

Metody dolování znalostí z dat

Jana Šarmanová

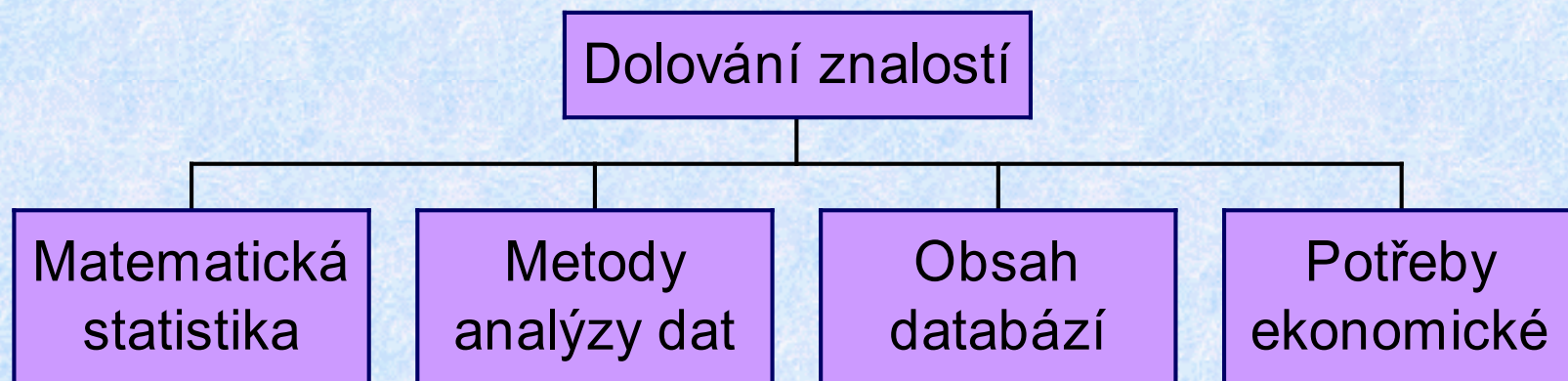
Katedra informatiky FEI VŠB-TU Ostrava

jana.sarmanova@vsb.cz

Obsah

- Dolování znalostí a jeho využití
- Předzpracování - filtrace a transformace
- Analýzy - metody, algoritmy, příklady, prezentace výsledků
- Interpretace a závěr

Dolování znalostí jako multioborová disciplína



Znalosti získané z dat a databází

Data dána maticí \mathbf{X} popisující množinu objektů $\mathbf{O} = \{O_1, O_2, \dots, O_n\}$ zadaných pomocí atributů $\mathbf{A} = \{A_1, \dots, A_m\}$ s doménami $\mathbf{D} = \{D_1, \dots, D_m\}$.

RČ	jméno	věk	váha	výška	skok_vys	...

Užitečné informace z dat získáme

- z dat účelově nasbíraných pro výzkum metodami statistiky, analýz
- z informačních systémů připravenými fcemi a příležitostnými dotazy
- z datových skladů připravenými typy fcí - globální, agregované údaje
- metodami dolování - rozptýlené další dosud nepoznané zákonitosti

Definice procesu dolování znalostí

Dolováním znalostí nazýváme proces netriviálního získávání implicitní, dříve neznámé a potencionálně užitečné informace z dat.

Rozdělení potřeb dolování z hlediska typů uživatelů:

průzkum - marketing, bankovníctví, výroba, pojišťovnictví, ...

(získání obchodních výhod)

výzkum – medicína, biologie, hutnictví, ...

(získání nových odborných znalostí, hypotéz)

sociologický průzkum – veřejné mínění, sčítání lidu, ...

(získání "politických" výhod)

Potřeby uživatelů metod dolování znalostí

Nároky na data, proces dolování, míru spolehlivosti výsledků, prezentaci

		Marketing	Výzkum	Sociolog
DATA	Zdroj	Databáze	Sběr + databáze	Sběr
	Rozsah	Velký	Menší	Malý
	Přírůstky	Časté	Řídké – žádné	Žádné
	Struktura	Stálá	Různá	Různá
ZPRAC	Předzpracování	Automatické	Na míru	Žádné
	Analýza potřeb	Jednorázová	Na míru	Standardní
	Rychlost	Vysoká-on line	Menší	Menší
	Úplnost, kvalita	Informativní	Vysoká	Informativní
	Výstupy	Graf, tabulka, text	Pracovní	Graf, tabulka, text
	Výsledky	Aktuální	Dlouhodobé	Aktuální
UŽIV	Primární	Manažer	Výzkumník	Sociolog
	Sekundární		Odborná veřejnost	Veřejnost

Životní cyklus procesu získávání znalostí z dat

- formulace problému
- datová a problémová analýza
- výběr (sběr) relevantních dat
- předzpracování
 - integrace do jednotného formátu,
 - transformace a odvozování dat
- dolování nových hypotéz
- interpretace výsledků
- využití a zhodnocení celého procesu

Metody předzpracování dat

filtrace, integrace

transformace

odvozování

Metody předzpracování dat

Filtrace a integrace dat

- výběr atributů vhodných k analýzám
- ošetření nebo vyloučení dat chybných, chybějících, redundandních, irelevantních, konstattních
- sjednocení formátů, měrných jednotek
- numerické zakódování některých dat, sjednocení kódování
- kategorizace a dichotomizace dat

Výsledkem jsou **data numerická**, formálně i věcně správná, konzistentní.

Numerická data – reálná, kategoriální.

