# Methods of Network Analysis Clustering and Blockmodeling 

\author{

1. Clustering <br> Vladimir Batagelj <br> University of Ljubljana, Slovenia
}

University of Konstanz, Algorithms and Data Structures
June 3, 2002, 10-12h, room F 426

## Clustering

- basic notions: cluster, clustering, feasible clustering, criterion function, dissimilarities, clustering as an optimization problem
- different (nonstandard) problems: assignment of students to classes, regionalization; general criterion function; multicriteria problems.
- complexity results about the clustering problem - NP-hardness theorems


## Basic notions

Let us start with the formal setting of the clustering problem. We shall use the following notation:

$$
\begin{aligned}
& \mathrm{X} \text { - unit } \\
& X \text { - description of unit } \mathrm{X} \\
& \mathcal{U} \text { - space of units } \\
& \mathbf{U} \text { - finite set of units, } \mathbf{U} \subset \mathcal{U} \\
& C \text { - cluster, } \emptyset \subset C \subseteq \mathbf{U} \\
& \mathbf{C} \text { - clustering, } \mathbf{C}=\left\{C_{i}\right\} \\
& \Phi \text { - set of feasible clusterings } \\
& P \text { - criterion function, } \\
& \quad P: \Phi \rightarrow \mathbb{R}_{0}^{+}
\end{aligned}
$$



## Clustering problem

With these notions we can express the clustering problem $(\Phi, P)$ as follows:
Determine the clustering $\mathbf{C}^{\star} \in \Phi$ for which

$$
P\left(\mathbf{C}^{\star}\right)=\min _{\mathbf{C} \in \Phi} P(\mathbf{C})
$$

Since the set of units $\mathbf{U}$ is finite, the set of feasible clusterings is also finite. Therefore the set $\operatorname{Min}(\Phi, P)$ of all solutions of the problem (optimal clusterings) is not empty. (In theory) the set $\operatorname{Min}(\Phi, P)$ can be determined by the complete search.

We shall denote the value of criterion function for an optimal clustering by $\min (\Phi, P)$.

## Units

real or imaginary objects of analysis
WORLD
UNITS

## DESCRIPTIONS

$$
\{\mathrm{X}\} \quad \longleftrightarrow \mathrm{X} \quad \longleftrightarrow \quad[\mathrm{X}]
$$

formalization
\{ produced cars T \} car T
operationalization

$$
[\text { seats=4, max-speed }=\ldots]
$$

Usually an unit X is represented by a vector/description $X \equiv[\mathrm{X}]=\left[x_{1}, x_{2}, \ldots, x_{m}\right]$ from the set $[\mathcal{U}]$ of all possible descriptions. $x_{i}=V_{i}(\mathrm{X})$ is the value of the $i$-th of selected properties or variables on X. Variables can be measured in different scales: nominal, ordinal, interval, rational, absolute (Roberts, 1976).

There exist other kinds of descriptions of units: symbolic object (Bock, Diday, 2000), list of keywords from a text, chemical formula, vertex in a given graph, digital picture,

## Clusterings

Generally the clusters of clustering $\mathbf{C}=\left\{C_{1}, C_{2}, \ldots, C_{k}\right\}$ need not to be pairwise disjoint; yet, the clustering theory and practice mainly deal with clusterings which are the partitions of $\mathbf{U}$

$$
\begin{gathered}
\bigcup_{i=1}^{k} C_{i}=\mathbf{U} \\
i \neq j \Rightarrow C_{i} \cap C_{j}=\emptyset
\end{gathered}
$$

Each partition determines an equivalence relation in $\mathbf{U}$, and vice versa.
We shall denote the set of all partitions of $\mathbf{U}$ into $k$ classes (clusters) by $P_{k}(\mathbf{U})$.

## Simple criterion functions

Joining the individual units into a cluster $C$ we make a certain "error", we create certain "tension" among them - we denote this quantity by $p(C)$. The criterion function $P(\mathbf{C})$ combines these "partial/local errors" into a "global error".

Usually it takes the form:

$$
\text { S. } \quad P(\mathbf{C})=\sum_{C \in \mathbf{C}} p(C)
$$

or
M. $\quad P(\mathbf{C})=\max _{C \in \mathbf{C}} p(C)$
which can be unified and generalized in the following way:
Let $(\mathbb{R}, \oplus, e, \leq)$ be an ordered abelian monoid then:

$$
\oplus . \quad P(\mathbf{C})=\bigoplus_{C \in \mathbf{C}} p(C)
$$

For simple criterion functions usually $\left.\min \left(P_{k+1} \mathbf{U}\right), P\right) \leq \min \left(P_{k}(\mathbf{U}), P\right)$ - we fix the value of $k$ and set $\Phi \subseteq P_{k}(\mathbf{U})$.

