SAS® GLOBAL FORUM 2018

USERS PROGRAM

An introduction to clustering techniques

-Xinghe Lu

April 8 – 11 | Denver, CO

#SASGF

An introduction to clustering techniques

Xinghe Lu

The Vanguard Group

ABSTRACT

Cluster analysis has been used in a wide variety of fields, such as marketing, social science, biology, pattern recognition etc. It is used to identify homogenous groups of cases to better understand characteristics in each group. There are two major types of cluster analysis- supervised and unsupervised. Unlike supervised cluster analysis, unsupervised cluster analysis means data is assigned to segments without the clusters being known a priori. Furthermore, it refers to partitioning a set of objects into groups where the objects within a group are as similar as possible and, on the other hand, objects in different groups are as dissimilar as possible. This paper provides overview on multiple techniques on unsupervised clustering analysis, including traditional data mining/ machine learning approaches and statistical model approaches. Hierarchical clustering, K-means clustering and Hybrid clustering are three common data mining/ machine learning methods used in big datasets; whereas Latent cluster analysis is a statistical model-based approach and becoming more and more popular. This paper also introduces other approaches: Nonparametric clustering method is suitable when the data has irregular shape and Fuzzy cluster (Q-technique) can be applied to data with relatively few cases.

Key Words:

K-means cluster analysis, Hierarchical cluster analysis, Hybrid cluster analysis, latent class analysis, Non-parametric cluster analysis, Fuzzy c cluster analysis, Discriminant analysis, SAS

METHODS

K means cluster analysis

- Assign each observation to the cluster iteratively until the distance between each observation and the center of the cluster or centroid is minimal.
- Number of clusters(K) has to be specified in the initial stage of modeling.
- Statistics such as Cubic Clustering Criterion(CCC) and Pseudo-F Statistic(PSF) from PROC FASTCLUS are used to decide number of clusters.
- Key SAS code example:

```
%macro k_means(dsn=,m=,n=);
    %do i = &m %to &n;
        proc fastclus data=&dsn maxclusters=&i least=2 outstat=stats(keep=_TYPE_ OVER_ALL) noprint out=&dsn._&i;
            var &input;
        run;
    %end;
%mend k_means;
```

METHODS CONTINUED

Hierarchical cluster analysis

- Refers to identifying homogeneous groups (clusters) based on the selected variables by using an algorithm that each observation starts its own cluster at the beginning and then combines clusters until all observations are combined into a big group.
- Several methods can be used in measurement of similarity within a cluster or between clusters, such as Wald's minimum variance, average linkage, centroid linkage etc.
- Cubic Clustering Criterion(CCC) and Pseudo-F Statistic(PSF) and Pseudo-T^2 (PST2) from PROC CLUSTER are the three common statistics used to decide the number of clusters.
- Key SAS code example:

```
ods graphics on;
ods output CccPsfAndPsTSqPlot=plotdata;
proc cluster data=bank_std method=ward ccc pseudo outtree=tree;
     var &input;
run;
ods graphics off;
```

Hybrid cluster analysis

- This method combines the strengths from the two previous approaches-efficiency from k-means cluster analysis and superior solution from hierarchical cluster analysis. More specifically, at first, centroids are generated and saved in the output data -preclus ; then the output data is fit into hierarchical model to decide the number of clusters; finally observations are assigned to different clusters by K-means cluster methods with adaptive training method(drift).
-

```
proc fastclus data = bank2_std maxc =300 outseed = preclus noprint;
     var &input;
run;

proc cluster data=preclus method=ward outtree=tree(keep=_ncl_ _ccc_) ccc pseudo;
     var &input;
run;

proc fastclus data=bank2_std maxclusters=&nc drift maxiter=20 out=results;
     var &input;
run;
```

An introduction to clustering techniques

Xinghe Lu

The Vanguard Group

METHODS CONTINUED

Nonparametric cluster analysis

- In nonparametric cluster analysis, a p-value is computed in each cluster by comparing the maximum density in the cluster with the maximum density on the cluster boundary, known as saddle density estimation.
- It is less sensitive to the shape of the data set and not required to have equal size in each cluster.
- No need to predefine the number of clusters.
- Key SAS code example:

```
proc modeclus data=bank_std method=1 r=0.2 join out=results;
     var &input;
run;
```

