Multivariátní porovnání dat

- **klastrová (shluková) analýza**
- bez apriorních předpokladů

Shluková analýza

✓ **Shluková analýza - cluster analysis**

- ***úvod - definice***
- ***princip***
- ***algoritmy***
- ***výsledky***

Shluková analýza

✓ **Úvod**

✓ **DEFINICE -zavedení pojmu**

➡ Cluster analysis classifies a set of observations into two or more mutually exclusive *unknown* groups based on combinations of interval variables.

✓ *Multivariate Statistics: Concepts, Models, and Applications*
David W. Stockburger, 1996

Shluková analýza

✓DEFINICE - zavedení pojmu

➡Cluster analysis is a technique for **grouping data** and **finding structures** in data. The most common application of clustering methods is to partition a data set into clusters or classes, where similar data are assigned to the same cluster whereas dissimilar data should belong to different clusters.

Shluková analýza

✓DEFINICE - zavedení pojmu

➡The term *cluster analysis* (first used by Tryon, 1939) actually encompasses a number of different classification algorithms. A general question facing researchers in many areas of inquiry is how to *organize* observed data into meaningful structures, that is, to develop taxonomies.

➡1984-2002 ELECTRONIC TEXTBOOK StatSoft

DEFINICE - zavedení pojmu

- Cluster Analysis is a **multivariate analysis technique** that seeks to organize information about variables so that relatively homogenenous groups, or "clusters," can be formed. The clusters formed with this family of methods should be **highly internally homogenous** (members are similar to one another) and **highly externally heterogenous** (members are *not* like members of other clusters.

Shluková analýza

- ✓ **Princip - postup**
- ➡ **shromáždění dat**
- ➡ **výběr proměnných**
- ➡ **volba metody pro vytvoření distanční matice (matice podobnosti)** /distance matrix, similarity matrix, proximities matrix/

Shluková analýza

- ✓ Princip - postup
- ➡ volba metody pro tvorbu dendrogramu /“dendrogram“/, hierarchické struktury /“hierarchical structure“/, stromového diagramu /“tree diagram“/

Shluková analýza

- ✓ Princip - postup
- ➡ volba výstupu
- ✏ grafický dendrogram
- ✏ textová diagnostika tříd
- ✏ histogram vzdáleností dat (statistická distribuce vzdáleností)

Shluková analýza

- ✓ shromáždění dat
- ➡ data vhodná k třídění dle podobnosti
- ✏ kompatibilita dat
  - rozsah nezávisle proměnné
  - hodnoty nezávisle proměnné
  - typ závisle proměnné (proměnných)

Shluková analýza

- ✓ **výběr rozsahu nezávisle proměnné**
  - výběr jednoho intervalu
  - výběr více oddělených intervalů nezávisle proměnné
  - výběr diskrétních hodnot nezávisle proměnné
- ✓ **výběr závisle proměnné (proměnných) užitých pro klasifikaci dat**

Shluková analýza

- ✓ **Výpočet distanční matice**
- ➡ **volba metody výpočtu**
- ✏ **standardní - Eukleidovské vzdálenosti** (příp. jejich kvadráty)
- ✏ **Chebychev**
- ✏ **City-block (Manhattan)**

Shluková analýza

- ✓ **Výpočet distanční matice**
- ➡ **volba metody výpočtu**
- ✏ **Pearsonova korelace**
- ✏ **škálovací s Pearsonovou korelací** (pro více dílčích rozsahů nezávisle proměnné)
- ✏ **s faktorem hladiny opakovatelnosti měření**

Shluková analýza

- ✓ **Výpočet distanční matice**
- ➡ **volba metody výpočtu**
- ✏ **Minkowski - více subvariant**
- ✏ **mocninný algoritmus**
  („power distance“)
- ✏ **procenta nesouhlasu**
  (nesouladu) *(„percent disagreement“)*

Distanční matice

- ✓ **Euklidovské vzdálenosti**
- ➡ **geometrické vzdálenosti v multidimenzionálním prostoru**
  - ✏ $\text{distance}(a,b) = \{\sum (a_i - b_i)^2 \}^{½}$
  - ✏ **používá se pro experimentální (neupravovaná) data**
  - ✏ **vzdálenosti mezi dvojicemi objektů nejsou ovlivněny dalšími objekty (přidáním objektů)**
  - ✏ **vzdálenosti významně závisí na volbě jednotek**

Distanční matice

- ✓ **kvadráty Eukleidovských vzdáleností**
- ➡ kvadáty geometrických vzdáleností v multidimenzionálním prostoru
- ✏ $\text{distance}(a,b) = \{\sum (a_i - b_i)^2 \}$
- ✏ **používá se pro experimentální (neupravovaná) data**
- ✏ **progresivně se zvyšuje vliv (význam) velkých Eukleidovských vzdáleností**

