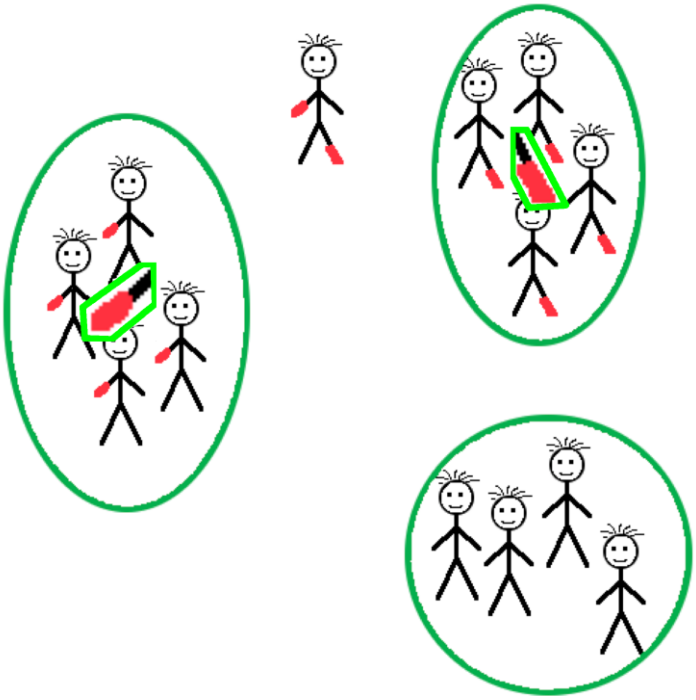
Státnice – odborné č. 20

Shlukování dat

Shlukování dat. Metoda k-středů, hierarchické (aglomerativní) shlukování, Kohonenova mapa SOM

Shlukování dat

Shluková analýza je snaha o seskupení objektů do skupin (shluků) na základě jejich vlastností tak, aby si byly podobné, a zároveň nebyly podobné objektům v jiných skupinách (shlucích). Jedná s o optimalizační problém s cílem optimalizovat počet shluků a přiřazení instancí do shluků.



Shluky instancí s podobnými vlastnostmi

Měřítkem podobnosti vlastností objektů je metrika poskytující číselný výsledek umožňující počítačové zpracování. Metrika musí splňovat základní podmínky použitelnosti:

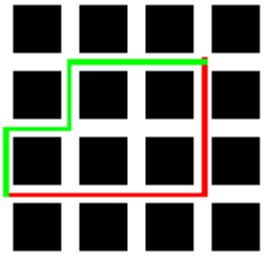
$d(x; y) >= 0$
$d(x; y) = d(y; x)$
$d(x; y) = 0 , \Leftrightarrow x = y$
$d(x; y) + d(y; z) >= d(x; z)$

Eukleidovská metrika pro $\Re^n$

$$\vec{x} = (x_1, x_2, \ldots, x_n), \vec{y} = (y_1, y_2, \ldots, y_n)$$

$$dist(\vec{x}, \vec{y}) = \sqrt{\sum_{i-1}^{n} (x_i - y_i)^2}$$

Manhattanská metrika:



Pokud znám souřadnice, vzdálenost spočítam takto:

$$dist(\vec{x}, \vec{y}) = |x_1 - y_1| + |x_2 - y_2| + \ldots + |x_n - y_n|$$

Kosinová metrika:

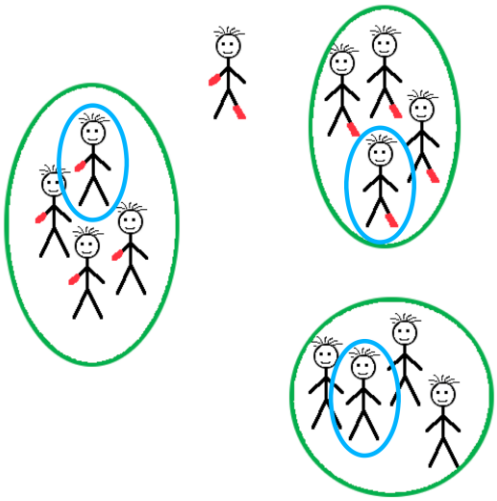
Vzdálenost dvou vektorů je úhel, který svírají.

