

Shlukování v souborech s odlehlými objekty pomocí metod k -průměrů

Marta Žambochová

Katedra matematiky a statistiky
Fakulta sociálně ekonomická
Univerzita J. E. Purkyně v Ústí nad Labem

Robust 2010
31. ledna – 5. února 2010
Králíky

Motivace

- vliv odlehlých objektů na shlukování
- potřeba možnosti detekce odlehlých objektů

Odlehlé objekty a shlukování

- vhodné předzpracování dat
- metody přímo zaměřené na data obsahující odlehlé objekty
- dvoufázový algoritmus k -průměrů
- modifikovaný filtrovací algoritmus
- varianta využívající algoritmus k -průměrů⁺⁺

Dvoufázový algoritmus k -průměrů

- v první fázi modifikovaný algoritmus k -průměrů rozdělí data do k' shluků, kde $k \leq k' \leq n$
- ve druhé fázi algoritmus za pomoci minimální kostry odhalí odlehlé objekty a vytvoří cílové rozdělení do požadovaných k shluků

První fáze

- ovlivněna algoritmem ISODATA
- na počátku k center
- heuristika: „pokud je vkládaný objekt velmi vzdálen od všech dosavadních center shluků, je zařazen do nově vzniklého shluku“

- minimální vzdálenost mezi dvěma libovolnými centry

$$\min(K) = \min \left\{ \sum_{l=1}^d (c_{il} - c_{jl})^2; i, j = 1, \dots, k', i \neq j \right\}$$

- vzdálenost objektu $\mathbf{x}$ od jemu nejbližšího centra z množiny K

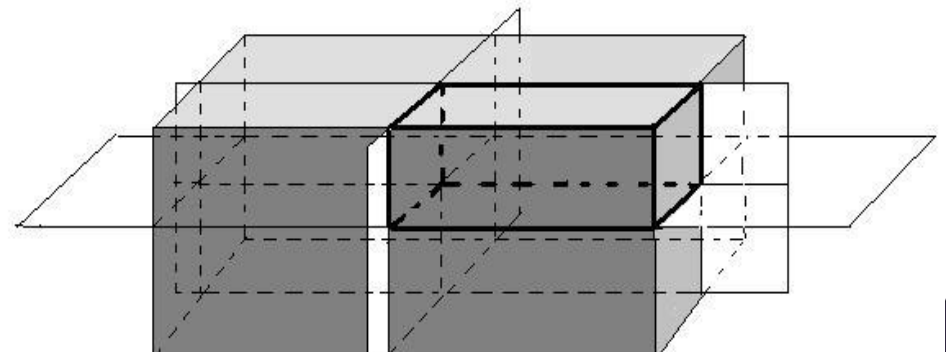
$$\min(\mathbf{x}, K) = \min \left\{ \sum_{l=1}^d (x_l - c_{il})^2; i = 1, \dots, k' \right\}$$

Druhá fáze

- k' center vzniklých v první fázi je považováno za objekty nového shlukování
- stanou se uzly ohodnoceného úplného grafu
- hrana je ohodnocena vzdáleností daných dvou center
- tvorba minimální kostry pro výše definovaný graf
- postupné rušení hrany s největším ohodnocením
- objekty obsažené v malých nově vzniklých komponentách označíme za odlehlé – vyloučíme je
- postup ukončen po vytvoření požadovaných k shluků