Distanční matice

- ✓ **City-block (Manhattan)**
- ➡ sumace absolutních hodnot rozdílů v multidimenzionálním prostoru
- ✏ $\text{distance}(a,b) = \{\sum |a_i - b_i| \}$
- ✏ **obvykle obdobné výsledky jako pro Eukleidovské vzdálenosti**
- ✏ **menší vliv JEDNOTLIVÝCH odlehlých bodů**

Distanční matice

- ✓ **Chebychev**
- ➡ **maximalizace vlivu rozdílu v jednom bodě**
- ✏ $\text{distance}(a,b) = \text{Maximum } |a_i - b_i|$
- ✏ **dva objekty odlišné, liší-li se v jednom bodě (dimenzi) - zásadní vliv odlehlých bodů - testování odlehlých bodů**

Distanční matice

- ✓ **Mocninný algoritmus**
- ➡ **zobecněné Eukleidovské vzdálenosti**

- ✏ $\text{distance}(a,b) = \{\sum |a_i - b_i|^p\}^{1/r}$
- ✏ **p,r - volitelné koeficienty**

  pro p = r = 2 jde o Eukleidovské vzdálenosti

  p - progresivita vlivu jednotlivých velkých vzdáleností bodů

  r - vliv na porovnávání celých objektů (a, b,...)

Distanční matice

- ✓ **procenta nesouhlasu**
- ➡ **počty odlišných bodů**
  - ✏ distance(a,b) = {Počet $a_i \neq b_i$ } / i
  - ✏ **vhodné pro porovnávání souborů diskrétních bodů**

Distanční matice

- ✓ **Pearsonova korelace**
- ➡ **využití Pearsonova korelačního koeficientu *r***

$$r = \sum (a_i * b_i)$$

- ✏ $\text{distance}(a,b) = (1 - r) * 1000$
- ✏ **provádí se pro vektorově normalizovaná data, rozsah vzdáleností pak je od 0 (identická data) do 2000 (maximálně odlišná)**

Distanční matice

- ✓ **škálovací s Pearsonovou korelací**
- ➡ **normalizace škály distancí pro sadu jednotlivých intervalů nezávisle proměnných**
- ✏ docílí se stejného rozsahu vzdáleností (od minimální po maximální) pro všechny sledované oblasti

Příprava dendrogramu

✓ **metody přípravy dendrogramu**

➡ **jednoduché propojení**

(single linkage, nearest neighbor)

vliv **nejbližších subobjektů**

ve dvou sousedních objektech

$D(r,i) = \min [D(p, i) , D(q,i)]$

kde r je nový objekt vzniklý z objektů p a q

(objektem se rozumí buď *"vstupní objekt"*, nebo "*vytvořený klastr*")

Příprava dendrogramu

✓ **metody přípravy dendrogramu**

➡ **kompletní propojení**

(complete linkage, furthest neighbor)

vliv **nejvzdálenějších subobjektů**

ve dvou sousedních objektech

$D(r,i) = \max [D(p, i) , D(q,i)]$

kde r je nový objekt vzniklý z objektů p a q

Příprava dendrogramu

✓ **metody přípravy dendrogramu**

➡ **průměrové propojení**

(unweighted pair-group average linkage - UPGMA - *unweighted pair-group method using arithmetic averages,*

Sneath and Sokal 1973)

vzdálenost mezi dvěma objekty je **aritmetickým průměrem** vzdáleností mezi všemi páry subobjektů obou objektů

Příprava dendrogramu

✓ **metody přípravy dendrogramu**

➡ **vážené průměrové propojení**

(weighted pair-group average linkage - WPGMA - *weighted pair-group method using arithmetic averages, Sneath and Sokal 1973*)

vážen počet subobjektů obou objektů

$D(r,i) = [n(p) * D(p,i) + n(q) * D(q,i)] / [n(p) + n(q)]$

n(p) a n(q) počty subobjektů v objektech p a q

Příprava dendrogramu

✓ **metody přípravy dendrogramu**

➡ **nevážené centroidové (těžišťové) propojení**

(unweighted pair-group centroid
UPGMC - *unweighted pair-group method using the centroid averages,*
Sneath and Sokal 1973)

porovnávána vzdálenost (poloha) těžišť objektů v multidimenzionálním prostoru

Příprava dendrogramu

✓ **metody přípravy dendrogramu**

➡ **vážené centroidové (těžišťové) propojení -** „median“ (weighted pair-group centroid - WPGMC - *weighted pair-group method using the centroid averages, Sneath and Sokal 1973*)

porovnávána vzdálenost (poloha) těžišť objektů v multidimenzionálním prostoru s vážením počtu subobjektů

Příprava dendrogramu

✓ **metody přípravy dendrogramu**

➡ **Wardova metoda -**

(Ward's method, *Ward 1963*)

