Pokročilé zpracování dat - shlukování dat

Zpracovaly : Tereza Mlynářová
Jarmila Zatyková

školní rok 2009/2010

Studijní skupina 3 70

Osnova

- **Shluková analýza**

 - Formulace úlohy

- **Objekty a znaky**

- Typy znaků
- Podobnost objektů
- Příklady ukazatelů

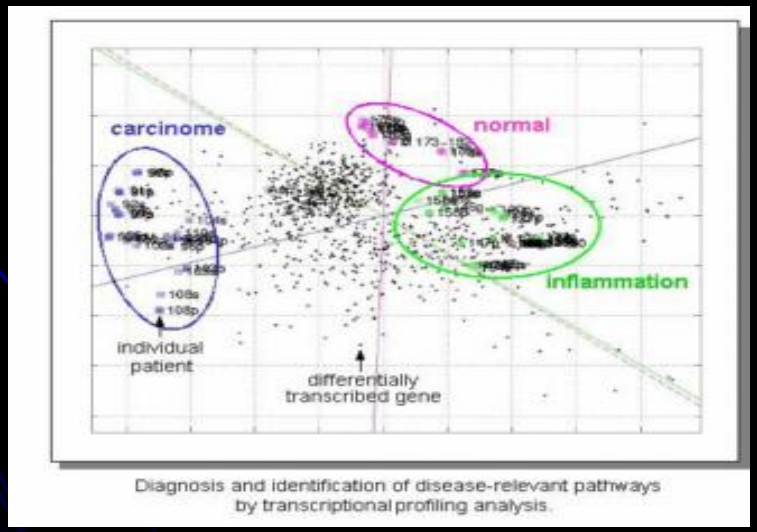
- **Typy metod shlukové analýzy**

- Hierarchické shlukování
- Nehierarchické shlukování
  - metoda K-means

- **Aplikace shlukové analýzy**

SHLUKOVÁ ANALÝZA

- Vícerozměrná statistická metoda
- Cíl = nalézt v dané množině objektů její podmnožiny – shluky objektů
- Členové shluku si musí být navzájem podobnější než objekty ze shluků ostatních
- Historicky se poprvé objevuje v 80.letech(těžké přímo vyčlenit,protože SA je spojena s dalšími disciplínami – např. biologie, psychologie, archeologie,…)



Diagnosis and identification of disease-relevant pathways by transcriptional profiling analysis.

- Formulace úlohy

- Nechť X značí množinu objektů n
- Rozklad Ω = {C1,C2,…,Cm} množiny X je množina disjunktních(nehierarchické),nebo nedisjunktních(hierarchické), neprázdných podmnožin tvořících dohromady X
- Ci ∩ Cj = { }

  C1 U C2 U … U Cm = X
- Každá množina Ci je komponentou rozkladu

OBJEKTY A ZNAKY

- Předměty(každý musí být popsán prostřednictvím stejného souboru znaků), nebo jevy(charakterizovány prostřednictvím určitého souboru objektů, nositelů těchto znaků)
- Každý objekt popsán několika stavy o několika znacích
- Stavům přiřazujeme číselné kódy=hodnoty znaků

- => objektem je několika-rozměrný vektor čísel

- Typy znaků

- Kvalitativní
  - konečná množina popisujících termínů(je jim přiřazen číselný kód)

→nominální

- např. barva (1-červená,2-žlutá,3-bílá)

→ordinální-dají se uspořádat

- např. zeleň listu (1-světlá,2-střední,3-tmavá)

  - binární(dichotomické) znaky

- např. mít modré oči (pravda/nepravda), pohlaví (muž / žena)

