K-medoids Algorithm
Cluster Analysis – Partitioning Approach

Ricco RAKOTOMALALA
Université Lumière Lyon 2
Outline

1. Cluster analysis – Concept of medoid
2. K-medoids Algorithm
3. Silhouette index
4. Possible extensions
5. Conclusion
6. References
Cluster analysis

Clustering, unsupervised learning
Cluster analysis

Also called: clustering, unsupervised learning, typological analysis

Input variables, used for the creation of the clusters
Often (but not always) numeric variables

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>cylindree</th>
<th>vitesse</th>
<th>longueur</th>
<th>largeur</th>
<th>hauteur</th>
<th>poids</th>
<th>co2</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>135</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>143</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>
</tr>
<tr>
<td>CITRONC2</td>
<td>61</td>
<td>1124</td>
<td>158</td>
<td>367</td>
<td>166</td>
<td>147</td>
<td>932</td>
<td>141</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>
</tr>
<tr>
<td>FIESTA</td>
<td>68</td>
<td>1399</td>
<td>164</td>
<td>392</td>
<td>168</td>
<td>144</td>
<td>1138</td>
<td>117</td>
</tr>
<tr>
<td>CLIQ</td>
<td>100</td>
<td>1461</td>
<td>185</td>
<td>382</td>
<td>164</td>
<td>142</td>
<td>980</td>
<td>113</td>
</tr>
<tr>
<td>P1007</td>
<td>75</td>
<td>1360</td>
<td>165</td>
<td>374</td>
<td>169</td>
<td>161</td>
<td>1181</td>
<td>153</td>
</tr>
<tr>
<td>MODUS</td>
<td>113</td>
<td>1598</td>
<td>188</td>
<td>380</td>
<td>170</td>
<td>159</td>
<td>1170</td>
<td>163</td>
</tr>
<tr>
<td>MUSA</td>
<td>100</td>
<td>1910</td>
<td>179</td>
<td>399</td>
<td>170</td>
<td>196</td>
<td>1275</td>
<td>146</td>
</tr>
<tr>
<td>GOLF</td>
<td>75</td>
<td>1968</td>
<td>163</td>
<td>421</td>
<td>176</td>
<td>149</td>
<td>1217</td>
<td>143</td>
</tr>
<tr>
<td>MERC_A</td>
<td>140</td>
<td>1991</td>
<td>201</td>
<td>384</td>
<td>177</td>
<td>160</td>
<td>1340</td>
<td>141</td>
</tr>
<tr>
<td>AUDIA3</td>
<td>102</td>
<td>1595</td>
<td>185</td>
<td>421</td>
<td>177</td>
<td>143</td>
<td>1205</td>
<td>168</td>
</tr>
<tr>
<td>CITRONC4</td>
<td>138</td>
<td>1997</td>
<td>207</td>
<td>426</td>
<td>178</td>
<td>146</td>
<td>1381</td>
<td>142</td>
</tr>
<tr>
<td>AVENSIS</td>
<td>115</td>
<td>1995</td>
<td>195</td>
<td>463</td>
<td>176</td>
<td>148</td>
<td>1400</td>
<td>155</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>
</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>
</tr>
<tr>
<td>LAGUNA</td>
<td>165</td>
<td>1998</td>
<td>218</td>
<td>458</td>
<td>178</td>
<td>143</td>
<td>1320</td>
<td>196</td>
</tr>
<tr>
<td>MEGANECC</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>
</tr>
<tr>
<td>P407</td>
<td>136</td>
<td>1997</td>
<td>212</td>
<td>468</td>
<td>182</td>
<td>145</td>
<td>1415</td>
<td>194</td>
</tr>
<tr>
<td>P307CC</td>
<td>180</td>
<td>1997</td>
<td>225</td>
<td>435</td>
<td>176</td>
<td>143</td>
<td>1490</td>
<td>210</td>
</tr>
<tr>
<td>PTCRUISER</td>
<td>223</td>
<td>2429</td>
<td>200</td>
<td>429</td>
<td>171</td>
<td>154</td>
<td>1595</td>
<td>235</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>
</tr>
<tr>
<td>MAZDAAR8</td>
<td>231</td>
<td>1308</td>
<td>235</td>
<td>443</td>
<td>177</td>
<td>134</td>
<td>1390</td>
<td>284</td>
</tr>
<tr>
<td>VELSATIS</td>
<td>150</td>
<td>2188</td>
<td>200</td>
<td>486</td>
<td>186</td>
<td>158</td>
<td>1735</td>
<td>188</td>
</tr>
<tr>
<td>CITRONCS</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>
</tr>
<tr>
<td>P607</td>
<td>204</td>
<td>2721</td>
<td>230</td>
<td>491</td>
<td>184</td>
<td>145</td>
<td>1723</td>
<td>223</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>1735</td>
<td>183</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>
</tr>
<tr>
<td>BMW530</td>
<td>231</td>
<td>2979</td>
<td>250</td>
<td>485</td>
<td>185</td>
<td>147</td>
<td>1495</td>
<td>231</td>
</tr>
</tbody>
</table>
Cluster analysis
Example into a two dimensional representation space