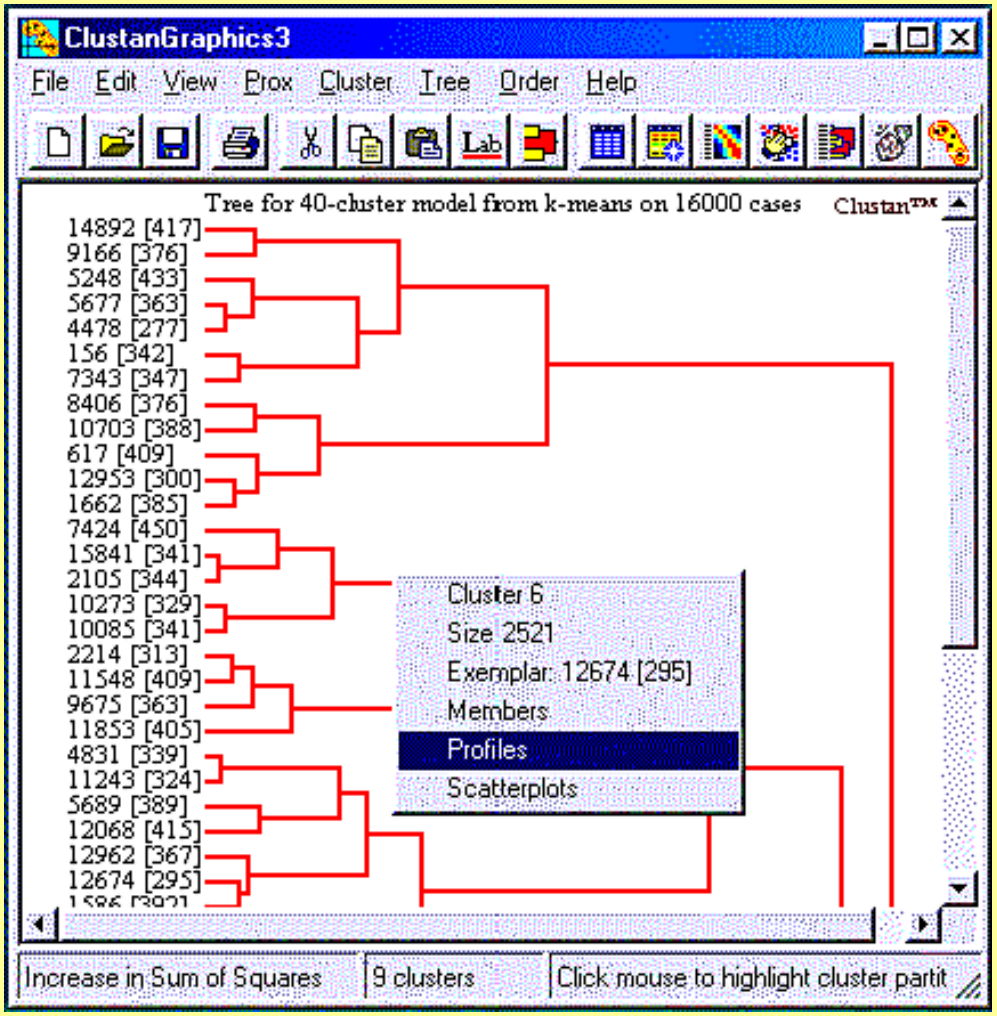
místo vzdáleností HETEROGENITA

hledá homogenní skupiny

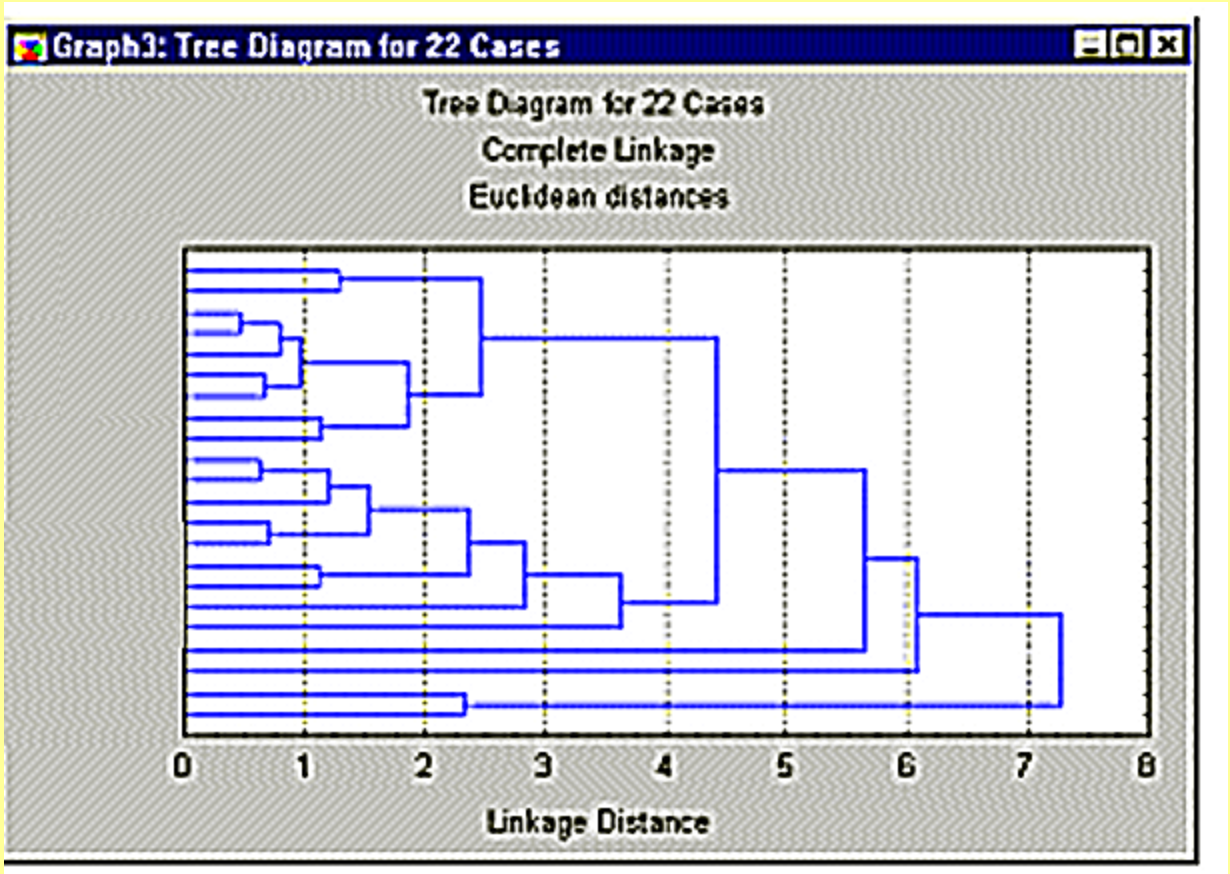
analýza - minimalizace sumy čtverců odchylek

pro všechny možné (i hypotetické) dvojice subobjektů v každém kroku tvorby hierarchické struktury

