Methods of Network Analysis

Clustering and Blockmodeling

1. Clustering

Vladimir Batagelj
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:

X – *unit*

X – *description* of unit X

$\mathcal{U}$ – *space* of units

$\mathbf{U}$ – finite *set of units*, $\mathbf{U} \subset \mathcal{U}$

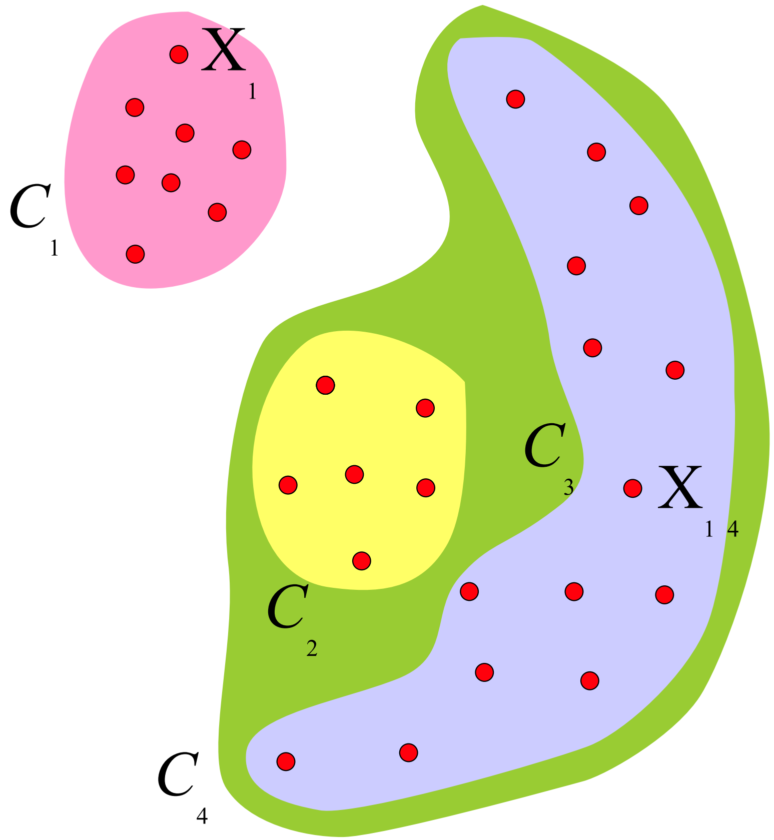
C – *cluster*, $\emptyset \subset C \subseteq \mathbf{U}$

$\mathbf{C}$ – *clustering*, $\mathbf{C} = \{C_i\}$

Φ – set of *feasible clusterings*

P – *criterion function*,

$$P : \Phi \to \mathbb{R}_0^+$$



Clustering problem

With these notions we can express the *clustering problem* (Φ, P) as follows:

Determine the clustering $\mathbf{C}^\star \in \Phi$ for which

$$P(\mathbf{C}^\star) = \min_{\mathbf{C}\in\Phi} P(\mathbf{C})$$

Since the set of units $\mathbf{U}$ is finite, the set of feasible clusterings is also finite. Therefore the set $\mathrm{Min}(\Phi, P)$ of all solutions of the problem (optimal clusterings) is not empty. (In theory) the set $\mathrm{Min}(\Phi, P)$ 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
{X}	$\longleftrightarrow$	X	$\longleftrightarrow$	[X]
	formalization		operationalization	
{ produced cars T }		car T		[seats=4, max-speed= ...]