$$similarity(\vec{x}, \vec{y}) = \frac{\sum_{i-1}^{n} (x_i * y_i)}{\sqrt{\sum_{i-1}^{n} (x_i^2) * \sum_{i-1}^{n} (y_i^2)}}$$

Výsledky této funkce jsou v rozmezí -1 ... +1. -1 znamená úplný opak, 0 nezávyslost a +1 naprostou shodu.
Aby výsledky vyhovovali definici metriky je potřeba podobnost odečíst od jedné.

$$dist(\vec{x}, \vec{y}) = 1 - similarity(\vec{x}, \vec{y})$$

Metoda k-středů (KMeans)



Jednotlivé shluky jsou zastoupeny reprezentantem s typickými vlastnostmi

Každá instance (vzor) v datech bude reprezentována jedním reprezentantem, který ponese vlastnosti typické pro danou skupinu (shluk). Tento reprezentant je instanci (vzoru) nejpodobnější – bude instanci nejblíže v dané metrice. Správná pozice reprezentantů mezi instancemi volena tak, aby součet všech vzdáleností mezi instancemi a jim příslušnými reprezentanty byla minimální (optimalizační problém). Optimalizace je prováděna iteračně.

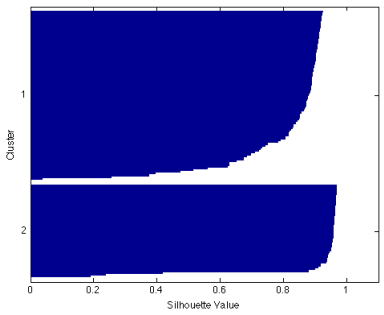
Algoritmus KMeans:

1. nastav reprezentanty do náhodných počátečních bodů,
2. najdi a přiřaď každé instanci jejího nejbližšího reprezentanta (reprezentanta s nejkratší metrikou),

   a pro každého reprezentanta vytvoř množinu naplněnou jeho nejbližšími instancemi,
3. přesuň reprezentanta doprostřed své množiny nejbližších instancí (minimalizuj součet metrik reprezentanta ke všem svým nejbližším instancím),
4. změnila-li se poloha aspoň jednoho reprezentanta, vrať se do bodu 2. Jinak skonči.

Vyhodnocení shluků vytvořených KMeans algoritmem:

jednou z možností je výpočet tzv. siluety



Silueta pro každou vstupní instanci spočítá jistotu zařazení instance do daného shluku.

$$s(x_k) = \frac{b(x_k) - a(x_k)}{\max(a(x_k), b(x_k))}$$

$a(x_k)$ je průměrná vzdálenost x_k od ostatních instancí shluku, ke kterému je přiřazena.

$b(x_k)$ je průměrná vzdálenost x_k od instancí v nejbližším dalším shluku.

Výsledné hodnoty jsou mezi -1 (x_k do shluku úplně nepatří) a +1 (úplně patří)

Pokud vypočítáme siluetu pro všechny instance a vykreslíme ji do grafu, lze si udělat představu, jak shlukování dopadlo (neměly by se vyskytovat žádné záporné hodnoty – instance mající blíže k instancím jiného než vlastního shluku). Lze rovněž shlukování hodnotit výpočtem průměrné siluety přes všechny instance (ideálně přes testovací data). Čím vyšší číslo, tím lépe shluky vytvořeny (počet a pozice reprezentantů).

Stabilitu výsledku shlukování lze testovat (náhodným) smazáním části (10%) testovacích dat a takto vygenerováním několika podmnožin testovacích dat, na nichž se shlukování otestuje opakovaně.

