Interpreting cluster analysis results

With numeric and categorical variables (active and/or illustrative)

Ricco RAKOTOMALALA
Université Lumière Lyon 2
Outline

1. Interpreting the cluster analysis results
2. Univariate characterization
   a. Of the clustering structure
   b. Of the clusters
3. Multivariate characterization
   a. Percentage of explained variance
   b. Distance between the centroids
   c. Combination with factor analysis
   d. Utilization of a supervised approach (e.g. discriminant analysis)
4. Conclusion
5. References
Cluster analysis

Clustering, unsupervised learning
Cluster analysis
Also called: clustering, unsupervised learning, typological analysis

“Active” input variables, used for the creation of the clusters. Often (but not always) numeric variables

“Illustrative” variables. Used only for the interpretation of the clusters. To understand on which characteristics are based the clusters.

Goal: Identifying the set of objects with similar characteristics

We want that:
1. The objects in the same group are more similar to each other
2. Than to those in other groups

For what purpose?
- Identify underlying structures in the data
- Summarize behaviors or characteristics
- Assign new individuals to groups
- Identify totally atypical objects

The aim is to detect the set of “similar” objects, called groups or clusters.
“Similar” should be understood as “which have close characteristics”.

<table>
<thead>
<tr>
<th>Modele</th>
<th>puissance</th>
<th>cylindre</th>
<th>vitesse</th>
<th>longueur</th>
<th>largeur</th>
<th>hauteur</th>
<th>poids</th>
<th>CO2</th>
<th>pnx</th>
<th>origine</th>
<th>carburant</th>
</tr>
</thead>
<tbody>
<tr>
<td>PANDA</td>
<td>54</td>
<td>1108</td>
<td>150</td>
<td>354</td>
<td>159</td>
<td>154</td>
<td>860</td>
<td>125</td>
<td>8070</td>
<td>Europe</td>
<td>Essence</td>
</tr>
<tr>
<td>TWINGO</td>
<td>60</td>
<td>1149</td>
<td>151</td>
<td>344</td>
<td>163</td>
<td>143</td>
<td>840</td>
<td>142</td>
<td>8950</td>
<td>France</td>
<td>Essence</td>
</tr>
<tr>
<td>CITRONC2</td>
<td>69</td>
<td>1124</td>
<td>128</td>
<td>367</td>
<td>166</td>
<td>147</td>
<td>912</td>
<td>141</td>
<td>10790</td>
<td>France</td>
<td>Essence</td>
</tr>
<tr>
<td>YARIS</td>
<td>65</td>
<td>998</td>
<td>155</td>
<td>364</td>
<td>166</td>
<td>150</td>
<td>880</td>
<td>134</td>
<td>10450</td>
<td>France</td>
<td>Essence</td>
</tr>
<tr>
<td>PIESTA</td>
<td>68</td>
<td>1399</td>
<td>164</td>
<td>382</td>
<td>168</td>
<td>144</td>
<td>1138</td>
<td>117</td>
<td>14150</td>
<td>Europe</td>
<td>Diesel</td>
</tr>
<tr>
<td>CORSA</td>
<td>70</td>
<td>1248</td>
<td>165</td>
<td>384</td>
<td>165</td>
<td>144</td>
<td>1035</td>
<td>127</td>
<td>13590</td>
<td>Europe</td>
<td>Diesel</td>
</tr>
<tr>
<td>GOLF</td>
<td>75</td>
<td>1368</td>
<td>163</td>
<td>421</td>
<td>176</td>
<td>149</td>
<td>1217</td>
<td>143</td>
<td>19140</td>
<td>Europe</td>
<td>Diesel</td>
</tr>
<tr>
<td>PEUGEOT</td>
<td>56</td>
<td>1336</td>
<td>165</td>
<td>374</td>
<td>169</td>
<td>161</td>
<td>1181</td>
<td>153</td>
<td>13600</td>
<td>France</td>
<td>Diesel</td>
</tr>
<tr>
<td>MUSA</td>
<td>100</td>
<td>1910</td>
<td>179</td>
<td>389</td>
<td>170</td>
<td>169</td>
<td>1275</td>
<td>146</td>
<td>17900</td>
<td>Europe</td>
<td>Diesel</td>
</tr>
<tr>
<td>Clio</td>
<td>100</td>
<td>1461</td>
<td>185</td>
<td>382</td>
<td>164</td>
<td>142</td>
<td>980</td>
<td>110</td>
<td>17800</td>
<td>Europe</td>
<td>Diesel</td>
</tr>
<tr>
<td>AUDIA3</td>
