Application of a Genetic Algorithm to Variable Selection in Fuzzy Clustering

Christian Röver and Gero Szepannek

Fachbereich Statistik
Universität Dortmund
roever@statistik.uni-dortmund.de
gero.szepannek@web.de

March 11, 2004

Overview

1. the problem
2. tackling the problem / methods
3. application to Dortmund data
4. conclusions

The Problem

- given: huge dataset (many **variables**)
  wanted: grouping of observations, clusters

- reduce dimensionality to
  - avoid **overfitting**
  - exclude **noise** and **redundant variables**
  - keep data **perceptible** and **interpretable**

- use **variable subsets** (instead of, e.g., linear combinations) for interpretability

➜ what is the **optimal** subset of variables?

Quality requirements

- needed: comparable quality measure for variable subsets of
  - different **scales** and
  - varying **subset size**
- **restriction**: variable subset should be **representative** of complete data

➜ quality measure?

➜ what makes a variable subset representative?

Quality measure

- focus on **fuzzy clustering**:
  no fixed cluster assignments, but membership scores:

	Cluster		
Observation	1	2	3
1	0.95	0.02	0.03
2	0.50	0.30	0.20
$\vdots$	$\vdots$	$\vdots$	$\vdots$

- compute a measure from **membership matrix** U

- classification entropy:

$$\mathrm{CE}(U) = -\frac{1}{N}\sum_{i=1}^{N}\sum_{j=1}^{k}(u_{ij}\cdot\log_2 u_{ij})$$

- $\mathrm{CE}(U) = 0$ if all $u_{ij} \in \{0, 1\}$ (most **crisp** partitioning)
  $\mathrm{CE}(U)$ greatest if all $u_{ij} = \frac{1}{k}$ (**fuzziest** partitioning)

- **minimize** $\mathrm{CE}(U)$ for 'optimal' subset

- number of clusters (k) was fixed and model-based clustering[1] (fitting of a normal mixture model to data) was applied

[1] Fraley, C. and Raftery, A.E. (2002): `mclust`: Software for model-based clustering, density estimation and discriminant analysis. *Technical Report, Department of Statistics, University of Washington*.
See `http://www.stat.washington.edu/mclust`.

Representativeness

- variable subset should reflect certain **aspects** of data
- define **subgroups** of variables having to appear in a subset
  - **manually** (by meaning) or
  - **systematically**
- systematical selection: groups of **correlated variables**
- motivation: subgroups have a common source of variability;
  by picking from different groups, different sources are covered

- cluster **variables** by their correlation

- define: **distance** between variables:

$$d(X, Y) = 1 - |\mathrm{Cor}(X, Y)|$$

  apply **agglomerative hierarchical clustering**

- **complete linkage**: (absolute) correlation *within* group is **bounded below**

- **single linkage**: correlation *between* groups is **bounded above**

Optimization

- problem: minimize function $f : \mathcal{M} \rightarrow \mathrm{I\!R}$
  where $\mathcal{M}$ has **varying dimension** and further **restrictions**

- use **genetic optimization algorithm**
  (applies principle of *survival of the fittest*):

$$
\begin{array}{rcl}
\text{fitness} & \longleftrightarrow & \text{objective function} \\
\text{genome} & \longleftrightarrow & \text{variable subset} \\
\text{mutation} & \longleftrightarrow & \text{change in subset} \\
\text{recombination} & \longleftrightarrow & \text{combination of 2 subsets} \\
\text{selection (survival)} & \longleftrightarrow & \text{comparison by objective function}
\end{array}
$$