## Cluster-error function / dissimilarities

The cluster-error $p(C)$ has usually the properties:

$$
p(C) \geq 0 \quad \text { and } \quad \forall \mathbf{X} \in \mathbf{U}: p(\{\mathrm{X}\})=0
$$

In the continuation we shall assume that these properties of $p(C)$ hold.
To express the cluster-error $p(C)$ we define on the space of units a dissimilarity $d: \mathcal{U} \times \mathcal{U} \rightarrow \mathbb{R}_{0}^{+}$for which we require D 1 and D 2 :

D1. $\forall \mathrm{X} \in \mathcal{U}: d(\mathrm{X}, \mathrm{X})=0$
D2. symmetric: $\quad \forall \mathrm{X}, \mathrm{Y} \in \mathcal{U}: d(\mathrm{X}, \mathrm{Y})=d(\mathrm{Y}, \mathrm{X})$
Usually the dissimilarity $d$ is defined using another dissimilarity $\delta:[\mathcal{U}] \times[\mathcal{U}] \rightarrow \mathbb{R}_{0}^{+}$ as

$$
d(\mathrm{X}, \mathrm{Y})=\delta([\mathrm{X}],[\mathrm{Y}])
$$

## Properties of dissimilarities

The dissimilarity $d$ is:
D3. even: $\forall \mathrm{X}, \mathrm{Y} \in \mathcal{U}:(d(\mathrm{X}, \mathrm{Y})=0 \Rightarrow \forall \mathrm{Z} \in \mathcal{U}: d(\mathrm{X}, \mathrm{Z})=d(\mathrm{Y}, \mathrm{Z}))$
D4. definite: $\quad \forall \mathrm{X}, \mathrm{Y} \in \mathcal{U}:(d(\mathrm{X}, \mathrm{Y})=0 \Rightarrow \mathrm{X}=\mathrm{Y})$
D5. metric: $\quad \forall \mathrm{X}, \mathrm{Y}, \mathrm{Z} \in \mathcal{U}: d(\mathrm{X}, \mathrm{Y}) \leq d(\mathrm{X}, \mathrm{Z})+d(\mathrm{Z}, \mathrm{Y}) \quad$ - triangle
D6. ultrametric: $\quad \forall \mathrm{X}, \mathrm{Y}, \mathrm{Z} \in \mathcal{U}: d(\mathrm{X}, \mathrm{Y}) \leq \max (d(\mathrm{X}, \mathrm{Z}), d(\mathrm{Z}, \mathrm{Y}))$
D7. additive, iff the Buneman's or four-point condition holds $\forall \mathrm{X}, \mathrm{Y}, \mathrm{U}, \mathrm{V} \in \mathcal{U}$ :

$$
d(\mathrm{X}, \mathrm{Y})+d(\mathrm{U}, \mathrm{~V}) \leq \max (d(\mathrm{X}, \mathrm{U})+d(\mathrm{Y}, \mathrm{~V}), d(\mathrm{X}, \mathrm{~V})+d(\mathrm{Y}, \mathrm{U}))
$$

The dissimilarity $d$ is a distance iff D4, D5 hold.
Since the description []: $\mathbf{U} \rightarrow[\mathbf{U}]$ need not to be injective, $d$ can be indefinite.

Dissimilarities on $\mathbb{R}^{m}$ / examples 1

| n | measure | definition | range | note |
| :---: | :---: | :---: | :---: | :---: |
| 1 | Euclidean | $\sqrt{\sum_{i=1} \sum_{m}^{m}\left(x_{i}-y_{i}\right)^{2}}$ | $[0, \infty)$ | $M(2)$ |
| 2 | Sq. Euclidean | $\sum_{\substack{i=1 \\ m}}^{m}\left(x_{i}-y_{i}\right)^{2}$ | $[0, \infty)$ | $M(2)^{2}$ |
| 3 | Manhattan | $\sum_{i=1}\left\|x_{i}-y_{i}\right\|$ | $[0, \infty)$ | $M(1)$ |
| 4 | rook | $\max _{i=1}^{m}\left\|x_{i}-y_{i}\right\|$ | $[0, \infty)$ | $M(\infty)$ |
| 5 | Minkowski | $\sqrt[p]{ } \sum_{i=1}^{m}\left(x_{i}-y_{i}\right)^{p}$ | $[0, \infty)$ | $M(p)$ |

