Machine learning a data mining 1

Jan Outrata

KATEDRA INFORMATIKY
UNIVERZITA PALACKÉHO V OLOMOUCI

přednášky

Shlukování

Shlukování

- $\sim$ seskupení/rozdělení dat do smysluplných, užitečných skupin (**shluků**) – zachycujících strukturu dat, usnadňujících jejich další zpracování (např. souhrny)
- použití a nesčetné aplikace v mnoha oborech a oblastech (DM/ML, informatiky):
- pro **porozumění datům** – lidé přirozeně seskupují/rozdělují (shlukování) a řadí (klasifikace) objekty do skupin/tříd na základě společných charakteristik $\rightarrow$ shluky = možné skupiny, shlukování = hledání skupin
  - biologie (taxonomie a jejich automatické vytváření, analýza genů), psychologie, medicína, sociální vědy (kategorizace a rozšíření nemocí, jevů), obchod (segmentace zákazníků), information retrieval (hierarchické seskupování výsledků vyhledávání) aj.
- pro **další zpracování dat** – abstrakce jednotlivých objektů celky (shluky), charakterizace reprezentativními objekty (**prototypy**) $\rightarrow$ shlukování = jejich hledání
  - např. sumarizace, vzorkování (místo celých velkých dat, srovnatelné výsledky), zvýšení výkonu, úspora dat (místo všech jednotlivých objektů, např. výpočet vzdáleností, přijatelná ztráta informace)

Shlukování (shluková analýza)

= nalezení/vytvoření skupin objektů záznamových dat, na základě popisu objektů (atributů), navzájem si co nejvíce **podobných (je mezi nimi vztah)** v rámci skupiny a různých (bez vztahu) od objektů mimo skupinu (v jiných skupinách)

- skupina = **shluk (klastr, cluster)**: nepřesně/nejednoznačně vymezený pojem – co tvoří shluk (a co už ne)??, např., viz dále typy shluků
- forma klasifikace: rozřazení objektů do kategorií/tříd (class labels) $\sim$ shluk
  - ale u *klasifikace* (nové) objekty rozřazeny na základě modelu (předem) vytvořeného z objektů se známými kategoriemi/třídami = **supervised klasifikace** (z příkladů)
  - u shlukování vytvořené pouze z dat $\sim$ **unsupervised klasifikace**

$\sim$! **segmentace/rozklad (partitioning)** – i mimo tradiční pojetí shlukování, např. u grafů, jednoduché dělení dat na části, typicky dle hodnot atributů objektů, např. u obrazu podle barev pixelů, u obchodních dat apod.

Shlukování (shluková analýza)

(a) Original points.

(b) Two clusters.

(c) Four clusters.

(d) Six clusters.

Typy shlukování

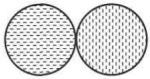
- způsoby seskupování/rozdělování objektů do shluků, **shlukování** = kolekce shluků
- **rozkladové** (partitional) = rozklad množiny objektů na nepřekrývající se podmnožiny (shluky), každý objekt v právě jedné, např.
- **hiearchické** (vnořené, nested) = množina vnořených stromově uspořádaných podmnožin objektů (shluků), každá kromě listů stromu sjednocením potomků (podshluků), kořen zahrnuje všechny objekty, listy často jeden, např., $\sim$ posloupnost rozkladových jako úrovní stromu
- **výlučné** = každý objekt v jednom shluku, např.
- **překrývající se** = objekt může být současně ve více shlucích, také „mezi“ shluky a v kterémkoliv z nich (než svévolně v jednom, lepší ale fuzzy), např.
- **fuzzy** = každý objekt v každém shluku ve stupni příslušnosti od 0 (vůbec není) do 1 (plně je), tj. shluky = fuzzy množiny, často navíc součet stupňů pro objekt roven 1 ($\sim$ **pravděpodobnostní**) a převod na výlučné (jen nejvyšší stupně)
- **úplné** = každý objekt ve shluku, např.
- **částečné** = objekt nemusí být ve shluku („nepatří“, šum, anomálie, „pozadí“), např.