We "perceive" the groups of instances (data points) into the representation space.

The clustering algorithm has to identify the “natural” groups (clusters) which are significantly different (distant) from each other.

2 key issues
1. Determining the number of clusters
2. Delimiting these groups by machine learning algorithm
Characterizing the partition

Within-cluster sum of squares (variance)

Give crucial role to the centroids

Huygens theorem

\[ \text{TOTAL.SS} = \text{BETWEEN - CLUSTER.SS} + \text{WITHIN - CLUSTER.SS} \]
\[ T = B + W \]
\[ \sum_{i=1}^{n} d^2(i,G) = \sum_{k=1}^{K} n_k d^2(G_k,G) + \sum_{k=1}^{K} \sum_{i=1}^{n_k} d^2(i,G_k) \]

Dispersion of the clusters' centroids around the overall centroid.
Clusters separability indicator.

Dispersion inside the clusters.
Clusters compacity indicator.

d() is a distance measurement characterizing the proximity between individuals. E.g. Euclidean distance or Euclidean distance weighted by the inverse of variance. Pay attention to outliers.

The aim of the cluster analysis would be to minimize the within-cluster sum of squares (W), to a fixed number of clusters (e.g. K-Means algorithm).

Note: Since the instances are attached to a group according to their proximity to their centroid, the shape of the clusters tends to be spherical.
The concept of “medoid”

Representative data point of a cluster

The centroid (\(o\)) may be totally artificial, it may not correspond to the real configuration of the dataset.

The concept of medoid (\(x\)) is more appropriate in some circumstances. This is an observed data point which minimizes its distance to all the other instances.

\[
M = \arg \min_{m} \sum_{i=1}^{n} d(i, m) \quad m = 1, \ldots, n; \text{each data point is candidate to be medoid.}
\]

\[
E = \sum_{k=1}^{K} \sum_{i=1}^{n_k} d(i, M_k)
\]

It can be used as measure for the quality of the partition, instead of the within cluster sum of squares.

\[
d(i, i') = \sum_{j=1}^{p} | x_{ij} - x_{i'j} |
\]

We are no longer limited to the Euclidean distance. The Manhattan distance for instance allows to dramatically reduces the influence of outliers.
Partitioning-based clustering

Generic iterative relocation clustering algorithm

Main steps

- Set the number of clusters K
- Set a first partition of the data
- **Relocation.** Move objects (instances) from one group to another to obtain a better partition
- The aim (implicitly or explicitly) is to optimize some objective function evaluating the partitioning
- Provides an unique partitioning of the objects (unique solution)

But can be depending on other parameters such as the maximum diameter of the clusters. Remains an open problem often.

Often in a random fashion. But can also start from another partition method or rely on considerations of distances between individuals (e.g., the K most distant individuals from each other).

By processing all individuals, or by attempting to have random exchanges (more or less) between groups.

The measure E will be used (see the previous slide).