Dissimilarities on $\mathbb{R}^{m} /$ examples 2

| n | measure | definition | range | note |
| ---: | :--- | :---: | :---: | :---: |
| 6 | Canberra | $\sum_{i=1}^{m} \frac{\left\|x_{i}-y_{i}\right\|}{\left\|x_{i}+y_{i}\right\|}$ | $[0, \infty)$ |  |
| 7 | Heincke | $\sqrt{\sum_{i=1}^{m}\left(\frac{\left\|x_{i}-y_{i}\right\|}{\left\|x_{i}+y_{i}\right\|}\right)^{2}}$ | $[0, \infty)$ |  |
| 8 | Self-balanced | $\sum_{i=1}^{m} \frac{\left\|x_{i}-y_{i}\right\|}{\max \left(x_{i} y_{i}\right)}$ | $[0, \infty)$ |  |
| 9 | Lance-Williams | $\frac{\sum_{i=1}^{m}\left\|x_{i}-y_{i}\right\|}{\sum_{i=1}^{m} x_{i}+y_{i}}$ | $[0, \infty)$ |  |
| 10 | Correlation c. | $\frac{\operatorname{cov}(X, Y)}{\sqrt{\operatorname{var}(X) \operatorname{var}(Y)}}$ | $[1,-1]$ |  |

## (Dis)similarities on $\mathbb{B}^{m} /$ examples

Let $\mathbb{B}=\{0,1\}$. For $X, Y \in \mathbb{B}^{m}$ we define $a=X Y, b=X \bar{Y}, c=\bar{X} Y, d=\overline{X Y}$. It holds $a+b+c+d=m$. The counters $a, b, c, d$ are used to define several (dis)similarity measures on binary vectors.

In some cases the definition can yield an indefinite expression $\frac{0}{0}$. In such cases we can restrict the use of the measure, or define the values also for indefinite cases. For example, we extend the values of Jaccard coefficient such that $s_{4}(X, X)=1$. And for Kulczynski coefficient, we preserve the relation $T=\frac{1}{s_{4}}-1$ by

$$
s_{4}=\left\{\begin{array}{ll}
1 & d=m \\
\frac{a}{a+b+c} & \text { otherwise }
\end{array} \quad s_{3}^{-1}=T= \begin{cases}0 & a=0, d=m \\
\infty & a=0, d<m \\
\frac{b+c}{a} & \text { otherwise }\end{cases}\right.
$$

We transform a similarity $s$ from $[1,0]$ into dissimilarity $d$ on $[0,1]$ by $d=1-s$.
For details see Batagelj, Bren (1995).

## (Dis)similarities on $\mathbb{B}^{m}$ / examples 1

| n | measure | definition | range |
| :---: | :--- | :---: | :---: |
| 1 | Russel and Rao (1940) | $\frac{a}{m}$ | $[1,0]$ |
| 2 | Kendall, Sokal-Michener (1958) | $\frac{a+d}{m}$ | $[1,0]$ |
| 3 | Kulczynski (1927), $T^{-1}$ | $\frac{a}{b+c}$ | $[\infty, 0]$ |
| 4 | Jaccard (1908) | $\frac{a}{a+b+c}$ | $[1,0]$ |
| 5 | Kulczynski | $\frac{1}{2}\left(\frac{a}{a+b}+\frac{a}{a+c}\right)$ | $[1,0]$ |
| 6 | Sokal \& Sneath (1963), un | $\frac{1}{4}\left(\frac{a}{a+b}+\frac{a}{a+c}+\frac{d}{d+b}+\frac{d}{d+c}\right)$ | $[1,0]$ |
| 7 | Driver \& Kroeber (1932) | $\frac{a}{\sqrt{(a+b)(a+c)}}$ | $[1,0]$ |
| 8 | Sokal \& Sneath (1963), un5 | $\frac{a d}{\sqrt{(a+b)(a+c)(d+b)(d+c)}}$ | $[1,0]$ |