- Kvantitativní
  - interval v reálných nebo celých číslech

- např. délka, teplota

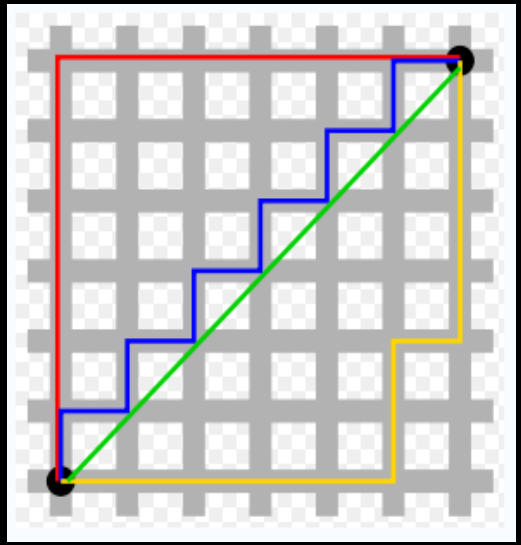
- Podobnost objektů

- Základní problém shlukové analýzy = pojetí vzájemné podobnosti(resp. vzdálenosti) objektů a její následné kvantitativní vyjádření
- Existuje mnoho metod sestrojení tohoto ukazatele
- Základní podmínky(pro vhodný předpis míry podobnosti):
  - nezápornost $d(O_i, O_j) \geq 0;$
  - oboustrannost (symetrie) $d(O_i, O_j) = d(O_j, O_i);$
  - shodné objekty by měly mít ukazatel vzdálenosti roven 0 (identita) $d(O_i, O_i) = 0$
  - trojúhelníková nerovnost $d(O_i, O_j) \leq d(O_i, O_h) + d(O_h, O_j)$

- Příklady ukazatelů

- **Metriky**

  - představují míru nepodobnosti objektů
  - vychází z geometrického modelu dat
  - metrika = fce definovaná na Ep x Ep přiřazující čtyři podmínky(viz.předchozí slide)
  - základní je eukleidovská vzdálenost(od toho odvozeny další – např. čtverec Eukleidovské vzdálenosti, metrika Manhattan,…)



- **Koeficienty asociace**

  - představují míru podobnosti

  - určeny pro hodnocení podobnosti určené tzv. dichotomickými znaky(ukazatele založené na počtu shod a počtu znaků)

  - 4 možné asociace dvou objektů = oba pravda, pravda/nepravda, nepravda/pravda, oba nepravda

  - operují s pozitivními(např. Jaccardův), nebo negativními shodami(Sokalův)

  - vztahují se k celkovému počtu znaků

- **Korelační koeficient**

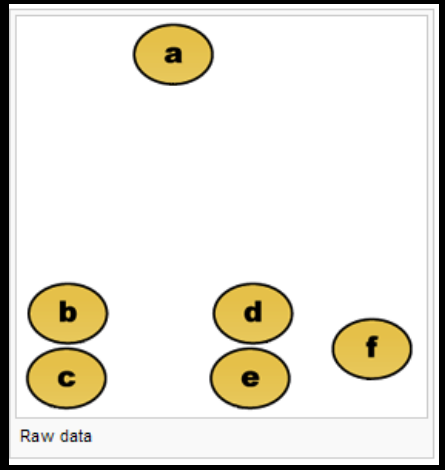
  - především pro shlukování proměnných

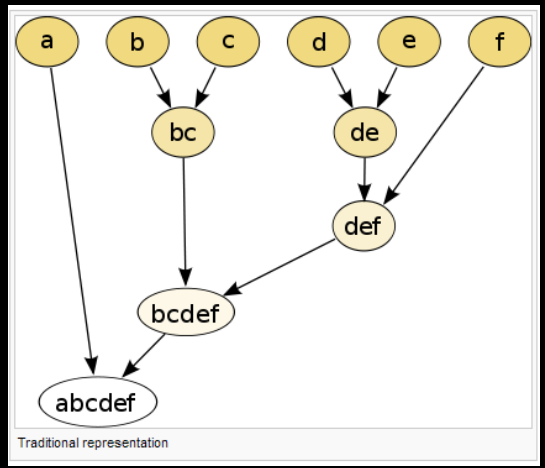
TYPY METOD SHLUKOVÉ ANALÝZY

- **HIERARCHICKÉ SHLUKOVÁNÍ**

-systém podmnožin
(průnik =prázdná množina, nebo jeden z nich)