We have a unique solution for a given value of K. And not a hierarchy of partitions as for HAC (hierarchical agglomerative clustering) for example.
K-medoids algorithm

Several possible approaches
K-Medoids Algorithm

A variant of K-Means Algorithm

A straightforward algorithm

Input: X (n obs., p variables), K #groups

Initialize K medoids $M_k$

REPEAT
  Assignment. Assign each observation to the group with the nearest medoid
  Update. Recalculate the medoids from individuals attached to the groups
UNTIL Convergence

Output: A partition of the instances in K groups characterized by their medoids $M_k$

The process minimizes implicitly the overall measure $E$

The complexity of this approach is especially dissuasive

It is necessary to calculate the matrix of pairwise distances between individuals $d(i,i')$, $i,i' = 1,...,n$

May be K instances selected randomly. Or, K instances which are nearest to the others.

The pairwise distance between the data points being calculated beforehand, it is no longer necessary to access to the database.

Inevitably, the dispersion $E_k$ between the cluster $C_k$ decreases (at least remains stable)

Fixed number of iterations
Or when $E$ no longer decreases
Or when the medoids $M_k$ are stable

Ricco Rakotomalala
Tutoriels Tanagra - http://tutoriels-data-mining.blogspot.fr/
PAM Algorithm
Partitioning around medoid (Kaufman & Rousseeuw, 1987)

**Input:** X (n obs., p variables), K #groups

Initialize K medoids $M_k$

**REPEAT**

Assign each observation to the group with the nearest medoid

*For Each* medoid $M_k$

Select randomly a non-medoid data point $i$

Check if the criterion $E$ decreases if we swap their role. If YES, the data point $i$ becomes the medoid $M_k$ of the cluster $C_k$

**UNTIL** The criterion $E$ does not decrease

**Output:** A partition of the instances in K groups characterized by their medoids $M_k$

Here again, it needs to calculate the matrix of pairwise distance $d(i,i')$.

The complexity of the approach remains excessive.

PAM Algorithm

PAM vs. K-Means on an artificial dataset

Because the shapes of the clusters are spherical, the medoids are almost equivalent to the centroids.

```r
library(cluster)
res <- pam(X, 3, FALSE, "euclidean")
print(res)
plot(X[,1], X[,2], type="p", xlab="X1", ylab="X2", col=c("lightcoral", "skyblue", "greenyellow")[res$clustering])
points(res$medoids[,1], res$medoids[,2], cex=1.5, pch=16, col=c("red", "blue", "green")[1:3])
```
PAM Algorithm

PAM vs. K-Means on an artificial dataset with outliers

Outliers. Source of trouble typically.

K-Means can be distorted.

PAM remains valid. The medoids are placed wisely.
PAM Algorithm

PAM on the Cars Dataset

Plotting the instances into the individuals factor map (principal components analysis). We distinguish the clusters. The medoids are highlighted (dotted circle).
CLARA Algorithm
Clustering Large Applications
(Kaufman & Rousseeuw, 1990)

Input: X (n obs., p variables), K #clusters

Draw S samples of size \( \eta \) (\( \eta \ll n \))

Apply PAM algorithm on each sample \( \Rightarrow S \) vectors of medoids

For Each vector of medoids
    Assign all the instances to its cluster
    Evaluate the quality of the partition E

Retain the solution which minimize E

Output: A partition of the instances in K groups characterized by their medoids \( M_k \)

In practice: \( S = 5 \) and \( \eta = 40 + 2 \times K \) are adequate \([default settings for clara()] in the R “cluster” package for Cluster Analysis\).

Only one single pass on the data is sufficient to evaluate all the configurations.

Ability to process large databases

The algorithm is heavily dependent on the size and representativeness of the samples
The CLARA Algorithm
Example for the "waveform" dataset (Breiman and al., 1984)

This is an artificial dataset. The "true" class (CLASSE) membership of the individuals is known.