Grafický dendrogram



Grafický dendrogram



Grafický dendrogram

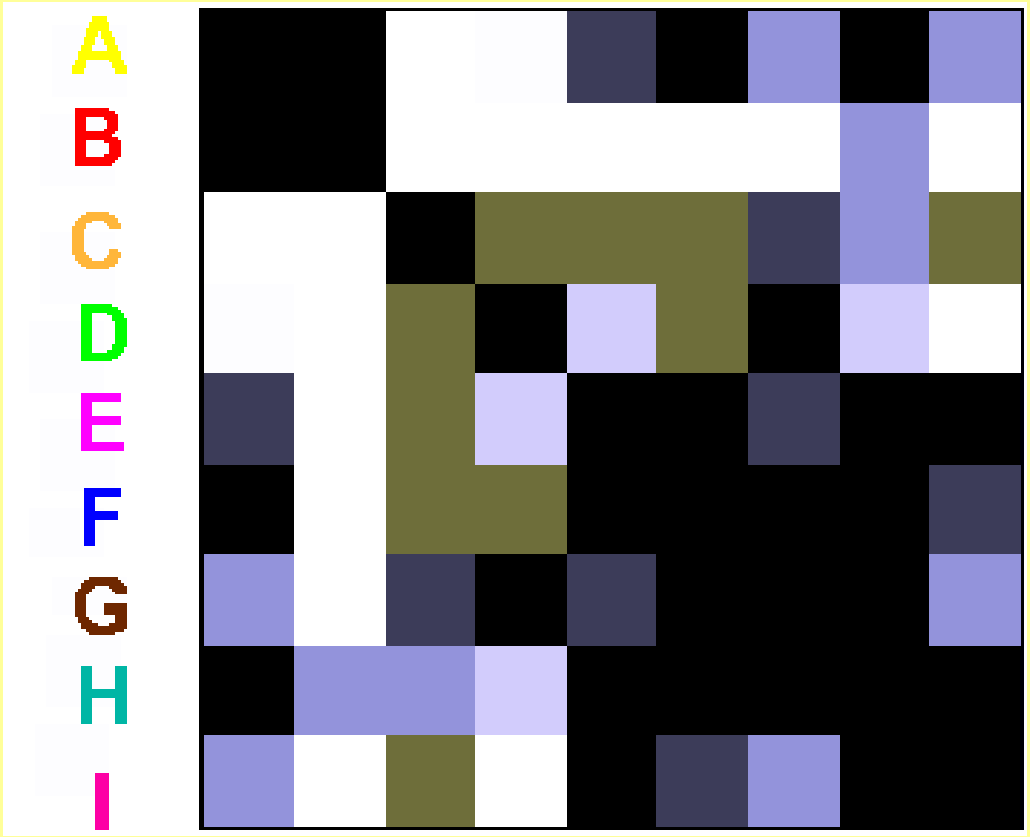
✓ **problém znázornění**

➡ **možnost převrácení větví ve stromu („volná otáčivost“ v bodech větvení)**

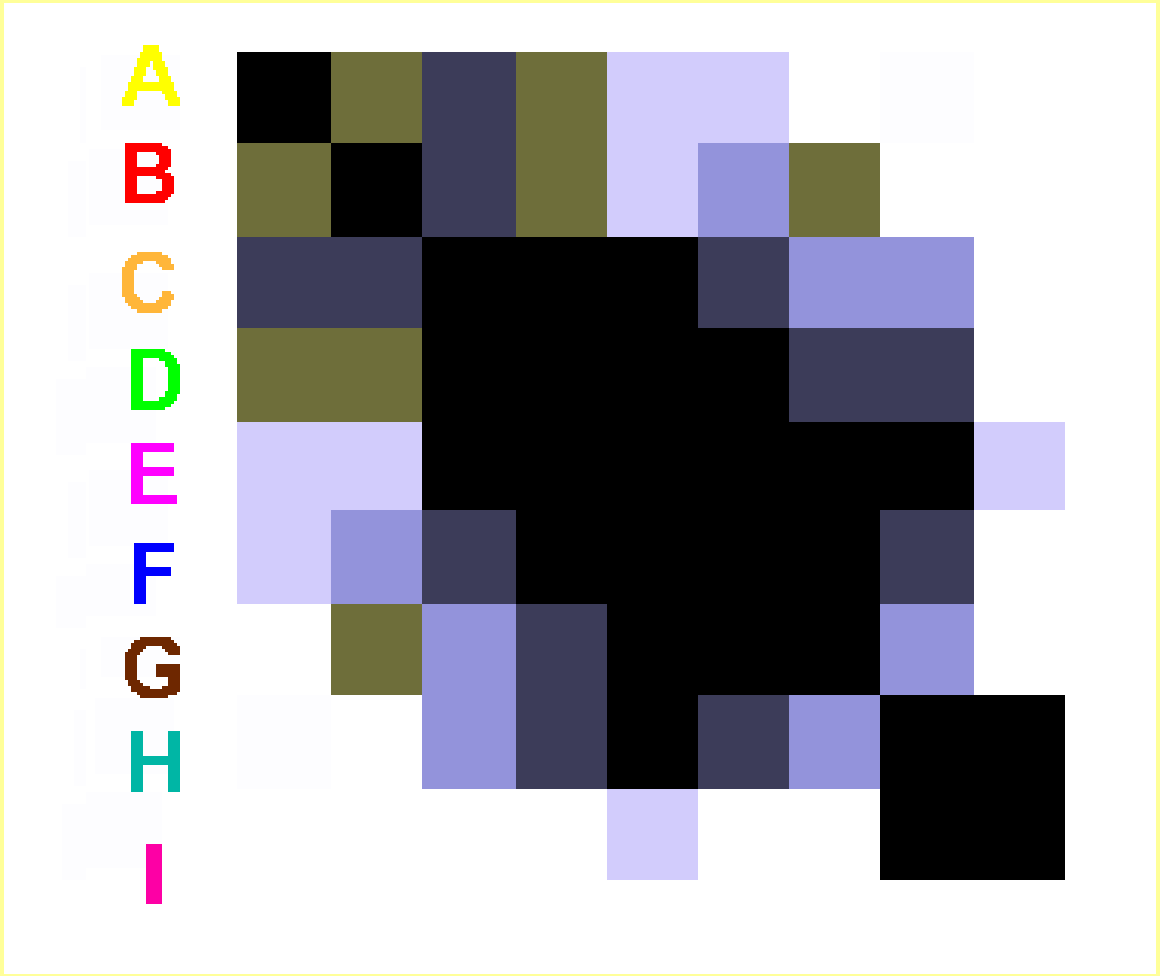
2^{n-2} možností zobrazení dendrogramu pro n vstupních objektů

vylepšení - **seřazovací algoritmy** („seriation“) (přeuspořádání distanční matice - snaha o seřazení dle rostoucích vzdáleností)

Znázornění matice vzdáleností



Znázornění matice vzdáleností - optimalizované



Textový klasifikační výstup

Range 1: 3117-2776
Range 2: 1791-802
Diagnosis for 12 classes:
Ward's

1. class has 38 members:

Last fusion occured at 0.905

Next nearest class is 2 at 1.938

1	A637L11D.1	A637L11D Praha Z. strom 637 le
32	A644P11H.1	A644P11H Praha Z. strom 644 pr
14	A640L11H.1	A640L11H Praha strom 640 le
102	P609L31H.1	P609L31H Pardubice strom 609
115	P610P11D.1	P610P11D Pardubice strom 610
2	A637L11H.1	A637L11H Praha Z. strom 637 le
199	U651P11D.1	U651P11D Uher. Hr. strom 651 p
114	P610L31H.1	P610L31H Pardubice strom 610
118	P610P21H.1	P610P21H Pardubice strom 610
13	A640L11D.1	A640L11D Praha strom 640 le
72	C531P31H.1	C531P31H Chrudim strom 531 p
106	P609P21H.1	P609P21H Pardubice strom 609
113	P610L31D.1	P610L31D Pardubice strom 610
198	U651L31H.1	U651L31H Uher. H. strom 651 le
200	U651P11H.1	U651P11H Uher. Hr. strom 651 p
232	U658L21H.1	U658L21H Uherske Hradiste strom 6
4	A637L21H.1	A637L21H Praha Z. strom 637 le
216	U652P31H.1	U652P31H Uherske H strom 652
11	A637P31D.1	A637P31D Praha-Zap strom 637
107	P609P31D.1	P609P31D Pardubice strom 609
63	C531L21D.1	C531L21D Chrudim strom 531 l
67	C531P11D.1	C531P11D Chrudim strom 531 p
70	C531P21H.1	C531P21H Chrudim strom 531 p
234	U658L31H.1	U658L31H Uherske Hradiste strom 6
229	U658L11D.1	U658L11D Uher. H. strom 658 le
6	A637L31H.1	A637L31H Praha zapad strom 637
182	S635L11H.1	S635L11H Strakonice strom 635
64	C531L21H.1	C531L21H Chrudim strom 531 l
68	C531P11H.1	C531P11H Chrudim strom 531 p
10	A637P21H.1	A637P21H Praha Z. strom 637 pr
111	P610L21D.1	P610L21D Pardubice strom 610
98	P609L11H.1	P609L11H Pardubice strom 609 l
150	S625L31H.1	S625L31H Strakonice strom 625
126	P612L31H.1	P612L31H Pardubice strom 612
12	A637P31H.1	A637P31H Praha-Zap strom 637
230	U658L11H.1	U658L11H Uher. H. strom 658 le
125	P612L31D.1	P612L31D Pardubice strom 612
154	S625P21H.1	S625P21H Strakonice strom 625 prava vetev, 2. vyhon, horni strana

2. class has 35 members:
Last fusion occured at 1.296
Next nearest class is 1 at 1.938

9	A637P21D.1	A637P21D Praha Z. strom 637 prava vetev, 2. vyhon, dolni strana
181	S635L11D.1	S635L11D Strakonnice strom 635 leva vetev, 1. vyhon, dolni strana
145	S625L11D.1	S625L11D Strakonice strom 625 leva vetev, 1. vyhon, dolni strana
65	C531L31D.1	C531L31D Chrudim strom 531 leva vetev, 3. vyhon, dolni strana
148	S625L21H.1	S625L21H Strakonice strom 625 leva vetev, 2. vyhon, horni strana
247	Z605P11D.1	Z605P11D Zdar. S. strom 605 prava vetev, 1. vyhon, dolni strana
51	C351L21D.1	C351L21D Chrudim strom 351 leva vetev, 2. vyhon, dolni strana
175	S633P11D.1	S633P11D Strakonice strom 633 prava vetev, 1. vyhon, dolni strana
37	A647L11D.1	A647L11D Praha strom 647 leva vetev, 1. vyhon, dolni strana
48	A647P31H.1	A647P31H Praha-Zap strom 647 prava vetev, 3. vyhon, horni strana
180	S633P31H.1	S633P31H Strakonice strom 633 prava vetev, 3. vyhon, horni strana
162	S628L31H.1	S628L31H Strakonnice strom 628 leva vetev, 3. vyhon, horni strana
178	S633P21H.1	S633P21H Strakonice strom 633 prava vetev, 2. vyhon, horni strana
179	S633P31D.1	S633P31D Strakonice strom 633 prava vetev, 3. vyhon, dolni strana
30	A644L31H.1	A644L31H Praha Zap. strom 644 leva vetev, 3. vyhon, horni strana
99	P609L21D.1	P609L21D Pardubice strom 609 leva vetev, 2. vyhon, dolni strana