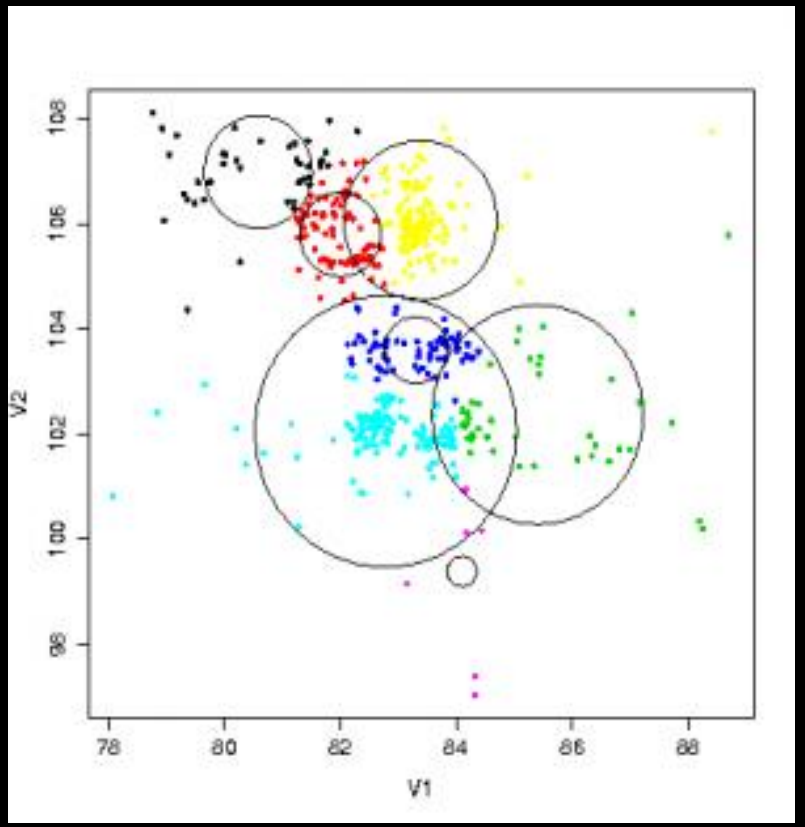
- zjemnění klasifikace, větvení
- Aglomerativní(shluky spojujeme), divizivní (z celku vytváříme shluky – dělíme ho)
- Dendrogram=binární strom znázorňující hierarchické shlukování
- Mnoho metod – metoda nejbližšího souseda, centroidní vzdálenost, párová vzdálenost,…





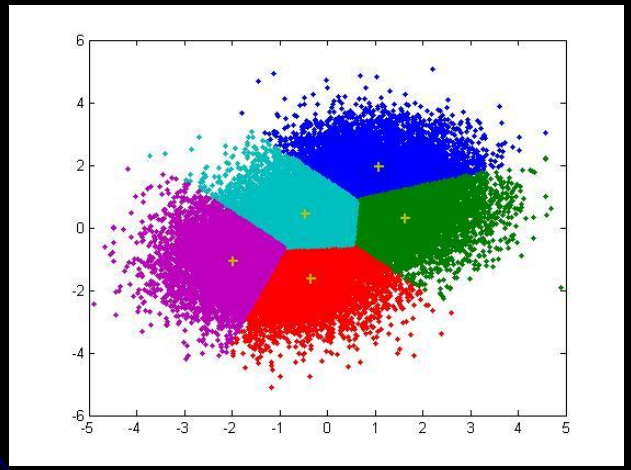
NEHIERARCHICKÉ SHLUKOVÁNÍ

- nepostupuje při shlukování hierarchicky – všechny podmnožiny jsou disjunktní
- Nalezení optimálního počtu shluků – velmi obtížné
- Mnoho metod – měnící i neměnící počet shluků