Usually an unit X is represented by a vector/description $X \equiv [\mathrm{X}] = [x_1, x_2, ..., x_m]$ 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 *scale*s: 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 $\mathbf{C} = \{C_1, C_2, \ldots, C_k\}$ need not to be pairwise disjoint; yet, the clustering theory and practice mainly deal with clusterings which are the *partition*s of $\mathbf{U}$

$$\bigcup_{i=1}^{k} C_i = \mathbf{U}$$

$$i \neq j \Rightarrow C_i \cap C_j = \emptyset$$

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:

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

or

M. $$P(\mathbf{C}) = \max_{C \in \mathbf{C}} p(C)$$

which can be unified and generalized in the following way:

Let $(\mathbb{R}, \oplus, e, \leq)$ be an ordered abelian monoid then:

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

For simple criterion functions usually $\min(P_{k+1}\mathbf{U}), P) \leq \min(P_k(\mathbf{U}), P)$ — 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:

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

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* $d : \mathcal{U} \times \mathcal{U} \to \mathbb{R}_0^+$ for which we require D1 and D2:

D1. $\forall \mathrm{X} \in \mathcal{U} : d(\mathrm{X}, \mathrm{X}) = 0$

D2. *symmetric*: $\forall \mathrm{X}, \mathrm{Y} \in \mathcal{U} : d(\mathrm{X}, \mathrm{Y}) = d(\mathrm{Y}, \mathrm{X})$

Usually the dissimilarity d is defined using another dissimilarity $\delta : [\mathcal{U}] \times [\mathcal{U}] \to \mathbb{R}_0^+$ 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*: $\forall \mathrm{X}, \mathrm{Y} \in \mathcal{U} : (d(\mathrm{X}, \mathrm{Y}) = 0 \Rightarrow \mathrm{X} = \mathrm{Y})$

D5. *metric*: $\forall \mathrm{X}, \mathrm{Y}, \mathrm{Z} \in \mathcal{U} : d(\mathrm{X}, \mathrm{Y}) \leq d(\mathrm{X}, \mathrm{Z}) + d(\mathrm{Z}, \mathrm{Y})$ – triangle

D6. *ultrametric*: $\forall \mathrm{X}, \mathrm{Y}, \mathrm{Z} \in \mathcal{U} : d(\mathrm{X}, \mathrm{Y}) \leq \max(d(\mathrm{X}, \mathrm{Z}), d(\mathrm{Z}, \mathrm{Y}))$

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 $[\,] : \mathbf{U} \to [\mathbf{U}]$ need not to be injective, d can be indefinite.

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

n	measure	definition	range	note
1	Euclidean	$\sqrt{\sum_{i=1}^{m}(x_i - y_i)^2}$	$[0, \infty)$	$M(2)$
2	Sq. Euclidean	$\sum_{i=1}^{m}(x_i - y_i)^2$	$[0, \infty)$	$M(2)^2$
3	Manhattan	$\sum_{i=1}^{m}\lvert x_i - y_i\rvert$	$[0, \infty)$	$M(1)$
4	rook	$\max_{i=1}^{m}\lvert x_i - y_i\rvert$	$[0, \infty)$	$M(\infty)$
5	Minkowski	$\sqrt[p]{\sum_{i=1}^{m}(x_i - y_i)^p}$	$[0, \infty)$	$M(p)$

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

n	measure	definition	range	note
6	Canberra	$\sum_{i=1}^{m} \frac{\lvert x_i - y_i \rvert}{\lvert x_i + y_i \rvert}$	$[0, \infty)$	
7	Heincke	$\sqrt{\sum_{i=1}^{m} (\frac{\lvert x_i - y_i \rvert}{\lvert x_i + y_i \rvert})^2}$	$[0, \infty)$	
8	Self-balanced	$\sum_{i=1}^{m} \frac{\lvert x_i - y_i \rvert}{\max(x_i, y_i)}$	$[0, \infty)$	
9	Lance-Williams	$\frac{\sum_{i=1}^{m} \lvert x_i - y_i \rvert}{\sum_{i=1}^{m} x_i + y_i}$	$[0, \infty)$	
10	Correlation c.	$\frac{\mathrm{cov}(X, Y)}{\sqrt{\mathrm{var}(X)\mathrm{var}(Y)}}$	$[1, -1]$	