Hierarchické (aglomerativní) shlukování

Základní myšlenkou je vytváření hierarchie shluků, vždy spojením dvou nejpodobnějších shluků (s nejmenší metrikou) do jednoho většího. Takto se pokračuje, dokud není vytvořen jeden mega-shluk.

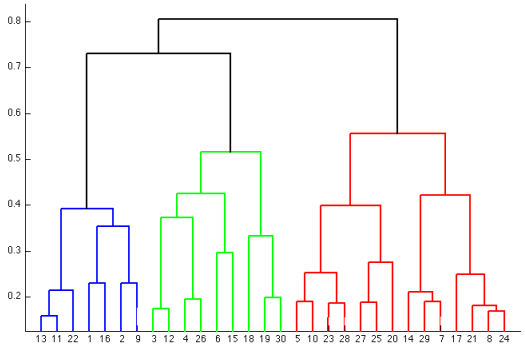
Základní algoritmus:

1. začít ze stavu, kdy každá jednotlivá instance je jedním shlukem,
2. najít dva nejbližší shluky,
3. spojit nalezené dva nejbližší shluky do jednoho,
4. zůstávají-li nějaké spojitelné shluky, vrátit se do bodu 2. Jinak skončit.

Vzdálenost shluků může být určena:

- „nejbližší sousedé“ – vzdálenost nejbližších instancí ve shlucích,
- „nejvzdálenější sousedé“ – vzdálenost nejvzdálenějších instancí ve shlucích,
- vzdálenost středů - center shluků,
- průměrná vzdálenost mezi všemi instancemi v obou shlucích.

Vizualizace postupu shlukování stromem - dendrogram



Vyhodnocení hierarchického shlukování

- lze použít siluetu,
- vypočítat CPCC (kofenetický koeficient korelace) – normovaná kovariance vzdáleností v původním prostoru a v dendrogramu. Čím vyšší je CPCC, tím nižší je ztráta informace způsobená slučováním instancí do shluků (instance shluk skutečně tvoří). Je-li CPCC menší než cca 0,8, patří všechny instance do jediného velkého shluku.

SOM (Self Organizing Map) – samoorganizující se mapa

Princip (kompetitivní učení): jedinci (reprezentanti, neurony) spolu soutěží, nepotřebují žádného arbitra (učitele), který by jim říkal, kam se mají přesunout. Každý jedinec to umí zjistit sám. Jedinci se učí z příkladů. Populace jedinců se v průběhu času samoorganizuje.

Míra optimalizace samoorganizace je vyjádřena kvantizační chybou – průměrnou vzdáleností mezi instancemi (vzory) a jejich reprezentanty. Minimalizace kvantizační chyby tlačí reprezentanty do míst vysoké hustoty instancí. Snaha aproximovat hustotu instancí pomocí menší hustoty reprezentantů.

$$\text{kvantizační chyba} = \frac{1}{\text{počet instancí}} \sum_{i=0}^{k} \sum_{x \in nearest(r_i)} dist(r_i, x)$$

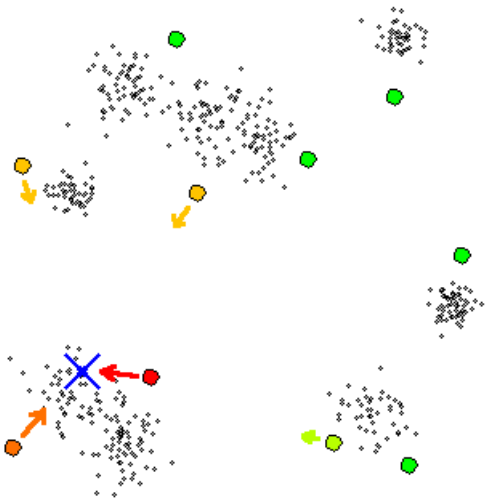
r_i je i-tý reprezentant. A $nearest(r_i)$ je množina instancí, které jsou mu nejblíž.

x je jedna z instancí.