Fuzzy cluster analysis

- In Fuzzy cluster analysis, each observation belongs to a cluster based the probability of its membership in a set of derived factors, which are the fuzzy clusters.
- Appropriate for data with many variables and relatively few cases.
- Eigenvalues, proportion of the common variances and scree plot from PROC FACTOR can be used in number of clusters determination.
- Key SAS code example:

```
ods graphics on;
title1 'Factor Loadings';
proc factor data=spear priors=smc method=principal plots=scree
     outstat=results;
     var obs:;
run;
ods graphics off;
```

METHODS CONTINUED

Latent class analysis

- A statistical approach for identifying unmeasured or latent class within a population based on observed characteristics .
- Independent variables can be either continuous variables or categorical .
- Number of clusters is based on AIC, BIC, CAIC, ABIC, G squared and Entropy.
- Key SAS codes example:

```
%Macro LCA(n_start=, n_end=);
     %do i=&N_start %to &n_end;
          proc LCA data=f12014_final outest=est_&i outpost=output_&i;
               id caseid;
               weight archive_wt;
               NCLASS &i;
               items alc_lt alc_yr alc_mo alc_drunk_lt alc_drunk_yr alc_drunk_mo alc_5plus_2wk;
               categories 2 2 2 2 2 2 2;
               seed 12345678;
               RHO PRIOR=1;
          run;
     %end;
%mend;
```

RESULTS

- The Bank Marketing dataset from https://archive.ics.uci.edu/ml/datasets.html was used in K-means, hierarchical ,combined, and nonparametric clustering for demonstration purposes, which contains 45,211 observations and 6 numeric variables. 5 out of 45,2111 observations were randomly selected for fuzzy clustering.
- The data used in latent class analysis is from the 2014 Monitoring the Future survey of high school seniors (http://www.monitoringthefuture.org/pubs/monographs/mtf-vol1_2014.pdf) with 7 selected alcohol behavior variables and 2,181 observations.

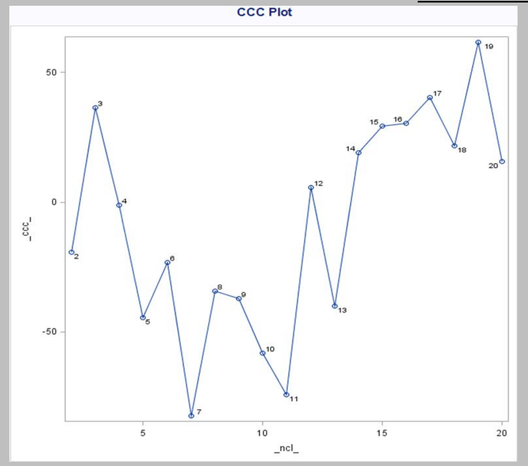
An introduction to clustering techniques