Typy shluků

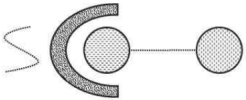
- smysluplnost, užitečnost shluků daná cíly analýzy $\Rightarrow$ různá pojetí/vymezení shluku
- **dobře oddělený (well-separated)** = všechny objekty ve shluku si navzájem podobnější (bližší) než jakémukoliv objektu mimo shluk – obvykle práh dostatečné podobnosti (blízkosti), např., nemusí být kulovitý, „ideál“ splňující jen přirozené oddělené shluky v datech
- **prototypový (prototype-based)** = každý objekt ve shluku podobnější (bližší) prototypovému/reprezentativnímu objektu definujícímu shluk než prototypu jiného shluku – pro numerické atributy typicky centroid (průměr objektů shluku), pro kategorické medoid (medián) $\sim$ **centrový** objekt a shluk, např., tendence kulovitého
- **grafový** = všechny objekty ve shluku „propojené“ (ne nutně navzájem) a nepropojené s objekty mimo shluk = komponenta grafu (uzly objekty, hrany propoje), propojení = do dané vzdálenosti = **styčný (contiguity)**, např., pro nepravidelné nebo provázané shluky, ale problém propojující šum
- **hustotový** = „hustá“ oblast objektů obklopená málo hustou oblastí (šum), hustota = např. počet objektů do dané vzdálenosti od objektu (centrová), např., pro nepravidelné nebo provázané shluky při šumu nebo anomáliích
- **konceptuální** = všechny objekty ve shluku sdílí nějakou vlastnost, např. společně tvoří nějaký specifický tvar shluku, např., $\rightsquigarrow$ pattern recognition

Typy shluků

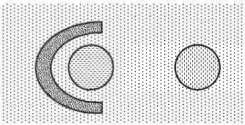
(a) Well-separated clusters. Each point is closer to all of the points in its cluster than to any point in another cluster.



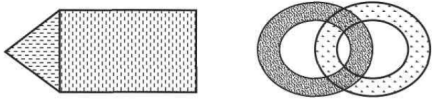
(b) Center-based clusters. Each point is closer to the center of its cluster than to the center of any other cluster.



(c) Contiguity-based clusters. Each point is closer to at least one point in its cluster than to any point in another cluster.



(d) Density-based clusters. Clusters are regions of high density separated by regions of low density.



(e) Conceptual clusters. Points in a cluster share some general property that derives from the entire set of points. (Points in the intersection of the circles belong to both.)

K-means

$=$ rozkladové shlukování s prototypovými shluky zadaného počtu K

- prototyp **centroid** $=$ (typicky) průměr objektů shluku s numerickými atributy, pro jiné (kategorické) atributy medián objektů shluku $=$ prototyp **medoid** (**K-medoids**) – pro objekty stačí jen míra blízkosti

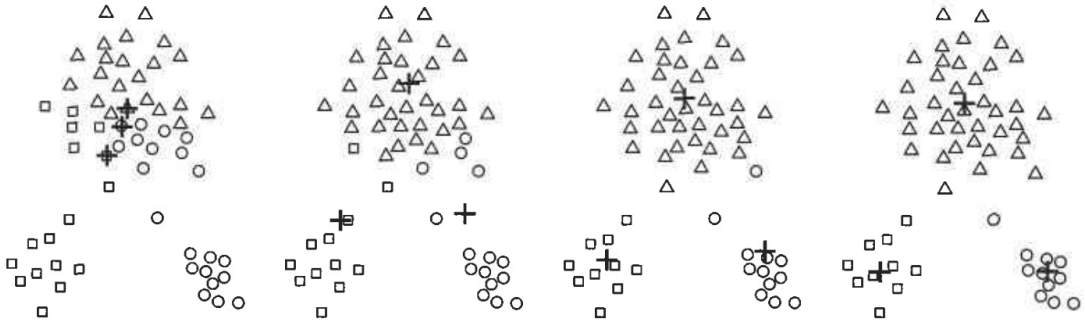
Základní algoritmus

- **zvolení** K počátečních prototypů – objektů z dat nebo i jiných

1. opakovaně přiřazení každého objektu k **nejbližšímu** prototypu $=$ vytvoření K shluků a
2. **aktualizace** prototypů dle objektů shluků, dokud se nezmění (prototypy a tedy i shluky)

- např. (centroid průměr)
- končí (konverguje, „neosciluje“) pro některé míry blízkostí a typy prototypů, např. euklidovská vzdálenost nebo kosinová podobnost a centroid jako průměr objektů
- ze začátku velké změny, ke konci minimální $\rightarrow$ konec při dosažení prahu minimálního počtu objektů přesunutých mezi shluky

K-means



(a) Iteration 1. (b) Iteration 2. (c) Iteration 3. (d) Iteration 4.

