Clustering with Mixed Type Variables and Determination of Cluster Numbers

Hana Řezanková, Dušan Húsek
Tomáš Löster

University of Economics, Prague
ICS, Academy of Sciences of the Czech Republic

Outline

- Motivation
- Methods for clustering with mixed type variables
- Implementation in software packages
- Proposal of new criteria for cluster evaluation
- Application
- Conclusion

Motivation

- Task: We are looking for groups of similar objects (e.g. respondents), i.e. we will concentrate on the problem of object clustering
- The objects are characterized by both quantitative and qualitative (nominal) variables (e.g. respondent opinions, numbers of actions)
- The number of clusters is unknown in advance – i.e. we should cope with appropriate number of clusters determination (assignment)

Methods for clustering with mixed type variables

- Using a specialized dissimilarity measure (Gower's coefficient, cluster variability based) and application of agglomerative hierarchical cluster analysis (AHCA)
- Clustering objects separately with quantitative and qualitative variables and combining the results by cluster-based similarity partitioning algorithm (CSPA)
- Latent class models

Implementation in software packages

- Specialized dissimilarity measures
  - are not implemented for AHCA
- Clustering objects with qualitative variables
  - is implemented only rarely (disagreement coef.)
- Cluster-based similarity partitioning algorithm
  - is not implemented but it could be realized
- LC Cluster models (Latent GOLD)
- Log-likelihood distance measure between clusters
  - implemented in two-step cluster analysis (SPSS)

Implementation in software packages

- Log-likelihood distance measure between clusters - implemented in two-step cluster analysis (SPSS)

$$D_{hh'} = \xi_{\langle h,h' \rangle} - (\xi_h + \xi_{h'})$$

$$\xi_g = n_g \left(\sum_{l=1}^{m^{(1)}} \frac{1}{2} \ln(s_l^2 + s_{gl}^2) + \sum_{l=1}^{m^{(2)}} H_{gl} \right)$$

$$H_{gl} = -\sum_{u=1}^{K_l} \frac{n_{glu}}{n_g} \ln \frac{n_{glu}}{n_g}$$ … entropy

Implementation in software packages

- Log-likelihood distance measure between objects - implemented in two-step cluster analysis (SPSS)

$$D_{hh'} = \xi_{\langle h,h' \rangle} - (\xi_h + \xi_{h'})$$

$$\xi_g = n_g \left(\sum_{l=1}^{m^{(1)}} \frac{1}{2} \ln(s_l^2 + s_{gl}^2) + \sum_{l=1}^{m^{(2)}} H_{gl} \right)$$

$$D(\mathbf{x}_i, \mathbf{x}_j) = \xi_{\langle \mathbf{x}_i, \mathbf{x}_j \rangle}$$

Evaluation criteria implemented in software packages

- BIC (*Bayesian Information Criterion)*
  AIC (*Akaike Information Criterion*)
  - implemented in two-step cluster analysis (SPSS)

$$I_{\mathrm{BIC}} = 2\sum_{g=1}^{k} \xi_g + w_k \ln(n)$$

… minimum

$$w_k = k\left(2m^{(1)} + \sum_{l=1}^{m^{(2)}} (K_l - 1)\right)$$

$$I_{\mathrm{AIC}} = 2\sum_{g=1}^{k} \xi_g + 2w_k$$

only for initial estimation of number of clusters

Proposed evaluation criteria

- Within-cluster variability for *k* clusters:

$$\xi(k) = \sum_{g=1}^{k} \xi_g = \sum_{g=1}^{k} n_g \left(\sum_{l=1}^{m^{(1)}} \frac{1}{2} \ln(s_l^2 + s_{gl}^2) + \sum_{l=1}^{m^{(2)}} H_{gl} \right)$$

- Variability of the whole data set:

$$\xi(1) = n \sum_{l=1}^{m^{(1)}} \frac{1}{2} \ln(2 s_l^2) + \sum_{l=1}^{m^{(2)}} H_l$$