Metody předzpracování dat

Transformace dat

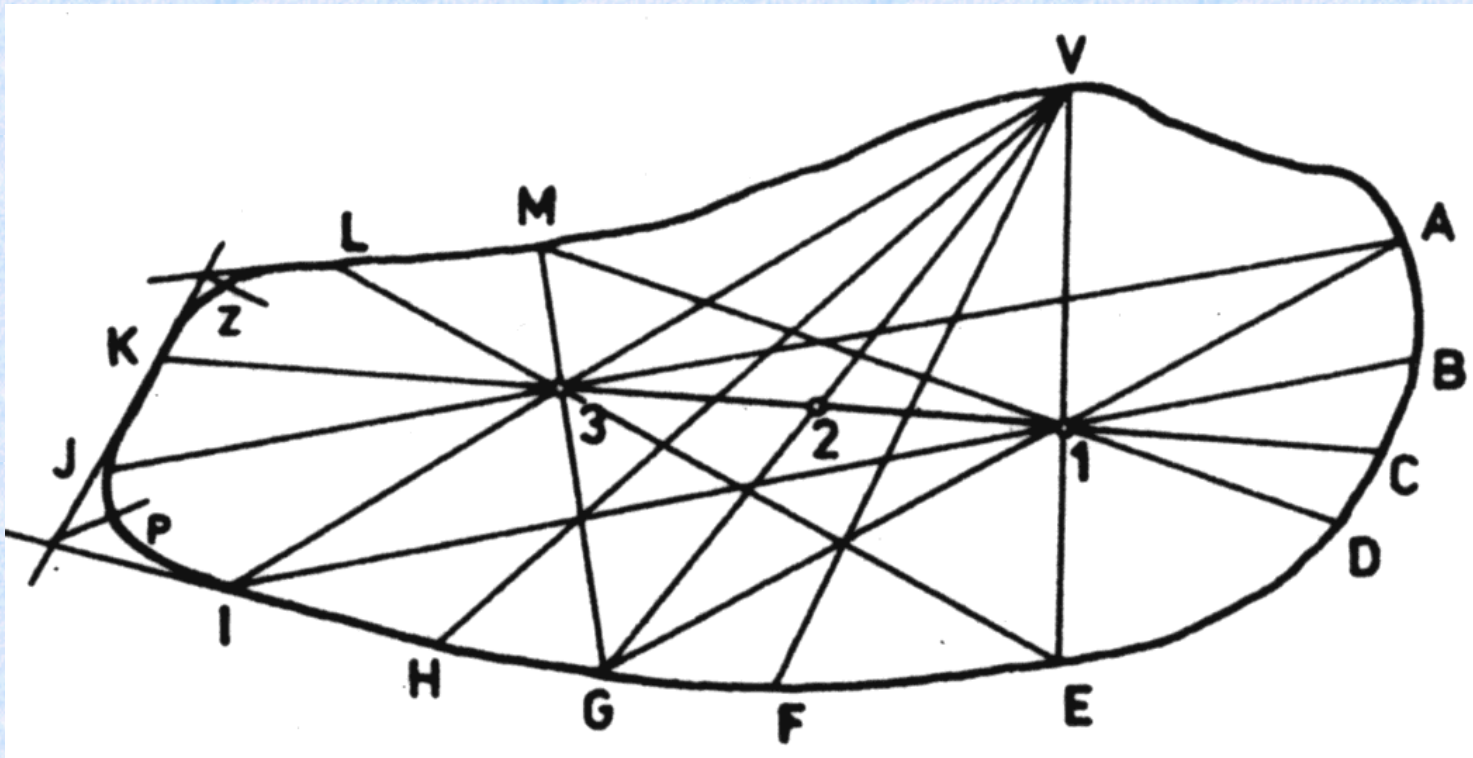
- standardizace atributů - odstranění závislosti reálných atributů na jednotkách měření
- normalizace objektů - odstranění závislosti na velikosti objektu
- hlavní komponenty

Odvozování dat

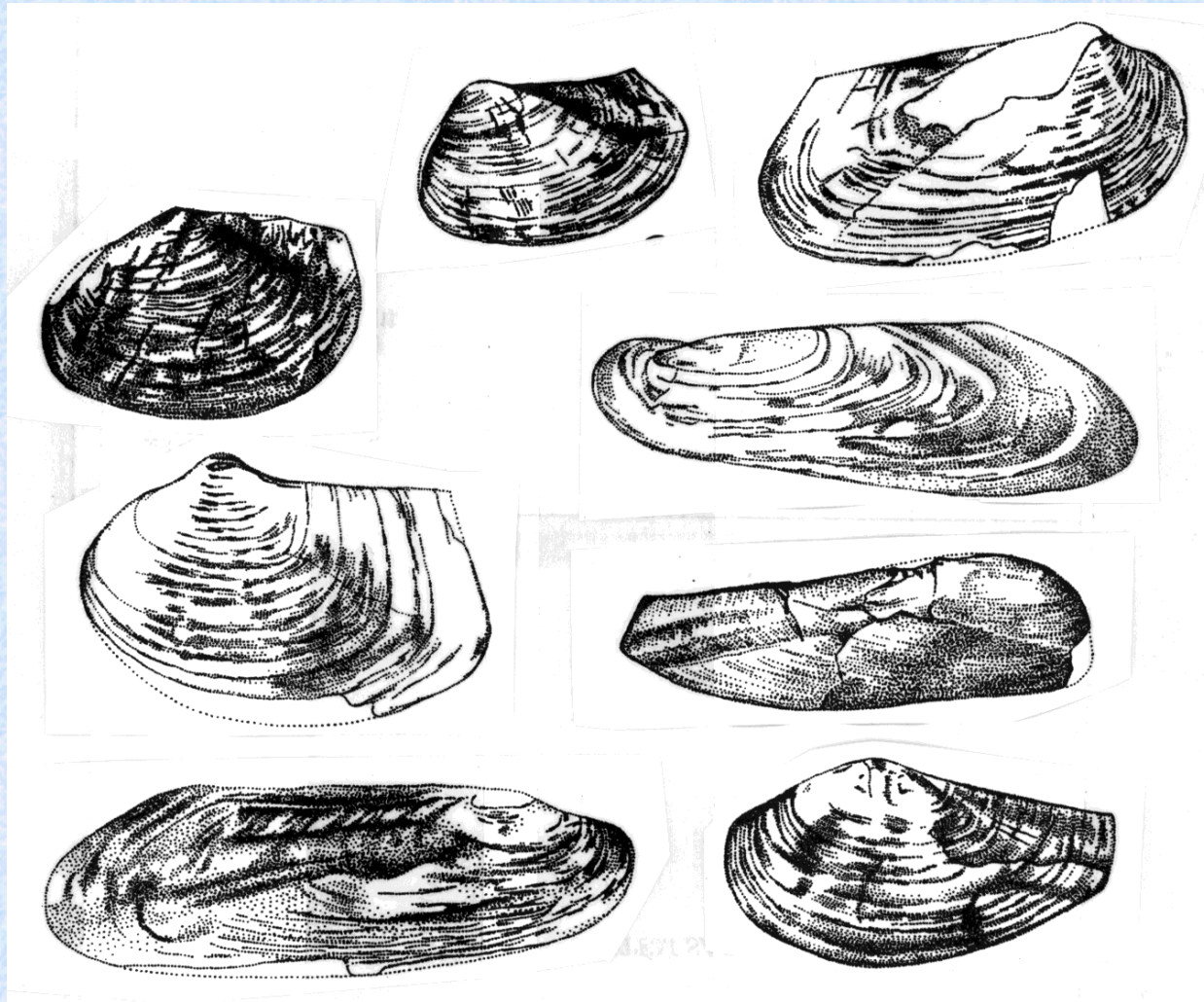
- odvozené atributy
- agregované údaje

Transformace dat - normalizace

Příklad: objekty jsou mlži několika rodů, hledají se tvarově blízké množiny objektů, popsaných 25 rozměry (v mm) dle obrázku.



Transformace dat - normalizace



Hlavní komponenty

Datová matice může obsahovat závislé atributy.

Předpoklad: dvě proměnné mohou korelovat proto, že obě měří tutéž skrytou společnou veličinu (latentní proměnná, společný faktor, hlavní komponenta).

Hlavní komponenty lze matematicky zkonstruovat.

Výsledky se mohou využít

1. ke klasifikaci atributů, spojenou s definováním skupinových proměnných aglomerováním původních proměnných,
2. k redukci počtu původních proměnných a tak ke zjednodušení popisu objektů,
3. ke zprostředkovanému měření nepřímo měřitelných proměnných a jejich odhadu,
4. k transformaci původních proměnných do výhodnějšího tvaru, spojené s jejich ortogonalizací.

Hlavní komponenty

Příklad: 104 plavkyně testovány ve 4 stylech a 2 délkách tratě. Úkolem zjistit, zda „plavecká schopnost“ je souhrn více nezávislých vlastností nebo jediná.

Data plav (50 kraul, 200 kraul, 50 prsa, ..., 50 znak, ... 50 delfín, ...)