Neuronový plyn je jiným způsobem minimalizace kvantizační chyby – používáno okolí (instance).

Algoritmus neuronového plynu:

1. náhodně rozmístit reprezentanty a zvolit velké okolí,
2. vybrat nějakou vstupní instanci,
3. spočítat vzdálenosti mezi zvolenou instancí a všemi reprezentanty,
4. upravit pozice reprezentantů v závislosti na jejich vzdálenosti od zvolené instance a okolí,
5. zmenšit okolí,
6. případně pokračovat bodem 2.



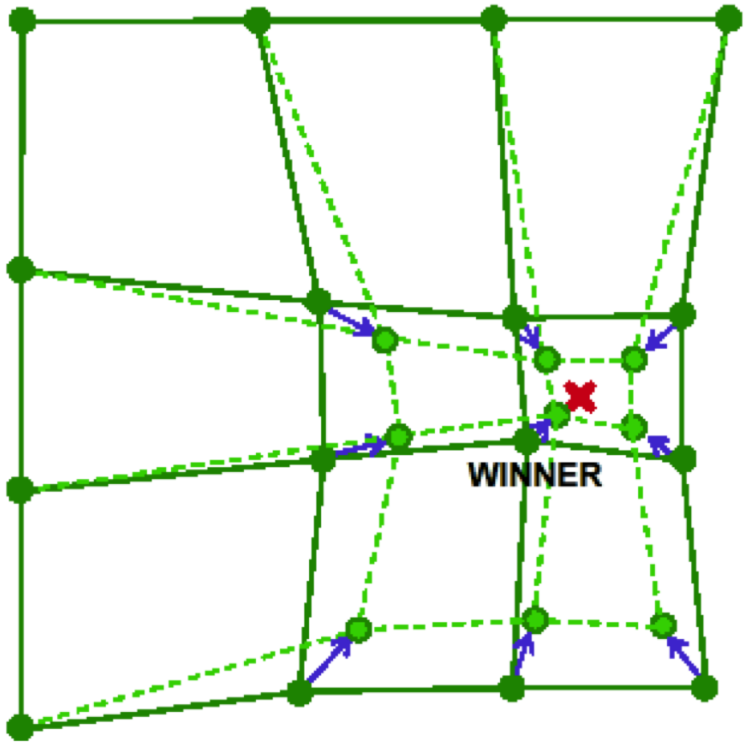
Vybraná instance a přesun reprezentantů v jejím okolí

Vylepšený neuronový plyn – vytvořeny „přátelské“ vztahy mezi sousedními reprezentanty, posouvají se jen „kamarádi“ vítězného reprezentanta. Vizualizací „přátelství“ lze získat pravidelnou mřížku (síť), typicky čtvercovou nebo šestiúhelníkovou.

Terminologie SOM (Kohonenova mapa): reprezentant je neuronem, jeho souřadnice jsou váhami.

Algoritmus SOM:

1. inicializovat váhy všech neuronů (souřadnice všech reprezentantů),
2. vybrat nějakou vstupní instanci,
3. spočítat vzdálenosti mezi vstupní instancí a všemi neurony,
4. určit nejbližší neuron BMU (best matching unit),
5. upravit váhy (pozici) BMU a jeho okolí,
6. případně pokračovat bodem 2.



Změna vah BMU a jeho okolí (vstupní instance značena x)

Vizualizace SOM (při více dimenzionálním problému):

- U-Matice - zobrazuje strukturu vzdáleností v prostoru dat (vzdálenost zachycena barvou),
- Analýza hlavních komponent – hledání nových os směry největšího rozptylu hodnot,
- Sammonova projekce – zachování vztahu mezi daty zobrazenými v novém prostoru získaném minimalizací kvadrátů vzdáleností objektů.