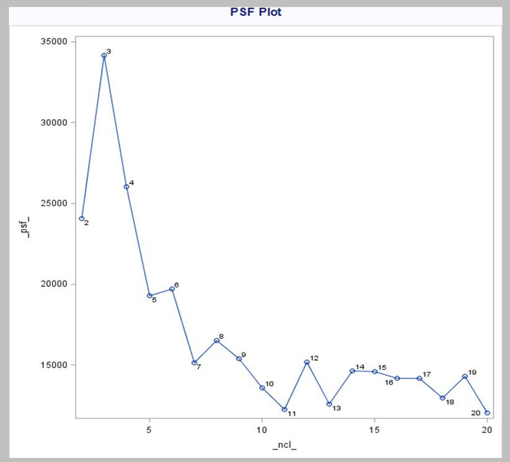
Xinghe Lu

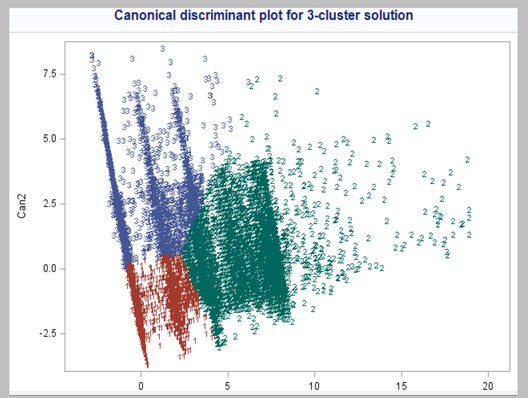
The Vanguard Group

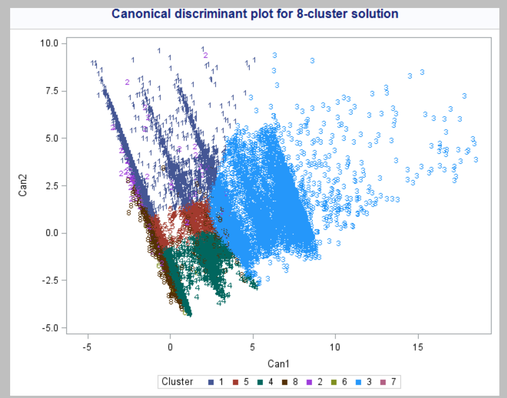
RESULTS CONTINUED

K means cluster analysis





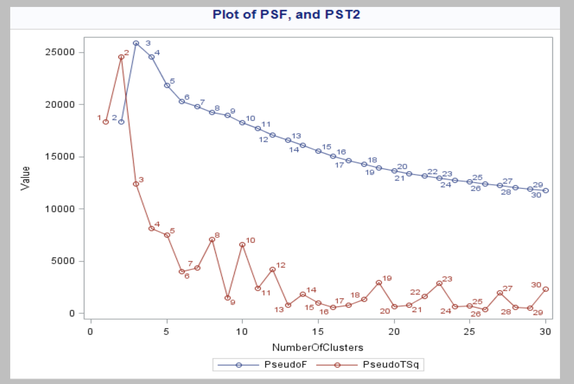


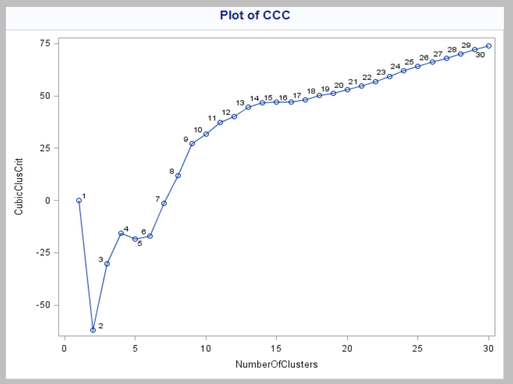


- In CCC and PSF plots, both CCC and PSF values have highest values at cluster 3 indicating the optimal solution is 3-cluster solution.
- Canonical discriminant plots further visualize that 3-cluster solution fits better than 8-cluster solution.