Výsledek

$$f1 = 0.88 x_1 + 0.82 x_2 + 0.27 x_3 + 0.20 x_4 + 0.77 x_5 + 0.76 x_6 + 0.70 x_7 + 0.57 x_8$$

$$f2 = 0.19 x_1 + 0.26 x_2 + 0.78 x_3 + 0.77 x_4 + 0.21 x_5 + 0.35 x_6 + 0.40 x_7 + 0.48 x_8$$

Pojmenování prvních dvou hlavních komponent:

f1 ...schopnost plavat převážně pomocí paží

f2 ...schopnost plavat převážně pomocí nohou

Metody dolování znalostí

asociace

shlukování

rozhodovací stromy

Asociace

Asociace

Asociacemi nazýváme vztahy mezi atributy:

- **klasické** - mezi dvěma podmnožinami atributů
- **transakční** - v rámci množiny atributů
- **agregované** - mezi podmnožinou atributů a jejich charakteristikami

Podstata rozdílu proti matematické statistice v automatizaci generování a testování všech možných případů.

Výsledkem jsou takové vygenerované hypotézy, pro něž platí

$S \geq \text{minconf}$ (minconf je zadaná minim spolehlivost, confidence)

$P \geq \text{minsupp}$ (minsupp je zadaná minim podpora, support)

kde

S ... spolehlivost vztahu definována vhodnými charakteristikami

P ... podpora je počet případů, pro něž je nalezená hypotéza splněna

Asociace základní

Asociacemi jsou vztahy mezi dvěma podmnožinami atributů (antecedentem a sukcedentem).

Je-li $A=a \wedge B=b \dots$, pak $X=x \wedge Y=y \dots$ se spol S a podp P

Hodn $A=a \wedge B=b \dots$ souvisí s $X=x \wedge Y=y \dots$ se spol S a podp P

kde A, B, \dots atributy jedné podmnožiny - antecedentu

X, Y, \dots atributy druhé podmnožiny - sukcedentu

a, b, \dots, x, y, \dots jsou hodnoty domén příslušných atributů

$S \dots$ spolehlivost vztahu definována vhodnými charakteristikami

$P \dots$ podpora je počet případů, pro něž je nalezená hypotéza splněna

Generovanými hypotézami jsou vztahy, pro něž platí

$S \geq \text{minconf}$ (minconf je zadaná minim spolehlivost, confidence)

$a \geq \text{minsupp}$ (minsupp je zadaná minim podpora, support)

Asociace základní

Příklad: Matky, které daly své dítě k adopci si to někdy později rozmyslí. 104 matek se 23 atributy (m-věk, m-povolání, m-stav, m-národ, o-stav, o-vztah, o-povol, o-národ, poč-těhot, poč-potrat, ... , kojení, ... ⇒ rozmys)

JE-LI	PAK	SPOL	PODP
m-věk < 30	⇒ rozm	66%	62
m-věk <30,40>	⇒ nerozm	100%	28
m-věk > 40	⇒ nerozm	91%	14
m-povol = dělnice, pomocná	⇒ nerozm	100%	19
o-povol = není ve vězení	⇒ nerozm	90%	86
o-vztah = ženatý jinde	⇒ nerozm	100%	67
poč-těhot > 3, poč-potrat = 0	⇒ nerozm	100%	22
útěk = ano	⇒ nerozm	100%	17
steril = ano	⇒ nerozm	100%	7
m-povol = prostituce	⇒ nerozm	100%	15
= ústavní výchova	⇒ nerozm	100%	21
= podvod s OP	⇒ nerozm	100%	5
...			

Algoritmy generující asociace

Algoritmy pro testování a generování hypotéz můžeme dělit na

- **triviální** - generují a testují všechny možné sentence jako kombinace hodnot atributů, délek ante- a sukce-, kombinací dvojic ante- a sukce- (odtud též kombinační analýza s exponenciální časovou složitostí, pro větší data nepoužitelná),
- **uspořádané generování pravidel** – například vhodně uspořádané postupné prodlužování délky sukce- a ante- snižuje počet kladných shod a , pokud $a < minsupp$, negenerují se další sentence s větší délkou –cedentu,
- **vzorkováním** - rozsáhlá data zpracují po částech (vzorcích) a hledají kandidáty pro hypotézy, pak testují přes celá data jen kandidáty.

Asociace transakční

Data jako nákupní košík: objektem je jeden (obchodní) případ, ten obsahuje několik atributů s pevnou strukturou (identifikace košíku), mnoho dalších atributů (obsah košíku).

A	B	...	X – seznam ($i \gg 1$)
a1	b1	...	$X1 = \{x1, x2, x3, x4, x5\}$
a2	b2	...	$X2 = \{x3, x8\}$
a3	b3
...			

Transakční asociací nazýváme jeden z typů tvrzení

$\{x1, x2, \dots\} \in Xi$ se spol V a podp P

Je-li $A=a \wedge B=b \wedge \dots$ **pak** $\{x1, x2, \dots\} \in Xi$ se spol V a podp P

tedy nalezené podmnožiny atributů, nacházejících se společně v košíku.

Asociace transakční

Příklad

Databáze o prodeji v supermarketu, vydolování "znalostí" typů

{jogurt, vločky} se spol 72% a podp 4533

je-li den = pátek \wedge období = léto, pak {pivo, buřty}
se spol 67% a podp 3367

V prvním případě se uvedená zboží umístí v prostoru co nejdále od sebe, ve druhém se v pátek vedle piva a buřtů umístí další víkendová lákadla.

Algoritmy

Speciální typy konstrukcí k-množin, ...

Asociace agregované

Agregovanou asociací nazveme neprůměrný vztah mezi hodnotami množiny antecedentových atributů $\{A, B, C \dots\}$ a skupinovými ukazateli $\{U_1, U_2, \dots\}$ vypočtenými z atributů sukcedentu $\{X, Y, \dots\}$ tvaru

Je-li $A=a \wedge B=b \dots$, pak $U_1=u_1(V_1) \wedge U_2=u_2(V_2) \dots$ s podp P

kde

A, B, \dots, a, b, \dots mají stejný význam jako u klasických asociací
 U_1, U_2, \dots identifikátory definovaných agregovaných ukazatelů,
 u_1, u_2, \dots vypočtené hodnoty ukazatelů pro testovanou skupinu,
 V_1, V_2, \dots označují pod- nebo nadprůměrnost ukazatelů U_i
 $P \dots$ počet výskytů skupiny

Asociace agregované

Příklad - porovnání výsledku klasických a agregovaných asociací

<u>Jestliže</u>	<u>pak</u>	<u>se spol</u>	<u>s podp</u>
TR provedl gynekol =A	výsledek gravid = netěh	87%	682
B	netěh	80%	298
C	netěh	87%	447
D	netěh	84%	659

<u>Jestliže</u>	<u>pak</u>	<u>s podp</u>
TR provedl gynekol =A	GR/ET = 13% -	682
B	28% .	298
C	13% .	447
D	23% ++	659

Algoritmy obdobné jako pro klasické asociace

Shluková analýza

Shluková analýza

- analyzuje, zda se množina objektů přirozeně rozpadá na výrazné podmnožiny (**shluky**) objektů si podobných a přitom nepodobných objektům podmnožin ostatních
- případně dále analyzuje, jestli existuje celá hierarchie takových rozkladů
- pokud shluky existují, čím jsou charakteristické
- jak se případné další objekty zařadí do již definovaných shluků

Shluková analýza

Shluková analýza **netvoří ucelenou teorii**, ale je to **řada metod** založených na různých principech (různorodost řešených problémů, požadovaných typů výsledků, velká data, neurčitost definice shluku).