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

Let $\mathbb{B} = \{0, 1\}$. For $X, Y \in \mathbb{B}^m$ we define $a = XY$, $b = X\overline{Y}$, $c = \overline{X}Y$, $d = \overline{XY}$. 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

$$s_4 = \begin{cases} 1 & d = m \\ \frac{a}{a+b+c} & \text{otherwise} \end{cases} \qquad s_3^{-1} = T = \begin{cases} 0 & a = 0,\ d = m \\ \infty & a = 0,\ d < m \\ \frac{b+c}{a} & \text{otherwise} \end{cases}$$

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	$\frac{1}{2}(\frac{a}{a+b} + \frac{a}{a+c})$	$[1, 0]$
6	Sokal & Sneath (1963), un_4	$\frac{1}{4}(\frac{a}{a+b} + \frac{a}{a+c} + \frac{d}{d+b} + \frac{d}{d+c})$	$[1, 0]$
7	Driver & Kroeber (1932)	$\frac{a}{\sqrt{(a+b)(a+c)}}$	$[1, 0]$
8	Sokal & Sneath (1963), un_5	$\frac{ad}{\sqrt{(a+b)(a+c)(d+b)(d+c)}}$	$[1, 0]$

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

n	measure	definition	range
9	Q_0	$\frac{bc}{ad}$	$[0, \infty]$
10	Yule (1927), Q	$\frac{ad-bc}{ad+bc}$	$[1, -1]$
11	Pearson, ϕ	$\frac{ad-bc}{\sqrt{(a+b)(a+c)(d+b)(d+c)}}$	$[1, -1]$
12	– bc –	$\frac{4bc}{m^2}$	$[0, 1]$
13	Baroni-Urbani, Buser (1976), S^{**}	$\frac{a+\sqrt{ad}}{a+b+c+\sqrt{ad}}$	$[1, 0]$
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)	$\frac{4(ad-bc)}{(a+d)^2+(b+c)^2}$	$[1, -1]$

Dissimilarities between sets

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

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

$$d_H(A, B) := \mathrm{card}(A \oplus B)$$

Usually we normalize it $d_h(A, B) = \frac{1}{M}\mathrm{card}(A \oplus B)$. One normalization is $M = \mathrm{card}(U)$; the other $M = m_1 + m_2$, where m_1 and m_2 are the first and the second largest value in $\{\mathrm{card}(X) : X \in \mathcal{F}\}$.

Other dissimilarities

$$d_s(A, B) = \frac{\mathrm{card}(A \oplus B)}{\mathrm{card}(A) + \mathrm{card}(B)} \qquad d_u(A, B) = \frac{\mathrm{card}(A \oplus B)}{\mathrm{card}(A \cup B)}$$

$$d_m(A, B) = \frac{\max(\mathrm{card}(A \setminus B), \mathrm{card}(B \setminus A))}{\max(\mathrm{card}(A), \mathrm{card}(B))}$$

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:

S. $$p(C) = \sum_{\mathrm{X},\mathrm{Y} \in C, \mathrm{X} < \mathrm{Y}} w(\mathrm{X}) \cdot w(\mathrm{Y}) \cdot d(\mathrm{X}, \mathrm{Y})$$

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

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

$$w(\{\mathrm{X}\}) = w(\mathrm{X}), \qquad \mathrm{X} \in \mathcal{U}$$

$$w(C_1 \cup C_2) = w(C_1) + w(C_2), \qquad C_1 \cap C_2 = \emptyset$$

Often $w(\mathrm{X}) = 1$ holds for each $\mathrm{X} \in \mathcal{U}$. Then $w(C) = \mathrm{card}(C)$.

M. $$p(C) = \max_{\mathrm{X},\mathrm{Y} \in C} d(\mathrm{X}, \mathrm{Y}) = \mathrm{diam}(C)$$ – diameter

T. $$p(C) = \min_{T \text{ is a spanning tree over } C} \sum_{(\mathrm{X}:\mathrm{Y}) \in T} d(\mathrm{X}, \mathrm{Y})$$

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

SM. $$P(\mathbf{C}) = \sum_{C \in \mathbf{C}} \max_{\mathrm{X},\mathrm{Y} \in C} d(\mathrm{X}, \mathrm{Y})$$

It is easy to prove:

Proposition 1.1 *Let* $P \in \{\mathrm{SS}, \mathrm{S\overline{S}}, \mathrm{SM}, \mathrm{MS}, \mathrm{M\overline{S}}, \mathrm{MM}\}$ *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 $P(\mathbf{C}) \geq \alpha_k^P(\mathbf{U}) \cdot \mathrm{MM}(\mathbf{C})$.

