5. SHLUKOVÁ ANALÝZA

Čas ke studiu: 6 hodin

Cíl Po prostudování této kapitoly budete umět

- popsat problémy shlukovací analýzy,
- popsat typy shluků,
- popsat typy shlukovacích metod a úloh jimi řešitelných,
- pro praktické problémy rozhodnout, zda a která metoda je pro shlukování vhodná,
- pro konkrétní data provést prakticky shlukovací analýzu dat.

Výklad

5.1. Co je shlukování

❑ Klasifikace a shlukování

Shluková analýza zkoumá, zda se množina objektů $\mathbf{O} = \{O_1, O_2, \ldots, O_m\}$ zadaných reálnými atributy $\mathbf{A} = \{A_1, A_2, \ldots, A_n\}$ přirozeně rozpadá na výrazné podmnožiny objektů si podobných a přitom nepodobných objektům shluků ostatních. Pokud takové podmnožiny existují, nazýváme je **shluky.**

Na první pohled jednoduchá úloha skrývá řadu problémů. Neexistuje jednoznačná definice podobnosti objektů a neexistuje ani jednoznačná definice shluku. Jak uvidíme, i zkušený analytik musí vyřešit řadu otázek souvisejících se zadanou množinou objektů, jejich atributy i důvodem shlukování, než vybere vhodnou shlukovací metodu. A stejně jako u jiných metod dolování je možno výsledky shlukování většinou formulovat jen jako hypotézy o klasifikaci zkoumaných objektů.

Shluková analýza tedy netvoří ucelenou teorii, ale skládá se z řady metod, založených na různých principech. Tato různorodost metod souvisí s různorodostí řešených problémů, požadovaných typů výsledků, časovou i prostorovou složitostí shlukovací úlohy pro velká data, neurčitostí definice shluku apod.

Než se začneme podrobněji zabývat shlukováním, zdůrazníme si rozdíl mezi různými typy rozdělování objektů do skupin:

- **klasifikací** nazveme úlohu, kdy je zadáno klasifikační kritérium (*roztřídit danou množinu lidí dle vzdělání*) nebo pravidla pro rozklad objektů,
- **shlukováním** nazýváme úlohu, kdy **kritéria klasifikace nejsou známa** nebo kdy i tato kritéria jsou předmětem našeho výzkumu; nejsou známy ani vzory objektů reprezentujících budoucí podmnožiny; pokud shluky existují, lze je interpretovat jako hledané klasifikační třídy.

Úlohy shlukování

Shluková analýza tedy řeší následující **typy úloh**:

- o množině objektů, zadaných svými atributy, není předem známo nic; úkolem je vyslovit hypotézy o jejich klasifikaci do shluků; toto je klasická úloha shlukové analýzy;
- případně dále rozpoznat, jestli existuje celá hierarchie takových rozkladů;
- pokud shluky existují, popsat, čím jsou charakteristické;
- formulovat pravidla, jak se případné další objekty zařadí do již definovaných shluků;
- jestliže je o množině objektů známo, že tvoří **k** podtříd, případně jsou známi i reprezentanti těchto podtříd, úkolem je nalézt optimální rozklad množiny do takových tříd; zde ne vždy jde o úlohu shlukování;
- pro danou množinu objektů je klasifikace předem známa; úkolem je nalézt tutéž klasifikaci algoritmicky; úloha slouží k výzkumu a ověřování samotných shlukovacích metod, případně k nalezení automatického klasifikátoru.

Historie shlukovací analýzy

Začátek oboru se uvádí do roku 1939, kdy R.C.Tryon, profesor psychologie v Kalifornii napsal první monografii "Shluková analýza".

Hlavní rozvoj oboru spadá do 60.-70. let, kdy byly publikovány mnohé monografie a články.

U nás existuje jediná monografie s roztříděním základních typů existujících obecných metod i s původními novými výsledky z roku 1985.

Průběžně se stále objevují nové algoritmy i aplikace. Aplikace užívají většinou dostupné programy, které jsou součástí velkých programových balíků.

Občas tyto programy nevyhovují, hledají se další metody, většinou si nekladou za cíl formulovat obecný algoritmus, ale jsou specializovány na užší třídu úloh (rozpoznávání objektů při analýze scény v 2D a 3D, objekty se šumem v datech, hledání podobnosti nenumerických dat apod.)

S rozvojem dalších oborů (např. neuronových sítí), se objevují i další přístupy k úloze shlukování.

S rozvojem datových skladů a dolování znalostí z nich nastává rozvoj nových algoritmů pro velmi rozsáhlá data.

Problémy úlohy shlukování

Na rozdíl od asociací nebo většiny jiných metod dolování není úloha shlukování jednoduše použitelná amatérským uživatelem. Množství existujících metod vydává různé typy výsledků. Bez jejich podrobnější znalosti může být představa uživatele o tom, co získal jako výsledek, velmi rozdílná od skutečnosti.

Než tedy přistoupíme k popisu konkrétních metod shlukovací analýzy, musíme nejprve zavést několik pojmů a formulovat, v čem jsou hlavní problémy shlukování.

Základní problémy, vznikající při řešení shlukovací úlohy jsou:

- výběr atributů charakterizujících podobnost objektů,
  měření podobnosti a nepodobnosti (vzdálenosti) objektů
  - koeficienty korelace, asociace
  - metriky
- pojem shluku a jeho geometrický model
- počet shluků rozkladu, počáteční rozklad

- pojem vzdálenosti shluků
- formulace řešené úlohy, důvod shlukování.

Postupně je probereme podrobněji.

❑ Výběr relevantních atributů a podobnost objektů

Chceme-li hledat skupiny podobných objektů v dané množině, musíme objekty nejprve dobře charakterizovat takovými **atributy, podle kterých se podobnost rozezná**. To je úloha společná pro experta a analytika.

Příklad 5.1.

Mějme data o množině lidí. Známe o nich mnoho údajů: o tělesném vzhledu – výška, váha, věk, barva pleti, očí, vlasů, ..., o jejich sportovních výsledcích – skoků, běhů, plavání, ..., o jejich studijních výsledcích – známkách z mnoha předmětů, o jejich zdravotním stavu – tlaku, rozboru krve, ... atd.

Co teď chápeme jako podobné lidi? Jsou si podobnější malí tlustí různě chytří a zdraví, nebo chytří různých pletí a sportovních výkonů, ...?

- *Známe-li o člověku jen jméno, můžeme hledat nejvýše skupiny podobných jmen, ale nebudeme nic vědět o podobnosti osob.*
- *Máme-li údaje o fyzickém vzhledu (věk, výška, váha, míry, barva očí, vlasů, typ pleti atd.), můžeme najít skupiny fyzicky si podobných lidí. Jméno nepotřebujeme, jen k identifikaci osob - stejného účinku dosáhneme přidělením jednoznačného čísla.*
- *Jiného rozdělení dosáhneme z údajů o vzdělání, prospěchu, znalostech, zájmech, výsledcích testů teoretických a praktických, výkonech. Skupiny podobných osob nebudou podobné fyzicky, ale svými schopnostmi.*
- *Prakticky nepoužitelného rozkladu dosáhneme při náhodné volbě atributů jméno, datum narození, barva vlasů a známka ze zeměpisu.*
- *Příliš mnoho aspektů – například všechny tělesné i duševní charakteristiky, jméno, adresa, adresa dědečka atd. rozdrobí osoby do velmi malých skupin a nebudou k použití.*

♦

Při výběru příliš mnoha atributů by skupin s podobnými všemi vlastnostmi bylo zřejmě příliš mnoho a příliš malých a neřešily by asi náš problém.

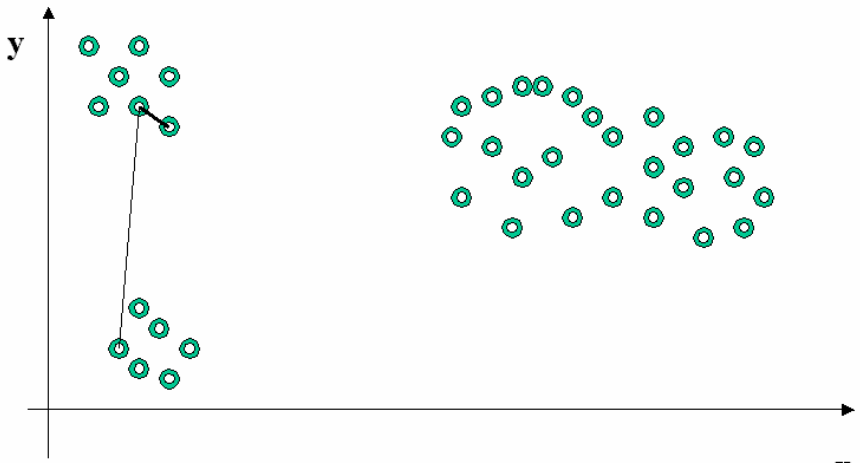
Prvním úkolem tedy je volba, **z jakého hlediska** se mají objekty klasifikovat. V souvislosti s tím se musí vybrat relevantní atributy – řekněme homogenní vzhledem ke zvolenému hledisku. Nad stejnou množinou objektů pak může vzniknout více zadání, každé nad jinou podmnožinou atributů. Výsledky spolu nemusejí souviset a pokud ano, neřeší ji shluková analýza, ale metody hledání asociací v datech.

❑ Koeficienty podobnosti a vzdálenosti

Při hledání podobných objektů vycházíme z úvahy, že dva **objekty jsou si tím podobnější, čím více atributů pro ně nabývá stejných nebo blízkých hodnot**.

Geometrický model

Pro představu o podobnosti objektů používáme geometrický model dat: každý objekt je zadán vektorem n číselných atributů. Vektor můžeme chápat jako bod v n-rozměrném prostoru a zadané objekty pak modeluje množina bodů (viz obr. x).



Obrázek 5.1. Geometrický model 2-rozměrných dat

Podobnost objektů můžeme chápat pomocí vzdálenosti bodů - čím větší vzdálenost, tím nepodobnější jsou si objekty. Dále budeme někdy místo objektů používat **body**, místo nepodobnosti pojmu vzdálenost.

Většina metod shlukové analýzy používá pojem **míry vzdálenosti** nebo naopak **míry podobnosti** objektů.

Pro vyjádření podobnosti objektů se používají buď koeficienty asociace nebo koeficienty korelace.

Pro měření vzdálenosti objektů (duální pojem k míře podobnosti) jsou používány metriky. Nejznámější a nejpoužívanější metrikou je Eukleidovská vzdálenost, ale používá se i řada dalších.

Míra vzdálenosti je používána častěji, než míra podobnosti.

❑ Měření podobnosti nebo nepodobnosti objektů

❑ Míra podobnosti objektů

Obecně pro vyjádření podobnosti objektů hledáme takový předpis, který by každé dvojici objektů (Oi, Oj) přiřadil číslo P(Oi, Oj), které bude splňovat požadavky:

$$P(Oi, Oj) \geq 0$$

$$P(Oi, Oj) = P(Oj, Oi)$$

$$P(Oi, Oi) = max$$

Problémem může být hodnota maxima pro shodné objekty.

Předpisy používané pro vyjádření podobnosti objektů můžeme rozdělit na dva základní typy: koeficienty asociace a koeficienty korelace.

Příklad 5.2.

Pro objekty s binárními znaky může být mírou podobnosti počet shodných znaků.

♦

❑ Koeficienty asociace

jsou používány pro objekty charakterizované pouze binárními znaky. Asociaci dvou objektů Oi, Oj charakterizujeme frekvenční tabulkou tvaru

O1 \ O2	1	0	
1	a	b	r
0	c	d	s
	k	l	m

kde frekvence a, b, c, d znamenají

a počet hodnot atributů, pro které je $O1(Ai) = O2(Ai) = 1$
b $O1(Ai) = 1 \wedge O2(Ai) = 0$
c $O1(Ai) = 0 \wedge O2(Ai) = 1$
d $O1(Ai) = O2(Ai) = 0$

a, d jsou frekvence shod, b, c jsou frekvence neshod.

Koeficienty asociace jsou definovány pomocí hodnot a, b, c, d. Vyskytuje se jich v literatuře řada, většinou nabývají hodnot z intervalu <0, 1>. Uvedeme si několik z nich:

Jaccardův koeficient

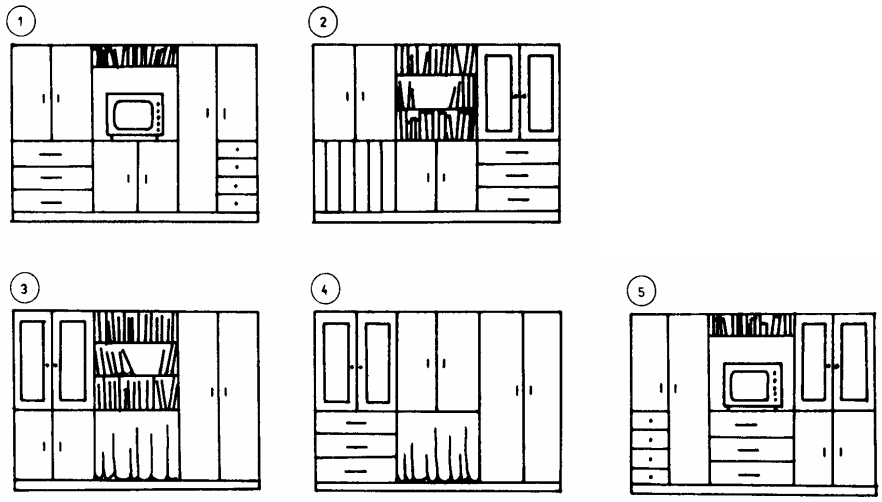
$$A_J = a / (a+b+c)$$

Sokalův koeficient

$$A_S = (a+d) / (a+b+c+d)$$

Příklad 5.3. [Luk13]

5 skříněk bytových stěn s různým vybavením je popsáno binárními atributy, znamenajícími obsahuje/neobsahuje příslušnou skříňku. Jsou to atributy: skříňka dolní, horní, šatník, zásuvky, televizor, knihovna, prádelník, příborník, pořadač, závěs.

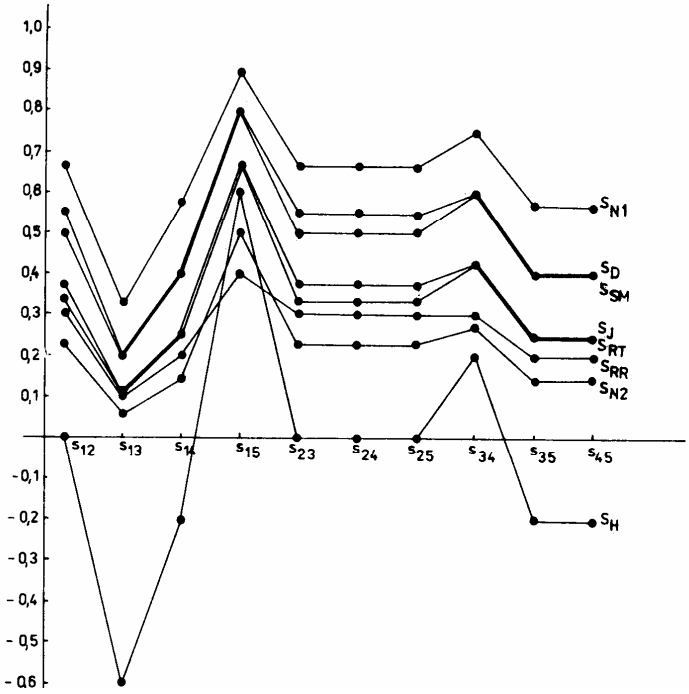


Obrázek 5.2. Skříňkové sestavy

Binární data charakterizující skříňky jsou v matici

skříň	do	hor	šat	zás	tele	kni	prá	pří	poř	záv
O1	1	1	0	1	1	0	1	0	0	0
O2	1	1	0	0	0	1	1	1	1	0
O3	1	0	1	0	0	1	0	1	0	1
O4	0	1	1	0	0	0	1	1	0	1
O5	1	0	0	1	1	0	1	1	0	0

Grafické porovnání hodnot koeficientů asociací skříněk. Vidíme, že hodnoty koeficientů se sice liší svými absolutními hodnotami, ale jejich vzájemné vztahy se téměř neliší.



Obrázek 5.3..Grafické porovnání asociací objektů různými koeficienty asociace

♦

❑ Koeficient korelace

Koeficient korelace objektů Oi = (x_{i1}, . . . , x_{in}), Oj = (x_{j1}, . . . , x_{jn}) s reálnými atributy se určí podle vzorce

$$r_{ij} = \frac{k_{ij}}{s_i \,.\, s_j} \qquad k_{ij} = \frac{1}{n}\sum_{l=1}^{n} x_{il} \,.\, x_{jl} - \bar{x}_i \,.\, \bar{x}_j$$

kde $\underline{x}_i$ a $\underline{x}_j$ jsou střední hodnoty objektů Oi, Oj,
s_i a s_j jsou směrodatné odchylky objektů.

Tento koeficient je použitelný jen v některých případech pro reálné znaky, kdy mají všechny stejnou měrnou jednotku, nebo po standardizaci. Jinak průměr hodnot znaků jednoho objektu nemá reálný smysl.

❑ Míra vzdálenosti objektů

Duálním pojmem k míře podobnosti objektů je míra nepodobnosti neboli **míra vzdálenosti** objektů.

Pro měření vzdálenosti objektů jsou používány metriky.

Metriky vychází z geometrického modelu dat, kde objekty o n znacích chápeme jako body v n-rozměrném Euklidovském prostoru $\mathbf{E_n}$. Pak podobnost objektů můžeme vyjadřovat pomocí vzdálenosti jim odpovídajících bodů.

Metrika V je funkce definovaná na $\mathbf{E_n \times E_n}$, která přiřazuje každé dvojici bodů (Oi, Oj) číslo V (Oi, Oj) takové, že platí:

$$V(Oi, Oj) = 0 \Leftrightarrow Oi = Oj$$
$$V(Oi, Oj) \geq 0$$
$$V(Oi, Oj) = V(Oj, Oi)$$
$$V(Oi, Oj) + V(Oj, Ok) \geq V(Oi, Ok)$$

Některé z používaných metrik:

Eukleidovská vzdálenost - nejznámější metrika

$$\mathbf{V(Oi, Oj) = \sqrt{\left(\sum_{i=1}^{n} (ai - bi)^2 \right)}}$$

Metrika V1 daná předpisem

$$\mathbf{V1(Oi, Oj) = \sum_{i=1}^{n} | ai - bi |}$$

Sokalova metrika

$$\mathbf{V_S(Oi, Oj) = Ö\ (V^2(Oi, Oj) / n)}$$

Sup - metrika V_∞

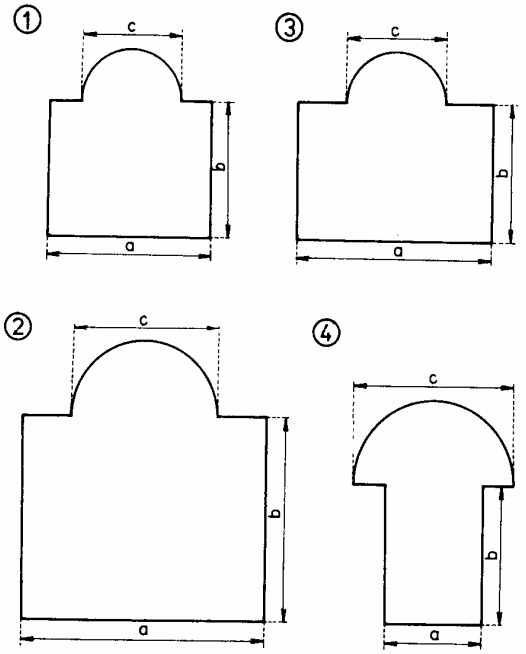
$$\mathbf{V_\infty(Oi, Oj) = \max_{i=1,\ldots,n} \{ | ai - bi | \}}$$

Příklad 5.4. [Lukxx]

K porovnání máme 4 jednoduché obrazce složené z obdélníka o stranách ***a, b*** *a půlkruhu o poloměru* ***c****.*

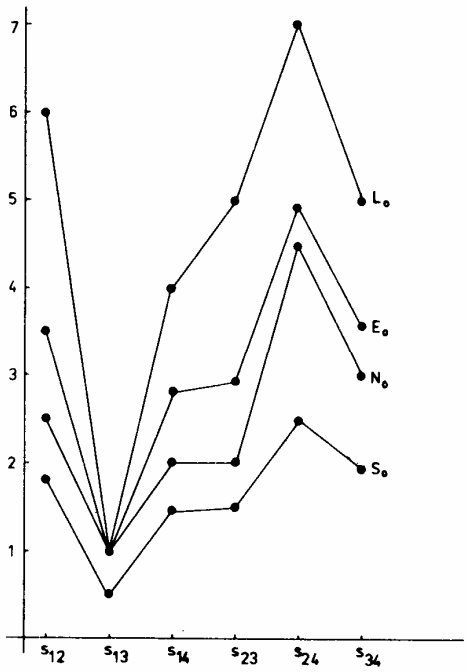
Jejich popis v datové matici je

obraz	a	b	c
O1	5.0	4.0	3.0
O2	7.6	6.0	4.4
O3	6.0	4.0	3.0
O4	3.0	4.0	5.0



Obrázek 5.4. Shlukované obrazce

Grafické porovnání hodnot různých metrik. Opět vidíme, že hodnoty se sice liší svými absolutními hodnotami, ale jejich vzájemné vztahy se téměř neliší.