Metody dle cíle shlukování

- **nehierarchické**, produkující prostý rozklad objektů na podmnožiny
- **hierarchické**, produkující hierarchii rozkladů, kde každý rozklad je zjemněním předcházejícího

Metody dle typu výsledných shluků

- shluky **kulové**, body soustředěné kolem svého těžiště
- shluky **obecné** tvoří souvislé husté oblasti nejrůznějších tvarů

Metody dle typu rozkladu

- shluky disjunktní
- shluky překrývající se

Shluková analýza

Problémy - řešení

- **výběr atributů charakterizujících podobnost**
- **podobnost a vzdálenost objektů**
 - koeficienty korelace, asociace
 - metriky (Eukleidovská vzdálenost)
- **pojem shluku – geometrický model**
- **počet shluků rozkladu**
- **počáteční rozklad – typické objekty**
- **pojem vzdálenosti shluků**

Shluková analýza

Definice 1: Je dána množina objektů $O = \{O_1, \dots, O_n\}$ a koeficient vzdálenosti objektů V . Shlukem nazveme takovou podmnožinu $X \subseteq O$, pro niž platí

$$\max_{O_i, O_j \in X} V(O_i, O_j) < \min_{O_i \in X, O_k \notin X} V(O_k, O_i)$$

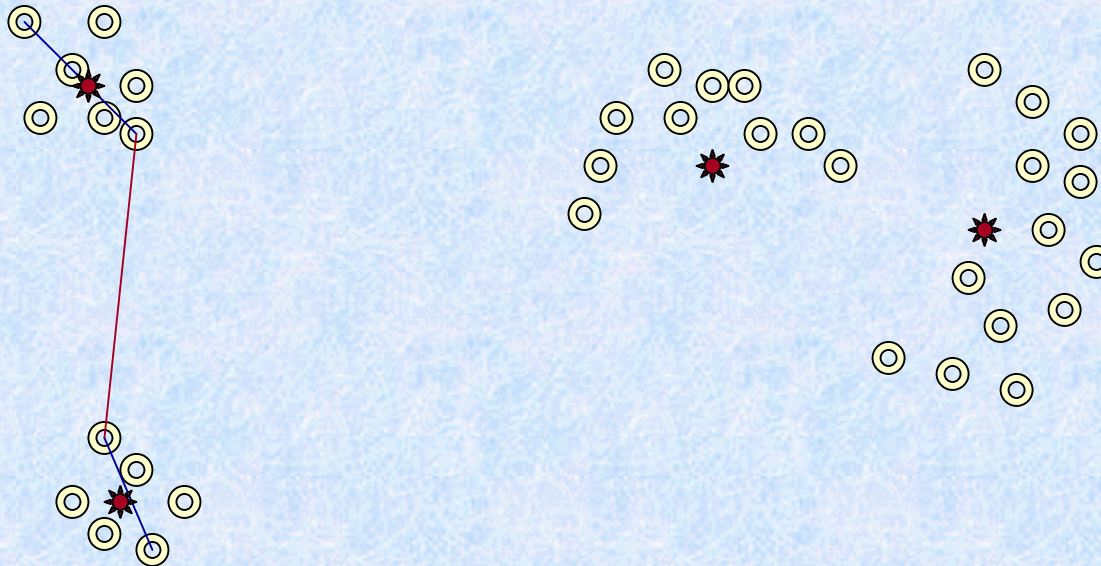
Definice 2: Je dána množina objektů $O = \{O_1, \dots, O_n\}$ a koeficient vzdálenosti objektů V . Objekt O_x nazveme α -souvislým s objektem O_y pro daný práh α , existuje-li řetěz objektů $O_x = O_1, O_2, \dots, O_k = O_y$, $k > 1$, takový, že $V(O_i, O_{i+1}) \leq \alpha$ pro $i = 1, \dots, k-1$.

α -souvislým shlukem (α -shlukem) nazveme takovou podmnožinu $X \subseteq O$, pro niž platí že

- každý pár objektů z X je α -souvislý,
- žádný objekt z $O - X$ není α -souvislý s žádným objektem z X .

Shluková analýza

Geometrický model 2-rozměrných dat



Shluková analýza - metody nehierarchické optimalizační

K-středové metody **optimalizační** hledají nejlepší rozklad množiny objektů iteračním způsobem.

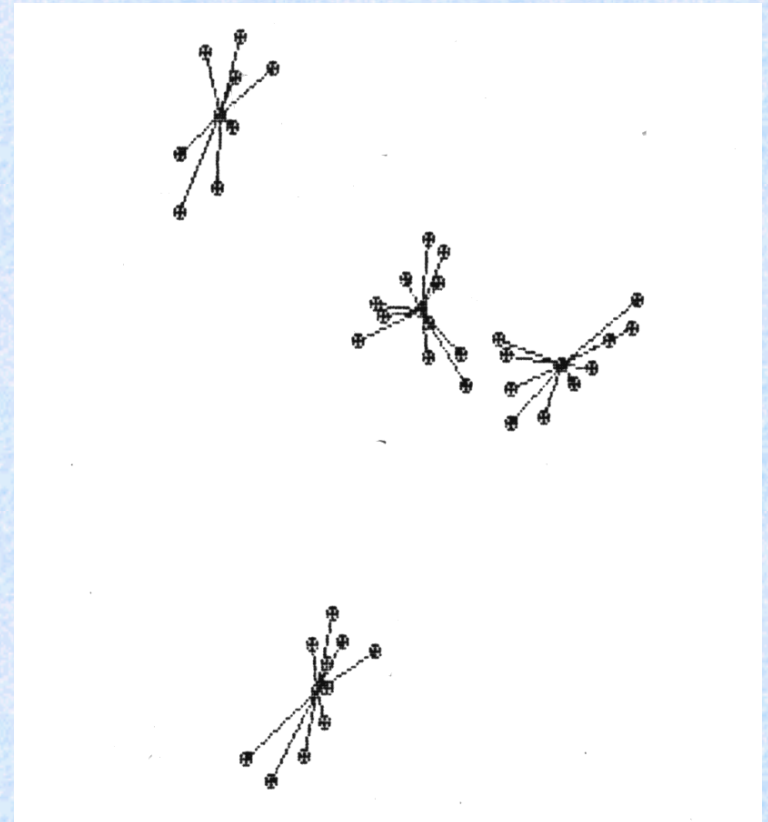
Počáteční rozklad (zadaný nebo vygenerovaný) zlepšují tak, že hledají rozklad s lepší hodnotou **kriteriální funkce**.