(Dis)similarities on $\mathbb{B}^{m}$ / examples 2

| n | measure | definition | range |
| ---: | :--- | :---: | :---: |
| 9 | $Q_{0}$ | $\frac{b c}{a d}$ | $[0, \infty]$ |
| 10 | Yule (1927), $Q$ | $\frac{a d-b c}{a d+b c}$ | $[1,-1]$ |
| 11 | Pearson, $\phi$ | $\frac{a d-b c}{\sqrt{(a+b)(a+c)(d+b)(d+c)}}$ | $[1,-1]$ |
| 12 | $-\mathrm{bc}-$ | $\frac{4 b c}{m^{2}}$ | $[0,1]$ |
| 13 | Baroni-Urbani, Buser (1976), $S^{* *}$ | $\frac{a+\sqrt{a d}}{a+b+c+\sqrt{a d}}$ | $[1,0]$ |
| 14 | Braun-Blanquet (1932) | $\frac{a}{\max (a+b, a+c)}$ | $[1,0]$ |
| 15 | Simpson (1943) | $\frac{a}{\min (a+b, a+c)}$ | $[1,0]$ |
| 16 | Michael (1920) | $\frac{4(a d-b c)}{(a+d)^{2}+(b+c)^{2}}$ | $[1,-1]$ |

## Dissimilarities between sets

Let $\mathcal{F}$ be a finite family of subsets of the finite set $U ; A, B \in \mathcal{F}$ and let $A \oplus B=$ $(A \backslash B) \cup(B \backslash A)$ denotes the symmetric difference between $A$ and $B$.

The 'standard' dissimilarity between sets is the Hamming distance:

$$
d_{H}(A, B):=\operatorname{card}(A \oplus B)
$$

Usually we normalize it $d_{h}(A, B)=\frac{1}{M} \operatorname{card}(A \oplus B)$. One normalization is $M=\operatorname{card}(U)$; the other $M=m_{1}+m_{2}$, where $m_{1}$ and $m_{2}$ are the first and the second largest value in $\{\operatorname{card}(X): X \in \mathcal{F}\}$.

Other dissimilarities

$$
\begin{gathered}
d_{s}(A, B)=\frac{\operatorname{card}(A \oplus B)}{\operatorname{card}(A)+\operatorname{card}(B)} \quad d_{u}(A, B)=\frac{\operatorname{card}(A \oplus B)}{\operatorname{card}(A \cup B)} \\
d_{m}(A, B)=\frac{\max (\operatorname{card}(A \backslash B), \operatorname{card}(B \backslash A))}{\max (\operatorname{card}(A), \operatorname{card}(B))}
\end{gathered}
$$

For all these dissimilarities $d(A, B)=0$ if $A=B=\emptyset$.

## Problems with dissimilarities

What to do in the case of mixed units (with variables measured in different types of scales)?

- conversion to a common scale
- compute the dissimilarities on homogeneous parts and combine them (Gower's dissimilarity)

Fairness of dissimilarity - all variables contribute equally. Approaches: use of normalized variables, analysis of dependencies among variables.

## Cluster-error function / examples

Now we can define several cluster-error functions:

$$
\begin{array}{ll}
\text { S. } & p(C)=\sum_{\mathrm{X}, \mathrm{Y} \in C, \mathrm{X}<\mathrm{Y}} w(\mathrm{X}) \cdot w(\mathrm{Y}) \cdot d(\mathrm{X}, \mathrm{Y}) \\
\overline{\mathrm{S}} . & p(C)=\frac{1}{w(C)} \sum_{\mathrm{X}, \mathrm{Y} \in C, \mathrm{X}<\mathrm{Y}} w(\mathrm{X}) \cdot w(\mathrm{Y}) \cdot d(\mathrm{X}, \mathrm{Y})
\end{array}
$$

where $w: \mathcal{U} \rightarrow \mathbb{R}^{+}$is a weight of units, which is extended to clusters by:

$$
\begin{gathered}
w(\{\mathrm{X}\})=w(\mathrm{X}), \quad \mathrm{X} \in \mathcal{U} \\
w\left(C_{1} \cup C_{2}\right)=w\left(C_{1}\right)+w\left(C_{2}\right), \quad C_{1} \cap C_{2}=\emptyset
\end{gathered}
$$

Often $w(\mathrm{X})=1$ holds for each $\mathrm{X} \in \mathcal{U}$. Then $w(C)=\operatorname{card}(C)$.
M. $\quad p(C)=\max _{\mathrm{X}, \mathrm{Y} \in C} d(\mathrm{X}, \mathrm{Y})=\operatorname{diam}(C) \quad$ - diameter
T. $\quad p(C)=\min _{T \text { is a spaning tree over } C} \sum_{(\mathrm{X}: \mathrm{Y}) \in T} d(\mathrm{X}, \mathrm{Y})$

