Clustering of categorical variables

Grouping categorical variables
Grouping categories of nominal variables

Ricco RAKOTOMALALA
Université Lumière Lyon 2
Outline

1. Clustering of categorical variables. Why?
   a. HCA from a dissimilarity matrix
   b. Deficiency of the clustering of categorical variables

2. Clustering categories of nominal variables
   a. Distance between categories – Dice’s coefficient
   b. HAC on the categories
   c. Interpretation of the obtained clusters

3. Other approaches for the clustering of categories

4. Conclusion

5. References
Clustering of categorical variables

Why? For what purpose?
Clustering of variables

Goal: grouping related variables

- The variables in the same group are highly associated together.
- The variables in different groups are not related (in the sense of association measure)

With what objective?

1. Identify the underlying structure of the dataset. Make a summary of the relevant information (the approach is complementary to the clustering of individuals).

2. Detect redundancies, for instance in order to selecting the variables intended for a subsequent analysis (e.g. supervised learning task)
   a. In a pretreatment phase, in order to organize the search space
   b. In a post-treatment phase, in order to understand the role of the removed variables in the selection process.
An example: **Vote dataset** (1984)

n = 435 individuals (US Congressmen)
p = 6 active variables

<table>
<thead>
<tr>
<th>Variable</th>
<th>Categories</th>
<th>Role</th>
</tr>
</thead>
<tbody>
<tr>
<td>affiliation</td>
<td>democrat, republican</td>
<td>illustrative</td>
</tr>
<tr>
<td>budget</td>
<td>yes, no, neither</td>
<td>active</td>
</tr>
<tr>
<td>physician</td>
<td>yes, no, neither</td>
<td>active</td>
</tr>
<tr>
<td>salvador</td>
<td>yes, no, neither</td>
<td>active</td>
</tr>
<tr>
<td>nicaraguan</td>
<td>yes, no, neither</td>
<td>active</td>
</tr>
<tr>
<td>missile</td>
<td>yes, no, neither</td>
<td>active</td>
</tr>
<tr>
<td>education</td>
<td>yes, no, neither</td>
<td>active</td>
</tr>
</tbody>
</table>

Political affiliation

Illustrative variable i.e. used for understanding the nature of the groups

Vote on each subject, 3 categories: yes (yea), no (nay), neither (not “yea” or “nay”)

Active variables

Identify the vote which are highly related
Establish their association with the political affiliation

We observe that a vote "yea" to a subject may be highly related to vote "nay" to another subject.
HAC from a dissimilarity matrix
Hierarchical agglomerative clustering

Using the Cramer’s V to measure the association between the nominal variables
Measure of association between 2 nominal variables

### Pearson's chi-squared statistic

\[
\chi^2 = \sum \sum \frac{(n_{kl} - e_{kl})^2}{e_{kl}}
\]

- \(e_{kl} = \frac{n_k \times n_l}{n}\)
- \# P(AB) observed
- \# P(A) x P(B)
- Under the independence assumption

### Cramer's \(v\)

\[
v = \sqrt{\frac{\chi^2}{n \times \min(K-1, L-1)}}
\]

- Symmetrical
- \(0 \leq v \leq 1\)

#### Example

<table>
<thead>
<tr>
<th>A \ B</th>
<th>(b_1)</th>
<th>(b_l)</th>
<th>(b_L)</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>(a_1)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>(a_k)</td>
<td>\cdots</td>
<td>(n_{kl})</td>
<td>\cdots</td>
<td>(n_k)</td>
</tr>
<tr>
<td>(a_K)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Total</strong></td>
<td>(n_{L})</td>
<td></td>
<td>(n)</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Nombre de budget</th>
<th>physician</th>
<th>Total général</th>
</tr>
</thead>
<tbody>
<tr>
<td>budget</td>
<td>n</td>
<td>neither y</td>
</tr>
<tr>
<td>n</td>
<td>25</td>
<td>146</td>
</tr>
<tr>
<td>neither</td>
<td>3</td>
<td>6</td>
</tr>
<tr>
<td>y</td>
<td>219</td>
<td>5</td>
</tr>
<tr>
<td><strong>Total général</strong></td>
<td>247</td>
<td>11</td>
</tr>
</tbody>
</table>

\(\chi^2 = 355.48\)

\(p.value < 0.0001\)

High association Significant at the 5% level

\(v = 0.639\)
#function for calculating Cramer's v
cramer <- function(y,x){
  K <- nlevels(y)
  L <- nlevels(x)
  n <- length(y)
  chi2 <- chisq.test(y,x,correct=F)
  print(chi2$statistic)
  v <- sqrt(chi2$statistic/(n*min(K-1,L-1)))
  return(v)
}