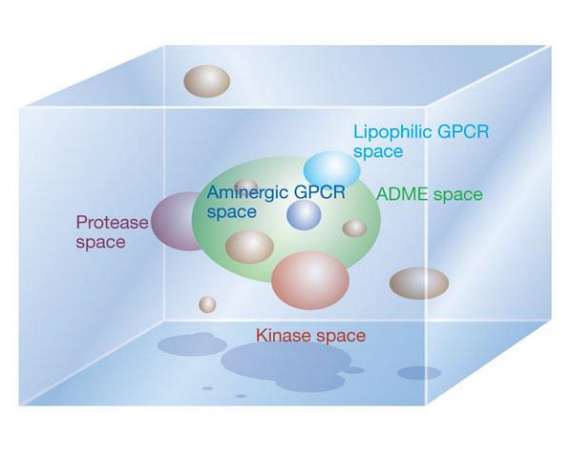


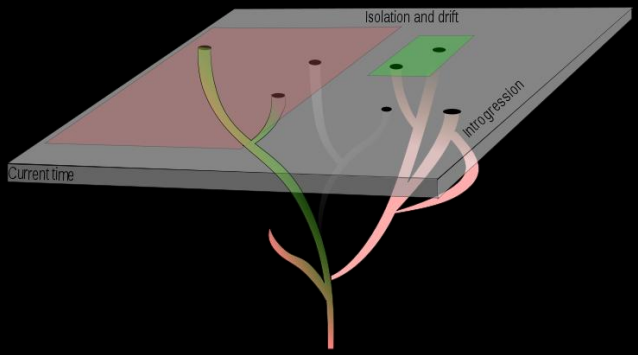
- Metoda K-means

- MacQueenova metoda
- Minimalizace střední odchylky mezi zadanou množinou dat a vektory(mající nejmenší eukleidovskou vzdálenost) – přiřazuje každý bod do shluku, jehož centrum(těžiště) je nejblíže
- Těžiště je průměrem všech bodů v daném shluku
- Rozdělení vektorů do předem daného počtu shluků