Přiřazení objektů k prototypu

- nejbližšímu na základě **míry blízkosti** ((ne)podobnosti) – např. (často) euklidovská (L_2 norm) vzdálenost, kosinová podobnost, také city block (Manhattan, taxicab, L_1 norm) vzdálenost, Jaccard koeficient
- opakované počítání blízkosti každého objektu každému prototypu $\rightarrow$ ušetření, např. půlící (bisecting) K-means

$=$ optimalizace hodnoty **objektivní funkce** (pro dané prototypy) – měří kvalitu shlukování (reprezentativnost prototypů pro objekty shluků), závisí na míře blízkosti, určuje typ prototypu $=$ aktualizaci $\Rightarrow$ optimalizační problém

Aktualizace prototypů

- dle objektů shluků pro optimalizaci hodnoty objektivní funkce (HOF) = **gradientní metoda** – nejlepší prototyp shluku = řešení nulové parciální derivace (pro daný prototyp/shluk) funkce, např.:
  - minimalizace **scatter** = **sumy squared error** shluků – error = míra blízkosti = vzdálenost d objektu x od (nejbližšího) prototypu p_i shluku C_i:

$$\sum_{i=1}^{K} \sum_{x \in C_i} d(x, p_i)^2$$

    $\Rightarrow$ pro d euklidovskou (L_2) vzdálenost nejlepší prototyp = centroid jako průměr objektů shluku
    pro **sumu absolute error** $\sum_{i=1}^{K} \sum_{x \in C_i} d(x, p_i)$ a city block (L_1) vzdálenost nejlepší prototyp = medoid jako medián objektů shluku (K-medoids)
  - maximalizace **koheze** shluků = sumy blízkostí = podobností s objektu od prototypu:

$$\sum_{i=1}^{K} \sum_{x \in C_i} s(x, p_i)$$

    $\Rightarrow$ pro s kosinovou podobnost nejlepší prototyp = centroid jako průměr objektů shluku

Aktualizace prototypů

! (pouze) lokální optimalizace – pro konkrétní prototypy $\Rightarrow$ klíčové počáteční!

Inkrementální aktualizace prototypů

= po přiřazení každého jednoho objektu k prototypu místo všech – aktualizace dvou nebo žádného prototypu

- možná rychlejší konvergence, ale i závislost shluků na pořadí objektů $\rightarrow$ náhodné

Zvolení počátečních prototypů

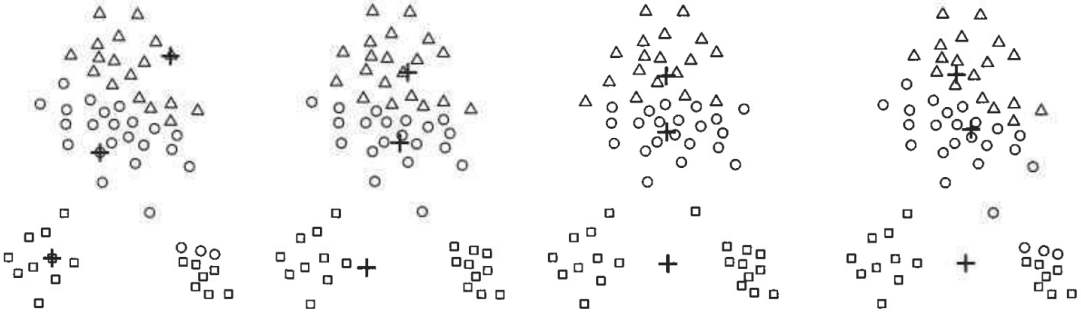
- běžně náhodné objekty, ale možné horší shluky (lokální optima) – i při „rovnoměrněji“ rozložených (v různých shlucích), např.