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

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

and is *α-sensitive* iff there exists an $\alpha_k^P(\mathbf{U}) > 0$ such that for each $\mathbf{C} \in P_k(\mathbf{U})$:

$$P(\mathbf{C}) \geq \alpha_k^P(\mathbf{U}) \cdot \mathrm{MM}(\mathbf{C})$$

Proposition 1.2 *Every α-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 α-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:

R. $$p(C) = \min_{\mathrm{L}\in\mathbf{F}} \sum_{\mathrm{X}\in C} w(\mathrm{X}) \cdot d(\mathrm{X},\mathrm{L})$$

where $\mathbf{F} \subseteq \mathcal{F}$ is the set of *representatives*. The element $\overline{C} \in \mathbf{F}$, 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. $[\mathbf{U}] \subseteq \mathbb{R}^2$ and $[\mathbf{F}] = \{(a,b,c) : ax + by = c, a^2 + b^2 = 1\}$. □

Example 2 In the case $[\mathbf{U}] \subseteq \mathbb{R}^m$, $[\mathbf{F}] = \mathbb{R}^m$, $d(\mathrm{X},\mathrm{L}) = d_2^2(\mathrm{X},\mathrm{L}) = \sum_{i=1}^m (x_i - l_i)^2$ there exists a uniquely determined representative – *center of gravity* $\overline{C} = \frac{1}{\mathrm{card}(C)} \sum_{\mathrm{X}\in C} X$. 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

$$p(C) = \sum_{X \in C} d_2^2(X, \overline{C}) = \frac{1}{2\,\mathrm{card}(C)} \sum_{X,Y \in C} d_2^2(X, Y)$$

replace the expression for $p(C)$ with

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

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

$$w(C_u \cup C_v) \cdot p(C_u \cup C_v) = w(C_u) \cdot p(C_u) + w(C_v) \cdot p(C_v) + \sum_{X \in C_u, Y \in C_v} w(X) \cdot w(Y) \cdot d(X, Y)$$

In Batagelj (1988) it is also shown how to replace $\overline{C}$ 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*

$$\begin{array}{lll} a) & p(C) + w(\mathrm{X}) \cdot d(\mathrm{X}, \overline{C \cup \mathrm{X}}) \leq p(C \cup \mathrm{X}), & \mathrm{X} \notin C \\ b) & p(C \setminus \mathrm{X}) + w(\mathrm{X}) \cdot d(\mathrm{X}, \overline{C}) \leq p(C), & \mathrm{X} \in C \end{array}$$

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

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

Therefore in case a):

$$p(C) = \sum_{\mathrm{Y} \in C} w(\mathrm{Y}) \cdot d(\mathrm{Y}, \overline{C}) \leq \sum_{\mathrm{Y} \in C} w(\mathrm{Y}) \cdot d(\mathrm{Y}, \overline{C \cup \mathrm{X}}) =$$

$$= \sum_{\mathrm{Y} \in C \cup \mathrm{X}} w(\mathrm{Y}) \cdot d(\mathrm{Y}, \overline{C \cup \mathrm{X}}) - w(\mathrm{X}) \cdot d(\mathrm{X}, \overline{C \cup \mathrm{X}}) =$$

$$= p(C \cup \mathrm{X}) - w(\mathrm{X}) \cdot d(\mathrm{X}, \overline{C \cup \mathrm{X}})$$

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

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

$$P(\mathbf{C}) = \bigoplus_{(C_1,C_2)\in\mathbf{C}\times\mathbf{C}} q(C_1, C_2), \quad q(C_1, C_2) \geq 0$$

is needed. We shall use it in blockmodeling.