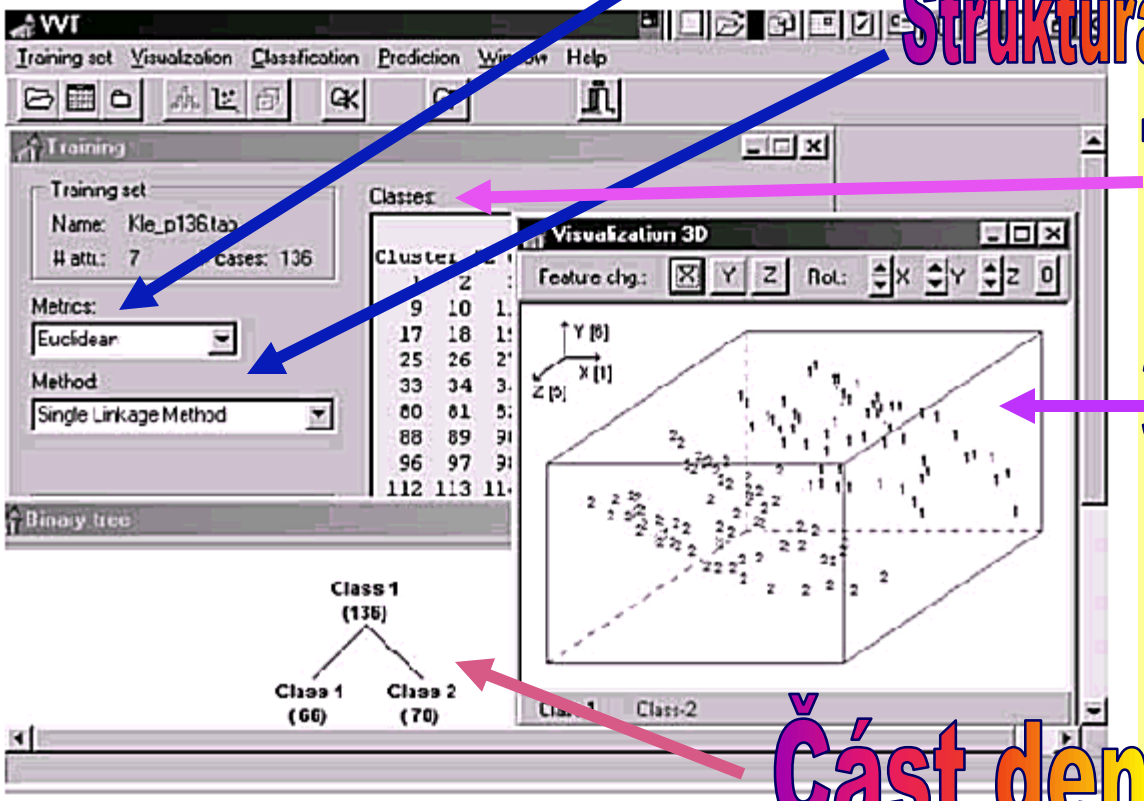
1. class has 38 members:

Last fusion occured at 0.905

Next nearest class is 2 at 1.938

1 A637L11D.1 A637L11D Praha Z. strom 637 leva vetev, 1. vyhon, dolni strana
32 A644P11H.1 A644P11H Praha Z. strom 644 prava vetev, 1. vyhon, horni strana
14 A640L11H.1A640L11H Praha strom 640 leva vetev, 1. vyhon, horni strana