21 descriptors

<table>
<thead>
<tr>
<th>V1</th>
<th>V2</th>
<th>V3</th>
<th>V4</th>
<th>V5</th>
<th>V6</th>
<th>V7</th>
<th>V8</th>
<th>V9</th>
<th>V10</th>
<th>V11</th>
<th>V12</th>
<th>V13</th>
<th>V14</th>
<th>V15</th>
<th>V16</th>
<th>V17</th>
<th>V18</th>
<th>V19</th>
<th>V20</th>
<th>V21</th>
</tr>
</thead>
<tbody>
<tr>
<td>-0.29</td>
<td>-2.24</td>
<td>-0.65</td>
<td>0.73</td>
<td>-0.15</td>
<td>1.37</td>
<td>3.21</td>
<td>1.31</td>
<td>1.94</td>
<td>2.06</td>
<td>1.86</td>
<td>2.67</td>
<td>1.42</td>
<td>4.4</td>
<td>3.99</td>
<td>4.17</td>
<td>1.57</td>
<td>1.28</td>
<td>2.34</td>
<td>2.32</td>
<td>-1.17 A</td>
</tr>
<tr>
<td>-1.82</td>
<td>1.89</td>
<td>1.1</td>
<td>0.76</td>
<td>2.42</td>
<td>4.59</td>
<td>3.54</td>
<td>3.55</td>
<td>3.07</td>
<td>4.81</td>
<td>1.74</td>
<td>1.65</td>
<td>2.31</td>
<td>2.64</td>
<td>2.73</td>
<td>2.2</td>
<td>1.85</td>
<td>0.33</td>
<td>0.04</td>
<td>-0.85</td>
<td>1.03 A</td>
</tr>
<tr>
<td>1.11</td>
<td>-0.42</td>
<td>1.4</td>
<td>-0.27</td>
<td>0.12</td>
<td>2.35</td>
<td>5.86</td>
<td>3.73</td>
<td>4.42</td>
<td>3.72</td>
<td>2.67</td>
<td>2.27</td>
<td>0.44</td>
<td>1.58</td>
<td>-0.02</td>
<td>2.48</td>
<td>0.58</td>
<td>1.04</td>
<td>0.46</td>
<td>1.55</td>
<td>-0.39 B</td>
</tr>
<tr>
<td>-1.57</td>
<td>0.52</td>
<td>0.55</td>
<td>1.67</td>
<td>4.56</td>
<td>2.15</td>
<td>0.04</td>
<td>5.24</td>
<td>2.94</td>
<td>1.15</td>
<td>0.48</td>
<td>1.64</td>
<td>0.2</td>
<td>0.26</td>
<td>1.37</td>
<td>3.03</td>
<td>2.03</td>
<td>1.28</td>
<td>0.53</td>
<td>1.07</td>
<td>0.23 A</td>
</tr>
<tr>
<td>-0.72</td>
<td>-0.44</td>
<td>-1.02</td>
<td>-0.49</td>
<td>-0.63</td>
<td>0.92</td>
<td>2.42</td>
<td>2.81</td>
<td>4.03</td>
<td>4.33</td>
<td>6.45</td>
<td>5.84</td>
<td>3.88</td>
<td>3.77</td>
<td>1.41</td>
<td>1.32</td>
<td>0.06</td>
<td>-1.22</td>
<td>0.28</td>
<td>-1.65</td>
<td>-0.42 C</td>
</tr>
</tbody>
</table>

The two partitions are almost equivalent. But, the computation time is dramatically reduced with CLARA.

PAM : 443 sec. (+ de 7 min)
CLARA : 0.04 sec.

With the external validation (knowing the real class membership), the various approaches provide similar performances.

Crosstab between CLUSTER vs. CLASSE.

<table>
<thead>
<tr>
<th>C1</th>
<th>C2</th>
<th>C3</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>9362</td>
<td>485</td>
</tr>
<tr>
<td>B</td>
<td>2</td>
<td>9147</td>
</tr>
<tr>
<td>C</td>
<td>852</td>
<td>153</td>
</tr>
</tbody>
</table>

Cramer’s V = 0.85

The three methods encounter the same difficulties on this dataset.
Silhouette analysis

A tool for selecting the number of clusters
Silhouette criterion

How well the object lies within its cluster

Rousseeuw (1987) provides a criterion which enables to evaluate a partition independently to the number of clusters (Silhouette).

\[
a(i) = \frac{1}{n_a - 1} \sum_{i' = 1}^{n_a} d(i, i')
\]