→ vícekrát náhodné a výběr nejlepších shluků (nejlepší optimum) – nemusí fungovat (ideálně počáteční prototyp v každém shluhu. . .)

→ prototypy shluků hierarchického shlukování z (náhodného) vzorku objektů – relativně malého (hierarchické shlukování je náročnější), ale většího než K

→ první náhodný objekt nebo prototyp všech objektů a každý další nejvzdálenější objekt od všech aktuálních – dobře oddělené, ale mohou být anomálie a výpočet → na (náhodném) vzorku objektů

Zvolení počátečních prototypů



(a) Iteration 1. (b) Iteration 2. (c) Iteration 3. (d) Iteration 4.

Postprocessing shluků

- pro další zlepšení HOF (často lokální optimalizace) – bez zvýšení K

→ střídavé rozdělování a spojování shluků – možný „únik“ z lokálního optima při zachování počtu shluků:

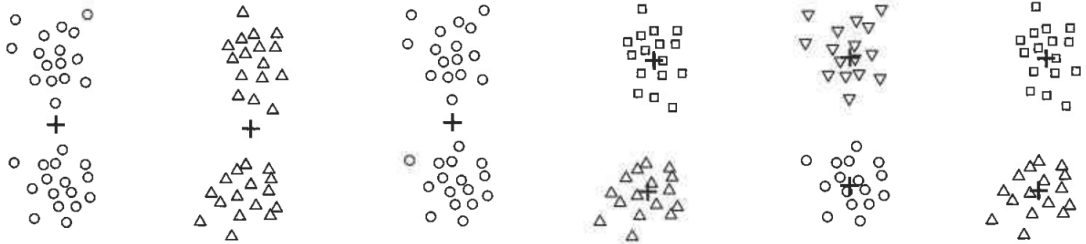
  - rozdělení shluku: s nejhorší HOF, nejvyšší směrodatnou odchylkou nějakého atributu aj., nový prototyp – nejvzdálenější objekt od prototypů, objekt nejvíc zhoršující HOF, náhodný aj.
  - spojení shluků: s nejbližšími prototypy, s nejmenším zhoršením HOF ($\sim$ prototypová a Wardova metoda v hierarchickém shlukování) aj., zrušení prototypu a přiřazení objektů k jiným – shluku s nejmenší HOF

Půlící (bisecting) K-means

= počínaje shlukem všech objektů, opakované **půlení zvoleného shluku** na dva pomocí základního K-means, dokud není K shluků

- např. (centroid průměr)
- vícekrát půlení a výběr toho s nejlepší HOF
- volba shluku: s nejvíce objekty, nejhorší HOF aj.
- postprocessing základním K-means s prototypy (z půlícího) jako počátečními – shluky (z půlícího) nemusí být lokální optimum objektivní funkce (základní K-means používáno „lokálně“ – půlení konkrétních shluků)
- méně závislé na počátečních prototypech (výběr z více půlení a jen dva prototypy v každém)
- posloupnost shlukování z iterací půlení = hierarchické shlukování

Půlící (bisecting) K-means



(a) Iteration 1. (b) Iteration 2. (c) Iteration 3.

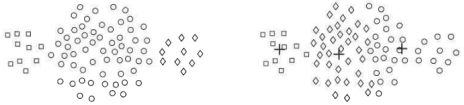
K-means

- jednoduché, rychlé (pokud K malé) – časová složitost lineární ve velikosti vstupních dat (počet iterací je omezený), obvykle ale potřeba více spuštění

Problémy

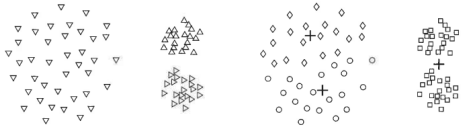
- prázdný shluk = jen prototyp $\rightarrow$ jiný prototyp: nejvzdálenější objekt od ostatních prototypů (nejvíce zhoršuje HOF) nebo objekt ze shluku s nejhorší HOF (= rozdělení shluku)
- **anomálie** („nepatří“ do žádného shluku) – prototypy mohou být horší reprezentanti shluku $\rightarrow$ nalezení a odstranění před nebo po shlukování (pokud nejsou zajímaví nebo potřební jako všechny objekty) – např. objekty nejvíc zhoršující HOF, malé shluky (anomálií), prototyp = medián objektů shluku (K-medoids)
- ne-dobře oddělené nebo ne-kulovité „přirozené“ shluky v datech nebo s (výrazně) rozdílnými velikostmi nebo hustotami objektů $\rightsquigarrow$ „smíchání“ částí shluků nebo spojení menších s částmi větších nebo dohromady, např. $\rightarrow$ vyšší K (části jako (pod)shluky), např.