<td>102</td>
<td>1595</td>
<td>185</td>
<td>421</td>
<td>177</td>
<td>145</td>
<td>1205</td>
<td>186</td>
<td>21600</td>
<td>Europe</td>
<td>Diesel</td>
</tr>
<tr>
<td>MODUS</td>
<td>111</td>
<td>1398</td>
<td>185</td>
<td>380</td>
<td>170</td>
<td>159</td>
<td>1170</td>
<td>183</td>
<td>16950</td>
<td>France</td>
<td>Diesel</td>
</tr>
<tr>
<td>AVENSIS</td>
<td>115</td>
<td>1985</td>
<td>195</td>
<td>463</td>
<td>176</td>
<td>148</td>
<td>1400</td>
<td>155</td>
<td>26400</td>
<td>France</td>
<td>Diesel</td>
</tr>
<tr>
<td>PEUGEOT</td>
<td>136</td>
<td>1979</td>
<td>212</td>
<td>488</td>
<td>182</td>
<td>145</td>
<td>1415</td>
<td>194</td>
<td>23400</td>
<td>France</td>
<td>Diesel</td>
</tr>
<tr>
<td>CITRONC4</td>
<td>138</td>
<td>1979</td>
<td>207</td>
<td>426</td>
<td>178</td>
<td>146</td>
<td>1381</td>
<td>142</td>
<td>23400</td>
<td>France</td>
<td>Diesel</td>
</tr>
<tr>
<td>MERC A</td>
<td>140</td>
<td>1993</td>
<td>201</td>
<td>384</td>
<td>177</td>
<td>160</td>
<td>1340</td>
<td>141</td>
<td>24550</td>
<td>France</td>
<td>Diesel</td>
</tr>
<tr>
<td>MONDEO</td>
<td>145</td>
<td>1999</td>
<td>215</td>
<td>474</td>
<td>194</td>
<td>143</td>
<td>1378</td>
<td>189</td>
<td>23100</td>
<td>Europe</td>
<td>Diesel</td>
</tr>
<tr>
<td>VECTRA</td>
<td>150</td>
<td>1910</td>
<td>217</td>
<td>460</td>
<td>180</td>
<td>146</td>
<td>1428</td>
<td>159</td>
<td>26550</td>
<td>Europe</td>
<td>Diesel</td>
</tr>
<tr>
<td>PASSAT</td>
<td>150</td>
<td>1781</td>
<td>221</td>
<td>471</td>
<td>175</td>
<td>147</td>
<td>1360</td>
<td>197</td>
<td>27740</td>
<td>Europe</td>
<td>Diesel</td>
</tr>
<tr>
<td>VELOSAIT</td>
<td>150</td>
<td>2188</td>
<td>200</td>
<td>486</td>
<td>186</td>
<td>158</td>
<td>1735</td>
<td>186</td>
<td>38200</td>
<td>Europe</td>
<td>Diesel</td>
</tr>
<tr>
<td>LAGUNA</td>
<td>165</td>
<td>1998</td>
<td>218</td>
<td>458</td>
<td>178</td>
<td>145</td>
<td>1520</td>
<td>196</td>
<td>25350</td>
<td>France</td>
<td>Diesel</td>
</tr>
<tr>
<td>MEGANESC</td>
<td>165</td>
<td>1998</td>
<td>225</td>
<td>436</td>
<td>178</td>
<td>141</td>
<td>1415</td>
<td>191</td>
<td>27800</td>
<td>Europe</td>
<td>Diesel</td>
</tr>
<tr>
<td>P407</td>
<td>180</td>
<td>1977</td>
<td>225</td>
<td>435</td>
<td>176</td>
<td>143</td>
<td>1490</td>
<td>210</td>
<td>28500</td>
<td>France</td>
<td>Diesel</td>
</tr>
<tr>
<td>MERC E</td>
<td>204</td>
<td>3222</td>
<td>243</td>
<td>482</td>
<td>183</td>
<td>146</td>
<td>1795</td>
<td>183</td>
<td>46450</td>
<td>Europe</td>
<td>Diesel</td>
</tr>
<tr>
<td>CITRONC5</td>
<td>210</td>
<td>2496</td>
<td>230</td>
<td>475</td>
<td>178</td>
<td>148</td>
<td>1589</td>
<td>238</td>
<td>33000</td>
<td>Europe</td>
<td>Diesel</td>
</tr>
<tr>
<td>PTCRUISER</td>
<td>233</td>
<td>2429</td>
<td>200</td>
<td>429</td>
<td>171</td>
<td>154</td>
<td>1595</td>
<td>235</td>
<td>27400</td>
<td>Europe</td>
<td>Diesel</td>
</tr>
<tr>
<td>MAZDA 6</td>
<td>233</td>
<td>1308</td>
<td>235</td>
<td>443</td>
<td>177</td>
<td>134</td>
<td>1390</td>
<td>284</td>
<td>34000</td>
<td>Europe</td>
<td>Diesel</td>
</tr>
<tr>
<td>BMW M3</td>
<td>233</td>
<td>2975</td>
<td>250</td>
<td>485</td>
<td>185</td>
<td>147</td>
<td>1495</td>
<td>231</td>
<td>46400</td>
<td>Europe</td>
<td>Diesel</td>
</tr>
<tr>
<td>ALFA 156</td>
<td>250</td>
<td>3179</td>
<td>250</td>
<td>443</td>
<td>175</td>
<td>141</td>
<td>1410</td>
<td>287</td>
<td>40800</td>
<td>Europe</td>
<td>Essence</td>
</tr>
</tbody>
</table>
Cluster analysis
Interpreting clustering results