Obrázek 5.5. Grafické porovnání vzdáleností objektů různými metrikami

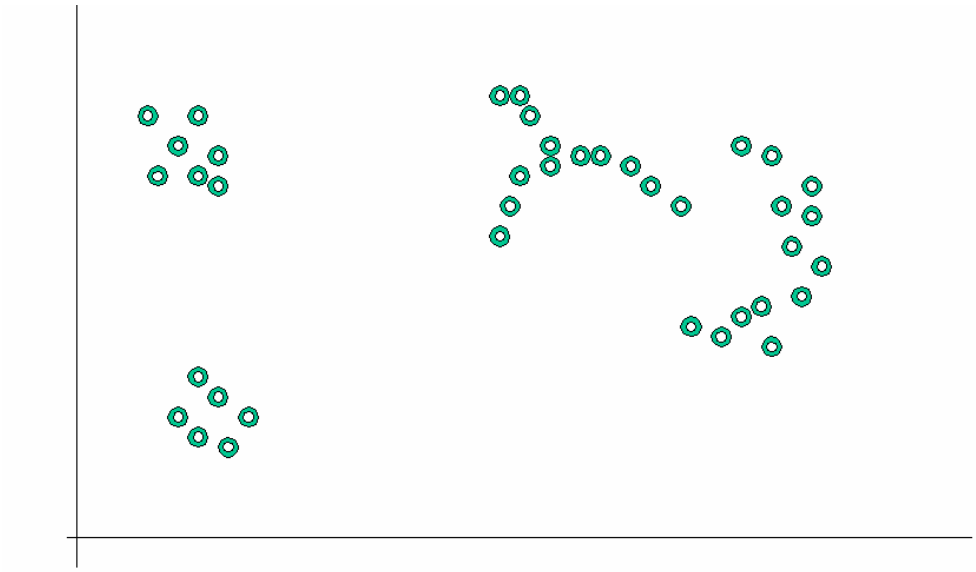
♦

Pojem shluku

Definovali jsme podobnost a vzdálenost dvou objektů, dále je nutné definovat, co je shluk objektů. V literatuře se často shluk uvádí jen vágním popisem, zachycujícím intuitivní představy analytika.

Příklad 5.5.

Podívejme se na několik bodů v rovině na obr. x, které chápeme jako obrazy dvourozměrných objektů s reálnými atributy. O dvou skupinách v levé části obrázku se zřejmě všichni shodnou, že tvoří shluky. Intuitivně chápeme jako shluk body blízko sebe a poměrně vzdálené bodům ostatním. Ovšem body v pravé části obrázku již tak jednoznačné nejsou. Tvoří dva „rozsochaté" útvary, ale není možné o nich říct, že v rámci jednoho útvaru jsou všechny blízko a vzdálené bodům útvarů ostatních. Jsou to tedy shluky nebo ne?



Obrázek 5.6. Skupiny bodů v rovině

Z mnoha definicí shluků v literatuře uváděných se uvedeme jako příklad dvě, reprezentující tato rozdílná chápání pojmu shluk.

Definice 5.1.

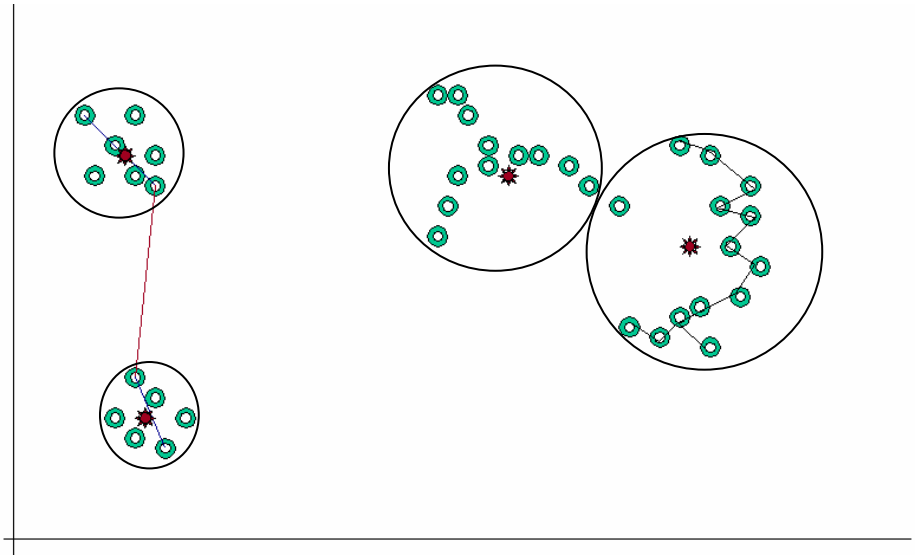
Je dána množina objektů O = {O1, . . . , Om} a koeficient vzdálenosti objektů V. Shlukem nazveme takovou podmnožinu $X \in O$, pro niž platí

$$\max_{Oi,Oj \in X} V(Oi, Oj) < \min_{Oi \in X,\, Ok \notin X} V(Ok, Oi)$$

Definice 5.2.

Je dána množina objektů O = {O1, . . . , Om} a koeficient vzdálenosti objektů V. Objekt Ox nazveme α-souvislým s objektem Oy pro daný práh α, existuje-li řetěz objektů Ox = O1,O2,...,Ok = Oy, k > 1, takový, že $V(Oi, Oi+1) \leq \alpha$ pro i = 1, . . . , m -1.

α-souvislým shlukem (**α-shlukem**) nazveme takovou podmnožinu $X \in O$, pro niž platí, že každý pár objektů z X je α-souvislý a žádný objekt z (O – X) není α-souvislý s žádným objektem z X.



Obrázek 5.7. Skupiny bodů tvořící shluky 2 různých typů

Definici 1 odpovídají skupiny bodů vlevo, ale ne dvě skupiny bodů vpravo. Všechny čtyři skupiny ale odpovídají definici 2.

Typy shluků

První typ shluku nazveme **kulovým** – protože jeho body jsou rozloženy okolo přirozeného středu shluku, okolo jeho těžiště.

Druhý typ shluku může nabývat libovolný „tvar“, za shluk jej považujeme proto, že existuje mezi každými dvěma objekty „cesta“, spojující sousední objekty vzdáleností dostatečně malou. Tyto shluky nazýváme **přirozenými** nebo obecnými.

Zřejmě kulové shluky jsou zvláštním případem shluků přirozených. Obecně nelze vyžadovat, aby shluky byly kulové, buď jsou, nebo nejsou. Reálné shluky mohou existovat v nejrůznějších seskupeních.

Jak uvidíme níže, ne všechny shlukovací metody chápou shluk ve stejném smyslu, a proto i jejich výsledky nad stejnou množinou objektů jsou různé. Našim úkolem bude umět se orientovat jednak v tom, jaký výsledek potřebujeme, jednak v tom, kterou metodu k jeho dosažení máme zvolit.

Existují metody, které najdou přirozené shluky jakýchkoliv tvarů. Jsou však také metody, které dávají jako výsledky „shluky“ vždy kulové, ať je skutečnost jakákoliv (viz. **obr. x**., kde se bod patřící k jednomu přirozenému shluku dostal do jiného kulového „shluku“).

Ovšem i toto řešení může být někdy užitečné – pokud hledáme optimální rozložení bodů vzhledem k několika středům. Není to však úloha shlukovací. Ještě se k tomuto problému vrátíme u konkrétních metod.

Charakteristiky shluků

Konečným cílem při shlukování obvykle je nalézt pravidla, podle kterých se objekty zařazují do jednotlivých nalezených shluků. K tomu potřebujeme každý shluk popsat nějakými údaji, které by shluk vhodně charakterizovaly. Které údaje to mohou být?

Přirozeně by se nabízelo těžiště shluku, tedy bod, jehož atributy tvoří průměrné hodnoty atributů bodů shluku. Ovšem již z jednoduchého **obr. x**. je vidět, že **těžiště není vhodná charakteristika shluku.** Někdy těžiště dokonce neleží „uvnitř“ shluku.

Jako další údaje se nabízí vzdálenosti vnitroshlukové (co nejmenší) a mezishlukové (co největší), jak by odpovídalo definici 1. Ovšem z obrázku je patrno, že to také nejsou vhodné charakteristiky.

Optimálně by měla definice shluku (a tedy i charakteristika kvality rozkladu a charakteristika příslušnosti objektu ke shluku) brát v úvahu současně

- vzájemnou podobnost bodů uvnitř shluku (minimální)
- vzájemnou vzdálenost shluků navzájem (maximální)
- tvar shluků (kulový – přirozený tvar)
- charakteristické vlastnosti jednotlivých shluků:
  počet bodů, těžiště, minimum, maximum, stand. odchylky, .,,, homogenita
- charakteristické vlastnosti celkového rozkladu na shluky:
  průměrná minimální vzdálenost shluků
  průměrná minimální vzdálenost těžišť
  vzájemné vzdálenosti těžišť
  celková Σ čtverců chyb

Taková charakteristika ovšem neexistuje, protože by musela obsahovat i protichůdné parametry. Proto neexistuje ani jednoznačný výsledek. Později si uvedeme, jakými údaji je možno shluky popsat pro uspokojivý popis.

❑ Typy shlukovacích metod

Již jsme zmínili, že shlukovou analýzu tvoří řada metod, založených na různých teoriích nebo jen na intuitivních představách autorů, jak se k vágnímu pojmu shluk dobrat. Do každého algoritmu je zabudována jakási „heuristická“ zkušenost konstruktéra metody, různé metody to provádí různou taktikou. Přesto mnozí autoři nezávisle na sobě definovali velmi podobné algoritmy a tak chápali „shluk“ stejně.

Existující metody je možno dělit podle několika kritérií, nejdůležitější jsou

dle cíle shlukování na

- **nehierarchické**, produkující prostý rozklad objektů na podmnožiny;
- **hierarchické**, produkující hierarchii rozkladů, kde každý rozklad je zjemněním předcházejícího; jsou vyvinuty různými autory, liší se většinou definicí podobnosti objektů a podobnosti shluků.

dle tvaru výsledných shluků v geometrické interpretaci na

- shluky **kulové**, body soustředěné víceméně pravidelně kolem svého těžiště,
- shluky **obecné** tvoří souvislé husté oblasti nejrůznějších tvarů, které není možno dost dobře charakterizovat těžištěm.

dle rozložení výsledných shluků na

- **disjunktní** shluky (klasické shlukování, odpovídající nepodobnosti objektů různých shluků)
- překrývající se shluky nebo též **fuzzy-shluky** pro zvláštní případ zadané úlohy.

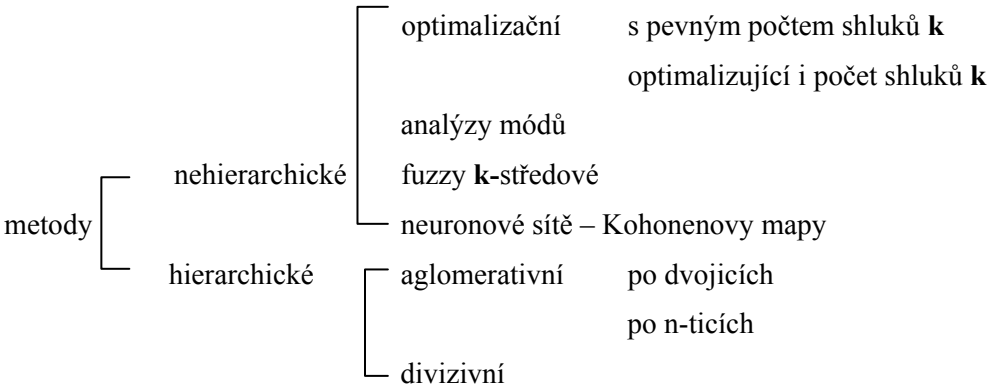
dle procesu shlukování (typu algoritmu)

- **sekvenční** - množinu objektů jedním (sekvenčním) průchodem rozdělují do shluků; obvykle se používaly pro velký počet objektů, které se nevešly do paměti, výsledky nebývaly kvalitní;
- **paralelní** - množina objektů je podle potřeby procházena vícekrát
- **přímé** - algoritmus přímo vytváří hledaný rozklad, může procházet objekty vícekrát;
- **optimalizační** - hledají optimum nějaké kriteriální funkce, obvykle iteracemi z nějakého zadaného či vygenerovaného počátečního rozkladu; průběžně je vytvářena řada zlepšujících se rozkladů;

- **rekurzivní** - hierarchické, které vytvářejí následující rozklad pomocí předcházejícího rozkladu, ne pomocí původních objektů;
- **heuristické** - v jistém smyslu všechny, zde však chápeme jako heuristické ty, které využívají k hledání rozkladu prohledávání prostoru všech možných rozkladů s využitím heuristické funkce.
- **neuronovou sítí** - proces nealgoritmický, při vhodném počátečním nastavení sítě rychlý,

Tomuto dělení odpovídají různé typy definic pojmu shluk. Každá metoda umí vyhledat jeden typ výsledku – tvaru a rozložení shluků, obvykle se to však v literatuře nijak nezdůrazňuje. Pro volbu metody vzhledem k dané úloze nebývají uvedena kritéria, vše záleží na intuitivní volbě či zkušenostech analytika. Také u nabízených SW systémů pro dolování dat nebývají tyto skutečnosti (zvláště tvar výsledných shluků) nijak uvedeny a nedostatečně poučený uživatel může dostat velmi zkreslené výsledky, aniž o tom ví.

Nejrozšířenější typy metod je možno rozdělit dle typu (hierarchie/ne), typu rozkladu (disjunktní/ne), částečně dle tvaru výsledku (kulový/přirozený) na následující metody. Mimo to mohou existovat mnohé další metody, které obvykle řeší specifické úlohy.

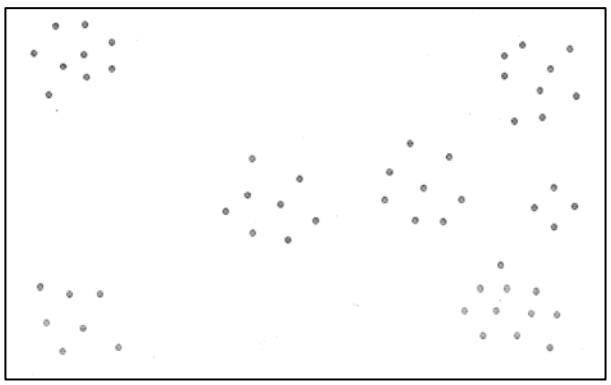


Charakteristiku uvedených typů a jejich základní algoritmy uvedeme postupně podrobně.

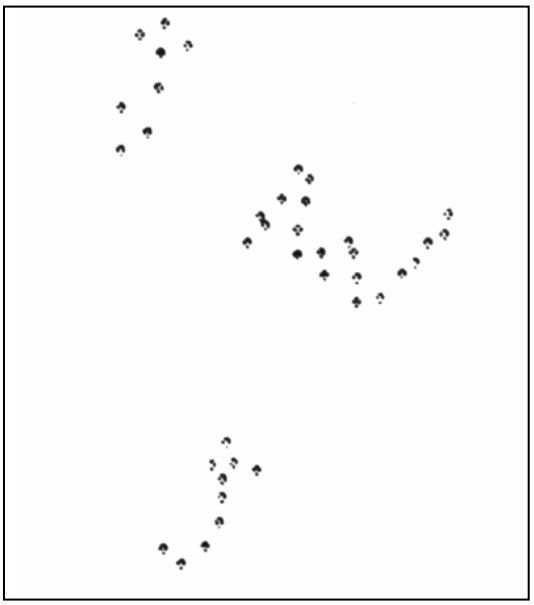
❑ Testovací data pro shlukování

V následujících kapitolách budeme používat pro prezentaci výsledků jednotlivých metod několik typů dvourozměrných dat. Data byla vytvořena pro testovací účely. Dvourozměrná jsou proto, že jejich zobrazením v rovině s vykreslením výsledku se snadno porovná „intuitivní“ představa uživatele o výsledku se skutečným výsledkem každé metody.

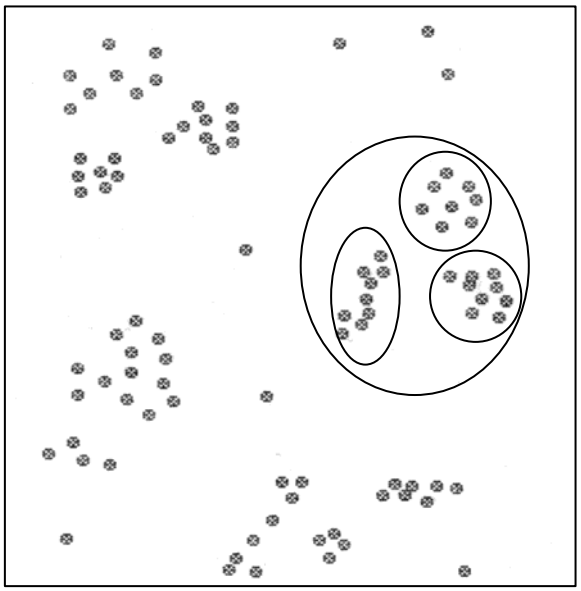
Data 1: Sedm zřetelných přirozeně „kulových“ shluků, případně 2 úrovně shlukování



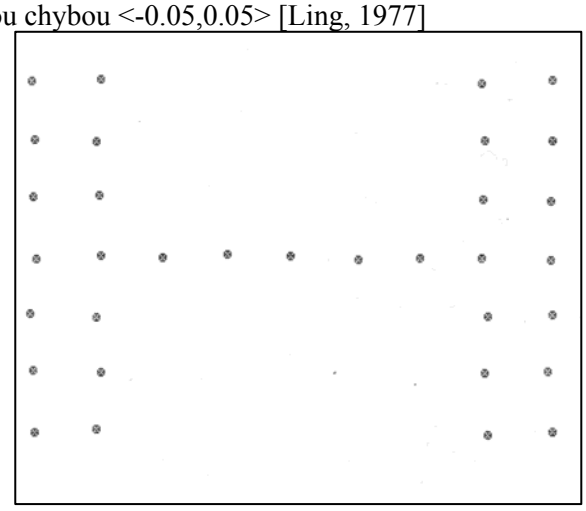
Data 2: Tři zřetelné obecné shluky



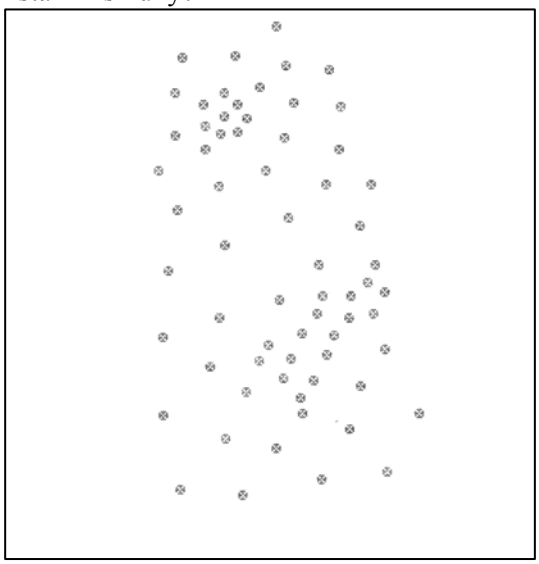
Data 3: shluky ve 2 shlukovacích úrovních + izolované body



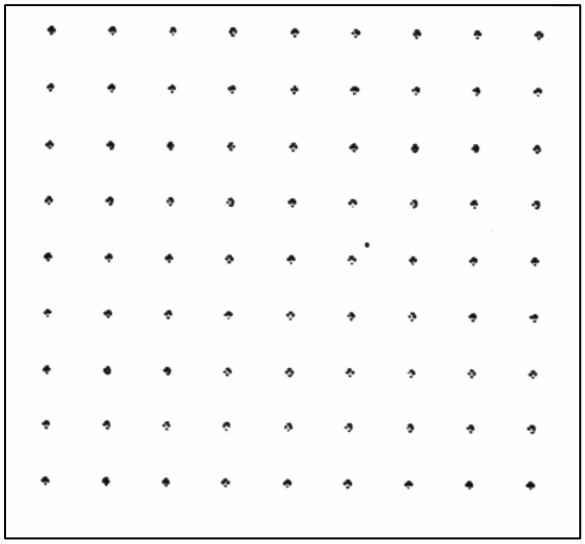
Data 4: H-data s náhodnou chybou <-0.05,0.05> [Ling, 1977]



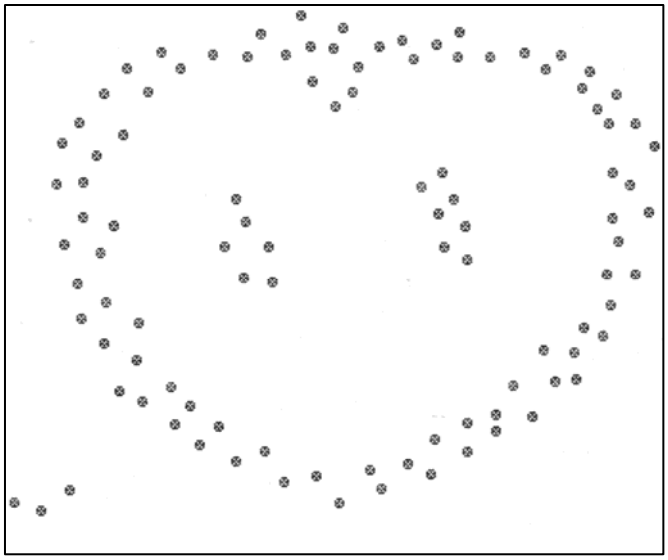
Data 5: 2 nevýrazná hustší místa ~ 2 shluky?