Average distance of a data point \(i\) with all the other data within the same cluster \(C_a\) of size \(n_a\).

\[
d(i, C_k) = \frac{1}{n_k} \sum_{i' = 1}^{n_k} d(i, i')
\]

Average distance of data point \(i\) with all the instances of the cluster \(C_k\) – other than \(C_a\) – of size \(n_k\).

\[
b(i) = \min_{k \neq a} d(i, C_k)
\]

Distance to the nearest cluster in the sense of \(d(i, C_k)\)

\[
s(i) = \frac{b(i) - a(i)}{\max\{a(i), b(i)\}}
\]

Level of membership to its cluster of the individual \(i\), by comparing the distance to its cluster with the distance to the nearest cluster. \(s(i)\) it is independent of \(K\) - the number of clusters - because we consider only the distance to the nearest cluster!

\(s(i) \to 1\) : the data point is well positioned within its cluster  
\(s(i) \approx 0\) : the individual is very close to the decision boundary between two neighboring clusters  
\(s(i) \to -1\) : the data point might be assigned to the wrong cluster
Silhouette criterion
Evaluation of the cluster and the partition

- $s(x) > s(o)$: (1) because « x » is near the central position (it is the medoid of the cluster) into $C_1$; (2) because « o » is closer to the cluster $C_2$.

Characterize both the **cohesion** of the cluster $C_k$ and its **separation** to the other clusters.

$$\bar{s}_k = \frac{1}{n_k} \sum_{i \in C_k} s(i)$$

**Average silhouette.** Characterize the overall quality of the partition in $K$ groups. As a rule of thumb:

- $S \in [0.71 ; 1]$ : strong separation
- $S \in [0.51 ; 0.70]$ : medium separation
- $S \in [0.26 ; 0.50]$ : low separation, may be questionable
- $S \in [0 ; 0.25]$ : the partition seems not meaningful
Determining the number of clusters is an open problem in cluster analysis. The silhouette criterion being independent to the number of clusters, we can choose the value $K$ which maximize the criterion.

$$\bar{s}_1 = 0.60$$

$$\bar{s}_2 = 0.53$$

$$\bar{s}_3 = 0.70$$

The cluster $C_3$ is the one which is furthest to the others.

$$S_{K=3} = 0.61$$

Overall quality of the partition in $K = 3$ groups.

Try various values of $K$ and identify the best solution (partition in $K$ clusters). Here $K = 2$ ($S_2 = 0.63$) and $K = 3$ ($S_3 = 0.61$) are competing. Why the solution $K = 2$ clusters appears always as the best one whatever the criteria used?
We observe on the one hand the cohesion of the cluster (the group has a higher value $s_k$ than the others); on the other hand, we observe the homogeneity of the situations within the cluster. For instance, for the red group, only few instances have a low value of silhouette $s(i)$.
Possible extensions

• The algorithm can be applied to a dataset with categorical variables (e.g. using the chi-squared distance).
• The tandem analysis approach (factor analysis + clustering) is also possible, we can process dataset with mixed data (with both numeric and categorical variables).
• Extension to fuzzy clustering is possible (“fanny” algorithm).
• The extension to the clustering of variables is also easy: $r^2$ can be used as similarity measure between variables, $(1 - r^2)$ is the distance measure [or respectively $r$ and $(1-r)$ if we want to take account the sign of the association].
Conclusion

• Partitioning methods are often simple. This is an advantage.

• The k-medoids approaches enables to alleviate the problem of outliers, by modifying the notion of representative points of the clusters, by using also appropriate distance measure (e.g. Manhattan distance).

• PAM (Partitioning Aroung Medoids) is a popular implementation of the approach. But the necessity to calculate distances between pairs of individuals is very expensive in computation time.

• We can dramatically improve the ability to handle large datasets by working on samples (CLARA method).

• A criterion for evaluating the partitions, insensitive to the number of clusters, is proposed: the criterion silhouette.

• We can use it to determine the right number of clusters. But it is a heuristic, the choice of the number of clusters must be supported by the interpretation.
Books and articles


Wikipedia, « k-medoids ».

Wikipedia, « Silhouette (clustering) ».