On which kind of information are based the results?

To what extent the groups are far from each other?

What are the characteristics that share individuals belonging to the same group and differentiate individuals belonging to distinct groups?

In view of active variables used during the construction of the clusters.

But also regarding the illustrative variables which provide another point of view about the nature of the clusters.
Cluster analysis
An artificial example in a two dimensional representation space

This example will help to understand the nature of the calculations achieved to characterize the clustering structure and the groups.
Univariate characterization

Interpretation using the variables taken individually
Characterizing the partition

Evaluate the importance of the variables, taken individually, in the construction of the clustering structure

The idea is to measure the proportion of the variance (of the variable) explained by the group membership

Huygens theorem

$$\text{TOTAL.SS} = \text{BETWEEN - CLUSTER.SS} + \text{WITHIN – CLUSTER.SS}$$

$$T = B + W$$

$$\sum_{i=1}^{n} (x_i - \bar{x})^2 = \sum_{g=1}^{G} n_g (\bar{x}_g - \bar{x})^2 + \sum_{g=1}^{G} \sum_{i=1}^{n_g} (x_i - \bar{x}_g)^2$$

The square of the correlation ratio is defined as follows:

$$\eta^2 = \frac{SCE}{SCT}$$

$$\eta^2$$ corresponds to the proportion of the variance explained ($0 \leq \eta^2 \leq 1$). We can interpret it, with caution, as the influence of the variable in the clustering structure.
Characterizing the partition
Quantitative variables – Cars dataset

The formation of the groups is based mainly on weight (poids), length (longueur) and engine size (cylindrée). But the other variables are not negligible (we can suspect that almost all the variables are highly correlated for this dataset).

About the illustrative variables, we observe that the groups correspond mainly a price differentiation.

Note: After a little reorganization, we observe that the conditional means increase from the left to the right (G1 < G3 < G2 < G4). We further examine this issue when we interpret the clusters.
Characterizing the partition
Categorical variables – Cramer’s V

A categorical variable leads also to a partition of the dataset. The idea is to study its relationship with the partition defined by the clustering structure.