We shall use the labels in front of the forms of (partial) criterion functions to denote types of criterion functions. For example:

$$
\text { SM. } \quad P(\mathbf{C})=\sum_{C \in \mathbf{C}} \max _{\mathrm{X}, \mathrm{Y} \in C} d(\mathrm{X}, \mathrm{Y})
$$

It is easy to prove:
Proposition 1.1 Let $P \in\{\mathrm{SS}, \mathrm{S} \overline{\mathrm{S}}, \mathrm{SM}, \mathrm{MS}, \mathrm{M} \overline{\mathrm{S}}, \mathrm{MM}\}$ then there exists an $\alpha_{k}^{P}(\mathbf{U})>0$ such that for each $\mathbf{C} \in P_{k}(\mathbf{U})$ :

$$
P(\mathbf{C}) \geq \alpha_{k}^{P}(\mathbf{U}) \cdot \max _{C \in \mathbf{C}} \max _{\mathrm{X}, \mathrm{Y} \in C} d(\mathrm{X}, \mathrm{Y})
$$

holds.
Note that this inequality can be writen also as $P(\mathbf{C}) \geq \alpha_{k}^{P}(\mathbf{U}) \cdot \operatorname{MM}(\mathbf{C})$.

## Sensitive criterion functions

The criterion function $P(\mathbf{C})$, based on the dissimilarity $d$, is sensitive iff for each feasible clustering $\mathbf{C}$ it holds

$$
P(\mathbf{C})=0 \Longleftrightarrow \forall C \in \mathbf{C} \forall \mathrm{X}, \mathrm{Y} \in C: d(\mathrm{X}, \mathrm{Y})=0
$$

and is $\alpha$-sensitive iff there exists an $\alpha_{k}^{P}(\mathbf{U})>0$ such that for each $\mathbf{C} \in P_{k}(\mathbf{U})$ :

$$
P(\mathbf{C}) \geq \alpha_{k}^{P}(\mathbf{U}) \cdot \operatorname{MM}(\mathbf{C})
$$

Proposition 1.2 Every $\alpha$-sensitive criterion function is also sensitive.
The proposition 1.1 can be reexpressed as:
Proposition 1.3 The criterion functions $\mathrm{SS}, \mathrm{S} \overline{\mathrm{S}}, \mathrm{SM}, \mathrm{MS}, \mathrm{M} \overline{\mathrm{S}}, \mathrm{MM}$ are $\alpha$-sensitive.

## Representatives

Another form of cluster-error function, which is frequently used in practice, is based on the notion of leader or representative of the cluster:

$$
\text { R. } \quad p(C)=\min _{\mathrm{L} \in \mathbf{F}} \sum_{\mathrm{X} \in C} w(\mathrm{X}) \cdot d(\mathrm{X}, \mathrm{~L})
$$

where $\mathbf{F} \subseteq \mathcal{F}$ is the set of representatives. The element $\bar{C} \in \mathbf{F}$, which minimizes the right side expression, is called the representative of cluster $C$. It is not always uniquely determined.

Example 1 The representation space need not be the same as the description space.

$$
[\mathbf{U}] \subseteq \mathbb{R}^{2} \text { and }[\mathbf{F}]=\left\{(a, b, c): a x+b y=c, a^{2}+b^{2}=1\right\} .
$$

Example 2 In the case $[\mathbf{U}] \subseteq \mathbb{R}^{m}, \quad[\mathbf{F}]=\mathbb{R}^{m}, d(\mathrm{X}, \mathrm{L})=d_{2}^{2}(\mathrm{X}, \mathrm{L})=$ $\sum_{i=1}^{m}\left(x_{i}-l_{i}\right)^{2}$ there exists a uniquely determined representative - center of gravity $\bar{C}=\frac{1}{\operatorname{card}(C)} \sum_{\mathrm{X} \in C} X$. In this case the criterion function SR is called Ward's criterion function (Ward, 1963).