Data 6: homogenní data bez shluků = 1 shluk



Data 7: „Srdce“ se třemi malými shluky uvnitř i vně



Shrnutí pojmů 5.1.

Klasifikace a shlukování. Úlohy shlukování.

Problémy úloh shlukování.

Výběr relevantních atributů a podobnost objektů.

Koeficienty podobnosti a vzdálenosti.

Míra podobnosti objektů. Koeficienty asociace. Koeficient korelace.

Míra vzdálenosti objektů.

Pojem shluku. Typy shluků.

Shluky kulové a shluky přirozené.

Charakteristiky shluků.

Typy shlukovacích metod.

Metody hierarchické a nehierarchické. Shluky disjunktní a shluky překrývající se.

Otázky 5.1.

1. Co rozumíme klasifikací objektů a co shlukováním?
2. Které úlohy musíme postupně vyřešit při shlukování objektů?
3. Podle čeho vybíráme atributy vhodné pro popis shlukovaných objektů?
4. Jak měříme podobnost objektů a které míry podobnosti znáte?
5. Co je míra vzdálenosti objektů, jak souvisí s podobností a k čemu se používá?
6. Které míry vzdálenosti objektů znáte?
7. Co je shluk?
8. Jaké typy shluků rozlišujeme?
9. Čím je možno shluky charakterizovat?
10. Které hlediska pro dělení a které typy shlukovacích metod znáte?
11. Jaké typy výsledků dostáváme metodami nehierarchickými a metodami hierarchickými?

Úlohy k řešení 5.1.

1. Jsou dána data BANKA, obsahující 1027 záznamů o poskytovaných úvěrech za minulých 5 let. Jejich struktura je
   BANKA pohlavi [muž / žena],
   vek [roku],
   svobodny [ano / ne],
   nezamestnany [ano / ne],
   cim_ruci [auto, dum, PC, kolo],
   problematicky_region [ano / ne],

mes_prijem [Kc],
hotovost_u_banky [Kc],
pocet_mesicu_splatky,
pocet_roku_u_souc_firmy,
dostal_uver [ano / ne],
splacel_bezproblemově [ano / ne])

Formulujte shlukovací úlohu (jak se dá využít shlukování nad těmito daty), vyberte z dat vhodné atributy, navrhněte metody jejich předzpracování, míru podobnosti či vzdálenosti a určete typ hledaných shluků.

2. Jsou dána data z evidence studentů sportovního GYMNÁZIA s atributy: školní rok, třída, učitel [jméno], jméno [studenta], pohlaví [chl/dív], věk, výška, váha, dále maximální výkony sportovní za aktuální rok ve skoku vysokém [cm], dalekém [cm], běhu_100m [12.3 sec], běhu-400m a závěrečné známky z češtiny, cizího jazyka [ruština a později angličtina], matematiky, fyziky, dějepisu a zeměpisu. Data jsou pořizována za dobu 30 let.

   Formulujte shlukovací úlohu (jak se dá využít shlukování nad těmito daty), vyberte z dat vhodné atributy, navrhněte metody jejich předzpracování, míru podobnosti či vzdálenosti a zvolte typ hledaných shluků.

5.2. Shlukování nehierarchické

Cíl Po prostudování této kapitoly budete umět

- rozlišit úlohu optimálního rozkladu od úlohy shlukovací,
- vybrat vhodnou metodu pro výpočet dané úlohy a nastavit její vstupní parametry,
- pro praktické problémy rozhodnout, zda a která metoda je pro shlukování vhodná.

Výklad

Metody nehierarchické hledají nejlepší prostý rozklad množiny objektů na shluky, tedy na disjunktní podmnožiny. Řada typů těchto metod dává různé typy výsledků.

Počet možných rozkladů m objektů na podmnožiny je dán rekurzivně vztahem

$$P[1] = 1 \qquad P[n+1] = 1 + \sum_{i=1}^{n} \left(\binom{n}{i} . P[i] \right)$$

výraz konverguje k

$$1 + \Sigma \left(P[n] * x^n \right) / n! \quad \rightarrow \quad \exp(\exp(n) - 1)$$

Příklad 5.6.

pro n =	10	*je počet podmnožin*	1.16 E 5
	20		5.17 E 13
	30		8.74 E 23
	40		1.57 E 35
	...		

♦

Odtud je zřejmé, že není možno procházet a testovat všechny možné rozklady, je nutné hledat algoritmy testující jen „pravděpodobnější“ rozklady, tedy **do algoritmu je zabudována zkušenost konstruktéra metody**. Různé metody to provádí různou taktikou.

❑ Metody optimalizační

hledají nejlepší rozklad množiny objektů iteračním způsobem. Počáteční rozklad (zadaný nebo vygenerovaný) zlepšují tak, že hledají rozklad s lepší hodnotou zadané kriteriální funkce. Metody se navzájem liší způsobem prohledávání množiny všech možných rozkladů. Jednodušší z nich hledají nejlepší rozklad pro daný počet shluků, obecnější hledají současně nejlepší rozklad i počet shluků.

Cílem je najít takový rozklad, pro který kriteriální funkce nabývá extrému. Po zvolení kriteriální funkce je úloha teoreticky definovaná a je možno optimální rozklad najít. Problémem je zvolit metodu, která vede rychle k cíli. Prohledávat všechny možné rozklady, byť jich je konečně mnoho, často není únosné ani na počítačích - pro jejich velký počet. Ovšem ani formulování kriteriální funkce není tak snadné, proto vzniklo mnoho metod.

Optimalizační metody vycházejí **z předem určeného počtu shluků** a iteračně zlepšují rozklad vzhledem k zadanému tvaru kriteriální funkce nebo k jinak definovanému konci algoritmu (někdy se tyto typy algoritmů nazývají horolezeckými).

❑ Problém počátečního rozkladu

Algoritmy optimalizační začínají zadáním nebo odvozením počátečního rozkladu množiny objektů. Optimální je případ, kdy uživatel má nějakou znalost nebo představu o datech předem, dovede určit počet shluků i jejich přibližné charakteristiky. Pokud ne, používá se buď náhodné zadání počátečních bodů, nebo se provádí jednoduché předzpracování, které rozmístí počáteční body rovnoměrně do zadané množiny bodů.

Počet bodů = počet shluků označíme **k**. Také tento počet uživatel buď zadává, nebo se provádí několik výpočtů s různými **k** a vybere se ten s nejlepší hodnotou kriteriální funkce.

Metody zadání počátečního rozkladu

TYP1 prvních **k** bodů zadané množiny dat

TYPN - **k** bodů náhodně vybraných ze zadaných
body náhodné se souřadnicemi náhodně vygenerovanými v rámci variačních intervalů souřadnic

TYPD - body vybrané uživatelem z množiny daných bodů
- body zadané uživatelem pomocí souřadnic (hodnot atributů)

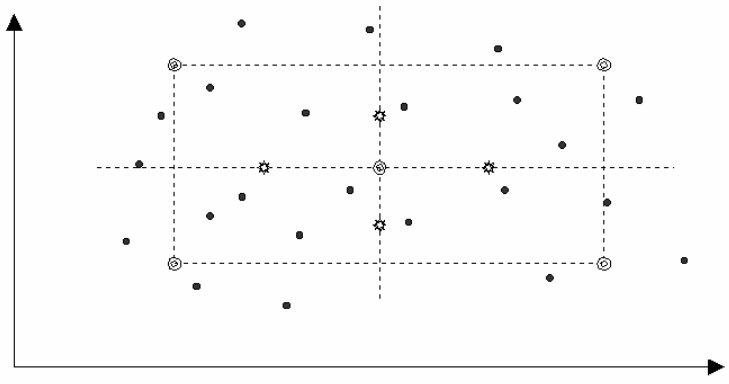
TYPF (Fromm)

k=1 + 2^s bodů v těžišti souboru a ve vrcholech s-rozměrného kvádru se středem v těžišti a stěnami rovnoběžnými se souřadnými rovinami a hranami rovnými dvojnásobku standardní odchylky příslušné souřadnice ...

$$y_{j\,i} = t_i + \mathrm{sgn}\left(\sin\left(2.\pi / 2^i - 2.\pi / 2^{i+1}\right)\right) s_i$$

TYPC

k bodů v těžišti a v bodech posunutých od těžiště v kladném a záporném směru podél jedné z prvních souřadných os o standardní odchylku příslušné souřadnice.



Obrázek 5.8. Analýza nasazení počátečních typických bodů dat

❑ Metody optimalizační k-středové s pevným počtem shluků

Nejužívanější „shlukovací“ metodou, vyskytující se ve většině SW systémů pro statistiku, analýzy dat i pro dolování znalostí, je k-středová metoda se zadaným pevným počtem shluků. Existuje několik jejích variant, které se jen málo liší iteračním algoritmem. Protože jde o velmi jednoduchý, přirozený a relativně rychle konvergující algoritmus, je jeho obliba vysoká.

První základní verzi formuloval Forgy.

Základní algoritmus FORG

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

Od tohoto algoritmu jsou pak odvozeny některé varianty, které však nemají výrazně odlišný ani průběh výpočtu, ani výsledek.

Varianta Janceyova

JANC

algoritmus jako u FORG, jen bod 4 definuje nové typické body jinak:

4. definování nových typických bodů do bodů souměrně sdružených s původními typickými body podle těžiště

Varianta MacQueenova

MACQ

od předcházejících dvou metod se liší tím, že přepočítává typický bod skupiny po každém přemístění bodu. To způsobuje závislost výsledného rozkladu na uspořádání množiny shlukovaných objektů. Tato metoda provádí přiřazení jen 2x, neiteruje rozklady až do ustálení bodů.

1. zadání **k** počátečních typických bodů,
2. přiřazení každého bodu k nejbližšímu typickému bodu a přepočtení typických bodů zmenšené i zvětšené skupiny.

Varianta Wishartova

WISH

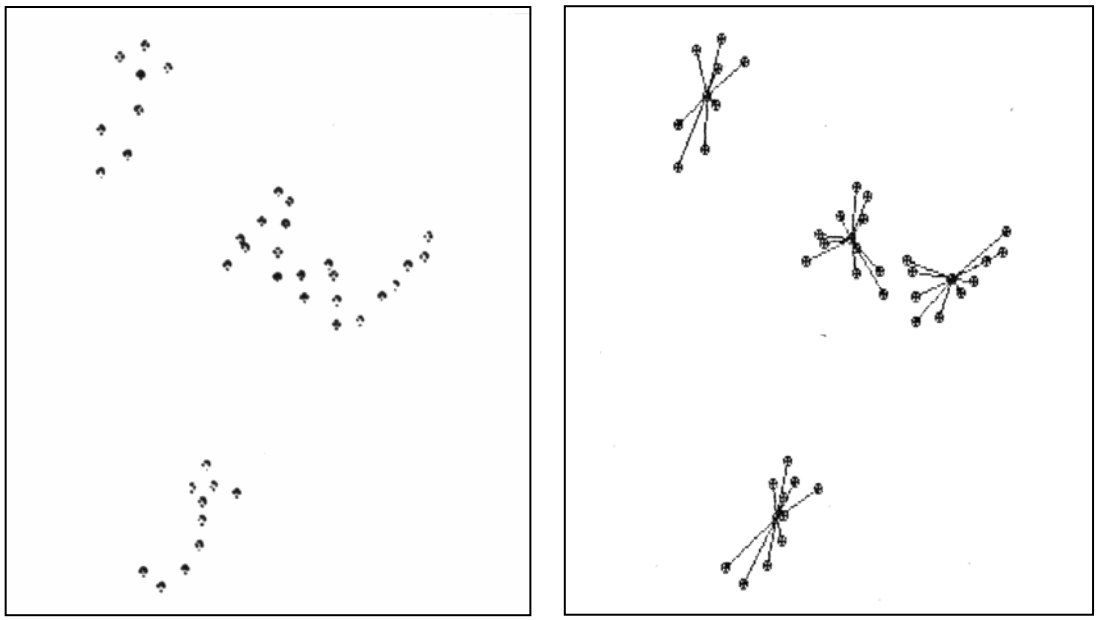
konvergentní varianta metody MacQ, po dvou bodech algoritmu MACQ následují

3. pokud došlo ke změně v přiřazení bodů shlukům, opakování od bodu
4. výpočet kriteriální funkce výsledného rozkladu

Protože jde o velmi rozšířené metody, podíváme se na jejich výsledky podrobně na testovacích datech.

Příklad 5.7.

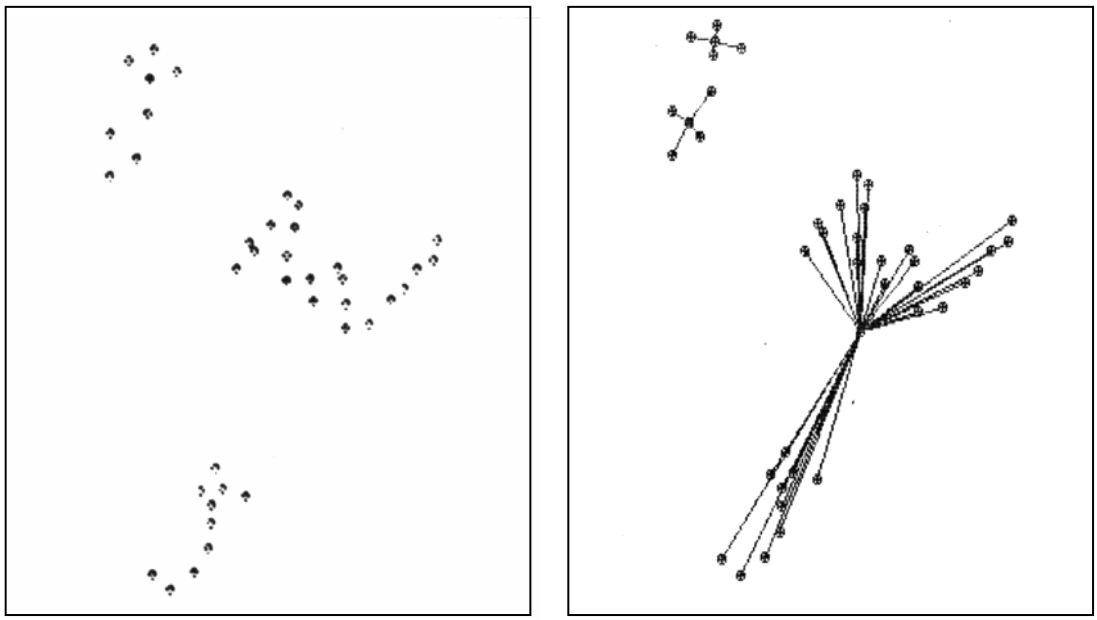
3 shluky, metoda k-středová, 4 typické body



♦

Příklad 5.8.

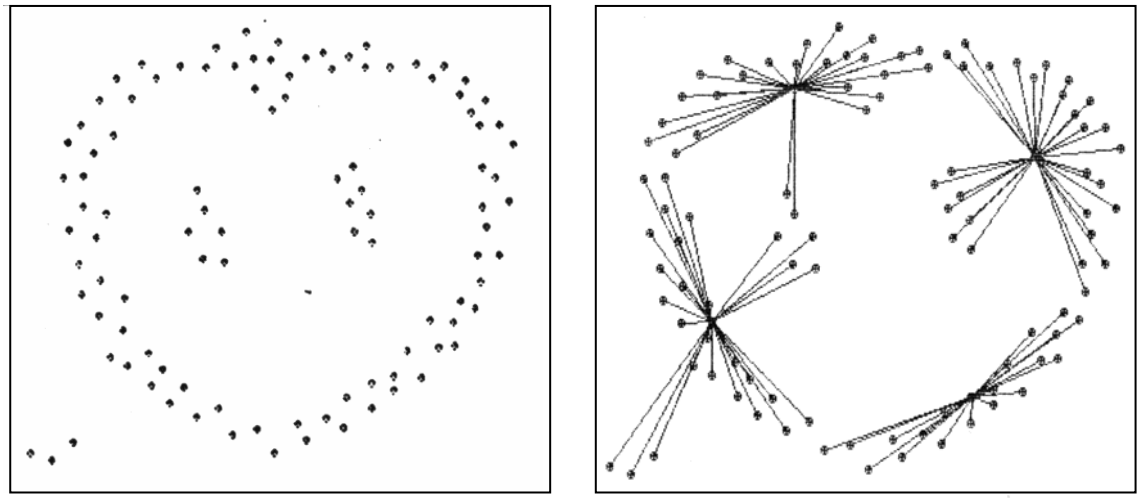
3 shluky, metoda K-středová, 3 náhodné typické body



♦

Příklad 5.9.

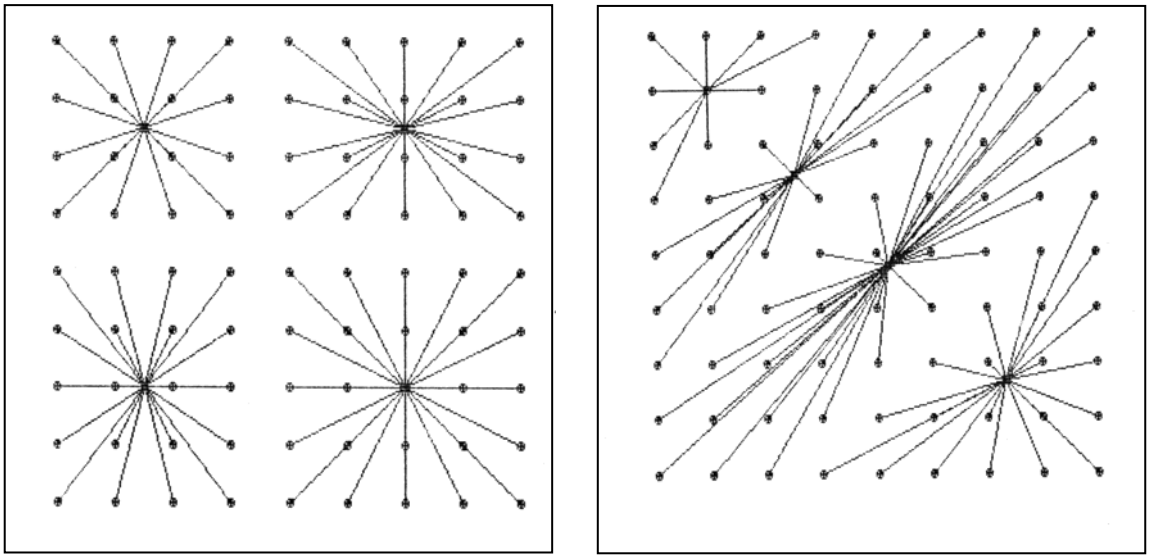
4 obecné shluky, metoda K-středová, 4 odvozené typické body



♦

Příklad 5.10.

homogenní množina, metoda K-středová, 4 náhodné typické body (2 x jiné)



♦

Závěr: metoda k-středová neshlukuje, výsledkem nejsou přirozené shluky, ale **k** kulových podmnožin objektů. Metoda nerozpozná, zda množina bodů skutečně obsahuje shluky a kolik, ale vykáže jako výsledek „shluky“ i v dokonale homogenní množině. Je vysoce závislá na počáteční volbě typických bodů vzhledem ke shlukované množině bodů a někdy dává velmi chybné výsledky i pro množiny se zřetelnými shluky. Pokud v datech existují izolované body (jednoprvkové shluky), mohou dále výrazně zkreslit charakteristiky výsledných shluků.