K-means



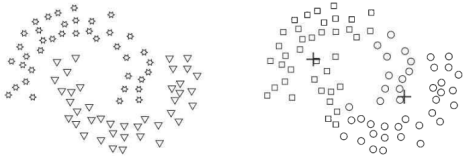
(a) Original points. (b) Three K-means clusters.

K-means with clusters of different size.



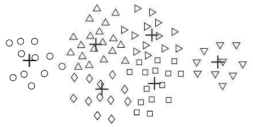
(a) Original points. (b) Three K-means clusters.

K-means with clusters of different density.

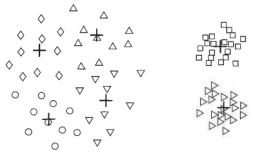


(a) Original points. (b) Two K-means clusters.

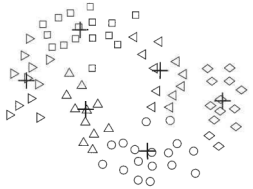
K-means with non-globular clusters.



(a) Unequal sizes.



(b) Unequal densities.



(c) Non-spherical shapes.

Hierarchické shlukování

1. **aglomerativní** = opakované **spojování dvou shluků** až do jednoho počínaje jednotlivými objekty jako jednoprvkovými shluky (singleton)
2. **divizivní** = opakované **rozdělování nějakého shluku** až na singletony počínaje jedním shlukem se všemi objekty

- grafické zobrazení pomocí **dendrogramu** = stromový diagram zobrazující (inkluzivní) vztahy mezi shluky (hrany) i pořadí jejich spojování/rozdělování (výška), např.
- pro aplikace vyžadující hierarchii shluků, např. taxonomie tříd objektů

Aglomerativní – základní algoritmus

1. každý objekt jako jednoprvkový shluk (singleton) a **blízkost shluků** = blízkost objektů
2. opakovaně spojení (sjednocení) dvou nejbližších shluků, dokud nezůstane jeden

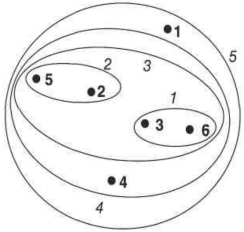
- $2n - 1$ shluků, n počet objektů

Blízkost shluků

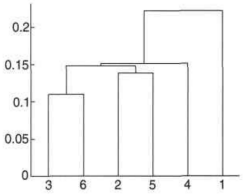
- grafové shluky:
  - **single link/min** = blízkost nejbližších dvou objektů z různých shluků (délka nejkratší hrany mezi dvěma uzly z různých komponent), např. $\rightsquigarrow$ styčné shluky, problém šum a anomálie; $\alpha_A = \alpha_B = 1/2, \beta = 0, \gamma = -1/2$
  - **complete link/max** = blízkost nejvzdálenějších dvou objektů z různých shluků, např., $\rightsquigarrow$ tendence kulovitých shluků; $\alpha_A = \alpha_B = \gamma = 1/2, \beta = 0$
  - **průměrná (group average)** = průměr blízkostí každých dvou objektů z různých shluků, např.; $\alpha_A = n_A/(n_A + n_B), \alpha_B = n_B/(n_A + n_B), \beta = \gamma = 0$
- prototypové shluky:
  - **prototypová (centroidní)** = blízkost prototypů shluků – možné **inverze** = bližší shluky spojené později (blízkosti spojovaných shluků nemusí tvořit neklesající posloupnost); $\alpha_A = n_A/(n_A + n_B), \alpha_B = n_B/(n_A + n_B), \beta = -n_A n_B/(n_A + n_B)^2, \gamma = 0$
  - **Wardova metoda** = zhoršení HOF po spojení shluků (= optimalizace objektivní funkce K-means), např. $\sim$ průměrná při druhé mocnině blízkosti objektů, použití pro zvolení počátečních prototypů K-means; $\alpha_A = (n_A + n_Y)/(n_A + n_B + n_Y), \alpha_B = (n_B + n_Y)/(n_A + n_B + n_Y), \beta = -n_Y/(n_A + n_B + n_Y), \gamma = 0$
- Lance-Williams formula: n_A počet objektů shluku A