Algoritmus

1. zadání k počátečních typických bodů
2. přiřazení každého bodu **k nejbližšímu typickému** bodu a jemu odpovídajícímu shluku
3. výpočet **těžiště** každého z k shluků
4. definování **nových typických bodů** ve vypočtených těžištích
5. pokud došlo ke změně v přiřazení bodů shlukům, opak. od bodu 2.
6. výpočet **kriteriální funkce** výsledného rozkladu

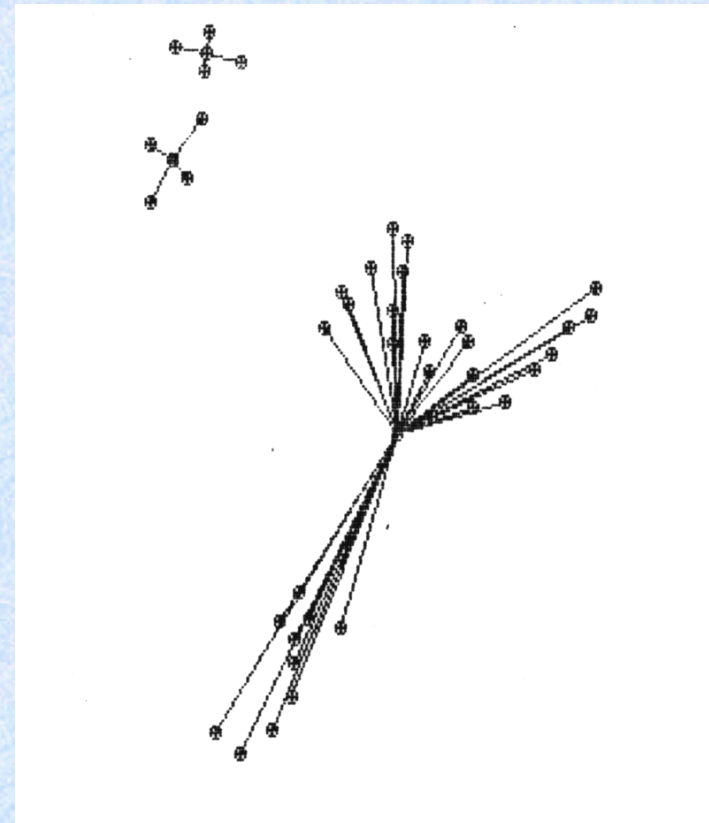
Shluková analýza

Příklad: 3 shluky, metoda K-středová, 4 typické body



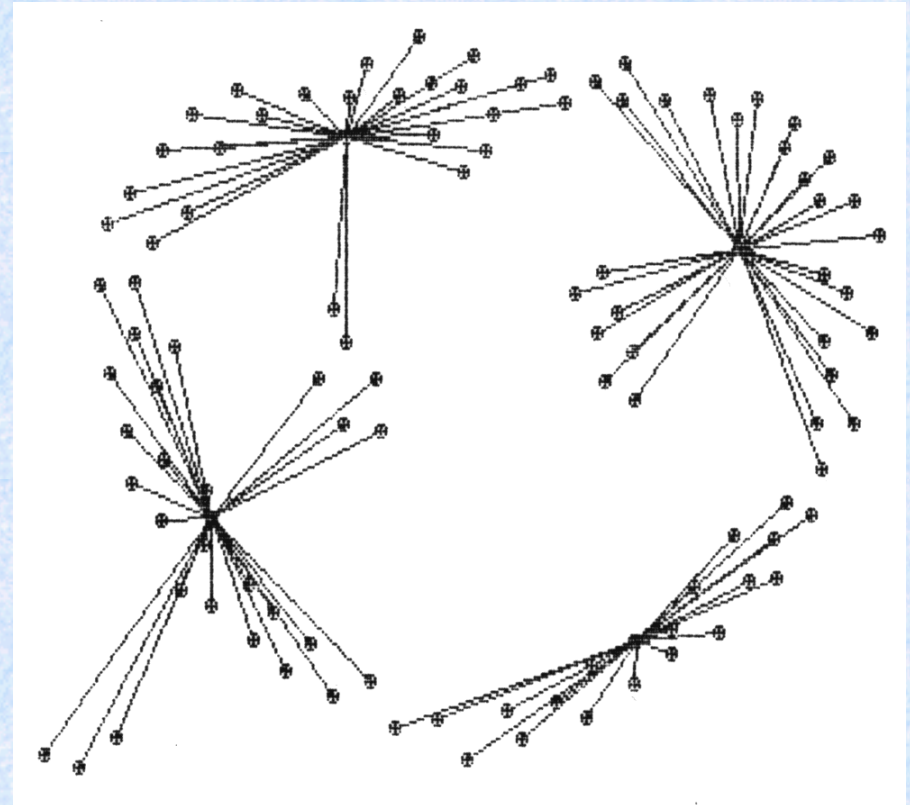
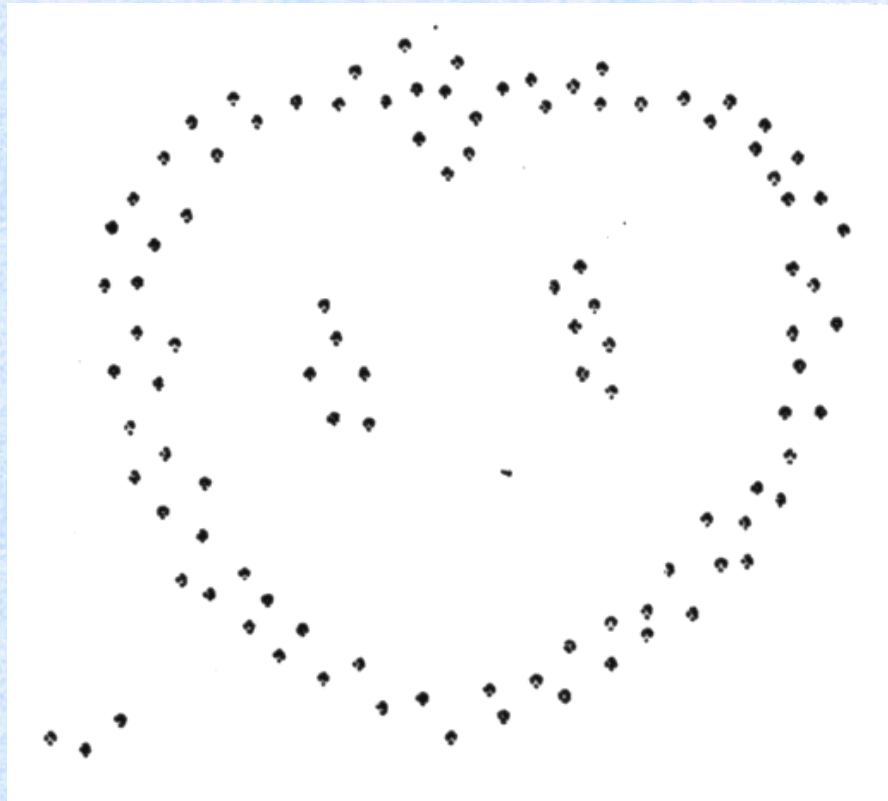
Shluková analýza