Metoda **hledá optimální rozklad množiny** (vzhledem k dané kriteriální fci) **na k podmnožin** “tak, aby všechny body měly ke svému středu přibližně stejně daleko”.

Příklad 5.11.

Vhodné využití k-středové metody.

Je dána množina obytných domů svými zeměpisnými souřadnicemi v rozsáhlém prostoru. Firma provádějící služby má v úmyslu zřídit 4 provozovny. Použitím k-means algoritmu se 4 typickými body najde optimální rozložení domů na 4 části, najde středy těchto „shluků" a tam umístí provozovny.

♦

❑ Metody optimalizační k-středové s proměnným počtem shluků

Velmi optimisticky zní zlepšená varianta k-středové metody, která navíc v průběhu iterací testuje, zda okamžité skupiny bodů „mají tendenci" se rozdělit nebo naopak sloučit a tím optimalizuje i původní zadaný počet shluků. První metoda tohoto typu byla pojmenovaná ISODATA. Dodnes se pod tímto názvem objevuje, i když jde obvykle o její zlepšení a automatizované nastavování parametrů, které byly původně zadávány uživatelem. Uvedeme si jednu takovou variantu metody ISODATA

CLASS

hledá optimální počet shluků i nejlepší rozklad;
pro podmínky slučování nebo rozdělování je definováno několik konstant:

THETAN = minimální počet bodů ve skupině, implicitně THETAN = 3
S0 = počáteční rozdělovací práh, implicitně S0 = 0.6
GAMA = maximální počet iterací, implicitně GAMA = 2*m (m je počet objektů)

Algoritmus CLASS

1. Zadání počtu **k** počátečních shluků a **k** počátečních typických bodů.
2. Počáteční rozdělení bodů metodou FORG

Dále je algoritmus iterační, v každém kroku se provádí tato tři rozhodnutí 3-5:

3. **Zmrazení malých shluků**: skupiny o méně než THETAN bodech, které se nezměnily v posledních dvou iteracích, jsou z další analýzy vyloučeny;
4. **Rozdělení shluků**: v m-té iteraci definujeme rozdělovací práh

$$S_m = S_{m-1} + \frac{1 - S0}{GAMA}$$

Pro každou skupinu se vypočtou nové souřadnice každého bodu jako odchylky od souřadnic těžiště skupiny. Pro j-tou souřadnici se vypočtou průměrné odchylky od těžiště shluku D_{j1} a D_{j2} bodů ležících vpravo a vlevo od těžiště; přitom k1 a k2 je počet těchto bodů vpravo a vlevo :

Podrobněji: V každé iteraci jsou teď nějaké shluky. Pro každý z těchto shluků postupně je spočítáno těžiště T [x, y, …], tedy pro shluk Si je to těžiště Ti [xi, yi, …].

Vezmeme atribut - **souřadnici x,** tedy **xi** tohoto těžiště a postupně všechny body shluku Si, těch je **h** (třeba 10). Pro každý bod (například bod Z [a, b, …]) shluku spočítáme jeho odchylku od těžiště, tedy (xi – a) ... označeno xij. Tyto odchylky sčítáme – zvlášť ty, pro něž ně a>xi = to jsou body ležící vpravo od těžiště (třeba jich je 6 ze všech 10), zvlášť pro ty, kde a<xi ... vlevo (třeba jsou 4 z 10). Každý součet vydělíme počtem bodů vpravo (= k1) a vlevo (= k2) a dostaneme **Dj1 a Dj2**.

$$D_{j1} = 1/k_1 . \sum_{i=1}^{k_1} x_{ij} \qquad D_{j2} = 1/k_2 . \sum_{i=2}^{k_2} x_{ij}$$

Současně se počítá maximum těch odchylek (a-xi) pro body vpravo a vlevo, označíme například max_P a max_L.

Totéž uděláme pro atributy **další y**, …: vždy sčítáme odchylky (yi – b), máme tedy Dj1_x, Dj2_x ... pro souřadnici x a Dj1_y, Dj2_y ... pro souřadnici y atd.

Dále se definují hodnoty

$$a1 = \max_j \frac{D_{j1}}{\max x_{ij}} \qquad a2 = \max_j \frac{D_{j2}}{\max x_{ij}}$$

pro body vlevo a vpravo od těžiště, j = 1, ... ,S

Když máme ty Dj1, ... atd, spočítají se 2 maxima – jedno pro pravé body, jedno pro levé body, ve jmenovateli zlomku jsou max_P a max_L; to se udělá pro všechny souřadnice (pro x, y, …) a z nich se vezme maximum, tedy a1 a a2 je největší z těch pravých (levých) zlomků.

Je-li v m-té iteraci počet skupin menší než 2k (k je počáteční počet shluků), a1 > Sm nebo a2 > Sm a zároveň je počet bodů větší než 2*(THETAN+1), rozdělí se shluk podle té j-té souřadnice, v níž a1 nebo a2 nabývá maximální hodnoty, a to na body vlevo a vpravo od j-té souřadnice těžiště shluku.

5. **Zrušení shluků**: v každém kroku se vypočte minimální vzdálenost skupin

$$TAU = 1/h \cdot \sum_{i=1}^{h} D_i$$

kde h je současný počet skupin, D je minimální vzdálenost těžiště i-té skupiny od těžišť ostatních skupin. Jestliže pro i-tou skupinu platí D_i < TAU a zároveň je počet skupin větší než K/2, zruší se tato skupina a každý její bod se přiřadí ke skupině s nejbližším těžištěm.

Po každé změně se provede FORG. Iterace se provádí do ustálení, maximálně GAMA-krát.

Jak se dá očekávat z použití metody FORG pro každý mezivýsledek, výsledkem jsou **opět kulové „shluky“**.

Uvedený algoritmus však není ideální ani pro optimalizaci rozkladu. Například kritérium pro rozdělování shluků testuje rozdělení jen podél souřadných os (= vždy podle jednoho atributu).

❑ Fuzzy C-means algoritmy

jsou variantou metod optimalizačních se zadaným pevným počtem shluků, ovšem na rozdíl od všech ostatních metod hledají „shluky“ překrývající se. Existují úlohy, kdy takový výsledek dává smysl, i když nejde o typickou shlukovací úlohu (viz příklad níže).

Fuzzy zobecnění minimální odchylky shlukování bylo definováno v roce 1973 [x] a byl vyvinut algoritmus minimalizace kriteriální funkce. Tato práce byla později zobecněna [x] pro nekonečné cílové funkce spolu s algoritmem minimalizace výsledného problému. V roce 1985 [x] byla formulována modifikovaná verze, která zaručuje konvergenci fuzzy shlukovacího algoritmu (podle c-typických bodů) buď k **lokálnímu minimu** nebo k **sedlovému bodu** cílové funkce.

Shlukovací úloha je definována jako minimalizace výrazu

$$J_m(W, Z) = \sum_{i=1}^{n}\sum_{j=1}^{c} w_{ij}^{m} \| x_i - z_j \|^2_A$$

s podmínkou, že $w_{ij} \in \langle 0,1 \rangle$, $\sum_{j=1}^{c} w_{ij} = 1$ pro $1 \le i \le n$,

$w_{ij} \geq 0$ pro $1 \leq i \leq n$ a $1 \leq j \leq c$,

$\sum_{j=1}^{n} w_{ij} > 0$ pro $1 \leq i \leq c$

kde ***n*** je počet bodů
c je pevný a zadaný počet shluků
m je skalár, tzv. váhový exponent ($m > 1$)
s je dimenze daného prostoru
A je jakákoliv kladná symetrická matice $s \times s$
$x_i \in R^S$, $1 \leq i \leq n$ jsou dané body v charakteristickém prostoru R^S
$z_j \in R^S$, $1 \leq j \leq c$ jsou středy shluků
$\| \ \|_A$ je norma skalárního součinu v R^S
w_{ij} představuje míru asociace vzoru (daného bodu) ***i*** se shlukem ***j***
$Z = [z_1, z_2, ... z_c]$ je matice středů shluků $s \times c$
$W = \{ w_{ij} \}$ je matice $n \times c$.

Fuzzy shlukovací algoritmus alternuje mezi počítáním Z a W dokud buď středy Z nebo členská funkce W se neopakují, potom algoritmus končí. Jestliže je m = 1, potom fuzzy shlukovací algoritmus konečně konverguje a zastaví se v lokálním minimu jestliže jako míra vzdálenosti je brán čtverec Euklidovské vzdálenosti.

Matematickým důkazem se zde nebudeme zabývat.

Fuzzy shlukovací algoritmus (pomocí c-typických bodů)

1. Zadej: ***n*** (je počet bodů)
c (je pevný a známý počet shluků, $2 \leq c \leq n$)
s (dimenze daného prostoru, počet atributů)
m (skalár, tzv. váhový exponent ,$1 \leq m < \infty$)
ε (chybová tolerance)
$\| . \|_A$ (norma skalárního součinu v R^S)
A (jakákoliv kladná symetrická matice $s \times s$)
$X = (x_1 , x_2 , ...,x_n)$ ($x_i \in R^S$, $1 \leq i \leq n$ jsou dané body v charakteristickém prostoru R^S)

Buď $W^{(0)} = \{ w^{(0)}_{ij} \}$ nebo $Z^{(0)} = [z^{(0)}_1 , z^{(0)}_2 ..... z^{(0)}_c]$,
kde $W^{(0)}$ je matice $n \times c$, $Z^{(0)}$ je matice středů shluků $s \times c$,
w_{ij} představuje míru asociace vzoru (daného bodu) ***i*** se shlukem ***j***
$z_j \in R^S$, $1 \leq j \leq c$ jsou středy shluků.

2. Opakuj
Jestliže v kroku 1 byla zadána matice $W^{(0)}$,
potom nejprve přepočítej středy shluků $\{z^{(t)}_j\}$ vzorcem

$$z_j = \frac{\sum_{i=1}^{n} (w_{ij})^m x_i}{\sum_{i=1}^{n} (w_{ij})^m} \qquad \text{pro } 1 \leq j \leq c \qquad (1)$$

a potom matici $W^{(t)}$ vzorcem

$$w_{ij} = \begin{cases} 1 / \left(\sum_{q=1}^{c} (\| x_i - z_j \|_A / \| x_i - z_q \|_A)^{2/(m-1)} \right) & \text{jestliže } z_q \neq x_i \text{ , } q = 1,z, ..., c \\ 1 & \text{jestliže } z_j = x_i \\ 0 & \text{jestliže } z_j = x_s \text{ , } s \neq i \text{ .} \end{cases} \qquad (2)$$

Jestliže v kroku 1 byly zadány středy shluků $\boldsymbol{Z}^{(0)}$,
potom nejprve přepočítej $\boldsymbol{W}^{(t)}$ podle vzorce (2) a potom $\boldsymbol{Z}^{(t)}$ podle vzorce (1).

dokud $\| \boldsymbol{W}^{(t+1)} - \boldsymbol{W}^{(t)} \| < \varepsilon$.

3. Výsledek je matice $\boldsymbol{W}$ a $\boldsymbol{Z}$.

Příklad 5.12.

Jednoduchý příklad využití fuzzy-shlukování je tento: máme dánu množinu 50 ovocných moštů neznámého složení (z jakého ovoce jsou vyrobeny). Dále je dána množina 5 moštů homogenních (jablečný, višňový, pomerančový, grepový, rybízový), ty tvoří pevný počet typických bodů. Otázkou je, ze kterých složek se skládají mošty analyzované.

*Zkoumaná množina moštů je matice **X**. Matice zadaných typických moštů je **Z**. Výsledkem výpočtu je matice **W,** v níž každému zkoumanému moštu je vypočteno 5 hodnot – váha příslušnosti k jednomu z pěti typických moštů. Jinam řečeno, kolik procent kterého moštu zkoumaný vzorek obsahuje.*

Je zřejmé, že součet těchto pěti čísel je roven 1 (=100%).

♦

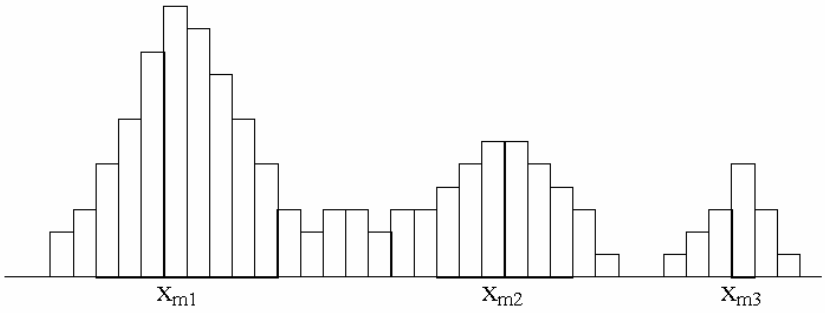
❑ Analýzy modů

Jiný přístup ke shlukování mají metody, hledající nejhustší místa prostoru - módy. **Modus** je hodnota a_m náhodné veličiny (atributu) A, v níž nabývá frekvenční funkce veličiny A lokálního maxima. To znamená, že této hodnoty nabývá maximum bodů. Předpokládá se, že poloha módu odpovídá poloze typického bodu shluku.

Prakticky se variační interval domény atributu rozdělí na třídní intervaly a v nich se sledují četnosti výskytu hodnot. Výsledné relativní četnosti se znázorní histogramem. Z něj je možno odhadovat existenci a polohu módů frekvenční funkce.

Příklad 5.13.

Představme si nejprve jednoatributové objekty s doménou atributu <a, b>. Rozdělíme interval na třídní intervaly, určíme četnosti v těchto třídních intervalech a dostaneme histogram z následujícího obrázku. V něm tvoří módy tři lokální maxima xm1, xm2, xm3. Jsou to hodnoty s (lokálně) nejvyšším počtem bodů, tedy nejhustší, tedy předpokládané „jádro“ shluku.



Obrázek 5.9. Jednorozměrné rozložení módů

Pokud jsou módy od sebe odděleny intervaly s nulovou četností (jako mezi módy xm2 a xm3), můžeme je považovat za různé shluky. Pokud ovšem takový přirozený předěl mezi módy neexistuje (jako mezi módy xm1 a xm2), volíme jistou minimální hranici četnosti pro určení předělu mezi shluky.

♦

Objekty ke shlukování jsou obecně popsány n atributy, jde tedy o populaci reprezentovanou n-rozměrným náhodným vektorem

$$\mathbf{A} = (A_1, A_2, \ldots, A_n)$$

Nyní hledáme shluky z existence a polohy módů mnohorozměrné frekvenční funkce. Odhad typu mnohorozměrné frekvenční funkce je zřejmě složitější, než v případě jednorozměrné veličiny. Obecné algoritmy pro její nalezení dosud neexistují.

Odhad polohy módů z frekvenční funkce v n-rozměrném prostoru

1. pro každý z n atributů určíme doménu a tím vymezíme n-rozměrný kvádr D (jako obal množiny bodů)
2. rozdělíme oblast D na n-rozměrné intervaly (reprezentované opět n-rozměrnými kvádry se zvoleným krokem délky hrany ve směru každé souřadné osy)
3. sledujeme četnosti výskytů sledovaných bodů v těchto kvádrech
4. intervaly s největší četnosti, oddělené vzájemně intervaly s malou četností, dávají odhady umístění módů v množině bodů.

N-rozměrný problém nelze znázornit histogramem. I sledování četností se pro větší n stává problémem: rozdělíme-li každý z n variačních intervalů na k úseků, dostaneme k^n n-rozměrných kvádrů. Rozhodnutí o tom, zda 1 bod patří do určitého intervalu, vyžaduje průměrně (pro jeden atribut) k testů, pro n atributů k * n testů. Obdobné množství testů by si vyžádalo porovnání výsledných četností v sousedních intervalech. Celkem by tedy takový logicky jednoduchý algoritmus měl exponenciální složitost.

Proto existuje v literatuře množství metod, odhadujících polohy módů, které jsou založeny na statistických, matematických a jiných přístupech, někdy také s vysokou časovou složitostí a pro větší data nepoužitelných.

Uvedeme si jeden z jednoduchých algoritmů - Kittlerův. Kittler nazývá svou metodu lokálně senzitivní (na rozdíl od metod globálně senzitivních, jako např. ISODATA), protože analyzuje data s ohledem na detaily struktury dat.

Algoritmus ANMO - Kittlerova metoda hledání módů

1. Zadáme α (pro α-souvislé shluky).
2. Ke každému bodu O_i množiny $\mathbf{O} = \{O_1, O_2, \ldots, O_n\}$ sestrojíme jeho α-okolí jako množinu všech bodů O_j takových, že pro něž platí $d(O_i, O_j) < \alpha$.
3. Sestrojíme množinu S takových bodů O_j, které mají ve svém α-okolí alespoň jeden bod rùzný od bodu O_i.
4. Dokud je množina S neprázdná, vybereme z ní libovolný bod L a zaznamenáme počet ostatních bodù v jeho α-okolí, jinak přejdeme k bodu 7.
5. Vybraný bod L vytiskneme i s počtem bodů jeho α-okolí.
6. Vyjmeme bod L z množiny S a zařadíme ostatní body jeho α-okolí do množiny P.
7. Dokud je množina P neprázdná, vybereme z ní (další) bod L s největším α-okolím a přejdeme k bodu 4., jinak přejdeme k bodu 3.
8. Vytiskneme zbývající izolované body (body s prázdným α-okolím).

Výsledkem metody je jedna nebo více posloupností dvojic čísel - pořadové číslo bodu a počet bodů v jeho α-okolí.

Na rozdíl od optimalizačních metod nacházejí analýzy módů přirozené shluky.

Výsledkem je nehierarchický rozklad, kde shluky tvoří seskupení objektů kolem předpokládaných modů, ostatní objekty tvoří izolované body.

Problémem bývá nastavení parametru α znamenajícího "jak vzájemně vzdálené" body ještě patří ke shluku a které již ne. Na tomto parametru závisí výsledný počet a velikost shluků. Obvykle nelze najít optimální rozklad jediným shlukováním.

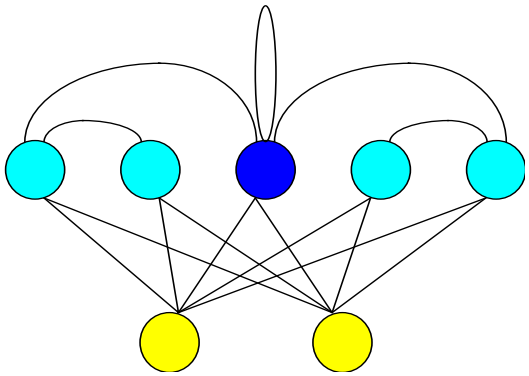
Závěr: Analýzy módů konstruují **přirozené α-shluky pro zadané α**. Pro praxi bývá užitečnější použít vhodně nastavenou hierarchicku metodu se strategií nejbližšího souseda (viz níže), která dává výsledky stejného typu a spolehlivěji najde optimální rozklad.

❑ Kohonenovy mapy

Zcela jiný přístup ke shlukování má jeden z typů neuronových sítí – Kohonenovy mapy. Přiřazují n-dimenzionálním vektorům na vstupu obvykle dvourozměrnou reprezentaci ve výstupní vrstvě neuronů tak, že podobné vektory jsou reprezentovány blízkými si neurony v dané topologii sítě. Na rozdíl od ostatních typů neuronových sítí jde o metodu, která nevyžaduje trénovací množinu, jde o adaptaci bez učitele. Při vhodně nastavených počátečních parametrech sítě je metoda rychlá. Výsledkem je příslušnost objektu ke shluku, charakteristiky se musí dopočítat.

Kompetitivní neuronová síť je tvořena dvěma vrstvami neuronů.

1. Spodní představuje vstupní jednotky, které jsou propojeny se všemi neurony vrstvy výstupní.
2. Vrchní vrstva je tvořena výstupními neurony, které jsou mezi sebou propojeny.



Obrázek 5.10. Kompetitivní síť

Propojení ve výstupní vrstvě je typické **sebeexcitující vazbou** a **inhibičními hranami** vzhledem k ostatním neuronům. Toto propojení vede k posilování neuronu, který byl na počátku nejvíce excitován. Je také **zachována topologie vstupních vektorů**. To znamená, že neurony, které jsou si blízko, rozpoznávají sobě blízké vektory.

Adaptace jednotlivých neuronů spočívá v tom, že při vstupu vstupního vektoru neurony soutěží v tom, který je vstupnímu vektoru nejblíž. Ten neuron který vyhraje, "zahoří". Neuron, který zahořel, sám sebe posiluje samoexcitující vazbou a ostatní potlačuje inhibičními hranami.