Similarity matrix (Cramer’s v)

<table>
<thead>
<tr>
<th></th>
<th>budget</th>
<th>physician</th>
<th>salvador</th>
<th>nicaraguan</th>
<th>missile</th>
<th>education</th>
</tr>
</thead>
<tbody>
<tr>
<td>budget</td>
<td>1</td>
<td>0.639</td>
<td>0.507</td>
<td>0.517</td>
<td>0.439</td>
<td>0.475</td>
</tr>
<tr>
<td>physician</td>
<td>0.639</td>
<td>1</td>
<td>0.576</td>
<td>0.518</td>
<td>0.471</td>
<td>0.509</td>
</tr>
<tr>
<td>salvador</td>
<td>0.507</td>
<td>0.576</td>
<td>1</td>
<td>0.611</td>
<td>0.558</td>
<td>0.470</td>
</tr>
<tr>
<td>nicaraguan</td>
<td>0.517</td>
<td>0.518</td>
<td>0.611</td>
<td>1</td>
<td>0.545</td>
<td>0.469</td>
</tr>
<tr>
<td>missile</td>
<td>0.439</td>
<td>0.471</td>
<td>0.558</td>
<td>0.545</td>
<td>1</td>
<td>0.427</td>
</tr>
<tr>
<td>education</td>
<td>0.475</td>
<td>0.509</td>
<td>0.470</td>
<td>0.469</td>
<td>0.427</td>
<td>1</td>
</tr>
</tbody>
</table>

Dissimilarity matrix (1-v)

<table>
<thead>
<tr>
<th></th>
<th>budget</th>
<th>physician</th>
<th>salvador</th>
<th>nicaraguan</th>
<th>missile</th>
<th>education</th>
</tr>
</thead>
<tbody>
<tr>
<td>budget</td>
<td>0</td>
<td>0.361</td>
<td>0.493</td>
<td>0.483</td>
<td>0.561</td>
<td>0.525</td>
</tr>
<tr>
<td>physician</td>
<td>0.361</td>
<td>0</td>
<td>0.424</td>
<td>0.482</td>
<td>0.529</td>
<td>0.491</td>
</tr>
<tr>
<td>salvador</td>
<td>0.493</td>
<td>0.424</td>
<td>0</td>
<td>0.389</td>
<td>0.442</td>
<td>0.530</td>
</tr>
<tr>
<td>nicaraguan</td>
<td>0.483</td>
<td>0.482</td>
<td>0.389</td>
<td>0</td>
<td>0.455</td>
<td>0.531</td>
</tr>
<tr>
<td>missile</td>
<td>0.561</td>
<td>0.529</td>
<td>0.442</td>
<td>0.455</td>
<td>0</td>
<td>0.573</td>
</tr>
<tr>
<td>education</td>
<td>0.525</td>
<td>0.491</td>
<td>0.530</td>
<td>0.531</td>
<td>0.573</td>
<td>0</td>
</tr>
</tbody>
</table>

We can use this matrix as input for the HAC algorithm
We get a vision of the structures of association between variables. e.g. "budget" and "physician" are related i.e. there is a strong coherence of votes \( (v = 0.639)\); budget and salvador are less related \( (v = 0.507)\), etc. but we do not know on what association of votes (yes or no) these relationships are based...
Other approaches for clustering categorical variables

ClustOfVar (Chavent and al., 2012)

“Centroid” (representative variable) of a group of variables = latent variable i.e. the group is scored as a single variable

\[ F = 1^{\text{st}} \text{ factor from the MCA} \]
\[ \eta(.) \text{ correlation ratio} \]
\[ \lambda \text{ Variation within the group} \]

\[ \lambda = \sum_{j=1}^{p} \eta^2 (X_j, F) \]

Various strategies for grouping are possible.

- HAC approaches: minimizing the loss of variation at each step
- K-Means approach: assign the variables to the closest "centroid" (in the sense of the correlation ratio) during the learning process

1. “ClustOfVar” can handle dataset with mixed numeric and categorical variables. The centroid is defined with first component of the factor analysis for mixed data
2. This is a generalization of the CLV approach (Vigneau and Qannari, 2003) which can handle numeric variables only and is based on PCA (principal component analysis)
ClustOfVar on the « vote » dataset

library(ClustOfVar)

arbre <- hclustvar(X.quali=vote.active)
plot(arbre)

mgroups <- kmeansvar(X.quali=vote.active,init=2,nstart=10)
print(summary(mgroups))