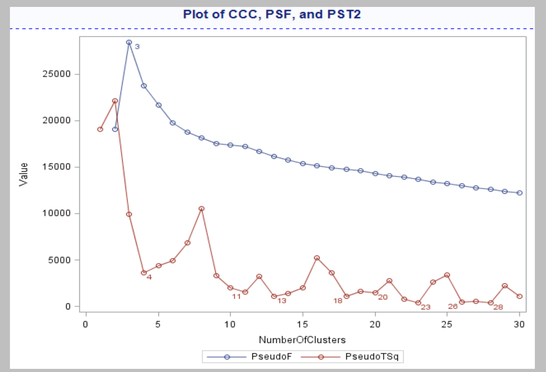
RESULTS CONTINUED

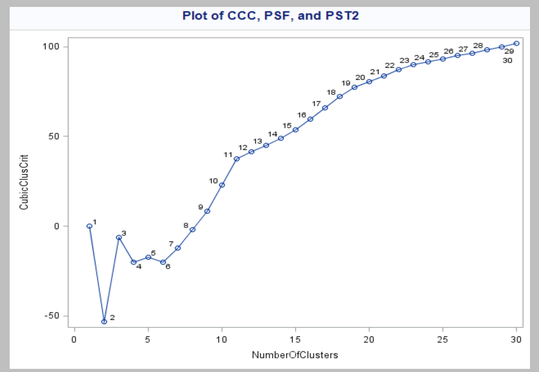
Hierarchical cluster analysis





- In CCC plot, peak value is shown at cluster 4. In PSF2(PseudoTSq) plot, the point at cluster 7 begins to rise. In PSF(PseudoF) plot, peak value is shown at cluster 3.
- The candidate solution can be 3 , 4 or 7 clusters based on the results.





- In combined method, CCC and PSF plots indicate 3 cluster fit the model the best, however PSF2 plot shows optimal number of clusters is 7.

An introduction to clustering techniques

Xinghe Lu

The Vanguard Group

RESULTS CONTINUED

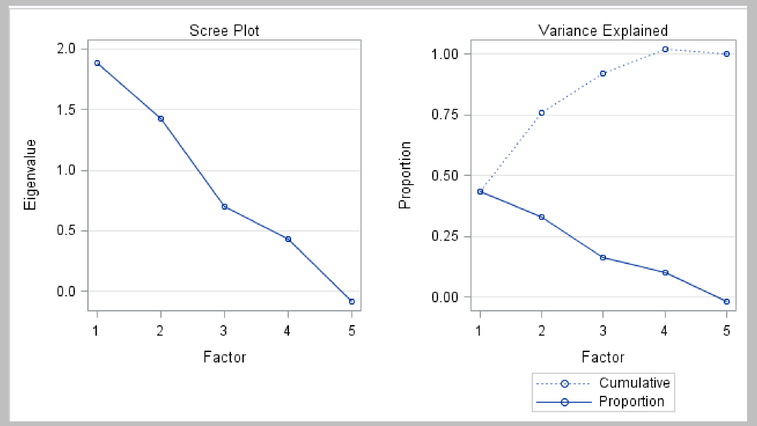
Nonparametric cluster analysis

Cluster Statistics					-Saddle Test: Version 92.7-				
Cluster	Frequency	Maximum Estimated Density	Boundary Frequency	Estimated Saddle Density	Mode Count	Saddle Count	Overlap Count	Z	Approx P-value
1	42015	69941.6944	1962	4780.91137	16340	1116	0	115.220	0
2	3137	6775.45451	2069	4956.39692	1582	1157	690	11.502	0

- Initially 59-clusters were created, then clusters were merged based on the density within each cluster to the density of its nearest neighbor by the saddle-density tests. The table above is the final stage of merging indicating 2-cluster is the optimal solution.