$$p(A \cup B, Y) = \alpha_A p(A, Y) + \alpha_B p(B, Y) + \beta p(A, B) + \gamma |p(A, Y) - p(B, Y)|$$

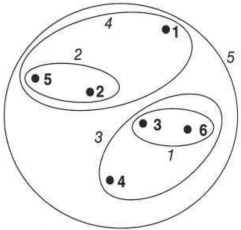
Blízkost shluků



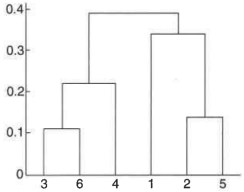
(a) Single link clustering.



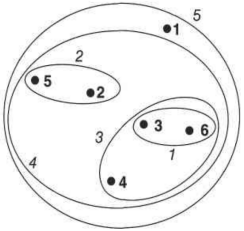
(b) Single link dendrogram.



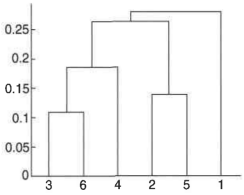
(a) Complete link clustering.



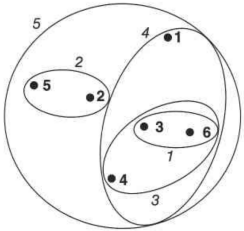
(b) Complete link dendrogram.



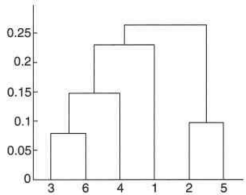
(a) Group average clustering.



(b) Group average dendrogram.



(a) Ward's clustering.



(b) Ward's dendrogram.

Blízkost shluků

- **nevážená**/**vážená** = uvažovaný/neuvažovaný počet objektů shluků $\Rightarrow$ stejné/různé váhy objektů různých shluků (např. různé třídy objektů), např. průměrná nevážená výše, vážená $\alpha_A = \alpha_B = 1/2, \beta = \gamma = 0$

Aglomerativní hierarchické shlukování

- jednoduché, ale časová složitost $O(N^2 \log N)$ ve velikosti N vstupních dat (potřeba blízkosti mezi každými dvěma objekty a zjištění/vyhledání pro shluky)

Problémy

- absence globální objektivní funkce – **lokální optimalizace** spojení shluků, ovšem s využitím blízkostí každých dvou objektů z různých shluků; shluky ale nejsou lokální optima např. (globální) optimalizační funkce K-means, ani při Wardově metodě (objekty shluku ani nemusí být nejblíže prototypu shluku)
- spojení shluků finální (nemění se): problém u vysokorozměrných dat se šumem, ale postprocessing např. přesun větví stromu hierarchie shluků pro optimalizaci nějaké globální objektivní funkce nebo preprocessing např. rozkladové shlukování (K-means) pro malé počáteční shluky místo jednotlivých objektů

Divizivní hierarchické shlukování

- více metod, např. i půlící (bisecting) K-means

Minimum spanning tree (Minimální kostra)

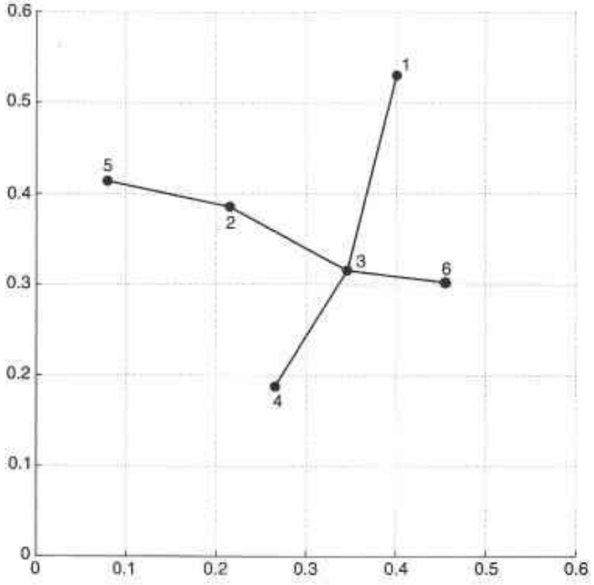
= neprázdný souvislý podgraf ohodnoceného grafu se všemi uzly, bez cyklů (strom) a s minimálním celkovým ohodnocením hran