Příklad: 3 shluky, metoda K-středová, 3 typické body



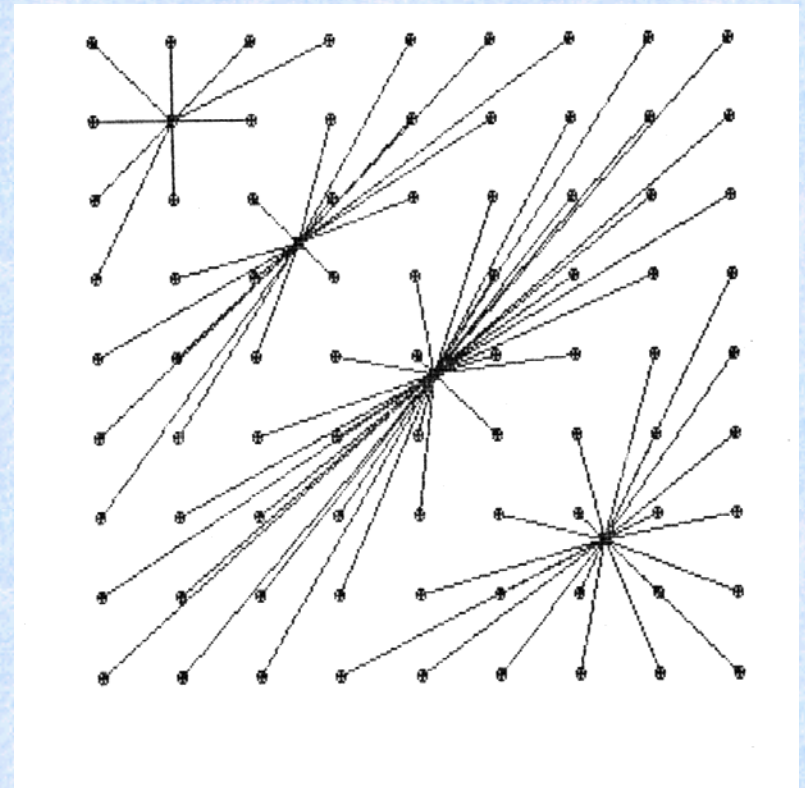
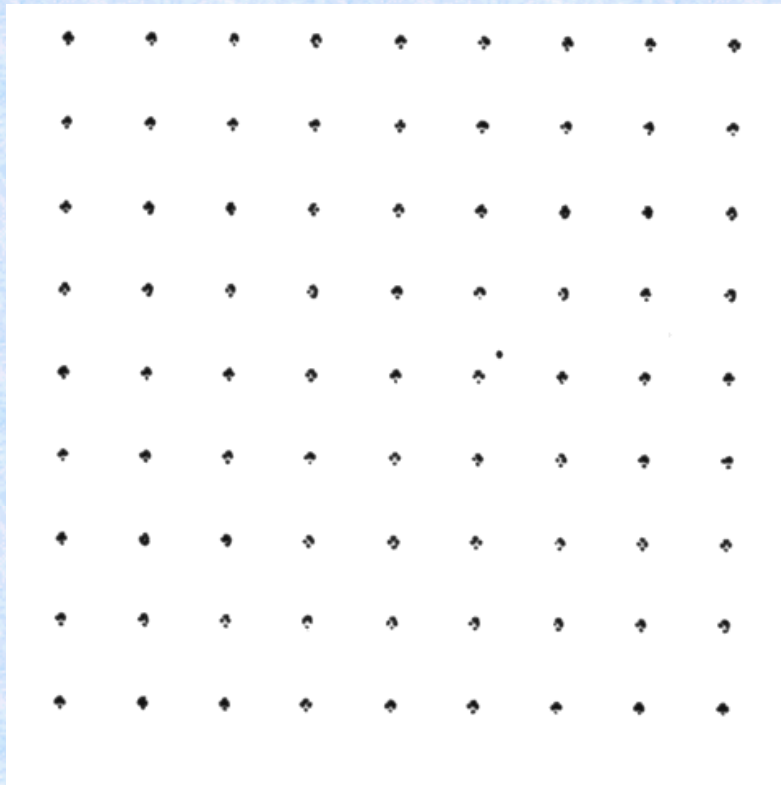
Shluková analýza

Příklad: 4 obecné shluky, metoda K-středová, 4 typické body



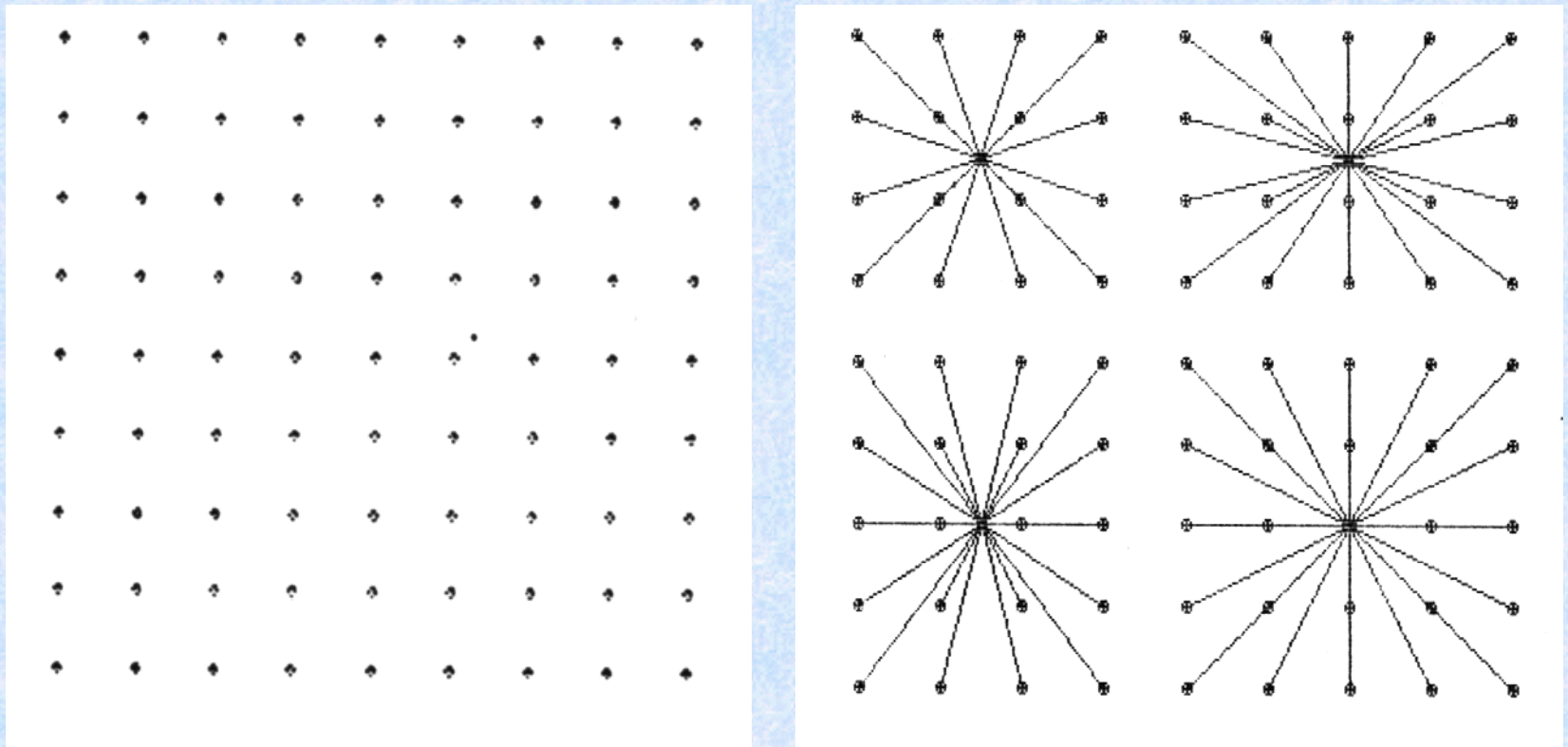
Shluková analýza

Příklad: homogenní množina, metoda K-středová, 4 typické body



Shluková analýza

Příklad: homogenní množina, metoda K-středová, 4 typické body



Shluková analýza – metody hierarchické

Algoritmus

1. výpočet matice podobnosti objektů, počáteční rozklad tvoří jednoobjektové shluky
2. nalezení nejmenší **vzdálenosti shluků** v aktuální „hladině“ hierarchie
3. spojení těchto nejbližších shluků do společného shluku vyššího stupně hierarchie, ostatní shluky zůstanou nezměněny
4. výpočet charakteristik shluků aktuální hladiny rozkladu
5. pokud existuje více než 1 shluk, opakování od bodu 2.