APLIKACE SHLUKOVÉ ANALÝZY



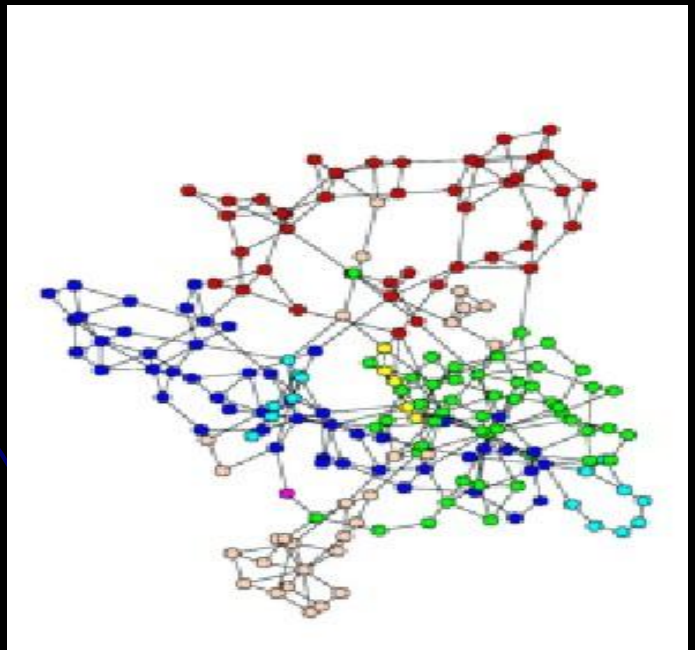


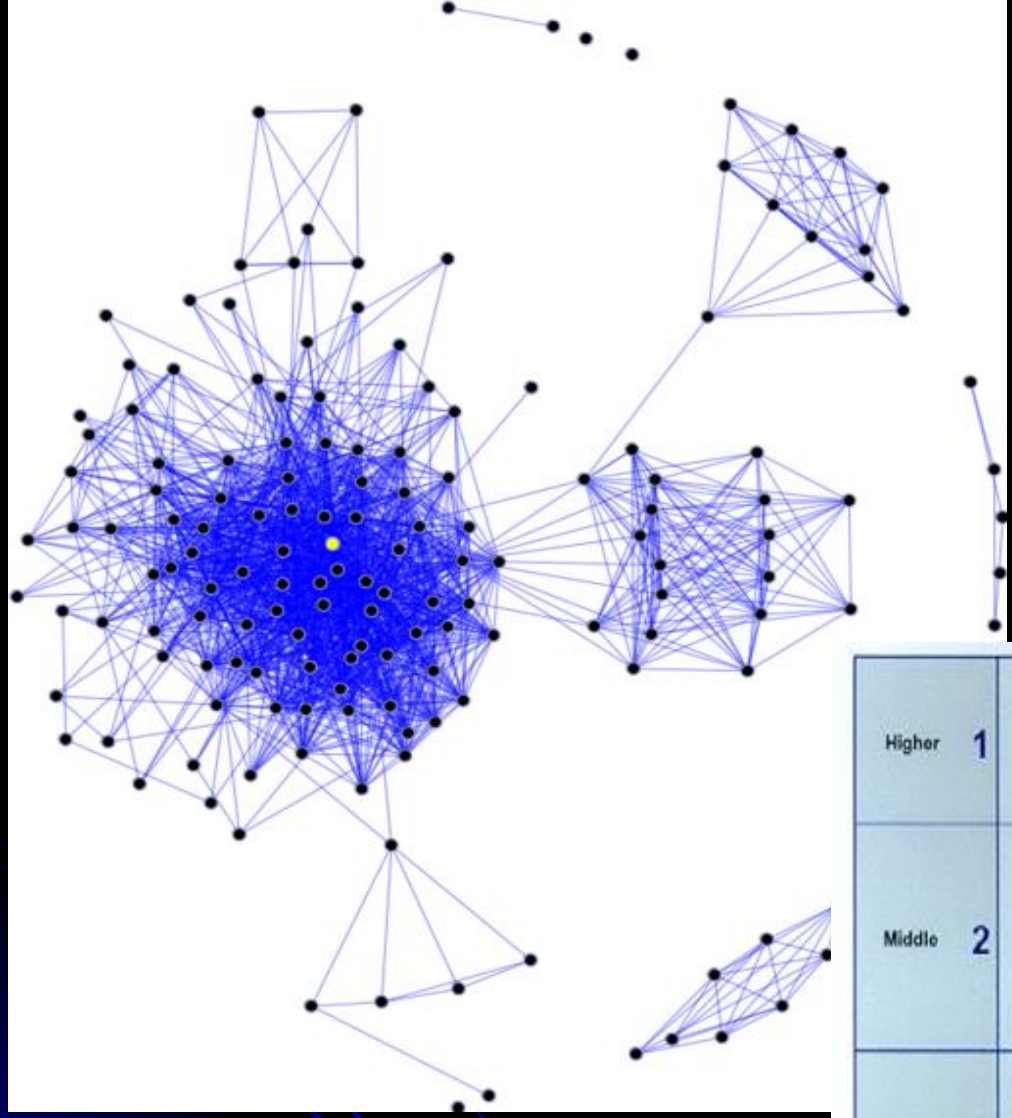
- Biologie – systém rostlin(vytváření umělých společenstev), zvířat(popsání rozdílů mezi různými společenstvími, vytváření skupin různých jedinců)