1. nalezení minimum spanning tree **grafu blízkostí (proximity graph)** = hrany mezi objekty jako uzly ohodnocené blízkostí (nepodobností) objektů, např., $\sim$ shluk se všemi objekty
2. opakovaně zrušení hrany ohodnocené nejmenší blízkostí $\sim$ rozdělení shluku na dva, dokud nezůstanou jen singletony

$\sim$ opakované ponechávání pouze hran grafu blízkostí mezi nejbližšími objekty (rozklad grafu)

$\rightarrow$ stejné shlukování (kolekce shluků) jako single link/min aglomerativní hierarchické shlukování

Divizivní hierarchické shlukování



Hustotové (density-based) shlukování

= vyhledávání „hustších“ oblastí objektů (= shluků) oddělených málo hustými oblastmi ($\sim$ anomálie/šum)

$\rightarrow$ částečné rozkladové shlukování s hustotovými shluky – libovolných tvarů a velikostí, odolné vůči šumu

- více definic **hustoty**

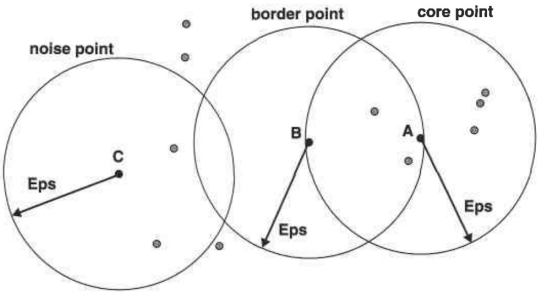
Centrová (center-based) hustota

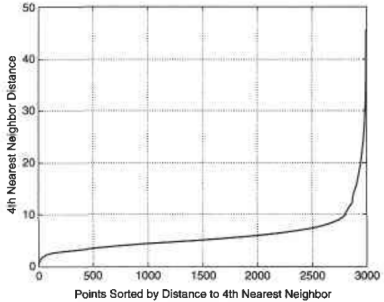
= pro daný objekt počet objektů do dané **vzdálenosti** (včetně daného objektu), např. – v extrémech rovna počtu všech objektů nebo 1

- core objekt (bod) = hustota nad **prahem** $\sim$ „uvnitř“ husté oblasti, např.
- border objekt = ne core, ale do dané vzdálenosti od (asociovaného) core objektu – možno více $\sim$ „na hranici“ husté oblasti, např.
- šumový objekt = ostatní, „mimo“ hustou oblast, např.

? jaká daná vzdálenost d a práh k hustoty: zlomová vzdálenost objektů k jejich k-tému nejbližšímu sousedu (pro objekty ve shluku je malá, pokud k není větší než počet objektů ve shluku, zlom pro ne příliš rozdílně husté shluky), např. – pro různá k se d příliš nemění, pro malá k i málo blízkých anomálií shluky, pro velká k malé shluky šum