Proposed evaluation criteria

- Within-cluster variability for *k* clusters:

$$\xi(k) = \sum_{g=1}^{k} \xi_g = \sum_{g=1}^{k} n_g \left(\sum_{l=1}^{m^{(1)}} \frac{1}{2} \ln(s_l^2 + s_{gl}^2) + \sum_{l=1}^{m^{(2)}} H_{gl} \right)$$

difference $$diff(k) = \xi(k-1) - \xi(k)$$

it should be maximal
for the suitable number of clusters

Evaluation criteria modified for qualitative variables

1. *Uncertainty index (R-square (RSQ) index)*

$$I_{\mathrm{U}}(k) = \frac{V_{\mathrm{B}}}{V_{\mathrm{T}}} = \frac{V_{\mathrm{T}} - V_{\mathrm{W}}}{V_{\mathrm{T}}} = \frac{\xi(1) - \xi(k)}{\xi(1)}$$

2. *Semipartial uncertainty index*
   (optimal number of clusters - minimum)

$$I_{\mathrm{SPU}}(k) = I_{\mathrm{U}}(k+1) - I_{\mathrm{U}}(k)$$

Evaluation criteria modified for qualitative variables

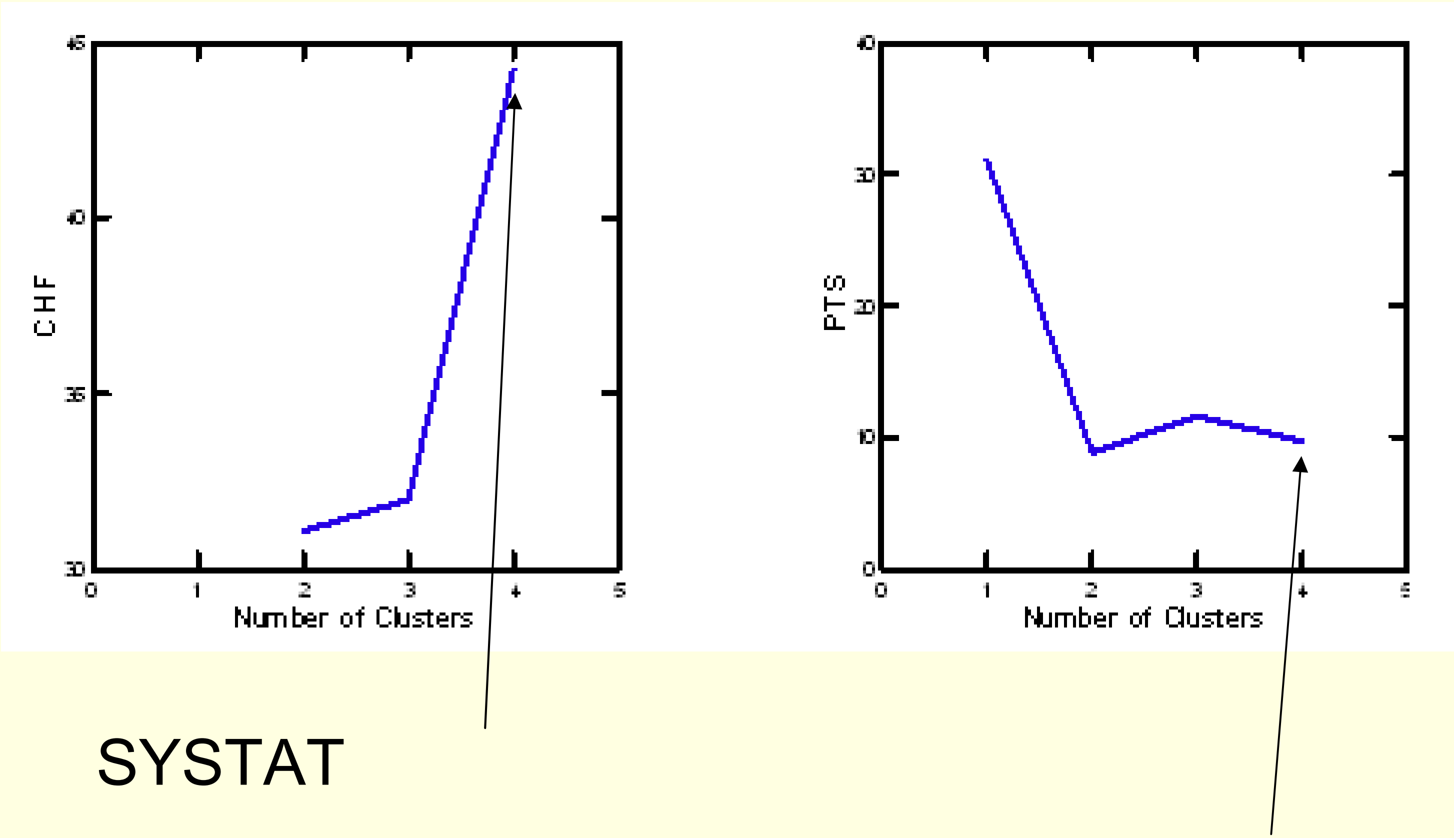
3. *Pseudo (Calinski and Habarasz) F index* – *PSF* (SAS), *CHF* (SYSTAT)