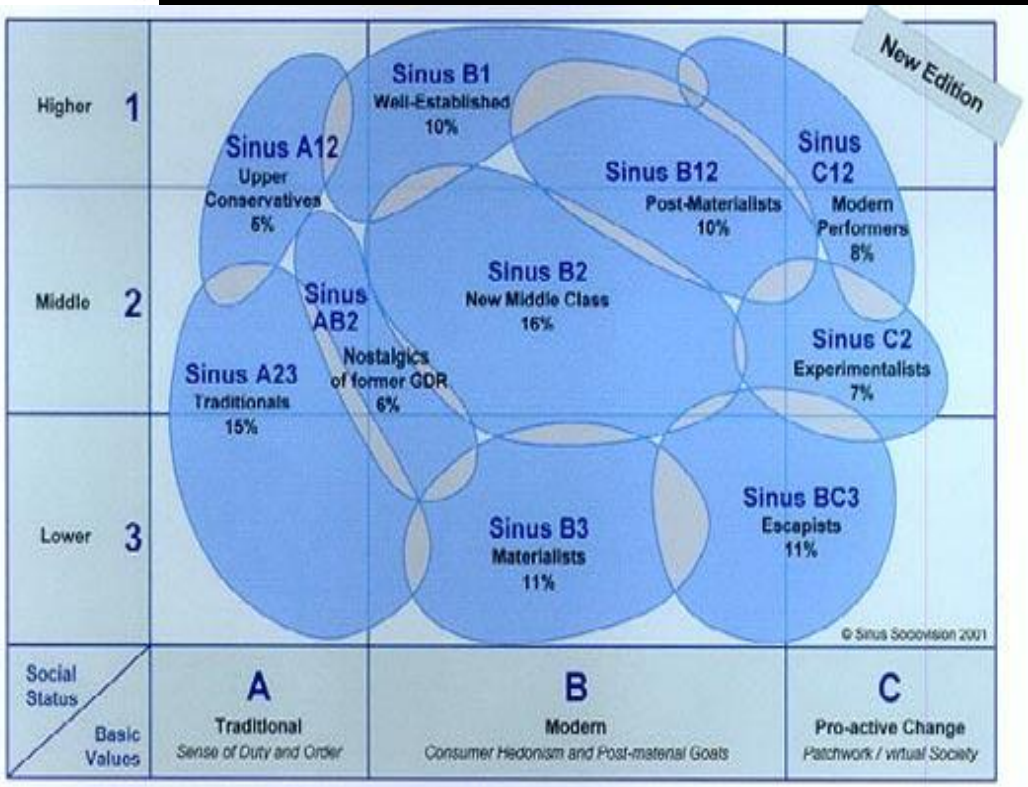
  - práce s geny - seskupení genů na základě určitých znaků-často se používá při seskupování bílkovin(enzymů)

  - seskupování stejných genů do rodin genů – velmi důležité ve vývoji genového inženýrství(např. pro klonování, rozmnožování,…)

- Medicína – rozbory krve, genové inženýrství
- Logistika – seskupování produktů v logistických centrech
- Chemie – nalézt podobnost struktur
- Trh – při výzkumech - rozdělení populace do určitých segmentů trhu(lepší porozumění vztahům mezi jednotlivými typy zákazníků)
  - při určování polohy výrobků
  - vývoj nových výrobků
- Sociální sítě
  - rozpoznání sociálních skupin v populaci






New Edition
Higher 1
Middle 2
Lower 3
Sinus A12
Upper Conservatives
5%
Sinus B1
Well-Established
10%
Sinus B12
Post-Materialists
10%
Sinus C12
Modern Performers
8%
Sinus AB2
Sinus B2
New Middle Class
16%
Nostalgics of former GDR
6%
Sinus A23
Traditionals
15%
Sinus C2
Experimentalists
7%
Sinus B3
Materialists
11%
Sinus BC3
Escapists
11%
© Sinus Sociovision 2001
Social Status
Basic Values
A
Traditional
Sense of Duty and Order
B
Modern
Consumer Hedonism and Post-material Goals
C
Pro-active Change
Patchwork / virtual Society

Zdroje

- Data clustering: A review; Jain, A.K. a další
- Shluková analýza;Kelbel,J.;Šilhán,D.
- http://en.wikipedia.org/wiki/Cluster_analysis#k-means_clustering