster solutions with either more or fewer clusters than the final k-group solution we seek). For our work, we leveraged CCA's standard approach of running multiple replicates under k-means (using different, intelligently drawn starting points) and we selected the one solution that was most reproducible as a possible final solution and candidate stopping point. We have found it useful to include a large number of cluster solutions in the ensemble, representing a wide variety of numbers of clusters. There doesn't seem to be any harm (overfitting) in including a very large number of runs in the ensemble. We have had good results using sixty or seventy cluster solutions in the ensemble, ranging from 2-group solutions clear up to 30-group solutions. And, we find the final clustering result is more stable (when employing different starting points seeds) if using large, diverse ensembles. Our software implementation seems very fast, with an ensemble analysis as just described typically requiring only about 30 seconds for 1000 respondents.

If several solutions are obtained by "clustering on cluster solutions (CC)", one can compute reproducibility across those replicates to ascertain how consistently one obtains the same result from different starting points. We might also consider the most reproducible of these as the best solution; however, it is not strictly necessary to introduce the notion of reproducibility. We can recode those replicates (now all on k-groups) using indicator coding and repeat the process (clustering on cluster solutions of cluster solutions (CCC)). This loop can continue indefinitely

(CCC…C), but we find that the process converges very quickly. When no respondents are reclassified in a subsequent step, we may take the previous candidate solution (the most reproducible one) as final. As far as we know, our approach is unique, though it owes a great deal to the notions set forth by Strehl and Ghosh.

The literature suggests that cluster ensembles which use diverse clusterers will be more robust to characteristics in the data which do not conform well to traditional k-means, such as elongated clusters. Even though we use k-means as our method to develop a consensus solution from the indicator coding matrix, the cluster solutions in our ensemble include hierarchical methods that add diversity and can yield more flexible final clusterings. However, our approach to ensemble construction and creating a consensus solution is based on the notion that clusters should be generally compact. For that reason, we have not employed single-linkage hierarchical clustering in the "clustering on clusters" consensus step. Therefore, our implementation should not be expected to work very well in recovering the sorts of artificial structures (spirals, rings, etc.) that other authors have used as a standard for prediction. But our approach should work well in detecting meaningful structure more commonly found in market and social research. And, if desired, one could use single-linkage hierarchical clustering to develop the consensus solution (rather than k-means), and this should do a creditable job of capturing data with very elongated or patterned structures.

Empirical Tests

We designed a series of tests to compare the standard CCA methodology versus the Ensemble approach described here. The first three tests were very tidy, but unrealistic, in that they assumed three groups with no overlap on the means:

	True Group Means:									
Group 1	1	2	3	1	2	3	1	2	3	1
Group 2	2	3	1	2	3	1	2	3	1	2
Group 3	3	1	2	3	1	2	3	1	2	3

We generated 1000 synthetic respondents by perturbing the mean vectors by normal error, with standard deviation either 1.5 or 2.0. We generated three test datasets:

Test #1: extreme group sizes, standard deviation of error=1.5
Group 1 = 100
Group 2 = 300
Group 3 = 600

Test #2: moderately different-sized groups, standard deviation of error=2
Group 1 = 200
Group 2 = 300
Group 3 = 500

Test #3: equal groups, standard deviation of error=2
Group 1 = 333
Group 2 = 333
Group 3 = 334

We ran CCA with 30 replicates (mixed starting point strategy) and let it choose the single replicate that had the highest reproducibility. For Ensemble analysis, we constructed the ensemble using a combination of k-means (mixed starting point strategy) and Hierarchical (complete linkage and average linkage) runs. We employed a large ensemble, with approximately 60 separate cluster solutions ranging from 2- to 30-groups.

The results are as follows:

Test #1:

	"Truth"	CCA	Ensemble
Group 1:	100	204	175
Group 2:	300	303	280
Group 3:	600	493	545
Hit rates:	100%	83.0%	85.8%
RMSE:	0.00	0.278	0.236

*RMSE is the root mean square error between the true group means versus the means for the observed groups resulting from the cluster approach.