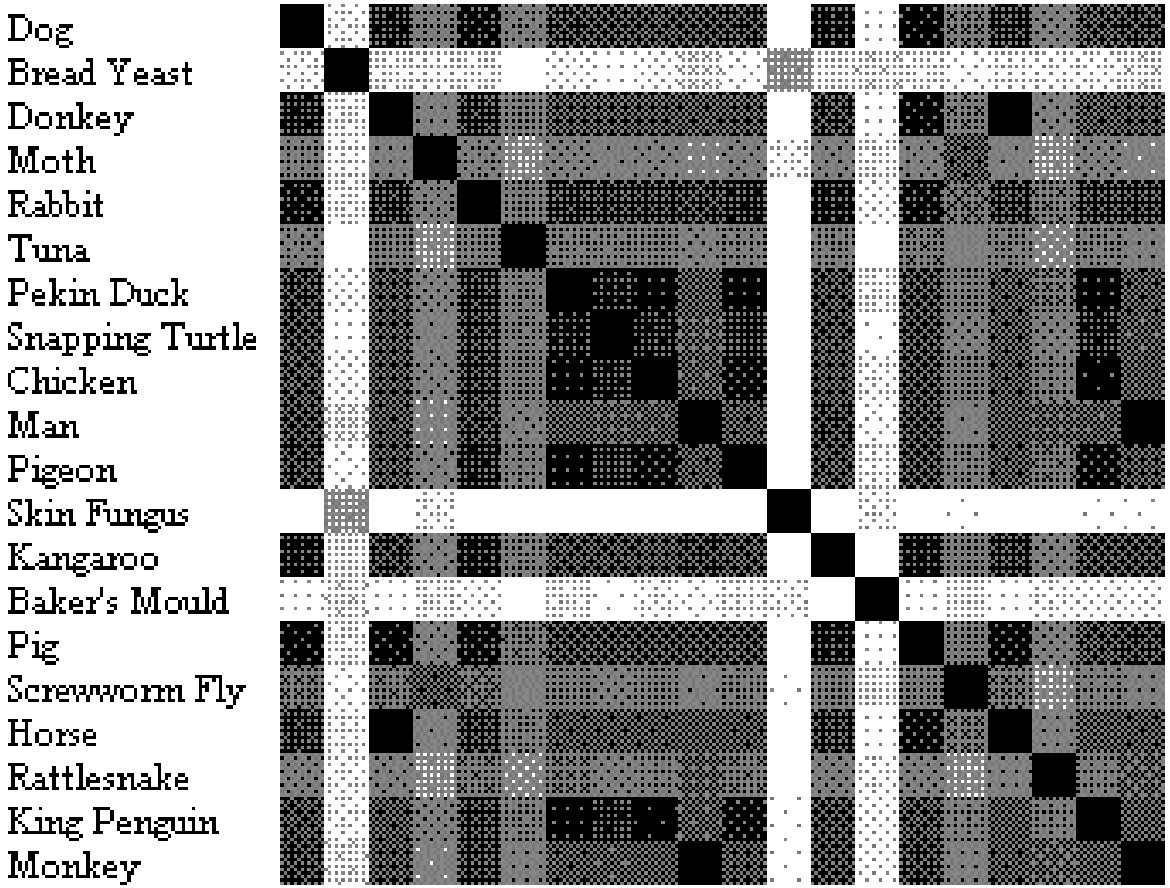
Kombinované znázornění



The results of the cluster analysis (table, part of dendrogram, and 3-d picture).