$$I_{\mathrm{CHFU}}(k) = \frac{\dfrac{V_{\mathrm{B}}}{k-1}}{\dfrac{V_{\mathrm{W}}}{n-k}} = \frac{(n-k)\cdot(\xi(1)-\xi(k))}{(k-1)\cdot\xi(k)}$$

4. *Pseudo T-squared statistic* – *PST2* (SAS) *PTS* (SYSTAT)

$$I_{\mathrm{PTSU}}(k) = \frac{\xi_{\langle h,h'\rangle} - (\xi_h + \xi_{h'})}{\dfrac{\xi_h + \xi_{h'}}{n_h + n_{h'} - 2}}$$

Evaluation criteria modified for qualitative variables



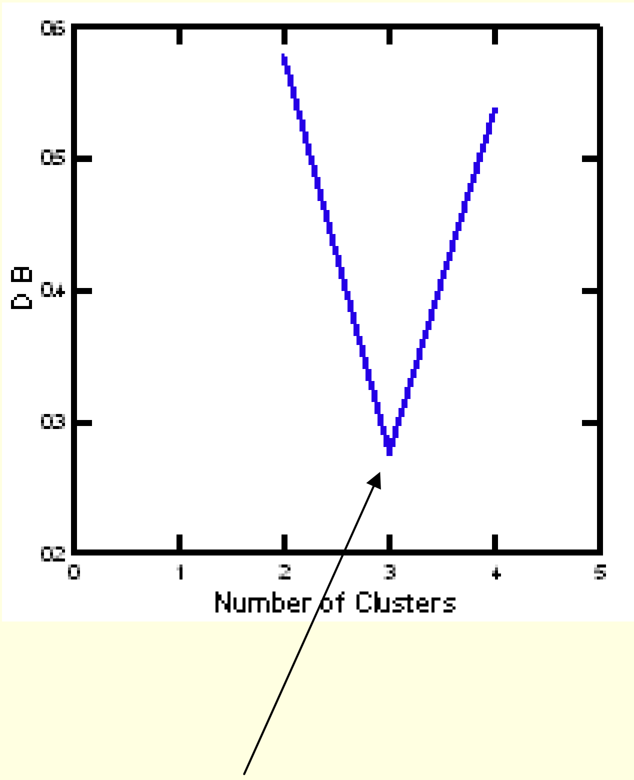
SYSTAT

Evaluation criteria modified for qualitative variables

5. *Modified Davies and Bouldin (DB) index*

$$I_{\mathrm{DB}}(k) = \frac{\sum_{h=1}^{k} \max_{h', h' \neq h} \left\{ \frac{s_{D,h} + s_{D,h'}}{D_{hh'}} \right\}}{k}$$

$$I_{\mathrm{DBU}}(k) = \frac{\sum_{h=1}^{k} \max_{h', h' \neq h} \left\{ \frac{\xi_h + \xi_{h'}}{\xi_{\langle h, h' \rangle} - (\xi_h + \xi_{h'})} \right\}}{k}$$



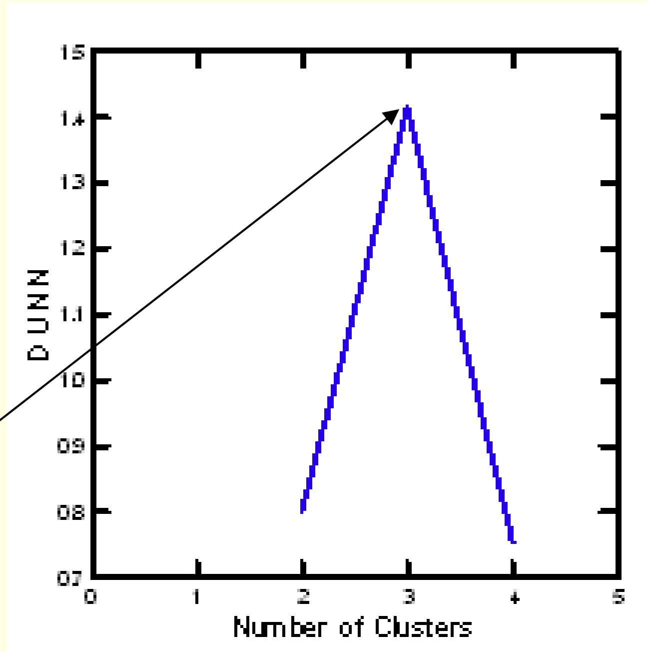
Evaluation criteria modified for qualitative variables