Fuzzy cluster analysis

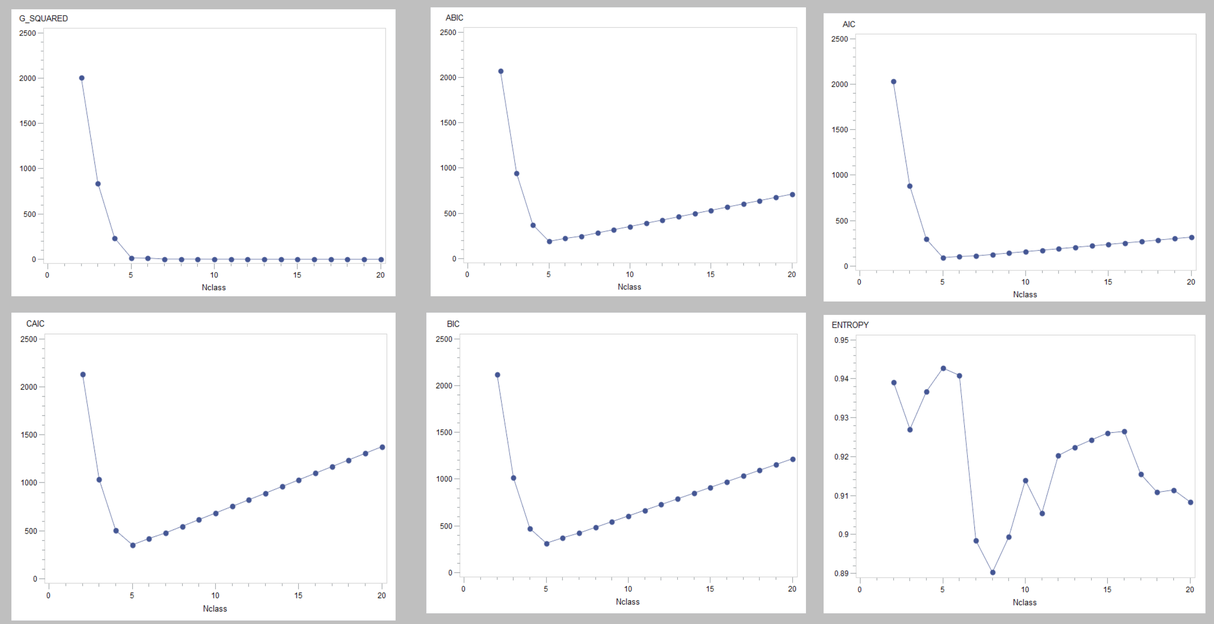
Eigenvalues of the Reduced Correlation Matrix: Total = 4.3615156 Average = 0.87230312				
	Eigenvalue	Difference	Proportion	Cumulative
1	1.88543977	0.45743614	0.4323	0.4323
2	1.42800363	0.72558207	0.3274	0.7597
3	0.70242156	0.26873705	0.1610	0.9207
4	0.43368451	0.52171837	0.0994	1.0202
5	-.08803387		-0.0202	1.0000



RESULTS CONTINUED

- Based on scree plot , eigenvalues (>=1) and proportion of the common variances (>=0.8), optimal number of clusters is either 2 or 3.

Latent class analysis



- AIC, BIC, CAIC, ABIC, G squared statistics all have the lowest value at cluster 5 and the peak value appears at cluster 5 in Entropy plot , so 5-cluster is the optimal number of clusters.

An introduction to clustering techniques

Xinghe Lu

The Vanguard Group

CONCLUSIONS

In the poster, several machine learning algorithms and statistical methods on unsupervised clustering analysis were introduced. They are summarized as follows,

	Strength	Weakness
K-means clustering	• Faster • Can deal with large dataset	• Sensitive to the initial seed • User has to specify number of cluster
Hierarchical clustering	• Provides more process details	• Time consuming process • Has to impute missing values
Combined clustering method	• Combines the strength from K-means and Hierarchical methods	• Sensitive to the initial seed
Nonparametric clustering	• Can handle the data with irregular shapes	• Not providing strong predictive power
Fuzzy clustering	• Applicable to data with few observations and many variables	• Results can be sensitive due to the small size of the data
Latent class analysis	• Applicable to data with categorical variables • The final result is generated based on statistical approach instead of machine learning approach	• Assuming latent structure among the variables in the data

REFERENCES

- Elayne Reiss, Sandra Archer, Robert Armacost, Ying Sun, and Yun (Helen) Fu.” Using SAS® PROC CLUSTER to Determine University Benchmarking Peers”. 2010 SESUG. http://analytics.ncsu.edu/sesug/2010/SDA10.Reiss.pdf
- Patricia A. Berglund. “Latent Class Analysis Using PROC LCA”. 2016, SAS Global Forum. http://support.sas.com/resources/papers/proceedings16/5500-2016.pdf
- SAS. 2010. Applied Clustering Techniques, SAS course note. Cary, NC, USA. SAS Institute Inc.
- Soni Madhulatha, T. (2012). An Overview on Clustering Methods. IOSR Journal of Engineering. 2. . 10.9790/3021-0204719725
- The Methodology Center at Penn State. https://methodology.psu.edu/ra/lca

SAS® GLOBAL FORUM 2018

April 8 – 11 | Denver, CO
Colorado Convention Center

#SASGF