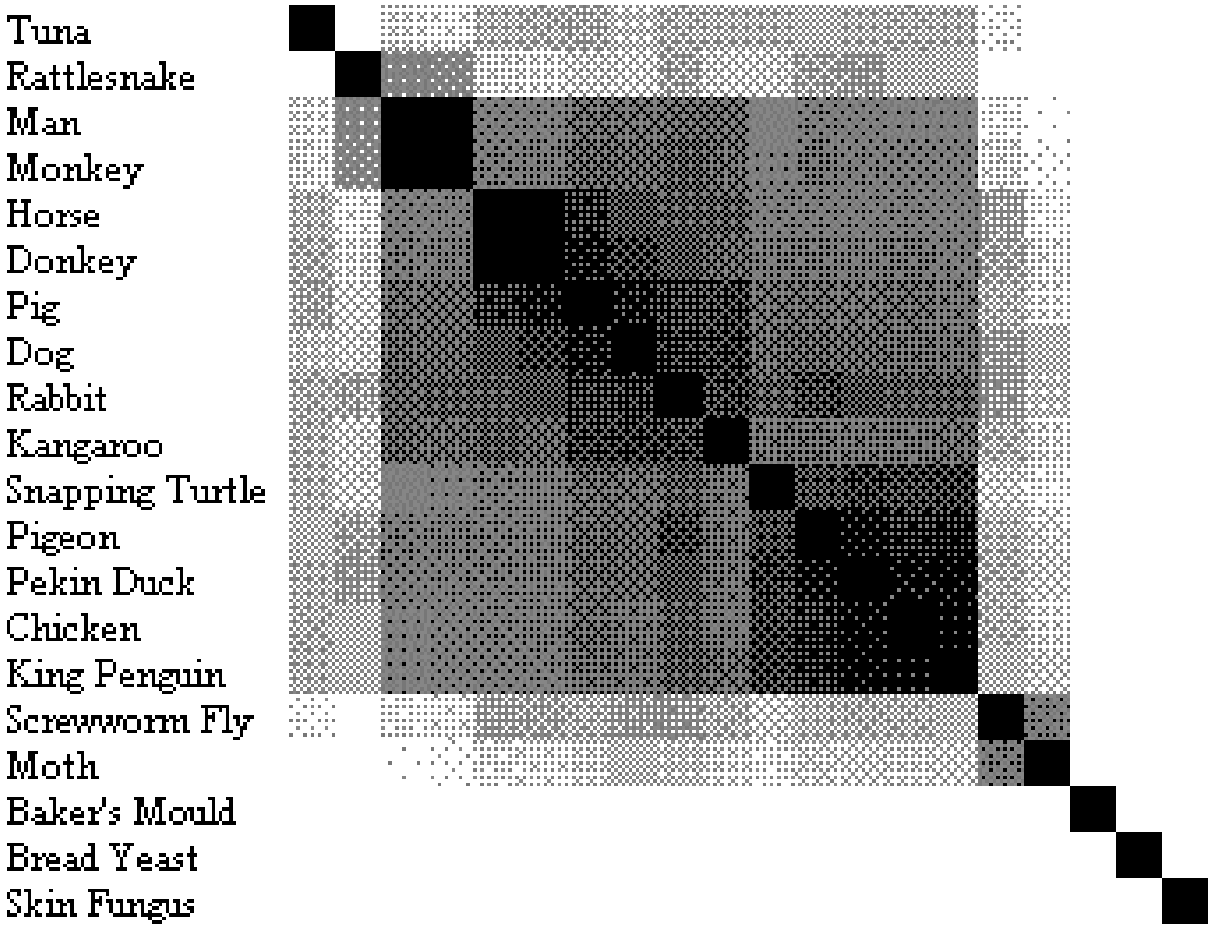
Aplikační příklad

Studium cytochromu-c



Aplikační příklad

Studium cytochromu-c



Aplikační příklad

Studium cytochromu-c