We obtain the same results as for the HAC on the (1-v) dissimilarity matrix
Issues for the interpretation of the results

The clustering of categorical variables gives a partial vision of the structure of the relationships among variables...
Interpreting a cluster – Ex. G2

### Main associations between the categories

<table>
<thead>
<tr>
<th>Budget = y</th>
<th>Physician = n</th>
<th>Education = n</th>
</tr>
</thead>
<tbody>
<tr>
<td>219</td>
<td>5</td>
<td>29</td>
</tr>
<tr>
<td>219</td>
<td>5</td>
<td>29</td>
</tr>
<tr>
<td>219</td>
<td>5</td>
<td>29</td>
</tr>
<tr>
<td>219</td>
<td>5</td>
<td>29</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Budget = n</th>
<th>Physician = y</th>
<th>Education = y</th>
</tr>
</thead>
<tbody>
<tr>
<td>25</td>
<td>10</td>
<td>133</td>
</tr>
<tr>
<td>25</td>
<td>10</td>
<td>133</td>
</tr>
<tr>
<td>25</td>
<td>10</td>
<td>133</td>
</tr>
<tr>
<td>25</td>
<td>10</td>
<td>133</td>
</tr>
</tbody>
</table>

This kind of analysis cannot be done manually.
Analyzing the illustrative variables
The illustrative variables are used to strengthen the interpretation of the results.

```r
#2 subgroups
groups <- cutree(tree,k=2)
print(groups)

#Cramer's v : affiliation vs. attributes
cv <- sapply(vote.active,cramer,x=vote.data$affiliation)
print(cv)

#mean of v for each group
m <- tapply(X=cv,INDEX=groups,FUN=mean)
print(m)
```

<table>
<thead>
<tr>
<th>Variable</th>
<th>Affiliation (Cramer’s v)</th>
<th>Mean (v)</th>
</tr>
</thead>
<tbody>
<tr>
<td>nicaraguan</td>
<td>0.660</td>
<td>0.667</td>
</tr>
<tr>
<td>missile</td>
<td>0.629</td>
<td></td>
</tr>
<tr>
<td>education</td>
<td>0.688</td>
<td></td>
</tr>
<tr>
<td>budget</td>
<td>0.740</td>
<td>0.781</td>
</tr>
<tr>
<td>physician</td>
<td>0.914</td>
<td></td>
</tr>
<tr>
<td>salvador</td>
<td>0.712</td>
<td></td>
</tr>
</tbody>
</table>

- The political affiliation has a little more influence for the votes in G2 than in G1 (why? the subjects are more sensitive in G2?)
- We do not know what are the votes of the democrats (republicans)?
Clustering the categories of categorical variables (1)

Identifying the nature of the association between the categorical variables
Distance between categories – Dice’s coefficient

Dice coefficient. Squared difference between the dummy coding 0/1 for each category of variables. Square of the Euclidean distance.

\[ \delta_{ij}^2 = \frac{1}{2} \sum_{i=1}^{n} (m_{ij} - m_{ij'}^2) \]

\( i \) is the individual no \( i \)
\( j \) is the \( j \)th category
\( m_{ij} \) is an indicator for the \( j \)th category

Transforming the initial data table into a table of indicator variables.

```
#dummy coding
library(ade4)
disj <- acm.disjonctif(vote.active)
print(head(vote.active))
print(head(disj))
```

Simple coding scheme
# Dice’s index

dice <- function(m1,m2){
  return(0.5*sum((m1-m2)^2))
}

# Dice’s index matrix
d2 <- matrix(0,ncol(dij),ncol(disj))
for (j in 1:ncol(disj)){
  for (jprim in 1:ncol(disj)){
    d2[j,jprim] <- dice(disj[,j],disj[,jprim])
  }
}
colnames(d2) <- colnames(disj)ownames(d2) <- colnames(disj)

#transform the matrix in a R ‘dist’ class
d <- as.dist(sqrt(d2))

## Distance matrix

The distance is high for the indicator variables coming from the same categorical variable.