Výsledkem je, že neurony, které často vyhrávají pro danou skupinu vzorů se stávají dominantními a ostatní jsou potlačeny. Tedy pro danou skupinu vstupních vektorů zahoří pouze jeden neuron (po naučení sítě).

Výhodou je, že síť se organizuje sama, bez žádného učitele.

Tento způsob je podobný postupu excitace a inhibice v mozku.Každý objekt pak reprezentuje nějaký objekt či třídu objektů ze vstupního prostoru. Tento neuron je pak schopen rozpoznat celou třídu si podobných vstupních vektorů.

Ve skutečném mozku se učení děje postupně tak, že při získávání znalostí se zabírají další a další části mozkové kůry, kde dochází k učení na jednotlivé vzory postupně, tak, jak přicházejí různé druhy vjemů. Výběr oblastí na různé typy vjemů však nejsou náhodné, protože umístění jednotlivých oblastí a základní vlastnosti jsou vrozené.

Výše uvedené principy samoorganizace a adaptace jsou aplikovány v Kohonenových mapách.

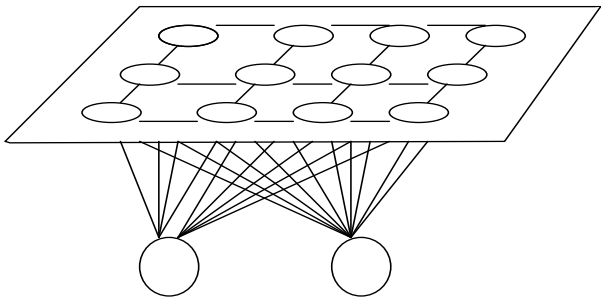
Hlavní ideou těchto neuronových sítí je nalézt prostorovou reprezentaci složitých datových struktur. Tedy aby třídy si podobných vektorů byly reprezentovány neurony blízkými si v dané topologii.

Tato vlastnost je typická i pro skutečný mozek, kde např. jeden konec sluchové části mozkové kůry reaguje na nízké frekvence, zatímco druhý konec reaguje na frekvence vysoké.

Tímto způsobem lze zobrazit mnohodimenzionální prostor do prostoru jednoduššího, nejčastěji dvoudimenzionálního.

Díky zkušenostem bylo zjištěno, že při vstupu zahoří vítězný neuron, ať už s nebo bez samoexcitujících a inhibičních hran. Důsledkem toho je, že tyto hrany není nutno implementovat.

Topologie může být například následující



Obrázek 5.11. Kohonenova mapa

Algoritmus učení Kohonenovy mapy

1. **Inicializace sítě**

   Definuj $w_{i,j}(t)(0 \le i \le n-1)$ jako váhu mezi vstupem i a neuronem j v čase t.

   Inicializuj tyto váhy malými náhodnými čísly. Nastav počáteční okolí kolem neuronu j, $N_j(0)$ na maximum (obvykle tak, aby toto okolí pokrývalo celou výstupní vrstvu).
2. **Předložení vstupního vektoru**

   Předlož vstup ve formě $x_0(t), x_1(t), \ldots, x_{n-1}(t)$, kde $x_i(t)$ je vstup uzlu i v čase t.
3. **Výpočet vzdáleností (stanovení vítěze kompetice)**

   Vypočti vzdálenosti d_j mezi vstupním vektorem a každým neuronem následovně :

$$d_j = \sum_{i=0}^{n-1} (x_i(t) - w_{i,j}(t))^2$$

4. **Výběr minimální vzdálenosti**

Určení vítězného neuronu prostřednictvím výše vypočteného d pro každý neuron. Vítězný neuron označme jako j^*.

5. **Úprava vah**

Úprava vah pro neuron j^* a jeho sousedy definované jeho okolím. Nové váhy jsou dány vztahem: $w_{i,j}(t+1) = w_{i,j}(t) + \alpha(t)\left(x_i(t) - w_{i,j}(t)\right)$, pro j patřící do okolí vítězného neuronu $j^*, 0 \le i \le n-1$, kde $0 \prec \alpha(t) \prec 1$, přičemž hodnoty $\alpha(t)$ a velikost okolí neuronu se postupně snižují v čase s cílem stabilizovat nastavené váhy a lokalizovat místa maximální aktivity.

6. **Opakování od bodu 2.**

Proces shlukování lze vysvětlit prostřednictvím **funkce hustoty pravděpodobnosti**. Tato funkce reprezentuje statistický nástroj popisující rozložení dat v prostoru. Pro daný bod prostoru lze tedy stanovit pravděpodobnost, že vektor bude v daném prostoru nalezen. Je-li dán vstupní prostor a funkce hustoty pravděpodobnosti, pak je možné dosáhnout takové organizace mapy, která se této funkci přibližuje (za předpokladu, že je k dispozici reprezentativní vzorek dat). Jinými slovy řečeno, jestli jsou vzory ve vstupním prostoru rozloženy podle nějaké distribuční funkce, budou v prostoru vah rozloženy vektory analogicky.

Příklad 5.14.

Máme data ze zdravotnictví, pro reprezentaci metody vzorek s 1100 pacienty se strukturou (tedy se 4-rozměrnými vstupními vektory):

- věk
- pohlaví
- druh placené položky
- částka

Použijeme jednoduchou Kohonenovu mapu se čtvercovou topologií, kde jsou analogicky rozloženy i jednotlivé neurony. Atributy byly předzpracovány standardizací. na hodnoty z <0, 1>.

Výsledky shlukování
Léky:
Obecně jsou velké náklady u všech věkových kategorií.
U žen jsou více zvýšeny náklady ve věku 20 - 60 let.
Individuální léky:
U mužů malé náklady v pokročilém stáří.
U žen sice "malé" náklady, ale zato ve všech věkových kategoriích.
Prostředky zdravotní techniky:
U mužů i u žen náklady nízké do věku asi 65 let, pak již jen ojediněle.
Často bývají spojeny s hospitalizací a výkonem.
Zubař:
Obecně malé náklady zvýšené u starších žen.
***Ambulantní paušál**:*
U mužů zřídka vysoké náklady ve všech věkových kategoriích a asi od 40 let výše je stabilní nižší úroveň.
V období od 60 let bývá často spojen s hospitalizací a výkonem.
U žen většinou na vysoké úrovni. Nejvyšší v obdobích 0-30let, 40-55let, 70 let a výše. Jinak na nižší úrovni.
Výkon
U mužů řídce na vyšší úrovni stále, jinak v obdobích 0-50let a 60-80 let stabilní nižší úroveň.
U žen Vysoké náklady prakticky stabilně.
Extrémně zvýšená úroveň ve věku 15-50 let a pak v pokročilém stáří.

Ve věku od 15 do 30 let jsou náklady nejvyšší a navíc bývají velice často spojeny s hospitalizací a dopravou.
Hospitalizační paušál:
Pořád na nízké úrovni.
U žen jsou náklady vysoké ve věku 15-30 let.
Návštěvní služba :
Využívají spíše ženy. Muži ne.
Stabilně nízká úroveň nákladů.

♦

Poznámka:

Pokud chceme testovat účinky jednotlivých parametrů, použitých v algoritmu, pak je vhodné vyzkoušet změnu všech tří:

1. **Koeficient učení,** poslední zadávaná položka. Jedná se o počáteční koeficient, který určuje jak moc se vítězný neuron a jeho okolí ztotožní se vstupním vektorem. Tento koeficient se postupem času snižuje na nulu, takže nakonec nemá žádný vliv. Nemá smysl zadávat tento koeficient větší než 1, protože účinný bude od 1 níže.
2. **Okolí** - počáteční okolí vítězného neuronu, ve kterém budou upravovány váhy. Tato položka se s časem opět snižuje. Je vhodné ji na počátku zvolit tak, aby první okolí pokrývalo celou plochu mapy.
3. **Počet neuronů na hranu** - rozměr čtverce výstupní mapy, kde se jedná o mapu s (počet * počet) neurony. Není vhodné volit tento počet příliš vysoký, neboť s růstem neuronů roste také výpočetní doba.

❑ Gravitační shlukovací metoda

Jen jako ukázku toho, že uvedené typy metod nejsou všechny, které se v literatuře objevují a že některé spočívají na zcela odlišných principech, si uvedeme tzv. gravitační metodu, která shlukuje na základě gravitačního zákona

$$F = \kappa * m_1 * m_2 / r^2$$

kde κ je gravitační konstanta, zadávaná pro výpočet analytikem, (0.0001)
m jsou hmotnosti přitahovaných těles, hmotnost tělesa je rovna součtu hmotností objektů, hmotnost objektu je 1
r je vzdálenost mezi tělesy
F je výsledná působící síla.

Dále jsou zadány parametry
R_{min}, udávající minimální vzdálenost, při které se objekty spojí a
P udávající, kdy se posun neuskuskuteční, P= $R_{min}/100$

Gravitační algoritmus shlukování:

V každém kroku se

1. spojí objekty se vzdáleností menší než R_{min},
2. vypočítají se posuny všech objektů vůči ostatním
3. objekty posunou.

kroky se opakují tak dlouho, až je matice posunů nulová.

Shrnutí pojmů 5.2.

Metody nehierarchické a jejich typy.

Metody optimalizační.

Problém počátečního rozkladu, typické body. Metody pro definování typických bodů počátečního rozkladu.

Náhodný výběr typických bodů, zadání typických bodů, řízený výpočet typických bodů podle rozložení množiny shlukovaných bodů.

Metody optimalizační k-středové s pevným počtem shluků a jejich varianty. Typ výsledku shlukování.

Metody optimalizační k-středové s proměnným počtem shluků. Typ výsledku shlukování.

Fuzzy C-means algoritmy a úlohy touto metodou řešené.

Analýzy modů, algoritmus Kittlerův. Typ výsledku shlukování.

Shlukování pomocí Kohonenovy mapy.

Otázky 5.2.

1. Které typy metod nehierarchických shlukovacích znáte?
2. Proč neexistuje jediná spolehlivá shlukovací metoda?
3. Čím jsou charakteristické shlukovací metody optimalizační, k-středové?
4. Jakými způsoby se určují počáteční typické body pro shlukování a jejich počet?
5. Uveďte základní algoritmus shlukovací metody k-středové s pevným počtem shluků.
6. Uveďte základní body algoritmu shlukovací metody k-středové s proměnným počtem shluků.
7. Uveďte princip fuzzy k-středového shlukování.
8. Uveďte princip shlukování pomocí metod analýzy módů.
9. Uveďte princip shlukování pomocí neuronových sítí.
10. Podle čeho se budete rozhodovat při výběru metody pro shlukování konkrétní množiny objektů?
11. Existují jiné metody shlukovací, než uvedené typy? Proč?

Úlohy k řešení 5.2.

1. Pro data z kapitoly 5.1./1 BANKA navrhněte, kterou metodou budete data shlukovat, případně navrhněte pro zvolenou metodu vhodné parametry.

2. Totéž proveďte pro data 5.1./2 GYMNÁZIUM.

3. Najděte alespoň 3 praktické úlohy, kdy je vhodné použít optimalizační k-středový algoritmus shlukování a zdůvodněte, proč.

5.3. Shlukování hierarchické

Cíl Po prostudování této kapitoly budete umět

- popsat princip hierarchického shlukování aglomerativního i divizivního,
- zdůvodnit, proč není vhodné používat metody shlukující jen dvojice shluků,
- rozhodnout, kdy je k řešení úlohy vhodné najít hierarchii shluků,
- určit pro úlohu vhodnou míru podobnosti a vhodnou shlukovací strategii,
- pro praktické problémy najít optimální shlukovací hladiny a interpretovat výsledek.

Výklad

Shlukování hierarchické

Hierarchické shlukovací metody hledají nejen prostý rozklad objektů na shluky, ale jako výsledek dávají celou hierarchii takových rozkladů. V dané množině objektů se může vyskytovat řada malých shluků, které opět mohou být vzájemně různě vzdálené. Tak mohou tvořit „shluky shluků“ a vytvářet méně shluků větších (viz například testovací data 3). Takových hierarchických stupňů může existovat několik.

Postupně byla vyvinuta různými autory skupina metod, vytvářejících hierarchii rozkladů daných **n** bodů. Základní dva přístupy jsou

- **divizivním,** shora dolů, rozdělující postupně celou množinu bodů vždy na dva shluky, až dojdou k jednotlivým bodům,
- **aglomerativním**, zdola nahoru, vycházejících od jednotlivých bodů jako jednobodových shluků a shlukujícím postupně vždy nejbližší shluky, až dojdou k jedinému shluku ze všech bodů.

Nevýhodou obou přístupů může být příliš velký počet rozkladů. Prakticky se totiž využívá většinou jen několik málo stupňů rozkladu. Existují však takové modifikace hierarchického shlukování, které umějí počet stupňů rozkladu významně snížit.

Shlukování aglomerativní základní = po dvojicích shluků

Je dána množina objektů **O** a koeficient vzdálenosti shluků **V**.

Princip algoritmu procedury aglomerativní

1. Počáteční rozklad tvoří jednoobjektové shluky {Oi}; zvolíme míru vzdálenosti objektů, vypočte se **matice vzdáleností** objektů.
2. Nalezne se nejmenší **vzdálenosti shluků** (tzv. shlukovací hladina hierarchie).
3. Spojením těchto nejbližších shluků do společného shluku se vytvoří vyšší stupeň hierarchie, ostatní shluky zůstanou nezměněny.
4. Vypočtou se charakteristiky shluků aktuální hladiny rozkladu.
5. Pokud existuje více než 1 shluk, opakuje se od bodu 2.

V algoritmu jsme použili dva nové pojmy, které nejprve vysvětlíme.

❑ Matice nepodobností - vzdáleností objektů

Na počátku algoritmu aglomerativního se chápe každý objekt jako samostatný shluk. Vzdálenost těchto jednoobjektových shluků je rovna vzdálenosti objektů. Metriky pro výpočet vzdáleností jsme definovali v úvodu kapitoly 5. Zvolíme tedy míru vzdálenosti pro náš výpočet (*například Euklidovskou vzdálenost*).

Pak maticí vzdáleností rozumíme čtvercovou matici se všemi vzájemnými vzdálenostmi objektů. Protože je to matice symetrická, stačí počítat jen její polovinu.

Příklad 5.15.

Je dáno 7 objektů – lidí, spočítaná matice vzdáleností je:

	O1	O2	O3	O4	O5	O6	O7
O1	0	100	100	50	33	25	20
O2		0	100	50	50	33	33
O3			0	100	33	25	20
O4				0	33	20	25
O5					0	100	100
O6						0	100
O7							0

♦

❑ Míra vzdálenosti shluků – shlukovací strategie

Další nový pojem v algoritmu je vzdálenost shluků. Když se v průběhu shlukování vytvoří shluky několika objektů a opět se hledají nejbližší shluky, je nutné tento vztah shluků definovat. V literatuře se za řadu let objevily různé míry pro vzdálenost shluků. Někdy se jim říká shlukovací strategie.

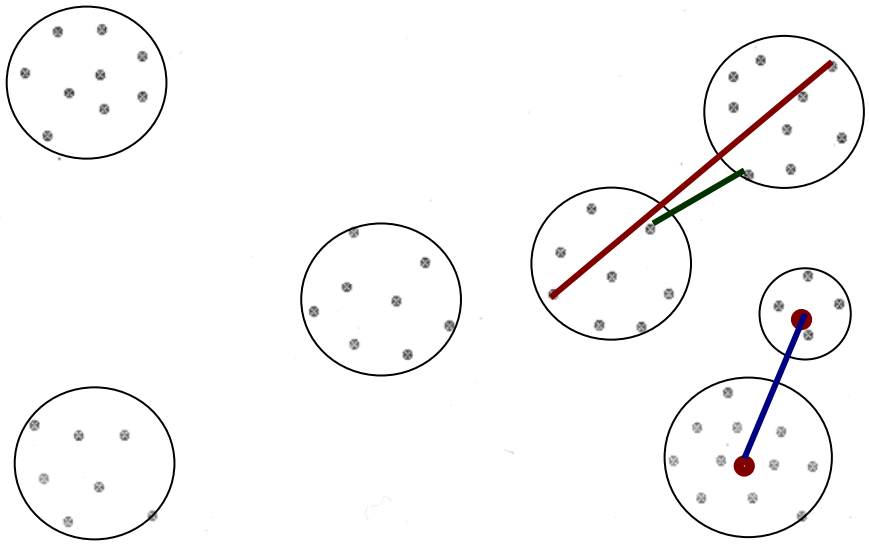
Nejznámější používané strategie jsou

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

Přesné vztahy pro jejich výpočet uvedeme níže. Protože je potřeba znát opět vzdálenosti všech shluků navzájem, jsou opět uspořádány do matice vzdáleností.

Příklad 5.16.

Testovací data 1 jsou v průběhu shlukování ve stavu, kdy již existuje 6 shluků podle následujícího obrázku. Vzdálenost shluků strategií nejbližšího souseda (zeleně), nejvzdálenějšího souseda (červeně) nebo centroidní (modře) je zobrazena na sousedních shlucích. Obdobně by se určila mezi všemi shluky navzájem.



♦

Některé z uvedených strategií bývají součástí SW balíků pro dolování. Dávají různé typy výsledků – shluky přirozené i shluky kulové.

Výhoda metody je v rekurzivitě algoritmu - není nutné počítat vzdálenosti shluků z původních souřadnic bodů, ale v každém kroku se pouze přepočítají vzdálenosti nově vzniklých shluků od ostatních. K tomu však je potřeba definice vzdálenosti shluků.

Ukazuje se, že všechny strategie se dají formulovat do jediného obecného schématu s několika parametry αi, αj, β, γ. Strategie se liší volbou těchto parametrů.

Obecné schéma pro výpočet vzdálenosti shluku U od shluku $R = P \cup Q$ je

$$V(U, R) = \alpha i \, .V\,(U, P) + \alpha j \, .V\,(U, Q) + \beta . \, V\,(P, Q) + \gamma \, . \, |V\,(U, P) - V\,(U, Q)|$$

Protože o něco níže ukážeme, že tento základní algoritmus má nedostatky, které jeho výsledky někdy zcela znehodnotí, nebudeme přesné vztahy pro jednotlivé strategie zatím uvádět. Uvedeme je až po zobecnění metody.

❑ Dendrogram

Problémem je, že všechny strategie produkují velké množství hierarchických stupňů a až „ručním“ výběrem analytik vybírá jednu nebo několik výsledných.

Výsledek hierarchie je vhodné vykreslit do grafu – shlukovacího stromu nazývaného **dendrogram.** Ovšem i dendrogram je jen jakýsi „polotovar“ k vyhodnocení. Podívejme se podrobněji, proč.

U klasické aglomerativní metody se v každém kroku – v každé hladině shlukují vždy 2 nejbližší shluky. Tedy pro **m** objektů existuje m-1 shlukovacích hladin. Je zřejmé, že tolik rozkladů není k ničemu užitečných. Je potřeba vybrat jen několik „nejlepších“. K tomu se využívá dendrogram s hodnotami shlukovacích hladin. Vyhodnocení provádí buď analytik, nebo je ho možno automatizovat. Z dendrogramu se vyhledávají

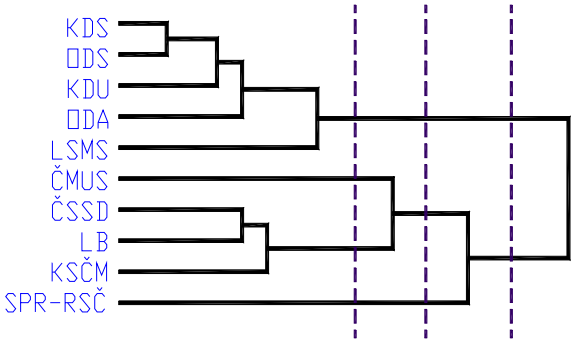
- největší rozdíly mezi sousedními hladinami … znamenají nejlepší hierarchické stupně
- hladina se zadaným počtem shluků … pokud existuje důvod nebo známe počet shluků předem
- zadání velikosti hladiny … pokud zadáme vzdálenost mezi shluky.

Obvykle nevyžadujeme příliš mnoho shluků, proto se dendrogram analyzuje „od konce“. Požadujeme-li kompletní „rozumnou“ hierarchii, najdeme několik málo (například 2 až 5) největších rozdílů mezi hladinami a rozklady na těchto stupních považujeme za výslednou hierarchii.

Příklad 5.17.

V letech 1995-96 bylo v parlamentu ČR celkem 10 stran a jejich 2049 hlasování bylo shluknuto následovně – pro zobrazení je použit dendrogram. Největší rozdíl mezi hladinami je mezi předposlední a poslední. To znamená, že nejlepší rozklad je na 2 shluky (vidíme, že odpovídají klasickému dělení politických stran na pravici a levici, s jedinou „výjimkou“ SRP-RSČ). Rozklad o hladinu níže tuto výjimku oddělí do samostatného jednoobjektového shluku.