Procedure

given: set of variables

↓

define: subgroups

↓

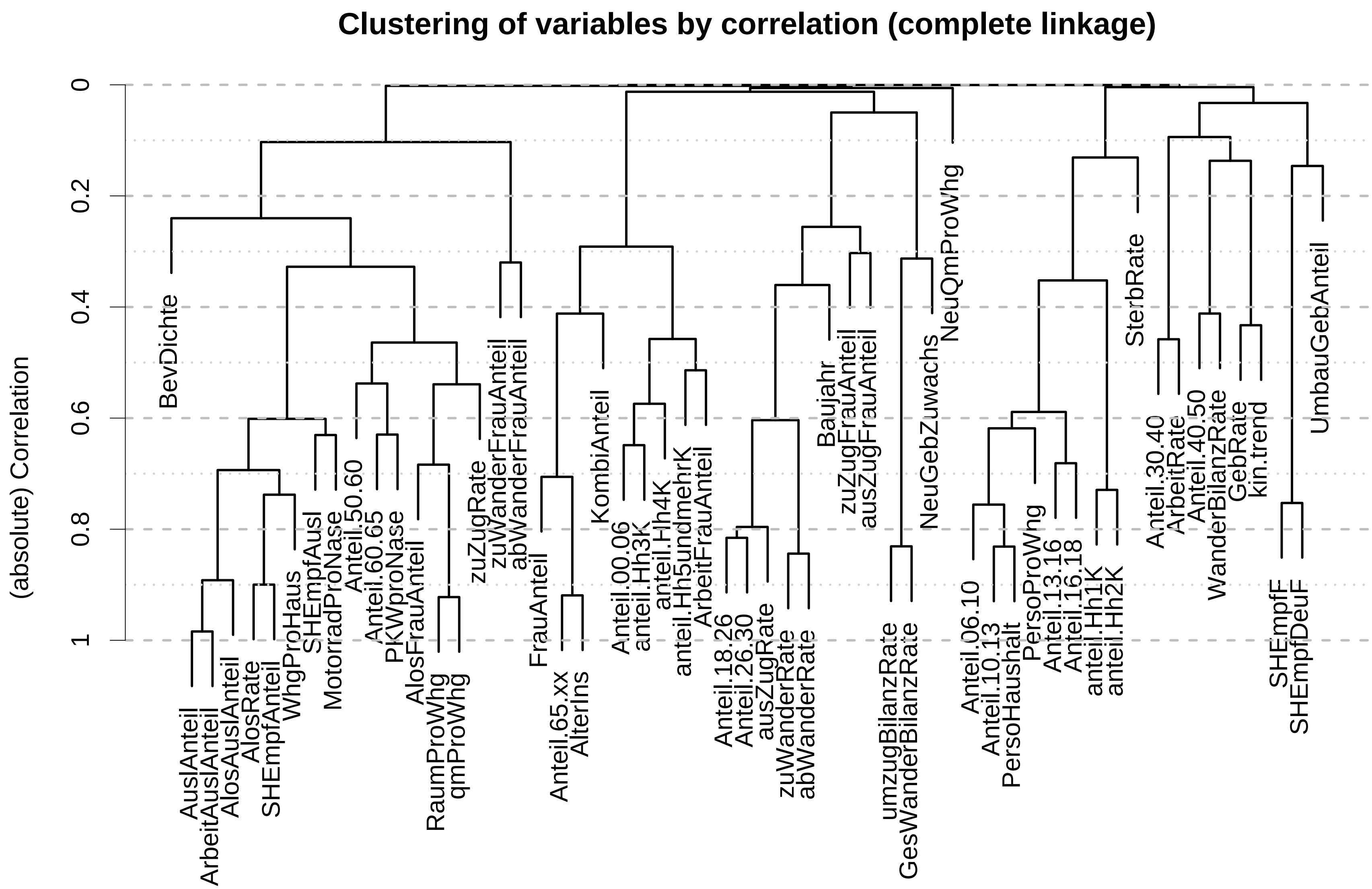
search: optimal composition out of subgroups

↓

return: best subgroup found

Application to Dortmund data

- **raw data**: 200 variables, 170 observations (subdistricts)
  constructed data set of 57 (scaled) variables
- 12 observations were considered **outliers**, e.g. districts containing
  - horse race track
  - steel plant being dismantled
  - university
  - ...
- **systematical selection** of variable subgroups proved to be **impractical**: either huge numbers of variable groups or correlation bounds of insignificant order



Clustering of variables by correlation (complete linkage)
(absolute) Correlation
0
0.2
0.4
0.6
0.8
1
BevDichte
AuslAnteil
ArbeitAuslAnteil
AlosAuslAnteil
AlosRate
SHEmpfAnteil
WhgProHaus
SHEmpfAusl
MotorradProNase
Anteil.50.60
Anteil.60.65
PKWproNase
AlosFrauAnteil
RaumProWhg
qmProWhg
zuZugRate
zuWanderFrauAnteil
abWanderFrauAnteil
FrauAnteil
Anteil.65.xx
AlterIns
KombiAnteil
Anteil.00.06
anteil.Hh3K
anteil.Hh4K
anteil.Hh5undmehrK
ArbeitFrauAnteil
Anteil.18.26
Anteil.26.30
ausZugRate
zuWanderRate
abWanderRate
Baujahr
zuZugFrauAnteil
ausZugFrauAnteil
umzugBilanzRate
GesWanderBilanzRate
NeuGebZuwachs
NeuQmProWhg
Anteil.06.10
Anteil.10.13
PersoHaushalt
PersoProWhg
Anteil.13.16
Anteil.16.18
anteil.Hh1K
anteil.Hh2K
SterbRate
Anteil.30.40
ArbeitRate
Anteil.40.50
WanderBilanzRate
GebRate
kin.trend
SHEmpfF
SHEmpfDeuF
UmbauGebAnteil

- variable groups:

  i. age distribution
  ii. births, deaths, migration
  iii. motoring
  iv. buildings, housing
  v. employment, welfare
  vi. some of above broken down by sex etc.