A low value points out a high association between the categories (e.g. budget = n and physician = y, …)
HAC of the categories, based on the Dice’s index

#cluster analysis on indicator variables

```r
arbre.moda <- hclust(d, method="ward.D2")
plot(arbre.moda)
```

3 groups now are highlighted. We distinguish clearly the relationships between the categories i.e. the votes which are related
HAC of categories under Tanagra

http://tutoriels-data-mining.blogspot.fr/2013/12/classification-de-variables-qualitatives_21.html

Linking criterion : « average linkage »

Dendrogram
(height : aggregation distance)

Association of the categories to the groups

Evolution of the aggregation distance
(an “elbow” gives an indication about the right number of groups)
HAC of categories, handling the illustrative variables

```r
# create 3 groups
dgroups <- cutree(arbre.moda,k=3)

# illustrative variable – dummy coding scheme
illus <- acm.disjonctif(as.data.frame(vote.data$affiliation))
colnames(illus) <- c("democrat","republican")

# distance to illustrative levels
dice.democrat <- sapply(disj,dice,m2=illus$democrat)
tapply(dice.democrat,dgroups,mean)

dice.republican <- sapply(disj,dice,m2=illus$republican)
tapply(dice.republican,dgroups,mean)
```

### Republican
- Budget = n
- Physician = y
- Salvador = y
- Nicaraguan = n
- Missile = n
- Education = y

### Democrat
- Budget = y
- Physician = n
- Salvador = n
- Nicaraguan = y
- Missile = y
- Education = n

We understand the influence of the political association on the votes.

For a category (from the illustrative variable), mean of the squared distance to the indicator variables of a group

<table>
<thead>
<tr>
<th>Variable = level</th>
<th>Cluster 1</th>
<th>Cluster 2</th>
<th>Cluster 3</th>
</tr>
</thead>
<tbody>
<tr>
<td>affiliation = republican</td>
<td>30.9</td>
<td>86.6</td>
<td>184.0</td>
</tr>
<tr>
<td>affiliation = democrat</td>
<td>186.6</td>
<td>130.9</td>
<td>33.5</td>
</tr>
</tbody>
</table>
Clustering the categories of categorical variables (2)

Using other measures of similarity and dissimilarity
Varclus of the « Hmisc » package for R

**Similarity measure**

\[ s_{jj'} = \frac{1}{n} \sum_{i=1}^{n} m_{ij} \times m_{ij'} \]

**Dissimilarity measure**

\[ d_{jj'} = 1 - s_{jj'} \]

Conjoint frequencies i.e. the proportion of the individuals which belong to the 2 categories (0: no instances belong simultaneously to the 2 studied categories; 1: all the instances have the two categories)

- Caution, this is not a distance (\( d_{jj'} \neq 0 \)), but this does not interfere with the hclust() procedure.
- \( d_{jj'} = 1 \) necessarily for 2 categories belonging to the same variable. Their merging is only possible at the end of the aggregation process (HAC).

```r
# loading the package
library(Hmisc)
# calling the "varclus" function
# see the help file for the parameters
v <- varclus(as.matrix(disj), type="data.matrix", similarity="bothpos", method="ward.D")
plot(v)
```

The partition into 3 groups is also obvious here.
Clustering the categories of categorical variables (3)
Tandem clustering

Two steps:
1. Calculating the coordinates of the categories in a new representation space.
2. Performing the clustering with the Euclidean distance.

Factor scores from the MCA (multiple correspondence analysis)

Individuals = categories. Performing the HAC (or another clustering approach) in the new representation space. We can use only a few number of factors. This can be viewed as a regularization strategy.

MCA: the two first factors seem enough here.

1 more disparate group (the votes are more scattered)
HAC from the factor scores – Euclidean distance

The individuals (categories) have not the same frequency (weight). If they are very different, we should take into account that in the clustering process (see. "members" parameter of `hclust`).

Coordinates of the categories of the illustrative variable in the factorial representation space.

The individuals (categories) have not the same frequency (weight). If they are very different, we should take into account that in the clustering process (see. "members" parameter of `hclust`).

L’association avec les groupes apparaît naturellement

The individuals (categories) have not the same frequency (weight). If they are very different, we should take into account that in the clustering process (see. "members" parameter of `hclust`).

L’association avec les groupes apparaît naturellement
Conclusion
Conclusion

The clustering of qualitative variables seeks to gather together variables into clusters: variables in the same group are strongly related each other; variables in different groups are weakly related.

The method is interesting if we try to detect redundancies, e.g. to help the variable selection process in a supervised learning task.

But it does not give indications about the nature of the associations between the variables.

In this context, it is more relevant to perform a clustering of categories of categorical variables.

The approach is mainly based on the definition of a similarity measure between categories.

But other approaches are possible e.g. a tandem clustering: in a first step, we calculate the scores of the categories in a new representation space; in a second step, we perform the clustering process using these new variables.
References
H. Abdallah, G. Saporta, « Classification d’un ensemble de variables qualitatives » (Clustering of a set of categorical variables), in Revue de Statistique Appliquée, Tome 46, N°4, pp. 5-26, 1998.