Problém

- **pojem vzdálenosti shluků**

Shluková analýza – metody hierarchické

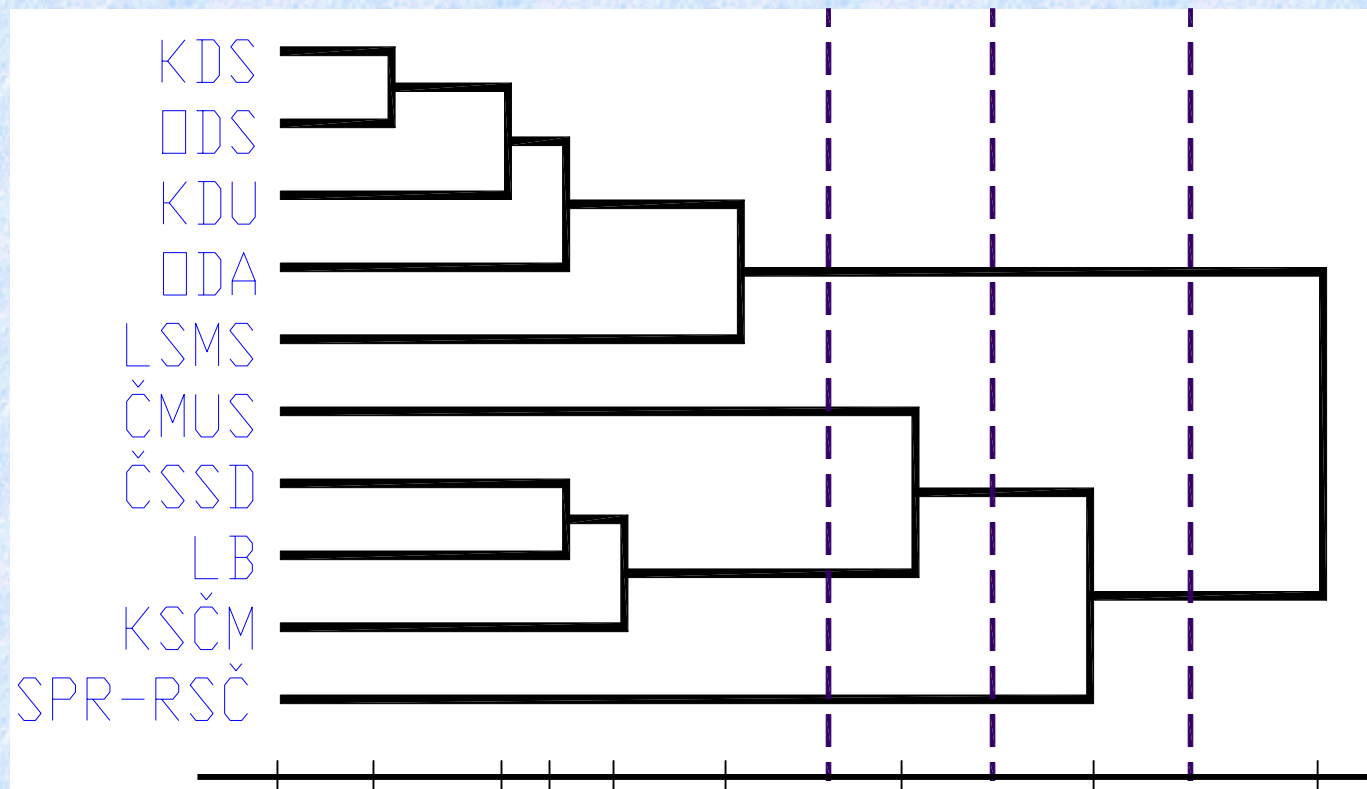
Vzdálenosti shluků

- **strategie nejbližšího souseda**
- **strategie nejvzdálenějšího souseda**
- **strategie průměrné vzdálenosti objektů**
- **strategie mediánová**
- **strategie centroidní**
- **strategie Ward-Wishartova**

Shluková analýza

Příklad zobrazení výsledku hierarchie shluků dendrogramem.

Hlasování parlamentu ČR v letech 1995-96, 10 stran, 2049 hlasování



Shluková analýza

Algoritmy

- **nehierarchické** **optimalizační k-středové**
analýzy módů
fuzzy k-středové
neuronové sítě – Kohonenovy mapy
- **hierarchické** **aglomerativní**
divizivní
- **vzorkování**

Rozhodovací stromy

Rozhodovací stromy

Množina objektů O_1, \dots, O_n zadaná atributy A_1, \dots, A_m s doménami D_i je rozdělena do klasifikačních tříd C_1, \dots, C_k . Třídy jsou charakterizovány hodnotami jednoho nebo více atributů, nazývaných předpovídáními či klasifikačními. Hledá se, při jakých hodnotách atributů předpovídajících či vstupních, případnou objekty do těchto tříd.

A	B	...	X	Y	...	C
a1	b1	...	x1	y1		c1
a2	b2	...	x2	y2		c2
a3	b3	...	x3	y3		c3
...						

RS je způsob reprezentace pravidel charakterizujících toto rozdělení typu

if $A=a$ **and** $B=b$ **and** ... **then** $O_i \in C_j$

Rozhodovací stromy

Konstrukce RS je založena na informačním zisku při rozšíření stromu o další uzly.

Označme $fr(C_j, D)$ počet výskytů objektů třídy C_j v množině D ,
 $|D|$ nebo $|D_i|$ počet objektů v množině D nebo D_i .

Entropie množiny D

$$info(D) = - \sum (fr(C_j, D) / |D| * \log_2 (fr(C_j, D) / |D|))$$

Vážený součet entropií množin rozkladu nad atributem A

$$info_A(D) = \sum |D_i| / |D| * info(S_i)$$

Informační zisk rozkladu nad atributem A

$$gain(A) = info(D) - info_A(D)$$

Pro rozklad se vybere atribut, jehož rozklad dává největší zisk informace.