Multicriteria clustering

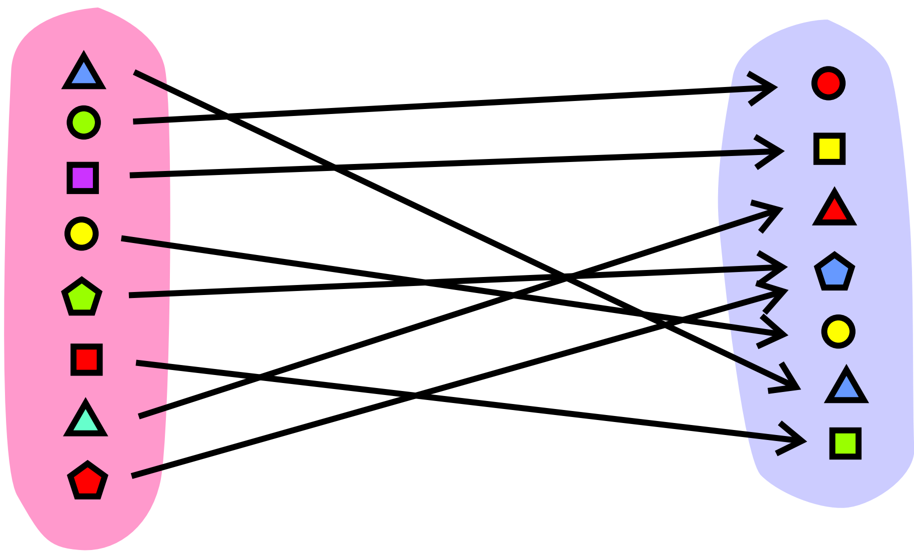
In some problems several criterion functions can be defined $(\Phi, P_1, P_2, \ldots, P_s)$.
See Ferligoj, Batagelj (1994).

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

$$P(\mathbf{C}) = \max_{\substack{\{C_1,C_2\}\in\mathbf{C}\times\mathbf{C}\\ \mathrm{card}(C_1)\geq\mathrm{card}(C_2)}} \;\min_{\substack{f:C_1\to C_2\\ f \text{ is surjective}}} \;\max_{\mathrm{X}\in C_1} d(\mathrm{X}, f(\mathrm{X}))$$

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:

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

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 Φ is finite the clustering problem (Φ, 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

$$\mathrm{card}(P_k) = S(n,k) = \frac{1}{k!}\sum_{i=0}^{k-1}(-1)^i\binom{k}{i}(k-i)^n, \quad 0 < k \le n$$

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 (P_2, MM) and (P_k, ST), 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 α-sensitive, then for each problem $(P_k(\mathbf{U}), P)$ there exists a problem $(P_{k+1}(\mathbf{U}'), P)$, such that $(P_k(\mathbf{U}), P) \propto (P_{k+1}(\mathbf{U}'), P)$.

Proof: Select a value P^* such that $P^* > \max_{\mathbf{C} \in P_k(\mathbf{U})} P(\mathbf{C})$, extend, $\mathbf{U}' = \mathbf{U} \cup \{\mathrm{X}^\bullet\}$, the set of units with a new unit $\mathrm{X}^\bullet$, and define the dissimilarities between it and the 'old' units such that $d(\mathrm{X}, \mathrm{X}^\bullet) > P^*/\alpha'$, for $\mathrm{X} \in \mathbf{U}$ and $\alpha' = \alpha^P_{k+1}(\mathbf{U}')$. We get a new clustering problem $(P_{k+1}(\mathbf{U}'), P)$.