## The generalized Ward's criterion function

To obtain the generalized Ward's clustering problem we, relying on the equality

$$
p(C)=\sum_{X \in C} d_{2}^{2}(X, \bar{C})=\frac{1}{2 \operatorname{card}(C)} \sum_{X, Y \in C} d_{2}^{2}(X, Y)
$$

replace the expression for $p(C)$ with

$$
p(C)=\frac{1}{2 w(C)} \sum_{X, Y \in C} w(X) \cdot w(Y) \cdot d(X, Y)=\overline{\mathrm{S}}(C)
$$

Note that $d$ can be any dissimilarity on $\mathcal{U}$.
From the definition we can easily derive the following equality: If $C_{u} \cap C_{v}=\emptyset$ then

$$
w\left(C_{u} \cup C_{v}\right) \cdot p\left(C_{u} \cup C_{v}\right)=w\left(C_{u}\right) \cdot p\left(C_{u}\right)+w\left(C_{v}\right) \cdot p\left(C_{v}\right)+\sum_{X \in C_{u}, Y \in C_{v}} w(X) \cdot w(Y) \cdot d(X, Y)
$$

In Batagelj (1988) it is also shown how to replace $\bar{C}$ by a generalized, possibly imaginary (with descriptions not neccessary in the same set as $\mathcal{U}$ ), central element in the way to preserve the properties characteristic for Ward's clustering problem.

## Representatives cluster error

Proposition 1.4 Let $p(C)$ be of type R then
a) $p(C)+w(\mathrm{X}) \cdot d(\mathrm{X}, \overline{C \cup \mathrm{X}}) \leq p(C \cup \mathrm{X}), \mathrm{X} \notin C$
b) $p(C \backslash \mathrm{X})+w(\mathrm{X}) \cdot d(\mathrm{X}, \bar{C}) \leq p(C), \quad \mathrm{X} \in C$

Proof: The definition of $\bar{C}$ can be equivalently expressed in the form:

$$
\forall \mathrm{L} \in F: p(C)=\sum_{\mathrm{Y} \in C} w(\mathrm{Y}) \cdot d(\mathrm{Y}, \bar{C}) \leq \sum_{\mathrm{Y} \in C} w(\mathrm{Y}) \cdot d(\mathrm{Y}, \mathrm{~L})
$$

Therefore in case a):

$$
\begin{array}{r}
p(C)=\sum_{\mathrm{Y} \in C} w(\mathrm{Y}) \cdot d(\mathrm{Y}, \bar{C}) \leq \sum_{\mathrm{Y} \in C} w(\mathrm{Y}) \cdot d(\mathrm{Y}, \overline{C \cup \mathrm{X}})= \\
=\sum_{\mathrm{Y} \in C \cup \mathrm{X}} w(\mathrm{Y}) \cdot d(\mathrm{Y}, \overline{C \cup \mathrm{X}})-w(\mathrm{X}) \cdot d(\mathrm{X}, \overline{C \cup \mathrm{X}})= \\
=p(C \cup \mathrm{X})-w(\mathrm{X}) \cdot d(\mathrm{X}, \overline{C \cup \mathrm{X}})
\end{array}
$$

In the similar way we can prove also inequality b).

## Other criterion functions

Several other types of criterion functions were proposed in the literature. A very important class among them are the "statistical" criterion functions based on the assumption that the units are sampled from a mixture of multivariate normal distributions (Marriott, 1982) .

## General criterion function

Not all clustering problems can be expressed by a simple criterion function. In some applications a general criterion function of the form

$$
P(\mathbf{C})=\bigoplus q\left(C_{1}, C_{2}\right), \quad q\left(C_{1}, C_{2}\right) \geq 0
$$

is needed. We shall use it in blockmodeling.

## Multicriteria clustering

In some problems several criterion functions can be defined $\left(\Phi, P_{1}, P_{2}, \ldots, P_{s}\right)$. See Ferligoj, Batagelj (1994).

## Example: problem of partitioning of a generation of pupils into a given number of classes

so that the classes will consist of (almost) the same number of pupils and that they will have a structure as similar as possible. An appropriate criterion function is

$$
P(\mathbf{C})=\max _{\substack{\left\{C_{1}, C_{2}\right\} \in \mathbf{C} \times \mathbf{C} \\ \operatorname{card}\left(C_{1}\right) \geq \operatorname{card}\left(C_{2}\right)}} \min _{\substack{f: C_{1} \rightarrow C_{2} \\ f \text { is surjective }}} \max _{\mathrm{X} \in C_{1}} d(\mathrm{X}, f(\mathrm{X}))
$$

where $d(\mathrm{X}, \mathrm{Y})$ is a measure of dissimilarity between pupils X and Y .