Čteme-li dendrogram z opačné strany, skutečně strany si nejbližší se později spojily.



♦

❑ Charakteristiky výsledných shluků

Dendrogram nám dává jistou informaci o rozložení objektů do shluků, ale necharakterizuje výsledné shluky. V předcházející kapitole jsme si uváděli, že charakterizovat shluky přirozené nelze jen průměrnými hodnotami atributů (těžištěm shluku). Je těžké najít stručný popis výsledných shluků. Pokusíme se o to následujícím způsobem. Do výsledku uvedeme:

Název dat, rozměry dat (počet objektů, počet atributů), názvy atributů.

Zvolená metoda a její parametry.

Globální charakteristiky dat – pro každý atribut jeho min, max, avg, std.

min:	0.0	0.0	11.0	0.4
max:	9.0	18.0	58.0	23.6
avg:	2.3	5.2	23.8	12.0
std:	...			

Počet výsledných shluků (případně v každé shlukovací hladině).

Pro každý shluk (případně v každé shlukovací hladině):

Shluk 1

počet objektů:	20			
objekty:	O1, O4, O8, ...			
součet čtverců chyb:	30.8			
průměrná vzdálenost bodu od těžiště:	21.5			
avg:	2.3	8.2	33.8	12.0
std:	...			
min:	2.3	8.0	21.2	0.4
max:	2.3	8.3	34.8	21.9

Shluk 2 ...

Při vyhodnocování výsledků pak sledujeme, které hodnoty atributů jsou uvnitř shluku málo rozptýlené. Tyto atributy pak jsou charakteristické pro shluk. Které atributy nabývají hodnot rozptýlených, nebudeme pro charakteristiku shluku užívat.

Příklad 5.18.

Jsou dána data o 156 pacientech s těmito 9 atributy:

pohlaví	[0=muž, 1=žena]
věk	[0–82]
stav	[0=svob, 1=ženatý/vdaná, 2=rozved, 3=vdov]
měsíc_přijetí	[1-12]
puls	
tlak	
zlozvyk	[ne_alkoh+nekouří, ne_alkoh+kouří, alk_málo+kouří, alk_hodně+kouří]
diagnóza	0=post_páteře, 1=nemoc_srdce, 2=cukrovka, 3=intoxikace, 4=cirhóza_jater, 5=rakovina, 6=epilepsie, 7=kolapsy_slabost, 8=chudokrevnost]
výsledek	[0=domů, 1=přeložen_jinam, 2=zemřel]

Z výsledných 5 shluků byly vyhodnoceny charakteristiky a z nich jsou provedeny tyto závěry (tučně jsou charakteristické hodnoty atributu pro shluk, tence netypické, uvedené jen pro úplnost):

Shluk 1: pohl=0.4, **věk=70**, **stav= vdov**, měs=5, zloz=4, puls=70, tlak=170, **diag=rak, výsl=zemřel**

Shluk 2: **pohl=0.14**, věk=48, **stav= {svob,rozv}**, měs=6, **zloz=alkohol, puls=72, tlak=181, diag=epil, výsl=jinam**

Shluk 3: **pohl=0.91**, věk=58, stav= {vdaná}, **měs=11**, zloz=, puls=68, tlak=140, **diag=ledviny**, výsl=domů

Shluk 4: pohl=0.22, věk=45, stav= {ženatý,vdov}, měs=2, zloz=alkoh, puls=70, tlak=170, diag=rak, výsl=zemřel

Shluk 5: pohl=0.62, věk=59, stav= rozv, měs=7, zloz={nekouří,málo_alk}, puls=75, tlak=172, diag=rak, výsl=domů

Interpretace výsledků:

První shluk jsou starší pacienti (muži i ženy), ovdovělí s rakovinou, zemřeli.
Druhý shluk tvoří převážně muži, žijící osaměle, holdující alkoholu, s epilepsií.
Zajímavý třetí shluk tvoří starší ženy, které na podzim mají potíže s ledvinami.
atd.

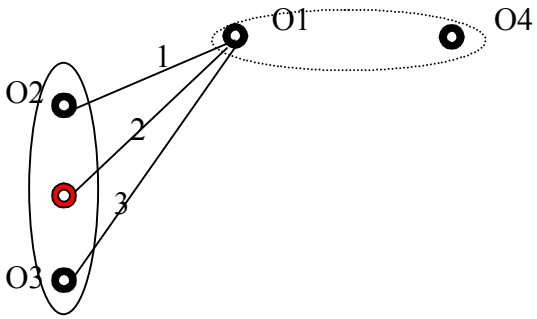
♦

❑ Shlukování aglomerativní definitní = po n-ticích shluků

Na tomto místě je nutné zdůraznit důležitý a ne obecně známý fakt. Původní hierarchické metody byly formulovány tak, že se v každém kroku shluknou **dva** nejbližší shluky vzhledem ke zvolené strategii shlukování. V případě, že není použit nejbližší soused a v některém hierarchickém kroku má stejnou nejmenší vzdálenost více dvojic či n-tic shluků, může dojít ke značné chybě výsledku.

Podívejme se podrobněji, proč. Na následujícím obrázku **x**. jsou tři černé body O1, O2, O3. Platí, že vzdálenosti V(O1, O2) = V(O2, O3). Shlukovací algoritmus, který bere vždy dva nejbližší shluky, si vybere body O2,O3 k vytvoření shluku. Potom se přepočítává vzdálenost tohoto nově vzniklého

shluku od všech ostatních shluků. Na obrázku jsou úsečkami zobrazeny tři vzdálenosti podle strategie nejbližšího (1), nejvzdálenějšího (3) souseda a mediánové (2). Je zřejmé, že vzdálenosti 2 a 3 jsou větší, než 1. Pokud v následujícím kroku shlukování (mediánovou strategií) bude existovat nejmenší vzdálenost menší, než 2, dojde ke vzniku nového shluku podle ní. Může se tak stát, že bod se dostane do shluku jiného, než ke svému nejbližšímu sousedovi.



Obrázek 5.13. Vzdálenost shluků strategií nejbližšího (1), nejvzdálenějšího (3) souseda a mediánovou (2).

Toto shlukování je tak závislé na pořadí objektů a může při různém pořadí čtení objektů dávat rozdílné výsledky.

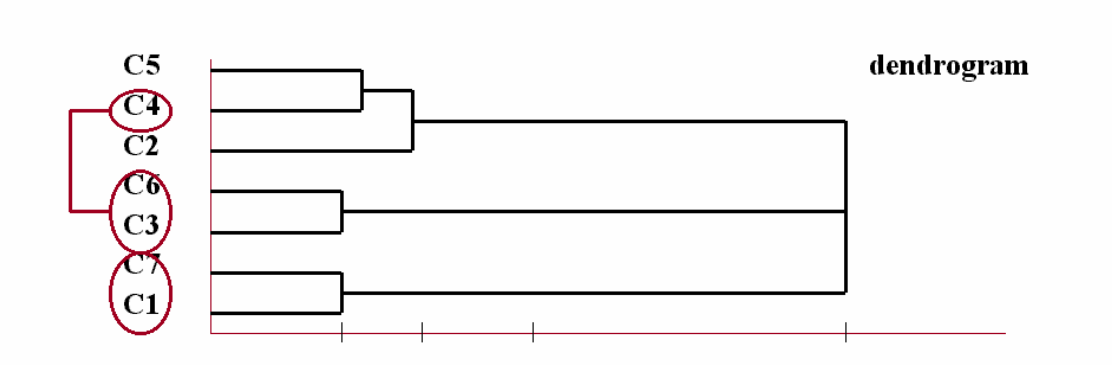
Některé SW systémy dosud používají chybný základní algoritmus a proto mohou dávat naprosto chybné výsledky!!!

Příklad 5.19.

Jsou dána data o 7 lidech. Ve vypočtené matici vzdáleností je nejmenší vzdálenost v 1. kroku shlukování rovna 20. Tato stejná nejmenší hodnota je mezi objekty O1-O7, O3-O6 a O6-O4.

	O1	O2	O3	O4	O5	O6	O7	
O1	**0**	**100**	**100**	**50**	**33**	**25**	**20**	**O1–O7-O3**
O2		**0**	**100**	**50**	**50**	**33**	**33**	
O3			**0**	**100**	**33**	**25**	**20**	
O4				**0**	**33**	**20**	**25**	**O6-O4**
O5					**0**	**100**	**100**	
O6						**0**	**100**	
O7							**0**	

Pro shlukování byl použit nejvzdálenější soused, výsledný dendrogram je

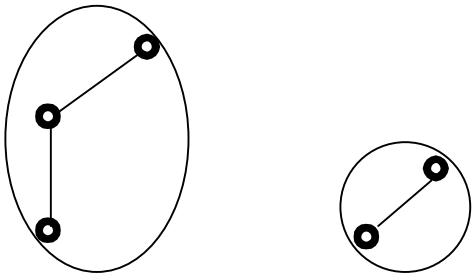


Obrázek 5.14. Dendrogram

Vidíme, že objekt O4, který má obecně nejblíže k O6, se díky zvolené strategii dostal dokonce do úplně jiného shluku! Přitom při jiném uspořádání objektů bychom mohli dostat jiný výsledek, protože jako první pro shlukování by se mohla vzít jiná dvojice bodů.

♦

Základní algoritmus byl později [x] zobecněn tak, že se v každém kroku shlukování spojují **všechny** shluky, které mají stejnou vzájemnou minimální vzdálenost. Přitom v jednom kroku nejen může vzniknout více shluků, ale může vzniknout nový shluk z více předcházejících – viz obrázek.



Obrázek 5.15. Stejné nejmenší vzdálenosti shluků

❑ Shlukovací strategie

Nyní si uvedeme vztahy pro současné shlukování všech n-tic se stejnou hodnotou minimální vzdálenosti:

Označíme:

$\mathbf{A} = \{A\}, \quad \mathbf{B} = \{B\}$	. . . jednoprvkové shluky
$\mathbf{U}, \mathbf{P}_1, \ldots, \mathbf{P}_t, \mathbf{Q}_1, \ldots, \mathbf{Q}_s$	. . . libovolné shluky
$\mathbf{P} = \mathbf{P}_1 \cup \ldots \cup \mathbf{P}_t, \quad \mathbf{Q} = \mathbf{Q}_1 \cup \ldots \cup \mathbf{Q}_s$	
$\lvert \mathbf{P} \rvert, \lvert \mathbf{Q} \rvert, \lvert \mathbf{U} \rvert$	. . . počty objektù shlukù
$D_{\_\_}$	. . . koeficient vzdálenosti shlukù

Pro všechny strategie platí

$V(A,B) = d(A,B)$

Strategie nejbližšího souseda

$$V(U,P) = \min_i \{ d(U,P_i) \}$$

$$V(U,P) = \min_i \left\{ \min_j \{ d(P_i,Q_j) \} \right\}$$

Strategie nejvzdálenějšího souseda

$$V(U,P) = \max_i \{ d(U,P_i) \}$$

$$V(U,P) = \max_i \left\{ \max_j \{ d(P_i,Q_j) \} \right\}$$

Strategie Ward-Wishartova

$$V(U,P_1 \cup P_2) = \frac{(|U|+|P_1|) \cdot V(U,P_1) + (|U|+|P_2|) \cdot V(U,P_2) - |U| \cdot V(P_1,P_2)}{|P_1 \cup P_2| + |U|}$$

$$V(U,P) = \frac{1}{|U|+|P|} \cdot \left(\sum_{i=1}^{t} (|U|+|P_i|) \cdot V(U,P_i) - \frac{|U|}{|P|} \cdot \sum_{i,j}^{\binom{t}{2}} (|P_i|+|P_j|) \cdot V(P_i,P_j) \right)$$

$$V(P,Q) =$$

$$= \frac{1}{|P|+|Q|} \cdot \left(\sum_{i,j}^{t \times s} (|P_i|+|Q_j|) \cdot V(P_i,Q_j) - \frac{|Q|}{|P|} \cdot \sum_{i,j}^{\binom{t}{2}} (|P_i|+|P_j|) \cdot V(P_i,P_j) - \frac{|P|}{|Q|} \cdot \sum_{i,j}^{\binom{s}{2}} (|Q_i|+|Q_j|) \cdot V(Q_i,Q_j) \right)$$

Strategie centroidní

$$V(U,P) = \frac{1}{|P|} \cdot \sum_{i=1}^{t} |P_i| \cdot V(U,P_i) - \frac{1}{|P|^2} \cdot \sum_{i,j}^{\binom{t}{2}} |P_i| \cdot |P_j| \cdot V(P_i,P_j)$$

$$V(P,Q) =$$

$$= \frac{1}{|P| \cdot |Q|} \cdot \left(\sum_{i,j}^{t \times s} (|P_i|+|Q_j|) \cdot V(P_i,Q_j) - \frac{1}{|P|^2} \cdot \sum_{i,j}^{\binom{t}{2}} (|P_i|+|P_j|) \cdot V(P_i,P_j) - \frac{1}{|Q|^2} \cdot \sum_{i,j}^{\binom{s}{2}} (|Q_i|+|Q_j|) \cdot V(Q_i,Q_j) \right)$$

Strategie průměrné vzdálenosti objektů

$$V(U,P) = \frac{\sum_{i=1}^{t} |P_i| \cdot V(U,P_i)}{|P|}$$

$$V(P,Q) = \frac{\sum_{i,j}^{t \times s} |P_i| \cdot |Q_j| \cdot V(P_i,Q_j)}{|P| \cdot |Q|}$$

Strategie mediánová

$$V(U,P) = \frac{\sum_{i=1}^{t} V(U, P_i)}{t} - \frac{\sum_{i,j}^{\binom{t}{2}} V(P_i, P_j)}{t^2}$$

$$V(P,Q) = \frac{\sum_{i,j}^{t\times s} \cdot V(P_i, Q_j)}{t \cdot s} - \frac{\sum_{i,j}^{\binom{t}{2}} V(P_i, P_j)}{t^2} - \frac{\sum_{i,j}^{\binom{s}{2}} V(Q_i, Q_j)}{s^2}$$

Tak může být sestaven jediný obecný algoritmus aglomerativních metod. Jednotlivé metody se v něm liší pouze výpočtem vzdáleností shluků.

❑ Algoritmus aglomerativní definitní = po n-ticích

Pro množinu objektů **O** se sestrojí posloupnost rozkladů Ωo, Ω1, . . ., Ωn-1, v ní se přiřadí každému shluku S reálné nezáporné číslo h (S) nazývané **shlukovací hladinou**, odpovídající jeho vzniku, takto:

1. Počáteční rozklad Ωo tvoří jednotlivé objekty = jednoprvkové shluky, h(S)=0
2. Rozklad Ωi = {Si1, Si2, ..., Sik} je rozklad v i-tém kroku procedury, v něm jsou shlukům Sij přiřazena čísla h (Sij) a

   $$\mu i = \min_{x\neq y,\ x,y=1,..,k} \mathbf{V}(Six, Siy)$$

   je minimální hodnota koeficientu vzdálenosti V. Potom každý μi-souvislý shluk shluků Sik_1, Sik_2, ..., Sik_v rozkladu Ωi přechází do následujícího rozkladu

   $$\Omega i+1 = \{Si+1,1, . . ., Si+1,k\}$$

   jako shluk $Si+1,l = Si+1,1 \cup . . . \cup Si+1,k$

   přičemž definujeme h(Si+1,l) = μi. Ostatní shluky zůstávají v novém rozkladu nezměněny.
3. V posledním kroku je Ωn = {Sn1}={**O**} o jediném shluku = celé množině objektů,

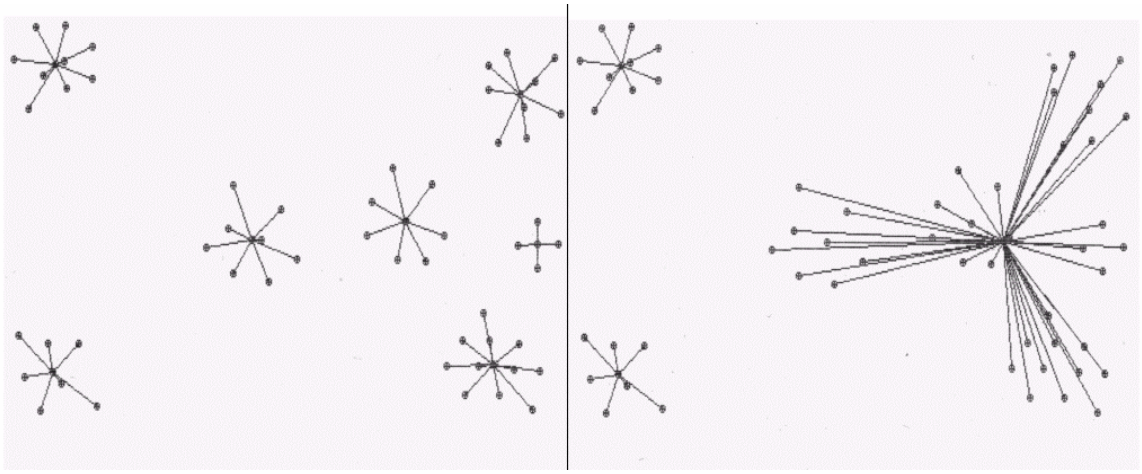
   $$h(Sn1) = \mu\ n-1, \text{ kde } \{Sn1\}$$

Výhodou tohoto algoritmu pro všechny strategie je především odstranění zmíněné možné chyby výsledného rozkladu, popsané u základního algoritmu, shlukujícího po dvojicích.

Další výhodou je menší počet shlukovačích hladin, tedy menší počet hierarchických úrovní. I v tomto případě jich však bývá často mnohem více, než je reálně potřebné. Výrazně se zrychlí i zkvalitní výsledek u dat s pravidelnější strukturou (*viz například data 6 – homogenní, která se shluknou v jediném kroku*).

Příklad 5.20.

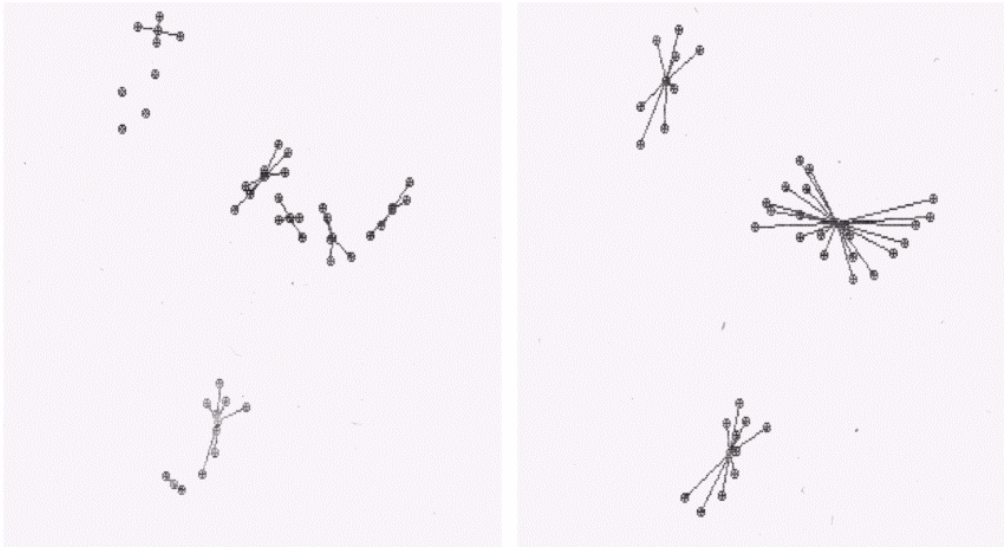
Nejbližší soused, 2 nejlepší rozklady



♦

Příklad 5.21.

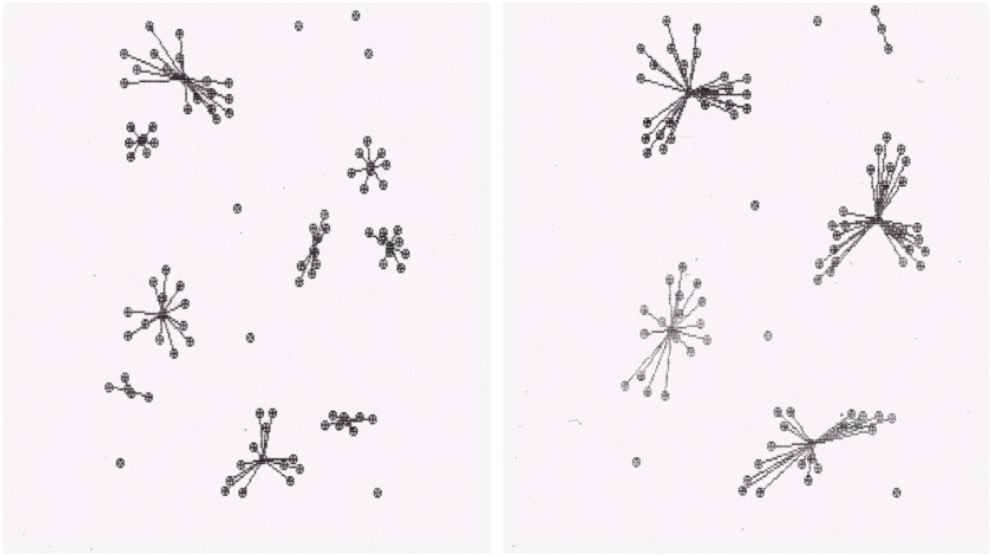
Nejbližší soused, 2 nejlepší rozklady



♦

Příklad 5.22.