We use a crosstab (contingency table)

<table>
<thead>
<tr>
<th>Nombre de Groupe</th>
<th>Étiquettes de lignes</th>
<th>Diesel</th>
<th>Essence</th>
<th>Total général</th>
</tr>
</thead>
<tbody>
<tr>
<td>G1</td>
<td>3</td>
<td>4</td>
<td>7</td>
<td></td>
</tr>
<tr>
<td>G2</td>
<td>4</td>
<td>8</td>
<td>12</td>
<td></td>
</tr>
<tr>
<td>G3</td>
<td>2</td>
<td>2</td>
<td>4</td>
<td></td>
</tr>
<tr>
<td>G4</td>
<td>3</td>
<td>4</td>
<td>7</td>
<td></td>
</tr>
<tr>
<td>Total général</td>
<td>12</td>
<td>18</td>
<td>30</td>
<td></td>
</tr>
</tbody>
</table>

\[
v = \sqrt{\frac{0.44}{30 \times \min(4-1,2-1)}} = 0.1206
\]

 Obviously, the clustering structure does not correspond to a differentiation by the fuel-type (carburant).

The chi-squared statistic enables to measure the degree of association.

The Cramer's \(v\) is a measure based on the chi-squared statistic with varies between 0 (no association) and 1 (complete association).

\[
v = \sqrt{\frac{\chi^2}{n \times \min(G-1, L-1)}}
\]
The rows and columns percentages provide often an idea about the nature of the groups.

The overall percentage of the cars that uses "gas" (essence) fuel-type is 60%. This percentage becomes 66.67% in the cluster G2. There is (very slight) an overrepresentation of the "fuel-type = gas" vehicles into this group.

44.44% of the vehicles "fuel-type = gas" (essence) are present in the cluster G2, which represent 40% of the dataset.

This idea of comparing proportions will be examined in depth for the interpretation of the clusters.
Characterizing the clusters
Quantitative variables – V-Test (test value) criterion

The samples are nested. We see in the denominator the standard error of the mean in the case of a sampling without replacement of \( n_g \) instances among \( n \).

\[
vt = \frac{\bar{x}_g - \bar{x}}{\sqrt{\frac{n - n_g}{n - 1} \times \frac{\sigma^2}{n_g}}}
\]

- \( \sigma^2 \) is the empirical variance for the whole sample
- \( n, n_g \) are respectively the size of whole sample and the cluster “g”

The test statistic is distributed approximately as a normal distribution (\(|vt| > 2\), critical region at 5% level for a test of significance).

Unlike for illustrative variables, the V-test for test of significance does not really make sense for active variables because they have participated in the creation of the groups. But it can be used for ranking the variables according their influence.
Characterizing the clusters

Quantitative variables – V-test – Example

We understand better the nature of the clusters.

The calculations are extended to illustrative variables.
Characterizing the clusters
Quantitative variables – V-test – Example

Instead of the computed value of V-TEST, it is the discrepancy and similarity between groups that must draw our attention.

There are 4 classes, but we realize that there are mainly two types of vehicles in the dataset ({G1, G3} vs. {G2, G4}). The height (hauteur) plays a major role in the distinction of the clusters.
Characterizing the clusters
Quantitative variables – Supplement the analysis

we can make pairwise comparisons.

The most important thing is to know how to read properly the results!!!

Or the comparison of one cluster vs. the others.
Characterizing the clusters
One group vs. the others – Effect size (Cohen, 1988)

The V-Test is highly sensitive to the sample size. E.g. If the sample size is multiplied by 100, the V-Test is multiplied by $10 = \sqrt{100}$

⇒ All the differences become “significant”

$$vt = \frac{\bar{x}_g - \bar{x}}{\sqrt{\frac{n - n_g \times \sigma^2}{n \times n_g}}} = \sqrt{n_g} \times \frac{\bar{x}_g - \bar{x}}{\sqrt{\frac{n - n_g \times \sigma^2}{n - 1 \times n_g}}}$$

The effect size notion allows to overcome this drawback. It is focused on the standardized difference, disregarding the sample size.

