An introduction to clustering techniques

-Xinghe Lu
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(dscr, n, d);
  %do i = 1 %to &n;
    proc fastclus data="&dscr" maxclust=&i least=0 outout=&i outseed=none noclprint outdim=&i; var &d; run;
  %end;
%mend k_means;
```

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 CcOfPsAndTsPlot=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).

Key SAS code example:
```
ods graphics on;
ods output CcOfPsAndTsPlot=plotdata;
proc fastclus data=bank2_std maxc=300 outseed=preclus noprin;
   var &input;
run;
ods graphics off;
```

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 CcOfPsAndTsPlot=plotdata;
proc cluster data=bank_std method=ward ccc pseudo outtree=tree;
   var &input;
run;
ods graphics off;
```
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 on 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=emo method=principal plots=acree
    outstat=results;
var obs; run;
ods graphics off;
```

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=2014_final outest=est_i outpost=output_i;
    id caseid;
    weight archive wt;
    NCLASS 4;
    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;
    seed 12345678;
    PRO PRIOR=1;
    %end;
  run;
%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 45,211 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

- 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

- 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

<table>
<thead>
<tr>
<th>Cluster</th>
<th>Frequency</th>
<th>Maximum Estimated Density</th>
<th>Boundary Frequency</th>
<th>Estimated Saddle Density</th>
<th>Mode Count</th>
<th>Saddle Count</th>
<th>Overlap Count</th>
<th>Z Approx P-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>42015</td>
<td>69941.6944</td>
<td>1962</td>
<td>4780.91137</td>
<td>16340</td>
<td>1116</td>
<td>0</td>
<td>115.220</td>
</tr>
<tr>
<td>2</td>
<td>3137</td>
<td>6775.4545</td>
<td>2069</td>
<td>4956.3692</td>
<td>1592</td>
<td>1157</td>
<td>690</td>
<td>11.502</td>
</tr>
</tbody>
</table>

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

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

RESULTS CONTINUED

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,

<table>
<thead>
<tr>
<th>Method</th>
<th>Strength</th>
<th>Weakness</th>
</tr>
</thead>
<tbody>
<tr>
<td>K-means clustering</td>
<td>• Faster</td>
<td>• Sensitive to the initial seed</td>
</tr>
<tr>
<td></td>
<td>• Can deal with large dataset</td>
<td>• User has to specify number of cluster</td>
</tr>
<tr>
<td>Hierarchical clustering</td>
<td>• Provides more process details</td>
<td>• Time consuming process</td>
</tr>
<tr>
<td></td>
<td></td>
<td>• Has to impute missing values</td>
</tr>
<tr>
<td>Combined clustering method</td>
<td>• Combines the strength from K-means and Hierarchical methods</td>
<td>• Sensitive to the initial seed</td>
</tr>
<tr>
<td>Nonparametric clustering</td>
<td>• Can handle the data with irregular shapes</td>
<td>• Not providing strong predictive power</td>
</tr>
<tr>
<td>Fuzzy clustering</td>
<td>• Applicable to data with few observations and many variables</td>
<td>• Results can be sensitive due to the small size of the data</td>
</tr>
<tr>
<td>Latent class analysis</td>
<td>• Applicable to data with categorical variables</td>
<td>• Assuming latent structure among the variables in the data</td>
</tr>
</tbody>
</table>

REFERENCES

• 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