Hustotové (density-based) shlukování





DBSCAN

- centrová hustota, pro původní algoritmus $k = 4$

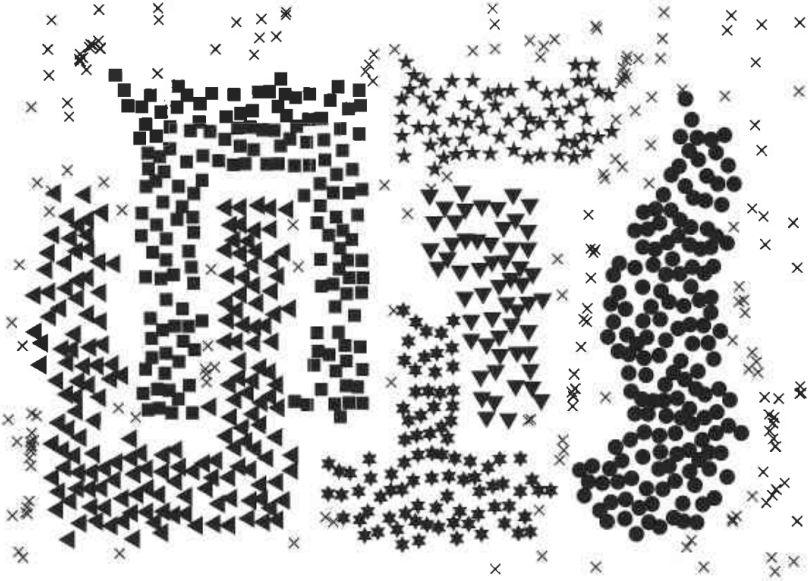
1. shluky = množiny core objektů do dané vzdálenosti d od sebe
2. zařazení border objektů do shluků s asociovanými core objekty – příp. vybraného

- např. $(k = 4, d = 10)$
- jednoduché, časová složitost kvadratická ve velikosti N vstupních dat (s méně atributy při výkonných datových strukturách pro vyhledávání objektů do dané vzdálenosti d od daného objektu, např. kd-stromech, $O(N \log N)$)

Problémy

- mnoho **podobných hustot shluků** – např. hustota méně hustých shluků podobná hustotě šumu kolem hustších shluků $\rightsquigarrow$ (pro dostatečně malou d pro oddělené méně husté shluky) hustší shluky spojeny se šumem kolem nich nebo (pro dostatečně velkou d pro oddělené hustší shluky) méně husté shluky šum
- vysokorozměrná data – náročnější definice hustoty, výpočet vzdáleností objektů

DBSCAN



Vyhodnocení shluků [SAMOSTUDIUM]

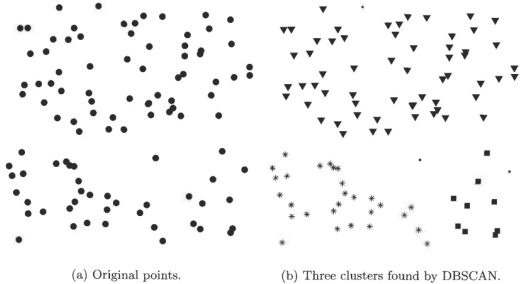
- v (supervised) klasifikaci vyhodnocení vytvořeného modelu dat součást jeho tvorby – ukazatele a vyhodnocení výkonnosti, řešení problému přeučení atd.
- ve shlukování ne – často jako část explorativní analýzy dat ($\Rightarrow$ vyhodnocení nadbytečné), různé typy shluků ($\Rightarrow$ různá vyhodnocení)

$\sim$ **validace shluků** – každá shlukovací metoda *nějaké* shluky v datech najde, i bez přirozených, např. (3 shluky z DBSCAN)

- *existence (tendence) shluků*? = nenáhodná struktura dat + počet shluků? + nenutnost externí informace = **unsupervised**: (interní) míry **koheze** (kompaktnosti, = blízkosti objektů ve shluku) a **separace** (izolace, = dobré oddělenosti) shluků, silhouette koeficient, kofenetický korelační koeficient – využití měr blízkosti objektů, např. objektivní funkce v K-means
- porovnání shluků s externí znalostí/strukturou dat, např. class labels objektů = **supervised**: (externí) míry např. entropie, purity, precision, recall, F – využití podílů class labels ve shlucích, podobnostní koeficienty – využití počtů párů objektů se stejnými/různými shluky a class labels
- porovnání shluků/shlukování mezi sebou = **relativní**: unsupervised i supervised

! problémy měr: použitelnost (např. jen pro dvou/třírozměrná prostorová data), interpretace (jaké hodnoty dobré? $\rightarrow$ statistické rozložení), složitost

Vyhodnocení shluků [SAMOSTUDIUM]



(a) Original points.

(b) Three clusters found by DBSCAN.

(c) Three clusters found by K-means.

(d) Three clusters found by complete link.

Koheze a separace shluků