$$es = \frac{\bar{x}_g - \bar{x}_{others}}{\sigma}$$

- The effect size is insensitive to the sample size.
- The value can be read as a difference in terms of standard deviation (e.g. $0.8 \Leftrightarrow$ the difference corresponds to 0.8 times the standard error). It makes possible a comparison between different variables.
- Interpreting the effect size as difference between probabilities is also possible (using the quantile of normal distribution).
Characterizing the clusters

One group vs. the others – Effect size – Interpreting the results

Under the assumption of normal distribution

\[ es = \frac{x_{rouge} - x_{autres}}{\sigma} = \frac{0.249 - 4.502}{2.256} = -1.885 \]

\( \Phi \) cumulative distribution function (cdf) of the standardized normal distribution

\[ U_3 = \Phi(es) = 0.03 \]

There are 3% of chance that the values of the “others” groups are lower than the median of the “red” group.

\[ U_2 = \Phi(|es|/2) = 0.827 \]

82.7% of the highest values of “others” are higher than 82.7% of the lowest values of “red”.

\[ U_1 = \frac{2U_2-1}{U_2} = 0.79 \]

79% of the two distributions are not overlapped.

Other kinds of interpretation are possible (e.g. CLES ‘Common Language Effect Size’ of McGraw & Wong, 1992)
Characterizing the clusters
Categorical variables – V-Test

Based on the comparison of proportions.
Proportion of a category in the studied group vs. its proportion in the whole sample.

Frequency of the category into the group of interest (e.g. proportion of ‘fuel-type: gas’ among G2 = 66.67%)
Frequency of the category into the whole sample (e.g. proportion of ‘fuel-type: gas’ = 60%)

\[
vt = \sqrt{n_g} \times \frac{P_{l/g} - P_l}{\sqrt{\frac{n-n_g}{n-1} \times P_l \times (1 - P_l)}}
\]

\[ vt = \sqrt{12} \times \frac{0.6667 - 0.6}{\sqrt{\frac{30 - 12}{30 - 1} \times 0.6 \times (1 - 0.6)}} = 0.5986 \]

\( vt \) is distributed approximately as a normal distribution. It is especially true for the illustrative variables. Critical value \( \pm 2 \) for a two-sided significance test at 5% level.

\( vt \) is very sensitive to the sample size. The effect size notion can be used also for the comparison of proportions (Cohen, chapter 6).
Multivariate characterization

Take into account the interaction between the variables
(which are sometimes highly correlated)
Characterizing the partition

Percentage of variance explained

$R^2 = \frac{4116.424}{4695.014} = 0.877$

<table>
<thead>
<tr>
<th>$G1$</th>
<th>$G2$</th>
<th>$G3$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$X$</td>
<td>$Y$</td>
<td>$Z$</td>
</tr>
</tbody>
</table>

Note: For ensuring that the measure is valid, the clusters must have a convex shape i.e. the centroids are approximately at the center of the clusters.

Huygens theorem

$T = B + W$

$\sum_{i=1}^{n} d^2(i, G) = \sum_{g=1}^{G} n_g d^2(g, G) + \sum_{g=1}^{G} \sum_{i=1}^{n_g} d^2(i, g)$

Dispersion of the conditional centroids in relation to the overall centroid.

Dispersion inside each cluster.

Multivariate generalization of the square of the correlation ratio.

$R^2 = \frac{B}{T}$

Proportion of variance explained

The $R^2$ criterion allows to compare the efficiency of various clustering structures, only if they have the same number of clusters.
Characterizing the partition

Evaluating the proximity between the clusters

Distance between the centroids (Squared Euclidean distance for this example)

The closeness between the centroids must confirm the results provided by the other approaches, especially the univariate approach. If not, there are issues that need deeper analysis.
Characterizing the clusters
In combination with factor analysis

A factor analysis (principal component analysis - PCA - here since all the active variables are numeric) allows to obtain a synthetic view of the data, ideally in a two dimensional representation space.

We observe that the clusters are almost perfectly separable on the first factor.

But to the difficulty to understand the clusters comes in addition the difficulty to interpret the factor analysis results.
Characterizing the clusters
Principal component analysis - Cars dataset

The first factor is dominated by the size of the cars (large cars have big engine, etc.). The 2nd factor is based on the height (hauteur) of cars. We have 86.83% of the information on this first two-dimensional representation space (71.75 + 15.08).