Nejbližší soused, 2 nejlepší rozklady, zřetelné jsou i izolované body

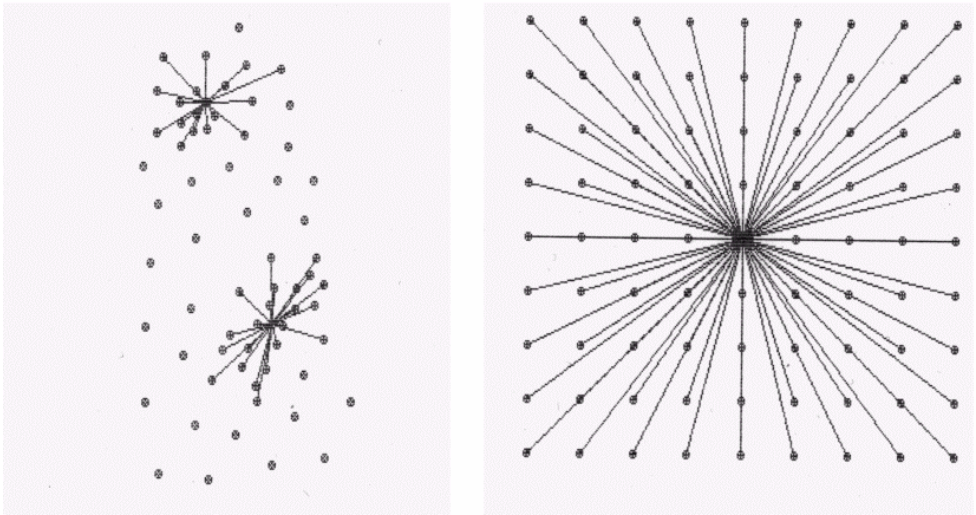


♦

Příklad 5.23.

Nejbližší soused na datech bez shluků, jen s hustšími místy.

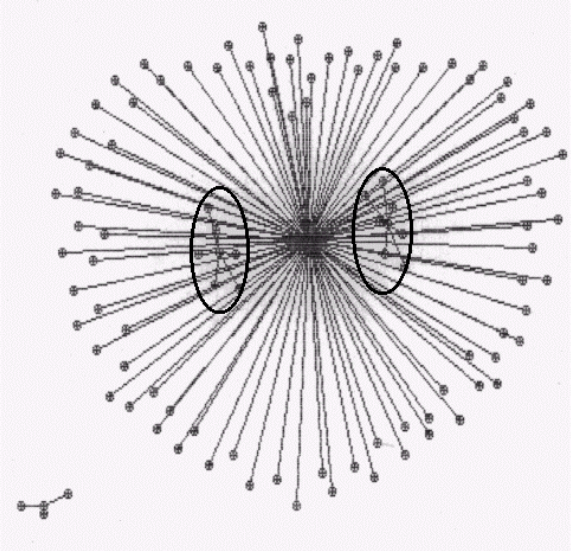
Nejbližší soused na homogenních datech – výsledek s jedinou shlukovací hladinou.



♦

Příklad 5.24.

Nejbližší soused, data odolávající mnoha algoritmům. Nejlepší hladina jasně oddělí od sebe jak 3 malé shluky vně i uvnitř, ale najde i shluk v uzavřeném tvaru srdce.



♦

❑ Tolerance při aglomerativním shlukování

Posledním příspěvkem ke zkvalitnění výpočtu je možnost využití zadané tolerance při určování nejmenší vzdálenosti shluků. Při velkém počtu shlukovaných objektů může být stále počet hierarchických stupňů vysoký. Přitom může docházet k tomu, že se vzájemné vzdálenosti od sebe liší jen málo a tento rozdíl nehraje pro rozklad žádnou podstatnou roli.

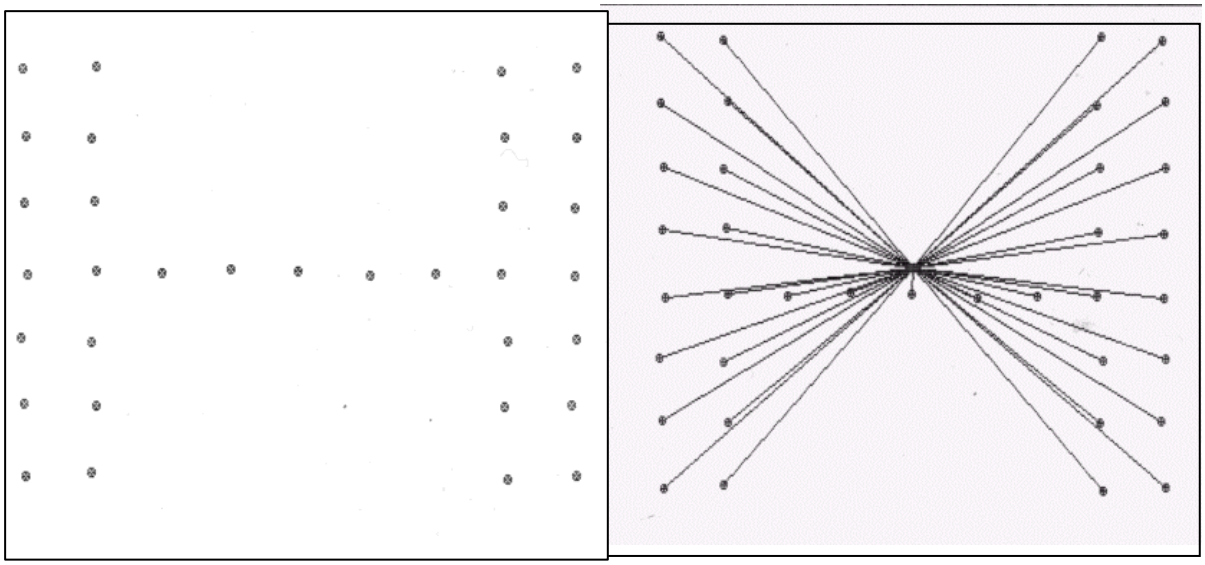
Pokud tedy používáme definitní aglomerativní algoritmus, můžeme urychlit shlukování takto: definujeme toleranci (například v procentech), o kolik se mohou lišit vzdálenosti shluků od nejmenší vzdálenosti, aby byly pořád ještě považovány za „stejné“. Tolerance může být zadána analytikem, nebo nastavována automaticky. Tak se na jedné hladině shlukne současně více shluků, počet hladin se zmenší a výpočet se zkrátí.

Použití tolerance je zvláště výhodné u experimentálních dat, kde se již při měření vyskytuje jistá chyba a přesný výpočet vzdáleností je dokonce zkreslující.

Příklad 5.25.

Pro otestování schopností shlukovacích metod byla vytvořena umělá data, tzv. H-data, kde jednotlivé body jsou zatíženy chybou ± 0.1. Až na tuto chybu data tvoří jeden homogenní přirozený shluk.

Použitím metody Ward-Wishartovy s tolerancí 20 % došlo k výpočtu jediné hladiny:

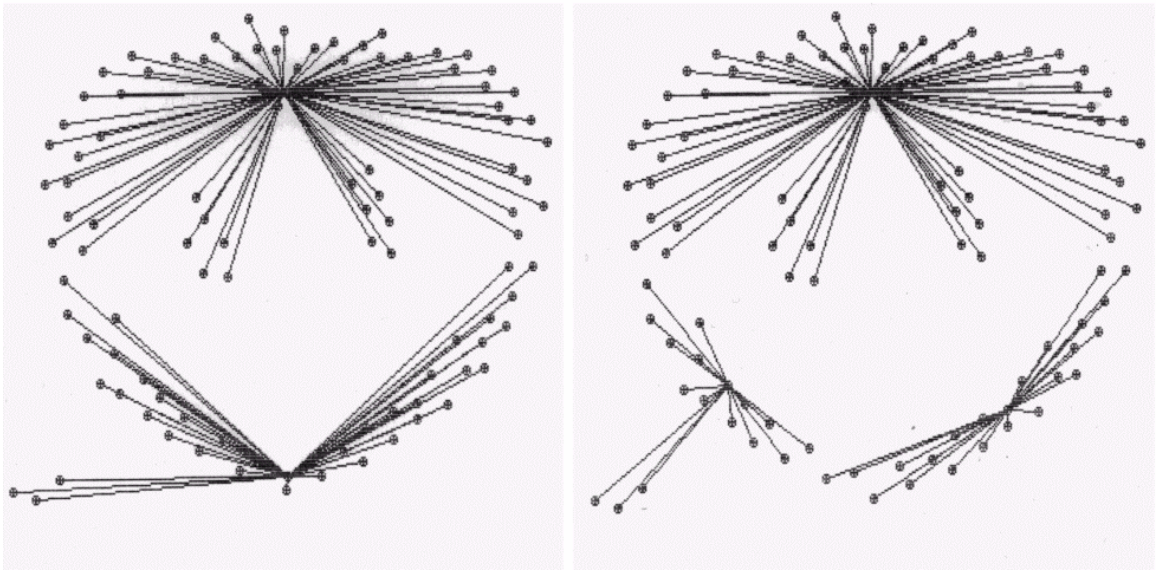


♦

Příklad 5.26.

Data srdce, metoda Ward-Wishartova, tolerance 10%.

Přesto, že metoda byla vyvinuta pro nalezení přirozených shluků a při vhodné toleranci na „stejné“ nejmenší vzdálenosti vykazuje dobré výsledky, tato data nezvládne. Výsledek má obdobný efekt, jako metody k-středové.



♦

❑ Shlukování divizivní

Divizivní hierarchické metody vycházejí z jednoho shluku, vytvořeného celou množinou objektů. Postupným rozdělováním dojdou až k jednoprvkovým shlukům. Vzájemně se liší způsobem rozdělování větších shluků na menší. Výsledkem je opět dendrogram, tentokrát od kořene (celá množina) k listům (jednotlivé objekty).

Dosud žádná z metod divizivních nemá zřetelnou výhodu proti aglomerativním. Rozdělení množiny na „podobnější si" objekty vyžaduje náročnější výpočty. Někdy se uvádí, že dělení „shora" najde rychleji několik základních shluků. Obvykle totiž nevyžadujeme celou hierarchii, ale vystačíme s několika shlukovacími úrovněmi. Ovšem dělení každého jednoho shluku se provádí vždy na 2 části, ty opět na 2 části atd. Vzniká proto obdobný efekt, jako při shlukování po dvojicích – při stejných shlukovacích hladinách se tyto informace ztrácí a může docházet ke zkreslení výsledku.

Uvedeme si algoritmus jedné z divizivních metod - metodu MacNaughtona-Smitha.

Algoritmus divizivní

1. Mějme shluk U.
2. Najdeme v něm bod K s největší průměrnou vzdáleností **vz** od ostatních bodů shluku. Tento bod tvoří základ nově vznikajícího shluku V, oddělujícího se od shluku U.
3. Hladina rozdělení v tomto okamžiku je rovna **vz/2**.
4. Dále ve shluku U najdeme bod, který má největší rozdíl vzdáleností G od zbývajících bodů shluku U a od všech bodů shluku V (obě vzdálenosti průměrné).
   Je-li G > 0, přemístíme tento bod do shluku V a proces opakujeme,
   G ≤ 0, rozdělování shluku U ukončíme a za hladinu rozdělení dosadíme poloviční součet vzdáleností prvního neodděleného bodu od bodů shluku U a V.
5. Rozdělujeme-li shluk o dvou objektech, přiřadíme rozdělovanému shluku hladinu o hodnotě poloviční vzdálenosti jeho objektů.
6. Rozdělování shluků se opakuje, dokud existují shluky víceprvkové.

❑ Postup při shlukování reálných dat

Shrňme si na závěr doporučený optimální postup shlukování, máme-li analyzovat reálná data:

1. vybereme z dat reálné a ordinální atributy
2. zvolíme hledisko, ze kterého budeme hledat podobnost objektů
3. vybereme atributy charakterizující tuto podobnost
4. podle typu atributů zvolíme míru podobnosti či míru vzdálenosti objektů
   - koeficienty korelace, asociace
   - metriky
5. mají-li reálná data různé měrné jednotky či rozdílné domény, standardizujeme je
6. podle typu řešené úlohy zvolíme pojem „shluku" jako optimalizovaný kulový nebo existující přirozený
7. podle typu řešené úlohy zvolíme typ rozkladu – prostý rozklad, hierarchie rozkladů, případně fuzzy-rozklad
8. hledáme-li kulové shluky a známe počet výsledných shluků,
   - zvolíme metodu k-středovou FORG se zadaným k
   - známe-li počáteční typické body, zadáme je, jinak pro ně volíme některou z metod TYPn

9. hledáme-li kulové shluky a neznáme počet výsledných shluků,
   - zvolíme metodu k-středovou CLASS, k zadáme přibližně
   - počáteční typické body zadáme některou z metod TYPn
10. hledáme-li přirozené shluky,
    - zvolíme aglomerativní metodu se strategií nejbližšího souseda; podporuje-li náš SW shlukování n-tic, případně zadání tolerance, zvolíme toleranci 10% ... dostaneme výsledek VYSL1
    - podle výsledného dendrogramu zvolíme jeden nebo několik optimálních rozkladů na shluky a reprezentanty těchto shluků jako jejich vnitřní body
    - zadáme známý počet shluků **k** i známé reprezentanty shluků jako typické body do metody k-středové FORG ... dostaneme výsledek VYSL2
    - na každé shlukovací úrovni porovnáme shluky obou výsledků
      - jsou-li oba výsledky stejné, množina našich objektů tvoří přirozené α-shluky dostatečně vzájemně vzdálené
      - liší-li se podstatně oba výsledky, množina našich objektů tvoří přirozené α-shluky, které nejsou od sebe příliš vzdálené nebo tvoří složité „nekulové“ tvary.
11. podle výsledků shlukovací analýzy a statistických charakteristik výsledných shluků interpretujeme charakteristické vlastnosti každého shluku; formulujeme je jako pravidla, podle kterých se objekt přiřadí do příslušného shluku
12. vyhodnotíme výsledek celé shlukovací analýzy

❑ Praktické příklady použití shlukovací analýzy

Příklad 5.27.

Během postgraduálního kurzu, kde byla mj. vykládána metoda GUHA, byla získána následující data. Objekty tvoří posluchači kurzu, data jsou získána z dotazníků, které posluchači vyplnili. Otázky dotazníku jsou několika druhů: jeden blok jsou obecné vlastnosti (pohlaví, věk, ...), druhý jejich psychologické vlastnosti (jste důvěřivý, ...), třetí pak jejich reakce na konkrétní hudební ukázky (autor dotazníku se mj. zabývá vztahem hudby a matematiky, zde výzkumem souvislostí mezi psychologickými vlastnostmi lidí a způsobem reakce na hudbu). Otázky jsou formulovány ano - ne, tedy data jsou binární. Posluchačů bylo 30, otázek 28. Hudební ukázky byly úryvky těchto skladeb: J.S.Bach: 25. variace z Goldbergovských variací pro cembalo, G.Mahler: začátek Adagia z 6.symfonie a A.Honegger: závěr 2. symfonie.

Otázky:	četnost A - N
1. Jste jedináček ?	6 - 24
2. Jste narozen před rokem 19xx ? (nad 35 let, pod 35 let)	26 - 4
3. Máte děti ?	20 - 10
4. Myslíte, že jste důvěřivý ?	16 - 14
5. Hrajete nebo hrál jste na nějaký hudební nástroj ?	17 - 13
6. Vadí vám neslušné vtipy ve společnosti ?	3 - 27
7. Věříte v nějaký vyšší řád nebo myslíte, že vše je dílem náhody ?	21 - 9
8. Máte rád rychlá rozhodnutí ?	19 - 11
9. Připadá vám ukázka Bacha radostná nebo smutná ?	8 - 22
10. Připadá vám ukázka Bacha vzrušená nebo klidná ?	21 - 9
11. Připadá vám ukázka Bacha blízká nebo cizí ?	26 - 9
12. Připadá vám ukázka Mahlera radostná nebo smutná ?	7 - 23

13. Připadá vám ukázka Mahlera vzrušená nebo klidná ?	7 - 23
14. Připadá vám ukázka Mahlera blízká nebo cizí ?	18 - 12
15. Připadá vám ukázka Honeggera radostná nebo smutná ?	26 - 4
16. Připadá vám ukázka Honeggera vzrušená nebo klidná ?	29 - 1
17. Připadá vám ukázka Honeggera blízká nebo cizí ?	10 - 20
18. Jste muž ?	27 - 3
19. Líbí se vám líčení u žen ?	17 - 13
20. Je víno vhodnější nápoj pro Čechy než pivo ?	10 - 20
21. Chodíte aspon jednou za dva měsíce na koncerty ?	6 - 24
22. Líbí se vám více červená než zelená ?	10 - 20
23. Líbí se vám více žlutá než modrá ?	9 - 21
24. Četl byste raději Tři muže ve člunu než Tři kamarády ?	16 - 13
25. Strávil byste volný večer raději sám doma nebo ve společnosti ?	16 - 13
26. Měl jste někdy při hudbě barevné představy ?	15 - 15
27. Měly souboje něco do sebe ?	25 - 5
28. Jste nápaditý ?	10 - 20

Shrnutí pojmů 73.

Typy shlukovacích metod hierarchických. Metody aglomerativní a divizivní.

Základní algoritmus shlukování aglomerativního po dvojicích shluků.

Matice nepodobností - vzdáleností objektů.

Míra vzdálenosti shluků – shlukovací strategie. Strategie nejbližšího souseda, nejvzdálenějšího souseda, mediánová, centroidní, Ward-Wishartova.

Dendrogram, shlukovací strom.

Výsledné charakteristiky shluků a jejich interpretace.

Chyba algoritmu shlukování nejbližších dvojic.

Shlukování aglomerativní definitní, po n-ticích shluků.

Tolerance výpočtu při aglomerativním shlukování.

Shlukování divizivní, základní princip metod divizivních.

Interpretace výsledků shlukování.

Kompletní řešení shlukovací úlohy a jeho etapy.

Otázky 5.3.

1. Co je shlukování hierarchické?
2. Které typy hierarchických metod znáte a čím se liší?
3. Uveďte základní algoritmus aglomerativního shlukování po dvojicích.
4. Co je a k čemu je matice vzdáleností?
5. Co je vzdálenost shluků a které míry vzdálenosti znáte?

6. V čem je rozdíl mezi jednotlivými mírami vzdáleností shluků?
7. Proč může být výsledek aglomerativního shlukování po dvojicích nejbližších shluků chybný?
8. Uveďte základní algoritmus aglomerativního shlukování definitního.
9. Co je výsledkem shlukování hierarchického?
10. Co je dendrogram a k čemu slouží?
11. Které charakteristiky výsledných shluků je vhodné znát pro interpretaci výsledků?
12. Jak se pomocí výsledných charakteristik shluků provádí interpretace výsledků?

Úlohy k řešení 5.3.

1. Řešte data z kapitoly 5.1./1 BANKA jako hierarchickou úlohu, navrhněte, kterou metodou budete data shlukovat, případně navrhněte pro zvolenou metodu vhodné parametry.

2. Totéž proveďte pro data 5.1./2 GYMNÁZIUM.

3. Najděte alespoň 3 praktické úlohy, kdy bude vhodné použít pro dolování znalostí hierarchickou shlukovací metodu. Zvolte ke každé z nich míru vzdálenosti objektů, shlukovací strategii, toleranci „shody“ nejmenších vzdáleností a zdůvodněte tyto volby.

6. KLASIFIKACE

Čas ke studiu: 1 hodina

Cíl Po prostudování této kapitoly budete umět

- charakterizovat úlohy vhodné pro řešení rozhodovacím stromem
- popsat princip algoritmu konstrukce RS
- řešit praktické úlohy pomocí RS

Výklad

Co je rozhodovací strom

Jistou kombinací hledání shluků objektů a nalezení asociací mezi atributy je klasifikace objektů pomocí rozhodovacího stromu.