- final variable subset shall **represent** groups **i**, **ii**, **iv** and **v** and have **at most 6** variables

- data exploration suggests presence of 4 clusters

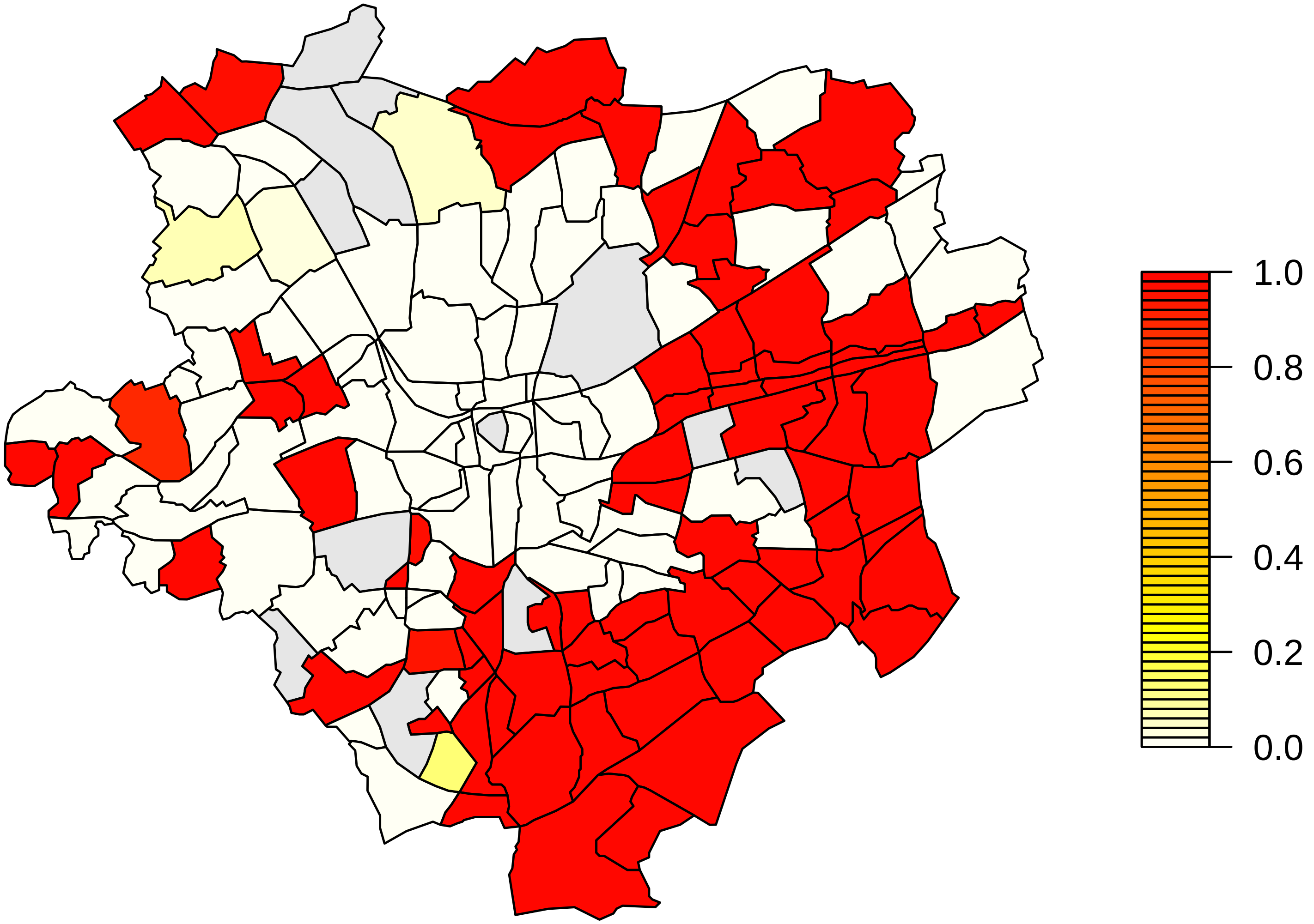
Results

- variable set and cluster means:

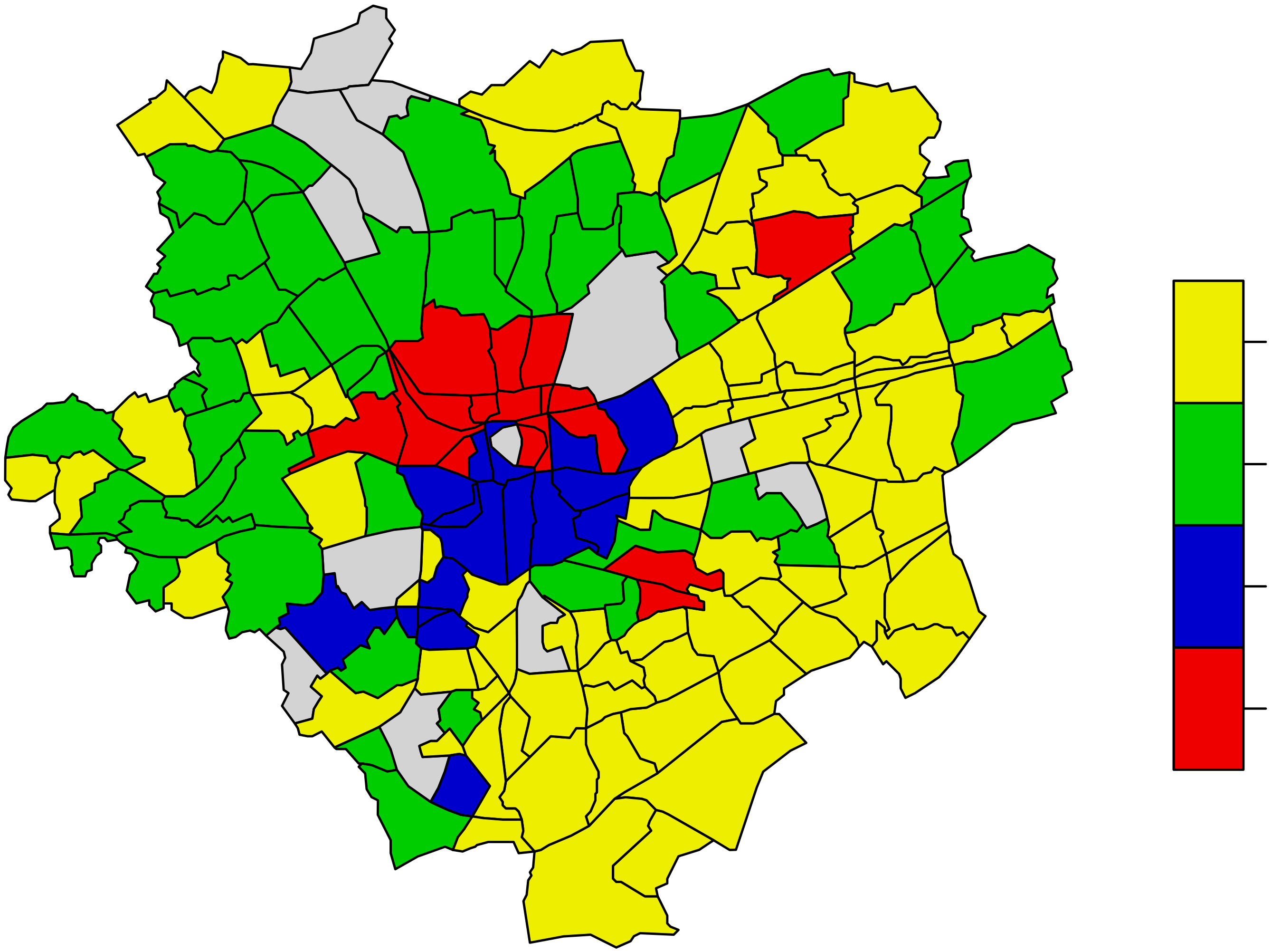
Variable	Group	Cluster 1	2	3	4
fraction of population of age 60–65	i.	**0.057**	0.065	0.064	**0.083**
moves to district per inhabitant	ii.	**0.075**	0.054	0.035	**0.025**
apartments per house	iv.	**7.831**	**5.331**	**3.367**	**2.524**
people per apartment	iv.	1.877	**1.676**	**2.216**	2.029
fraction of welfare recipients	v.	**0.129**	0.031	0.066	**0.023**
fraction of immigrants of employed people	vi.	**0.274**	0.073	0.086	**0.032**

minimum, **maximum**

Fuzzyness (cluster 4)



Spatial distribution of the 4 clusters



- **cluster 1** (*center N*) is most different from **cluster 4** (*suburbs SE*): cluster 1 has
  - few old inhabitants
  - many immigrants
  - many welfare recipients
  - much migration
  - many apartments per house

  while cluster 4 takes opposite extreme values

- **clusters 2** and **3** lie mostly between these extremes and differ by their housing situation: cluster 3 (*suburbs NW*) has
  - less apartments per house
  - most people per apartment

  while cluster 2 (*center S*) has the least people per apartment.

Conclusions

➜ **variable selection** problem was expressed as a **minimization problem** by introducing a quality measure and certain restrictions

➜ an appropriate **optimization algorithm** was utilized to search for an optimal subset

➜ automatical **generation of restrictions** proved to be impractical for Dortmund data

➜ **variable selection** worked well, resulted in an interpretable variable set