PT Cruiser and Vel Satis are sedan cars, but they have also high height. We had not seen this information anywhere before.
Characterizing the clusters
Using supervised approach – E.g. Discriminant Analysis

We predict the clusters membership using a supervised learning algorithm. We have an overall point of view about the influence of the variables.

1st step: we are lucky (because the clusters provided by the K-Means algorithm are convex), we have a perfect discrimination. The discriminant analysis allows to recreate perfectly the clusters for our dataset.

Predicted clusters (linear discriminant analysis - LDA)

<table>
<thead>
<tr>
<th></th>
<th>G3</th>
<th>G1</th>
<th>G2</th>
<th>G4</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>G3</td>
<td>4</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>4</td>
</tr>
<tr>
<td>G1</td>
<td>0</td>
<td>7</td>
<td>0</td>
<td>0</td>
<td>7</td>
</tr>
<tr>
<td>G2</td>
<td>0</td>
<td>0</td>
<td>12</td>
<td>0</td>
<td>12</td>
</tr>
<tr>
<td>G4</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>7</td>
<td>7</td>
</tr>
<tr>
<td>Total</td>
<td>4</td>
<td>7</td>
<td>12</td>
<td>7</td>
<td>30</td>
</tr>
</tbody>
</table>

2nd step: interpretation of the LDA coefficients

<table>
<thead>
<tr>
<th>Attribute</th>
<th>G1</th>
<th>G3</th>
<th>G2</th>
<th>G4</th>
<th>F(3,20)</th>
<th>p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>puissance</td>
<td>0.688092</td>
<td>0.803565</td>
<td>1.003939</td>
<td>1.42447</td>
<td>8.37255</td>
<td>0.001</td>
</tr>
<tr>
<td>cylindree</td>
<td>-0.033094</td>
<td>-0.027915</td>
<td>-0.019473</td>
<td>0.004058</td>
<td>8.19762</td>
<td>0.001</td>
</tr>
<tr>
<td>vitesse</td>
<td>3.101157</td>
<td>3.33956</td>
<td>2.577176</td>
<td>1.85096</td>
<td>9.84801</td>
<td>0.000</td>
</tr>
<tr>
<td>longueur</td>
<td>-1.618533</td>
<td>-1.87907</td>
<td>-1.383281</td>
<td>-1.205849</td>
<td>6.94318</td>
<td>0.002</td>
</tr>
<tr>
<td>largeur</td>
<td>12.833058</td>
<td>13.640492</td>
<td>13.2026</td>
<td>13.311159</td>
<td>1.21494</td>
<td>0.330</td>
</tr>
<tr>
<td>hauteur</td>
<td>19.56544</td>
<td>21.647641</td>
<td>19.706549</td>
<td>20.206701</td>
<td>16.09182</td>
<td>0.000</td>
</tr>
<tr>
<td>poids</td>
<td>-0.145374</td>
<td>-0.122067</td>
<td>-0.130198</td>
<td>-0.118567</td>
<td>0.43201</td>
<td>0.732</td>
</tr>
<tr>
<td>constant</td>
<td>-2372.594203</td>
<td>-2816.106674</td>
<td>-2527.437401</td>
<td>-2689.157002</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Classification functions: Statistical Evaluation

These results seem consistent with the previous analysis. Comforting!

On the other hand, this is a very strange result. The speed (vitesse) seems to influence differently the clusters. We know that it is not true in the light of the PCA conducted previously.

Why these variables are not significant?

To the difficulty of recreating exactly the clusters is added the weakness of the supervised method. In this example, clearly, the coefficients of some variables are distorted by the multicollinearity.
Conclusion

- Interpreting the clustering results is a vital step in cluster analysis.

- Univariate approaches have the advantage of the simplicity. But they do not take into account the joint effect of the variables.

- Multivariate methods offer a more global view but the results are not always easy to understand.

- In practice, we have to combine the two approaches to avoid missing out important information.

- The approaches based on comparisons of means and centroids are relevant only if the clusters have convex shape.
References

Books


Tanagra Tutorials