6. Dunn's index

$$I_{\mathrm{D}}(k) = \min_{1 \le h \le k} \left\{ \min_{1 \le h' \le k} \frac{D_{hh'}}{\max\limits_{1 \le g \le k} diam_g} \right\}$$

$$D_{hh'} = \min_{\mathbf{x}_i \in C_h,\, \mathbf{x}_j \in C_{h'}} D(\mathbf{x}_i, \mathbf{x}_j)$$

$$diam_g = \max_{\mathbf{x}_i, \mathbf{x}_j \in C_g} D(\mathbf{x}_i, \mathbf{x}_j)$$



Modified evaluation criteria

- Cluster variability based on the variance and Gini's coefficient of mutability

$$G_g = n_g \left(\sum_{l=1}^{m^{(1)}} \frac{1}{2} \ln(s_l^2 + s_{gl}^2) + \sum_{l=1}^{m^{(2)}} G_{gl} \right)$$

$$G_{gl} = 1 - \sum_{u=1}^{K_l} \left(\frac{n_{glu}}{n_g} \right)^2$$

Gini's coefficient of mutability

$$I_{\mathrm{BGC}} = 2 \sum_{g=1}^{k} G_g + w_k \ln(n)$$

$$G(k) = \sum_{g=1}^{k} G_g$$

Evaluation criteria modified for qualitative variables

1. ***Tau index (RSQ index)***

$$I_{\tau}(k) = \frac{V_{\mathrm{B}}}{V_{\mathrm{T}}} = \frac{V_{\mathrm{T}} - V_{\mathrm{W}}}{V_{\mathrm{T}}} = \frac{G(1) - G(k)}{G(1)}$$

2. ***Semipartial tau index***
   (optimal number of clusters - minimum)

$$I_{\mathrm{SP}\tau}(k) = I_{\tau}(k+1) - I_{\tau}(k)$$

Application to a real data file

- Data from a questionnaire survey (for the participants of the chemistry seminar)
- 7 qualitative and 1 quantitative (count) variables
- Two-step cluster analysis for clustering of respondents (experiments for the numbers of clusters from 2 to 4)
- LC Cluster model (experiments for the numbers of clusters from 2 to 6) – the quantitative variable was recoded to 5 categories

Application to a real data file

Criteria based on the entropy (TSCA in SPSS)

Measure	Number of clusters			
	1	2	3	4
Within-cluster variability	273.92	241.17	206.39	186.51
Variability difference	-	32.75	**34.78**	19.88
I_U	0	0.12	0.25	0.32
I_{SPU}	0.12	0.13	**0.07**	-
I_{CHFU}	0	6.52	**7.69**	7.19
I_{BIC}	590.85	568.41	**541.88**	545.15

Application to a real data file

Criteria based on the Gini's coefficient (TSCA in SPSS)

Measure	Number of clusters			
	1	2	3	4
Within-cluster variability	185.41	162.57	137.83	127.86
Variability difference	-	22.84	**24.74**	9.97
I_{τ}	0	0.12	0.26	0.31
$I_{SP\tau}$	0.12	0.13	**0.05**	-
$I_{CHF\tau}$	0	6.74	**8.11**	6.90
I_{BGC}	413.85	411.20	**404.75**	427.84

Application to a real data file

Comparison of BIC

Method	Number of clusters			
	1	2	3	4
Two-step CA	590.85	568.41	**541.88**	545.15
LC Cluster Model	1397.01	1059.24	**1019.18**	1036.90

Conclusion

- If the distance between objects, distance between clusters, within-cluster variability and the total variability are defined for the case when objects are characterized by mixed-type variables, then the evaluation criteria for quantitative variables can be modified.
- One possibility is an application of log-likelihood distance measure based on the entropy
- Another possibility is to use the analogous measure with using of Gini's coefficient

Thank you for your attention