Consider a clustering $\mathbf{C}' \in P_{k+1}(\mathbf{U}')$. There are two possibilities:

a. $\mathrm{X}^\bullet$ forms its own cluster $\mathbf{C}' = \mathbf{C} \cup \{\{\mathrm{X}^\bullet\}\}$, $\mathbf{C} \in P_k(\mathbf{U})$. Then

$$P(\mathbf{C}') = P(\mathbf{C}) \oplus p(\{\mathrm{X}^\bullet\}) = P(\mathbf{C}) \leq \max_{\mathbf{C} \in P_k(\mathbf{U})} P(\mathbf{C}) < P^*$$

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

$$P(\mathbf{C}') \geq \alpha' \cdot \max_{C \in \mathbf{C}'} \max_{\mathrm{X},\mathrm{Y} \in C} d(\mathrm{X}, \mathrm{Y}) \geq \alpha' \cdot \max_{\mathrm{X},\mathrm{Y} \in C^\bullet} d(\mathrm{X}, \mathrm{Y}) =$$

$$= \alpha' \cdot \max_{\mathrm{X} \in C^\bullet \setminus \{\mathrm{X}^\bullet\}} d(\mathrm{X}, \mathrm{X}^\bullet) > P^*$$

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

$$\mathbf{C}' \in \mathrm{Min}(P_{k+1}(\mathbf{U}'), P) \Leftrightarrow \mathbf{C} \in \mathrm{Min}(P_k(\mathbf{U}), P)$$

$\square$

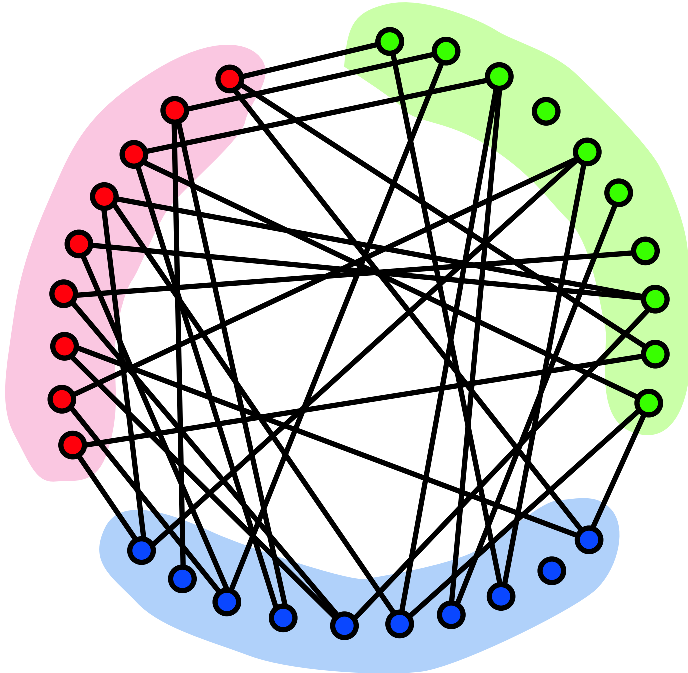
Complexity results 1

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

$$3 - \text{COLOR} \propto (P_3, P)$$

Proof: Let $G = (V, E)$ be a simple undirected graph. We assign to it a clustering problem $(P_3(V), P)$ 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** $\min(P_3(V), P) = 0$.
Let $\mathbf{C} = \{C_1, C_2, C_3\}$ then $P(\mathbf{C}) = 0$ **iff** $c : V \to \{1, 2, 3\} : (c(v) = i \Leftrightarrow v \in C_i)$ is a vertex coloring.

□

Complexity results 2

Polynomial	NP-hard	note
(P_2, MM)	(P_3, MM)	Theorem 1.6
	(P_3, SM)	Theorem 1.6
	(P_2, SS)	MAX-CUT $\propto (P_2, \mathrm{SS})$
	$(P_2, \mathrm{S\overline{S}})$	$(P_2, \mathrm{SS}) \propto (P_2, \mathrm{S\overline{S}})$
	(P_2, MS)	PARTITION $\propto (P_2, \mathrm{MS})$
$(\mathbb{R}_2^m, \mathrm{S\overline{S}})$		
$(\mathbb{R}_k^1, \mathrm{S\overline{S}})$		
$(\mathbb{R}_k^1, \mathrm{SM})$		
$(\mathbb{R}_k^1, \mathrm{MM})$		

Note that, by the Theorem 1.5, (P_k, MM), $k > 3$ are also NP-hard ...

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