## Example: Regionalization

The motivation comes from regionalization problem: partition given set of territorial units into $k$ connected subgroups of similar units - regions.

Suppose that besides the descriptions of units $[\mathbf{U}]$ they are related also by a binary relation $R \subseteq \mathbf{U} \times \mathbf{U}$.

In such a case we have an additional requirement - relational constraint on clusterings to be feasible. The set of feasible clusterings can be defined as:

$$
\begin{aligned}
\boldsymbol{\Phi}(R)= & \{\mathbf{C} \in P(\mathbf{U}): \text { each cluster } C \in \mathbf{C} \text { is a subgraph }(C, R \cap C \times C) \text { in the } \\
& \operatorname{graph}(\mathbf{U}, R) \text { with the required type of connectedness }\}
\end{aligned}
$$

If $R$ is nonsymmetric we can define different types of sets of feasible clusterings for the same relation (Ferligoj and Batagelj, 1983).

## Complexity of the clustering problem

Because the set of feasible clusterings $\Phi$ is finite the clustering problem $(\Phi, P)$ can be solved by the brute force approach inspecting all feasible clusterings. Unfortunately, the number of feasible clusterings grows very quickly with $n$. For example

$$
\operatorname{card}\left(P_{k}\right)=S(n, k)=\frac{1}{k!} \sum_{i=0}^{k-1}(-1)^{i}\binom{k}{i}(k-i)^{n}, \quad 0<k \leq n
$$

where $S(n, k)$ is a Stirling number of the second kind. And to get an impression:

$$
\begin{aligned}
S(20,8) & =15170932662679 \\
S(30,11) & =215047101560666876619690 \\
S(n, 2) & =2^{n-1}-1
\end{aligned}
$$

For this reason the brute force algorithm is only of theoretical interest.
We shall assume that the reader is familiar with the basic notions of the theory of complexity of algorithms (Garey and Johnson, 1979) .

## Complexity results

Although there are some polynomial types of clustering problems, for example $\left(P_{2}, \mathrm{MM}\right)$ and $\left(P_{k}, \mathrm{ST}\right)$, it seems that they are mainly NP-hard.

Brücker (1978) showed that ( $\propto$ denotes the polynomial reducibility of problems) :
Theorem 1.5 Let the criterion function

$$
P(\mathbf{C})=\bigoplus_{C \in \mathbf{C}} p(C)
$$

be $\alpha$-sensitive, then for each problem $\left(P_{k}(\mathbf{U}), P\right)$ there exists a problem $\left(P_{k+1}\left(\mathbf{U}^{\prime}\right), P\right)$, such that $\left(P_{k}(\mathbf{U}), P\right) \propto\left(P_{k+1}\left(\mathbf{U}^{\prime}\right), P\right)$.

Proof: Select a value $P^{*}$ such that $P^{*}>\max _{\mathbf{C} \in P_{k}(\mathbf{U})} P(\mathbf{C})$, extend, $\mathbf{U}^{\prime}=\mathbf{U} \cup\left\{\mathrm{X}^{\bullet}\right\}$, the set of units with a new unit $\mathrm{X}^{\bullet}$, and define the dissimilarities between it and the 'old' units such that $d\left(\mathrm{X}, \mathrm{X}^{\bullet}\right)>P^{*} / \alpha^{\prime}$, for $\mathrm{X} \in \mathbf{U}$ and $\alpha^{\prime}=\alpha_{k+1}^{P}\left(\mathbf{U}^{\prime}\right)$. We get a new clustering problem $\left(P_{k+1}\left(\mathbf{U}^{\prime}\right), P\right)$.
Consider a clustering $\mathbf{C}^{\prime} \in P_{k+1}\left(\mathbf{U}^{\prime}\right)$. There are two possibilities:
a. $\mathrm{X}^{\bullet}$ forms its own cluster $\mathbf{C}^{\prime}=\mathbf{C} \cup\left\{\left\{\mathrm{X}^{\bullet}\right\}\right\}, \mathbf{C} \in P_{k}(\mathbf{U})$. Then

$$
P\left(\mathbf{C}^{\prime}\right)=P(\mathbf{C}) \oplus p\left(\left\{\mathrm{X}^{\bullet}\right\}\right)=P(\mathbf{C}) \leq \max _{\mathbf{C} \in P_{k}(\mathbf{U})} P(\mathbf{C})<P^{*}
$$