Test #2:

	"Truth"	CCA	Ensemble
Group 1:	200	266	236
Group 2:	300	332	319
Group 3:	500	402	445
Hit rates:	100%	76.9%	78.3%
RMSE:	0.00	0.232	0.212

Similar pattern of findings here as Test #1, and the consensus solution provides modest improvement on all fronts.

Test #3:

	"Truth"	CCA	Ensemble
Group 1:	333	361	352
Group 2:	333	307	312
Group 3:	334	332	336
Hit rates:	100%	77.2%	76.9%
RMSE:	0.00	0.223	0.218

Test #3 achieves very similar results for CCA and Ensemble.

Test 4:

In this test, we modified group 3's vector, so that it has a lot of overlap with groups 1 and 2. Groups 1 and 2 are unique with respect to each other. This is probably more realistic of what is seen in practice with human respondents, rather than groups of respondents who lack any similarity with respect to their means on basis variables.

```
                  True Group Means:
Group 1 (n=100)  1  2  3  1  2  3  1  2  3  1
Group 2 (n=300)  2  3  1  2  3  1  2  3  1  2
Group 3 (n=600)  2  2  1  1  3  3  2  2  1  1
```

The vectors were perturbed with normal random error, with standard deviation=1.5. The results were:

Test #4:

	"Truth"	CCA	Ensemble
Group 1:	100	248	215
Group 2:	300	335	333
Group 3:	600	417	452
Hit rates:	100%	70.9%	73.1%
RMSE:	0.00	0.227	0.190

Test 5:

This test is just like Test 4, except we switched the sizes of groups 1 and 3.

Test #5:

	"Truth"	CCA	Ensemble
Group 1:	600	439	540
Group 2:	300	298	275
Group 3:	100	263	185
Hit rates:	100%	79.2%	87.7%
RMSE:	0.00	0.299	0.170

This result is a strong win for Ensemble analysis. For all three measures of success, Ensemble exceeds CCA.

Test 6:

This sixth test uses simulated data, based on patterns observed in a real respondent dataset. There were four true clusters, with sample sizes 50, 100, 150, and 200. Group means were as observed in the data, for 25 basis variables. The true means were disturbed using a pattern of covariances observed in the real data set. This dataset was a difficult one for both CCA and Ensemble to consistently get right. Both methods tended to flip between good and bad solutions, depending on the random starting point; but Ensemble more consistently got it right, and its good solutions were superior.

Hit rates when using 10 different random starting seeds were as follows:

	CCA	Ensemble
Seed = 1	91.2%	95.2%
Seed = 2	73.2	74.6
Seed = 3	91.6	95.2
Seed = 4	73.2	95.0
Seed = 5	73.2	95.2
Seed = 6	74.0	94.8
Seed = 7	73.0	95.4
Seed = 8	74.0	95.0
Seed = 9	72.8	95.4
Seed = 10	91.6	95.2
Average:	78.8	93.1
Max:	91.6	95.4
Min:	72.8	74.6

We also tried this data set with a higher degree of noise, and found that both methods performed equally poorly in terms of respondent classification.

Test 7:

For this test, true means and group sizes were generated randomly, as follows:

```
                 True Group Means:
Group 1 (n=300):  6  4  4  1 10  4  6  1  7  1
Group 2 (n=50 ):  4  5  8  5  5  8  7  3  5  2
Group 3 (n=100): 10  4  4  2  5 10  7  3  4  8
Group 4 (n=200):  5  2  2  8  8  5  2  4  3  1
Group 5 (n=150):  2  3  4  9  2  5  5 10  4 10
Group 6 (n=200):  2  5 10  6  7 10  9  9  3  4
```

We created five separate datasets for this test, disturbing the data by normal random error with standard deviation of 1, 2, 3, 4 or 5.

Hit rates by level of error disturbance were:

	CCA	Ensemble
Error = 1	100.0%	100.0%
Error = 2	99.0	99.1
Error = 3	89.1	90.0
Error = 4	73.4	76.1
Error = 5	62.3	70.0

One of the benefits of CCA software has been the ability to use the reproducibility figures to help identify the true number of groups in the data set. One also obtains reproducibility from Ensemble Analysis in the first clustering step (clustering on clusters), as we repeat the k-means clustering from different starting points.