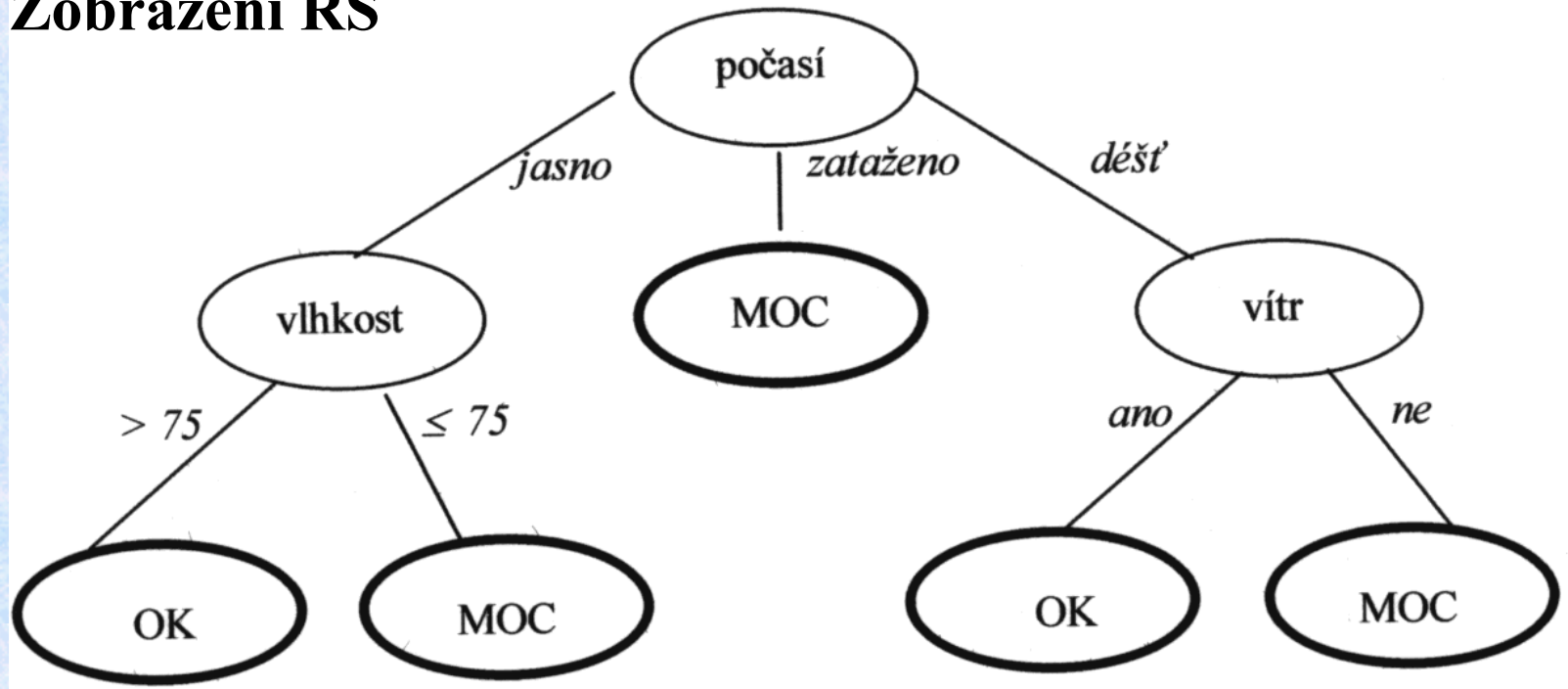
Rozhodovací stromy

Údaje o počasí, vlhkosti a větru jako předpovídající atributy, koncentraci oxidu siřičitého v ovzduší jako předpovídáný atribut. Hledáme pravidla, pomocí nichž budeme na základě znalosti hodnot počasí, vlhkosti a větru předpovídat koncentraci SO₂. Podle normy je maximální povolená hodnota SO₂ = 150 μg/m³, vyšší hodnota znamená překročení normy. Na základě toho určíme z předpovídáné hodnoty koncentrace SO₂ dvě třídy MOC a OK.

předpovídající			předpovídáný	klasifikační třída
počasí	vlhkost [%]	větr	koncentrace SO ₂	
zataženo	78	ne	220	MOC
déšť	80	ne	195	MOC
jasno	76	ne	130	OK
jasno	65	ano	200	MOC
déšť	70	ano	115	OK
déšť	80	ano	110	OK
jasno	77	ano	120	OK

Rozhodovací stromy

Zobrazení RS



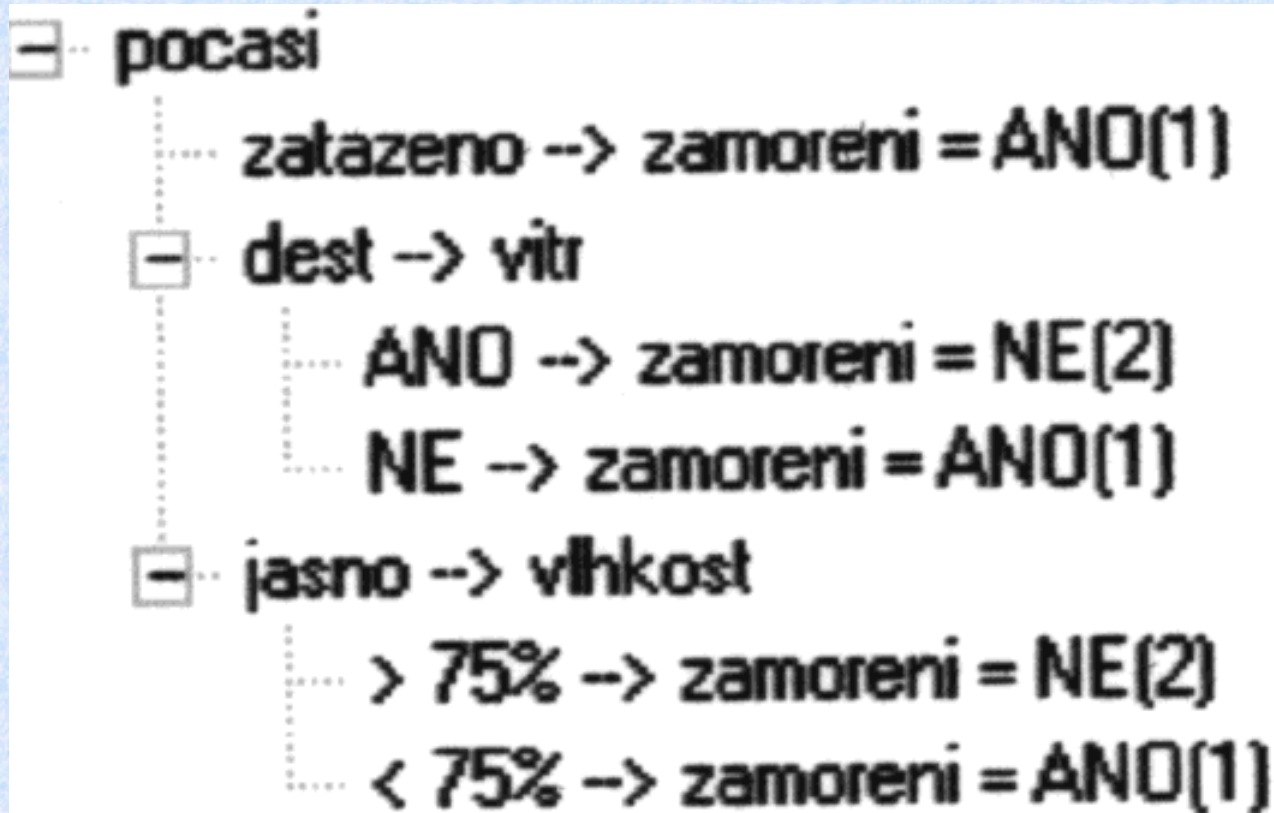
Implicitní třída = MOC

IF počasí=jasno AND vlhkost>75 THEN třída = OK

IF počasí=jasno AND vlhkost>75 THEN třída = OK

Rozhodovací stromy

Jiné zobrazení RS



Optimistický výhled místo závěru

Data numerická

- data rozsáhlá, přírůstky
- vzorky – dvoufázové, reprezentativní výběr
- fuzzy-data

Data nenumerická, multimediální

- textová – strukturovaná, fulltexty, hypertexty, ...
- jednorozměrná grafická, zvuková – signály, sekvence,...
- rovinné obrazy - fotografie, ...
- ... ↓
 data numerická

Děkuji za pozornost

jana.sarmanova@vsb.cz