Na rozdíl od shlukování, které hledá shluky bez apriorní znalosti klasifikačních tříd, jsou zde tyto třídy dány. Hledají se podmínky ve formě hodnot vstupních atributů, za kterých padne objekt do jednotlivých klasifikačních tříd.

Na rozdíl od asociací nehledá, ale konstruuje optimální implikace mezi množinou vstupních atributů a výslednou klasifikací.

Rozhodovacím stromem se nazývá tato metoda proto, že výsledná pravidla se s výhodou zobrazují formou stromu. Uzly zobrazují atributy, podle nichž se právě dělí, hrany charakterizují jednotlivé hodnoty atributů.

Konstrukce rozhodovacích stromů je metodou získávání znalostí z dat, která v datech **hledá charakteristický popis zadaných tříd pomocí kombinací hodnot atributů.**

Patří k metodám induktivního učení s učitelem. To znamená, že z úplných dat (včetně hodnot klasifikační třídy) odvodíme pravidla. Ty pak můžeme použít i pro data, u nichž klasifikační třídu neznáme. Podle odvozených pravidel ji předpovíme.

Základní pojmy

Je dána matice **X** o **m** objektech **O**={O_1 ,…, O_m} a **n** atributech **A**={A_1 ,…, A_n} s doménami **D**j={Dj0,…Djh}, množina tříd **C** = {C1, … Ck}; třídy jsou charakterizovány hodnotami jednoho nebo několika atributů **T** ⊂ **A**. Jedna nebo několik kombinací hodnot atributů z **T** definuje jednu třídu Ci.

Trénovací množinou T pro konstrukci konkrétního rozhodovacího stromu nazýváme takovou množinu objektů **O** = {O1, O2, …, Om} s atributy **A** = {A1, A2, …, An }, kde každý objekt je ohodnocen některou hodnotou klasifikačního atributu C.

Nazveme

atributy Ai ⊂ (**A-T**) **předpovídajícími** (~ vstupní, antecedenty)
atributy Aj ⊂ **T předpovídanými** (~ klasifikační, výstupní, sukcedenty)

X

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

Klasifikačním atributem nazýváme kategoriální atribut C, jehož hodnota určuje třídu objektu. Klasifikační atribut nemusí být primárně zadaným atributem, ale může být odvozen z kombinací hodnot předpovídaných atributů.

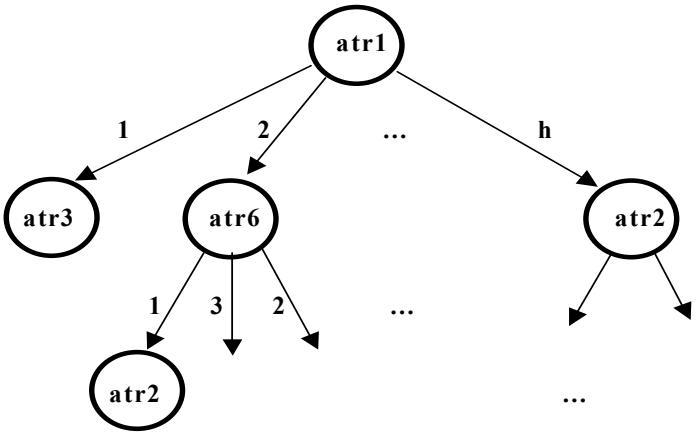
Rozhodovací strom je orientovaný acyklický souvislý graf – strom. Každý vnitřní uzel spolu s hranami z něj vycházejícími reprezentuje rozdělení trénovací množiny.

Každý list má přiřazenu hodnotu klasifikačního atributu, reprezentující nejpočetnější třídu objektů z podpůrné množiny listu. Nazýváme jej **klasifikační třídou listu.**

Podpůrná množina uzlu P_U je taková podmnožina trénovací množiny, do které patří objekty splňující podmínky rozdělení reprezentované vnitřními uzly a hranami na cestě od kořenu stromu k tomuto uzlu. Podpůrnou množinou kořene stromu je celá trénovací množina.

List je prostý, pokud všechny objekty jeho podpůrné množiny patří do stejné třídy.

Chyba listu je počet objektů podpůrné množiny listu, které nepatří do klasifikační třídy listu.



Obrázek 8.1. Rozhodovací strom

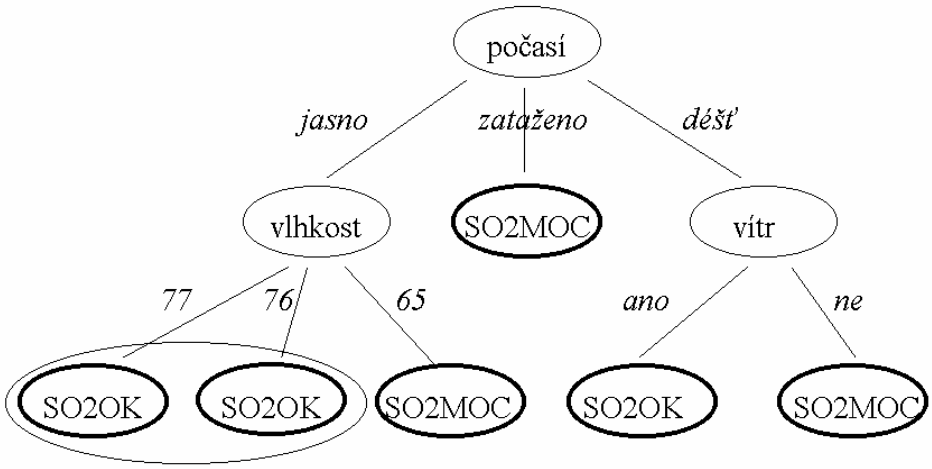
Příklad 8.1.

Použijeme často citovaný příklad Počasí [x]. *V datech z následující tabulky zaznamenané údaje o počasí, vlhkosti a větru chápeme jako předpovídající atributy, koncentraci oxidu siřičitého v ovzduší jako předpovídaný atribut. Jinak řečeno hledáme pravidla, pomocí nichž budeme na základě znalosti hodnot počasí, vlhkosti a větru předpovídat koncentraci SO2. Podle normy je maximální povolená hodnota SO2 = 150μg/m3, vyšší hodnota znamená překročení normy. Na*

základě toho určíme (odvodíme) z předpovídané hodnoty koncentrace SO2 dvě třídy SO2MOC a SO2OK.

předpovídající			předpovídané	klasifik.. třída
počasí	vlhkost[%]	vítr	koncentrace SO2	
zataženo	78	ne	220	SO2MOC
déšť	80	ne	195	SO2MOC
jasno	76	ne	130	SO2OK
jasno	65	ano	200	SO2MOC
déšť	70	ano	115	SO2OK
déšť	80	ano	110	SO2OK
jasno	77	ano	120	SO2OK

Pomocí algoritmu popsaného níže dostaneme následující rozhodovací strom:



Obrázek 8.2. Strom předpovídání koncentrace SO2 z počasí

♦

❑ Princip metody

Máme matici **X** o **m** objektech a **n** atributech a množinu tříd **C** = {C1, … Ck}.

Můžeme rozlišit dva případy:

- Všechny objekty z **X** patří do stejné třídy Ci (mají konstantní hodnotu posledního sloupce C). Pak rozhodovací strom pro **X** obsahuje jediný list, reprezentující třídu Ci.
- Objekty z **X** patří do více než jedné třídy C1, …, Ct. Pak je rozhodovací strom konstruován následujícím postupem. Vybere se jeden z antecedentů Ai, ten tvoří uzel stromu. Pro něj se provede rozklad τ uzlu tak, že každé hodnotě Dj atributu Ai je přiřazena jedna hrana označená hodnotou Dj, $j \in \{j1,\ldots, jk\}$. Rozklad rozdělí množinu **X** do podmnožin $\mathbf{X}_{j1}, \ldots, \mathbf{X}_{jk}$ tak, že objekt Oi patří do podmnožiny **X**j právě tehdy, když Oi má v rozkladu τ výstup Dj.

Další růst stromu je rekurzivní aplikací bodů 1 a 2 na podmnožiny $\mathbf{X}_{j1}, \ldots, \mathbf{X}_{jk}$. V bodě 2 se vybírá pro rozklad takový antecedent, který ještě nebyl použit v cestě od kořene stromu k aktuálnímu uzlu. Když není možno aplikovat bod 2, růst stromu se zastaví.

Dosud jsme blíže nespecifikovali, který z dosud nepoužitých atributů bude v konkrétním okamžiku vybrán pro další rozklad. Přitom tento výběr je rozhodující pro správnost klasifikace objektů.

Jeden z možných postupů je založen na informačním zisku rozhodovacího stromu při přechodu z jednoho stavu budování stromu do druhého, rozšířeného o další uzel. Informační zisk je měřen rozdílem entropie rozkládané podmnožiny objektů a váženého součtu entropií množin představujících výsledky rozkladu.

❑ Konstrukce optimálního rozhodovací stromu

Nechť **X** je počáteční učící množina objektů a $X \subset \mathbf{X}$ je podmnožina, která má být v nějaké fázi budování rozhodovacího stromu rozložena rozkladem pomocí atributu $A \subset \mathbf{A}$. Nechť D1, …, Dk jsou výstupy tohoto rozkladu a platí pro ně:

$$Di \subseteq \mathbf{D}, \quad \cup Di = \mathbf{D}, \quad Di \cap Dj = \varnothing, \quad i \neq j, \; 1 \leq i, j \leq k$$

Označme frek (Cj, D) počet výskytů objektů třídy Cj v množině D,

|D| nebo |Di| počet objektů v množině D nebo Di.

Pro rozkládanou množinu **D**, předpovídající atribut **A** s doménou {D0, ... Dh} a klasifikační třídu **C** s doménou {C1, … Ck} je frekvenční tabulka

C \ A	**D_0**	**D_1**	**D_2**	**…**	**D_h**	
C_0	**$\|D_{00}\|$**	**$\|D_{01}\|$**	**$\|D_{02}\|$**		**$\|D_{0h}\|$**	**frek(C_0, D)**
C_1	**$\|D_{10}\|$**	**$\|D_{11}\|$**	**$\|D_{12}\|$**		**$\|D_{1h}\|$**	**frek(C_1, D)**
C_2	**$\|D_{20}\|$**	**$\|D_{21}\|$**	**$\|D_{22}\|$**		**$\|D_{2h}\|$**	**frek(C_2, D)**
…						
C_k	**$\|D_{k0}\|$**	**$\|D_{k1}\|$**	**$\|D_{k2}\|$**		**$\|D_{kh}\|$**	**frek(C_k, D)**
	$\|D_0\|$	**$\|D_1\|$**	**$\|D_2\|$**		**$\|D_h\|$**	**|D|**

Potom definujeme:

Entropie množiny D je dána vztahem (suma pro j = 1, …, k)

$$\text{info}(D) = -\sum (\text{frek}(Cj, D) / |D| \; * \log_2 (\text{frek}(Cj, D) / |D|))$$

Vážený součet entropií množin představujících výstupy rozkladu nad atributem A je

$$\text{info}_A(D) = \sum |Di| / |D| * \text{info}(Si)$$

Informační zisk rozkladu nad atributem A označujeme gain (A) a je měřen rozdílem entropií před rozkladem a po něm.

$$\text{gain}(A) = \text{info}(D) - \text{info}(Di)$$

Pro rozklad se vybere ten atribut, jehož použití vede k největšímu zisku informace.

❑ Převod rozhodovacího stromu na pravidla

Místo reprezentace výsledku pomocí rozhodovacího stromu je možno použít ekvivalentního zápisu pomocí množiny **produkčních pravidel**. Levé strany pravidel jsou tvořeny konjunkcemi hodnot atributů v cestě od kořene stromu k listu, který označuje jednu třídu z pravé strany pravidla. Majoritní třída (ta, která koresponduje s největším počtem listů) je prohlášena za **implicitní třídu** a tedy pravidla pro ni se neuvádí.

Výsledný rozhodovací strom lze mechanicky převést na pravidla typu

Jestliže α pak β se spolehlivostí ε

kde α je podmínka tvaru elementární konjunkce typu A1=a1 ∧ A2 = a2 ∧ …

β je určení příslušnosti ke klasifikační třídě C= ci,

ε je chyba listu (pravidla).

Příklad 8.2.

Tentýž výsledek z dat POČASÍ zapsaný formou produkčních pravidel:

implicitní třída = *SO2MOC*
if počasí = *jasno* ∧ vlhkost > 75 ⇒ **třída = *SO2OK***
if počasí = *déšť* ∧ vítr = *ano* ⇒ **třída = *SO2OK***

♦

❑ Algoritmy pro konstrukci rozhodovacích stromů

Algoritmů pro konstrukci RS existuje řada, většina z nich je variantami níže uvedeného základního algoritmu. Rodina těchto algoritmů bývá označována jako TDIDT (Top-Down Induction of Decision Trees), nejznámější z nich jsou ID3, C4.5 a C5 [3], [4].

Základní algoritmus je rekurzivní, konstruuje RS shora dolů.

1. Pokud všechny body trénovací (pod)množiny patří do stejné třídy, pokračuje se bodem 4.
2. Vybere se nejvhodnější atribut pro rozdělení trénovací (pod)množiny.
3. Objekty trénovací (pod)množiny se rozdělí na podmnožiny podle hodnot vybraného atributu.
4. Pro všechny vytvořené podmnožiny se opakuje postup od bodu 1.
5. Vytvoří se list a přiřadí se mu třída, do níž patří objekty podmnožiny.

Ačkoliv je algoritmus přirozeně rekurzivní, není obvykle možno použít rekurzivní implementaci. Pro rozsáhlejší data (s větším počtem atributů a jejich hodnot) se totiž velmi brzy vyčerpá paměť pro mezivýsledky rekurze. Proto je nutné algoritmy implementovat bez rekurze, což je výrazně náročnější.

Místo reprezentace výsledku pomocí rozhodovacího stromu je možno použít ekvivalentního zápisu pomocí množiny **produkčních pravidel**. Levé strany pravidel jsou tvořeny konjunkcemi hodnot atributů v cestě od kořene stromu k listu, který označuje jednu třídu z pravé strany pravidla. Majoritní třída (ta, která koresponduje s největším počtem listů) je prohlášena za **implicitní třídu** a tedy pravidla pro ni se neuvádí.

Rozhodovací stromy nad velkými daty

Základní algoritmus je možno modifikovat například pro nekonzistentní nebo neúplná data nebo k potlačení upřednostňování atributů, jejichž test vede k triviálním výstupům.

Cílem konstrukce rozhodovacího stromu je získat optimální strom takový, že

1. svou strukturou nejlépe reprezentuje charakter zdrojových dat reprezentovaný objekty trénovací množiny,
2. obsahuje co nejmenší počet uzlů.

Důvody jsou zřejmé. Bez přesnosti by konstrukce stromu nedávala smysl a menší strom je přehlednější, srozumitelnější, lépe se interpretuje.

Základní algoritmus obě zásady dodržuje volbou optimálního atributu pro dělení v každém uzlu stromu. Přesto u rozsáhlých dat může být výsledný strom příliš rozsáhlý. Nejen se tak prodlužuje výpočet, ale výsledek je nepřehledný, špatně se interpretuje.

U rozsáhlých dat se proto používají techniky k omezení růstu stromu. Existují dvě základní techniky pro ovlivnění přesnosti a růstu stromu:

1. **omezení růstu stromu** během konstrukce; do algoritmu se vloží dodatečné podmínky, které rozhodují o dalším dělení uzlu
   a) definováním **minimální velikosti uzlu** pro dělení (malé uzly se již nedělí, i když nejsou homogenní vůči klasifikační třídě),
   b) definováním **minimálního zisku z dělení**,
2. pomocí **vícefázové konstrukce** výsledného stromu minimalizovat strom na optimální velikost pro interpretaci. V první fázi se zkonstruuje celý strom, ve druhé, případně dalších fázích se výsledný strom redukuje na optimální (z hlediska velikosti a přesnosti) velikost
   a) prořezáváním stromu, a to buď nahrazením podstromu listem, nebo zvednutím podstromu,
   b) optimalizací stromu převedeného na pravidla,
   c) konstrukcí a testováním; v případě velkého počtu zdrojových dat rozdělením dat na dvě části; první se použije ke konstrukci stromu, druhá pak k testování, zda strom rozhoduje správně; v případě chybného zařazení se strom opraví (například nahrazením listu podstromem).

Využití rozhodovacích stromů

- Klasifikační strom slouží k odhalení závislostí klasifikačního atributu na vstupních atributech.
- Výsledná pravidla umožní automaticky zařazovat nové objekty do klasifikačních tříd.
- Výsledný strom či pravidla umožní rozlišit důležitost předpovídajících atributů, případně vybrat jen atributy významné pro zařazování do klasifikačních tříd; tak je možno snížit rozměr dat při zachování stejné informace.
- Výsledný strom umožní rozpoznat existenci shluků v datech (charakterizovaných výstupní třídou). Stromová struktura umožní shluky lépe interpretovat.

❑ Výhody a nevýhody konstrukce rozhodovacích stromů

Závěrem si přehledně zopakujeme výhody a nevýhody metody konstrukce rozhodovacích stromů.

Výhody

- Metoda je široce použitelná na mnoho typů dat.
- Základní algoritmus je jednoduchý, což umožňuje využití v mnoha oborech.
- Výsledky u menších stromů mají jednu z nejnázornějších reprezentací.
- Výsledek umožňuje klasifikovat i objekty v době konstrukce stromu neznámé.
- Rozhodovací strom lze snadno převést na pravidla.

Nevýhody

- Ne vždy jsou stromy schopny modelovat složitější reálné situace.
- U větších dat může velikost stromu komplikovat jejich interpretaci.
- Kvalita výsledných znalostí je silně závislá na tom, jak dobře a úplně pokrývá trénovací množina celou problematiku úlohy. Pokud trénovací množinu vytváří expert, dá se očekávat systematické pokrytí celé problematiky úlohy. Pokud máme k analýze observační data, nedá se obecně předpokládat nic, pokud data a jejich původ dobře neznáme. Obecně pak platí „čím více, tím lépe“, při větší trénovací množině je vyšší pravděpodobnost budoucí správné klasifikace nových objektů.
- Výsledek klasifikace je citlivý na neúplná data, kvalita se tím snižuje.
- Jedním z důvodů chybného výsledku může být také provedená kategorizace původních reálných dat – rozdělení do třídních intervalů napříč některými hodnotami klasifikační třídy.

Shrnutí pojmů 8.

Objekty a atributy. Atributy předpovídající a předpovídané. Klasifikační třída.

Trénovací množina.

Rozhodovací strom. Klasifikační třída listu. Podpůrná množina uzlu.

List prostý. Chyba listu.

Entropie, vážený součet entropií. Informační zisk.

Algoritmus konstrukce rozhodovacího stromu.

Výsledek výpočtu formou rozhodovacího stromu a formou rozhodovacích pravidel.

Implicitní třída.

Optimalizace rozhodovacího stromu pro rozsáhlá data.

Omezení růstu stromu.

Vícefázová konstrukce stromu. Prořezávání.

Konstrukce a testování stromu.

Otázky 8.

1. Co je rozhodovací strom a jakou úlohu nad daty řeší?
2. Jak se dělí atributy pro využití rozhodovacího stromu?
3. Co je entropie množiny objektů?
4. Co je vážený součet entropií množin představujících výstupy rozkladu nad atributem A?
5. Co je informační zisk?
6. Uveďte princip algoritmu pro konstrukci rozhodovacího stromu.
7. Jak se vybírají optimální atributy v průběhu konstrukce rozhodovacího stromu?
8. Jak se zobrazují výsledky výpočtu konstrukce rozhodovacího stromu?
9. Jak se výsledky zapisují formou produkčních pravidel?
10. Co je a jak se určí implicitní třída?
11. Kterými typy metod se řeší problémy s rozsáhlými daty a rozsáhlým výsledným stromem?
12. Jakými způsoby se omezuje růst stromu?
13. Co je vícefázová konstrukce stromu?
14. Jak je možno využít velkou trénovací množinu k urychlení konstrukce stromu?

Úlohy k řešení 8.

1. Jsou dána data PACIENTI se strukturou Jsou dána data PACIENTI s atributy
   INFARKT (ano/ne),
   ANG_PECTORIS (ano / ne),
   BERC_VRED (ano / ne),
   VAHA (kg),
   VYSKA (cm),
   KURAK (ano / ne),
   POHLAVI (muž / žena),
   VEK (roků),
   MESTO (ano / ne),
   DUCHODCE (ano / ne),
   STRES (ano / ne).
   Data jsou pořízena za 10 let praxe praktického lékaře a obsahují 2300 objektů.
   Navrhněte shlukovací metodu a její parametry pro nalezení přirozeného rozložení pacientů do shluků.