Using the data disturbed by standard deviation = 3, reproducibilities were as follows, for five different starting seeds:

Reproducibility for CCA by Different Starting Seeds

	Groups						
	2	3	4	5	6	7	8
Seed=1	99	69	80	100	88	77	67
Seed=2	99	71	80	100	87	74	69
Seed=3	99	68	78	99	85	76	67
Seed=4	99	76	86	100	87	75	72
Seed=5	99	72	77	100	87	76	69
Average:	99	71	80	100	87	76	69

Reproducibility for Ensemble Analysis by Different Starting Seeds

	Groups						
	2	3	4	5	6	7	8
Seed=1	75	75	70	93	92	85	77
Seed=2	80	76	70	93	93	84	79
Seed=3	75	71	74	91	93	83	78
Seed=4	77	68	72	94	91	83	82
Seed=5	76	75	76	98	91	85	80
Average:	77	73	72	94	92	82	78

CCA suggests either a 2-group or 5-group solution, and reproducibility is nearly 100% in either case. The 6-group solution (the true number of groups for this data set), has high reproducibility;

but not nearly as high as for 2 and 5 groups. Cluster ensemble suggests either a 5-group or 6-group solution. Both have about equal reproducibility.

We repeated the reproducibility analysis, this time with more error disturbance (standard deviation = 4):

Reproducibility for CCA by Different Starting Seeds

	Groups						
	2	3	4	5	6	7	8
Seed=1	99	68	81	98	79	71	62
Seed=2	99	63	80	97	79	70	61
Seed=3	99	74	83	98	82	68	63
Seed=4	99	66	76	98	81	67	62
Seed=5	99	66	81	96	82	72	61
Average:	99	67	80	97	81	70	62

Reproducibility for Ensemble Analysis by Different Starting Seeds

	Groups						
	2	3	4	5	6	7	8
Seed=1	69	69	76	90	86	80	77
Seed=2	65	69	80	93	85	81	79
Seed=3	66	79	75	97	90	83	81
Seed=4	67	69	76	92	87	81	76
Seed=5	70	69	77	88	90	82	81
Average:	67	71	77	92	88	81	79

Again, CCA finds strong evidence for a 2-group as well as a 5-group solution. Cluster ensemble analysis points to a 5-group solution, with the 6-group solution as a next-likely candidate. For some reason, it cannot consistently partition the data into just two groups.

This analysis (and similar analyses using the other data sets in this paper) suggests that the reproducibility resulting from our implementation of meta clustering (clustering on clusters) potentially does a better job than the traditional method offered in CCA for diagnosing the true number of clusters, for a data set with known group structure.

Conclusions

Our implementation of ensemble analysis generally performs better than CCA's approach of choosing the most reproducible replicate. The ensemble approach seems especially useful when the true sizes of the groups are quite different (which is often true in practice) and when groups have differing degrees of overlap with respect to each other on the basis variables (again more likely in practice). In those cases, it achieves significantly better hit rates, better fit to true group means, and better estimates of the true group sizes. With equal-sized groups that are completely

unique with respect to their means on the basis variables, it seems to perform just as well as CCA's approach. Like CCA, our ensemble method provides a measure of reproducibility, which can be used to help determine how many groups provides a good characterization of the data structure. The reproducibility statistic for our ensemble method seems to perform just as well or better than the similar statistic in CCA for indicating the correct number of groups.

We haven't evaluated other methods of forming consensus solutions for ensembles, and thus cannot comment on the relative performance of our method versus others described in the literature. This remains an avenue for future research.

References:

Retzer, J. and M. Shan (2007), "Cluster Ensemble Analysis and Graphical Depiction of Cluster Partitions," Proceedings of the 2007 Sawtooth Software Conference, Sequim WA.

Sawtooth Software (1998), "CCA System," Evanston, IL.

Strehl, A. and J. Ghosh (2002), "Cluster Ensembles — A Knowledge Reuse Framework for Combining Multiple Partitions," *Journal on Machine Learning Research (JMLR)*, 3:583-617, December 2002.