b. $\mathrm{X}^{\bullet}$ belongs to a cluster $C^{\bullet}$ with $\operatorname{card}\left(C^{\bullet}\right) \geq 2$. Then

$$
\begin{gathered}
P\left(\mathbf{C}^{\prime}\right) \geq \alpha^{\prime} \cdot \max _{C \in \mathbf{C}^{\prime}} \max _{\mathrm{X}, \mathrm{Y} \in C} d(\mathrm{X}, \mathrm{Y}) \geq \alpha^{\prime} \cdot \max _{\mathrm{X}, \mathrm{Y} \in C} d(\mathrm{X}, \mathrm{Y})= \\
=\alpha^{\prime} \cdot \max _{\mathrm{X} \in C^{\bullet} \backslash\{\mathrm{X} \bullet\}} d\left(\mathrm{X}, \mathrm{X}^{\bullet}\right)>P^{*}
\end{gathered}
$$

We see that all optimal solutions of the problem $\left(P_{k+1}\left(\mathbf{U}^{\prime}\right), P\right)$ have the form $\mathbf{a}$. Since in this case $P\left(\mathbf{C}^{\prime}\right)=P(\mathbf{C})$

$$
\mathbf{C}^{\prime} \in \operatorname{Min}\left(P_{k+1}\left(\mathbf{U}^{\prime}\right), P\right) \Leftrightarrow \mathbf{C} \in \operatorname{Min}\left(P_{k}(\mathbf{U}), P\right)
$$

## Complexity results 1

Theorem 1.6 Let the criterion function $P$ be sensitive then

$$
3-\mathrm{COLOR} \propto\left(P_{3}, P\right)
$$

Proof: Let $G=(V, E)$ be a simple undirected graph. We assign to it a clustering problem $\left(P_{3}(V), P\right)$ as follows. We define a dissimilarity $d$ (on which $P$ is based) by

$$
d(u, v)= \begin{cases}1 & (u: v) \in E \\ 0 & (u: v) \notin E\end{cases}
$$

Since $P$ is sensitive it holds: the graph $G$ is 3-colorable iff $\min \left(P_{3}(V), P\right)=0$.
Let $\mathbf{C}=\left\{C_{1}, C_{2}, C_{3}\right\}$ then $P(\mathbf{C})=0$ iff $c: V \rightarrow\{1,2,3\}:\left(c(v)=i \Leftrightarrow v \in C_{i}\right)$ is a vertex coloring.


Complexity results 2

| Polynomial | NP-hard | note |
| :--- | :--- | :--- |
| $\left(P_{2}, \mathrm{MM}\right)$ | $\left(P_{3}, \mathrm{MM}\right)$ | Theorem 1.6 |
|  | $\left(P_{3}, \mathrm{SM}\right)$ | Theorem 1.6 |
|  | $\left(P_{2}, \mathrm{SS}\right)$ | MAX-CUT $\propto\left(P_{2}, \mathrm{SS}\right)$ |
|  | $\left(P_{2}, \mathrm{~S} \overline{\mathrm{~S}}\right)$ | $\left(P_{2}, \mathrm{SS}\right) \propto\left(P_{2}, \mathrm{~S} \overline{\mathrm{~S}}\right)$ |
|  | $\left(P_{2}, \mathrm{MS}\right)$ | PARTITION $\propto\left(P_{2}, \mathrm{MS}\right)$ |
| $\left(\mathbb{R}_{2}^{m}, \mathrm{~S} \overline{\mathrm{~S}}\right)$ |  |  |
| $\left(\mathbb{R}_{k}^{1}, \mathrm{~S} \overline{\mathrm{~S}}\right)$ |  |  |
| $\left(\mathbb{R}_{k}^{1}, \mathrm{SM}\right)$ |  |  |
| $\left(\mathbb{R}_{k}^{1}, \mathrm{MM}\right)$ |  |  |

Note that, by the Theorem $1.5,\left(P_{k}, \mathrm{MM}\right), k>3$ are also NP-hard $\ldots$

## Consequences

From these results it follows (it is believed) that no efficient (polynomial) exact algorithm exists for solving the clustering problem.

Therefore the procedures should be used which give "good" results, but not necessarily the best, in a reasonable time.

The most important types of these procedures are:

- local optimization
- hierarchical (agglomerative, divisive and adding)
- leaders and the dynamic clusters method
- graph theory methods