Použití mrkd-stromů

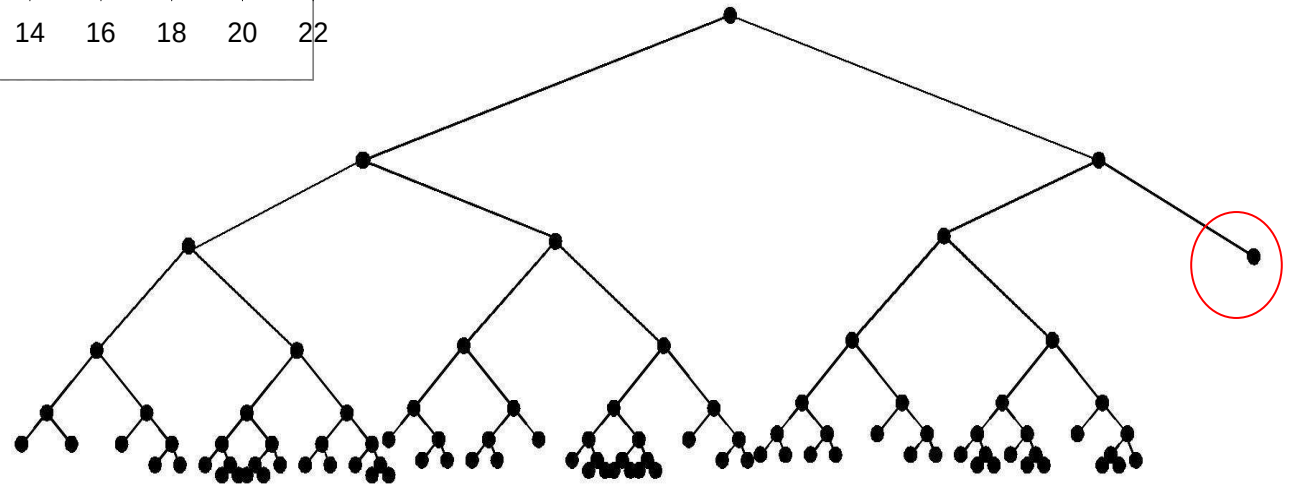
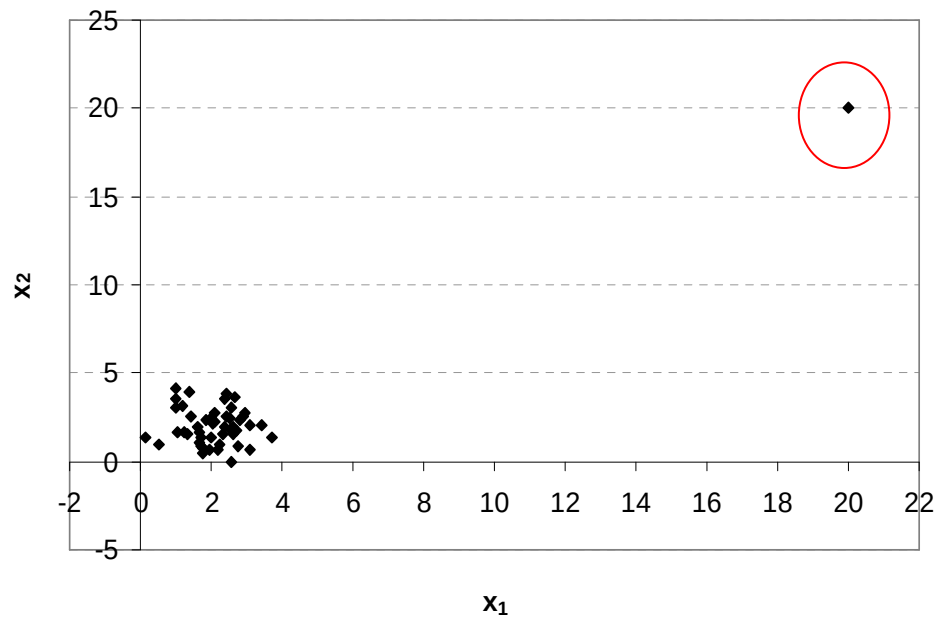
- Pelleg, D., Moore, A. (1999)
- Kanungo, T., Mount, D. M., Netanzahu, N. S., Piatko, CH. D., Silverman, R., Wu, A. Y. (2002)
- kd-strom (kvadrantový strom) je datová stromová struktura, která reprezentuje rekurzivní dělení konečné množiny bodů z d -dimenzionálního prostoru na k částí, pomocí $d-1$ dimenzionálních ortogonálních nadrovin (např. pomocí rozdělení ortogonálně k nejdelší straně hyperkvádrů na úrovni mediánu ze všech bodů hyperkvádrů)
- mrkd-stromy jsou binární formou kd-stromů
- každý z vrcholů obsahuje informaci o všech bodech z příslušného hyperkvádrů



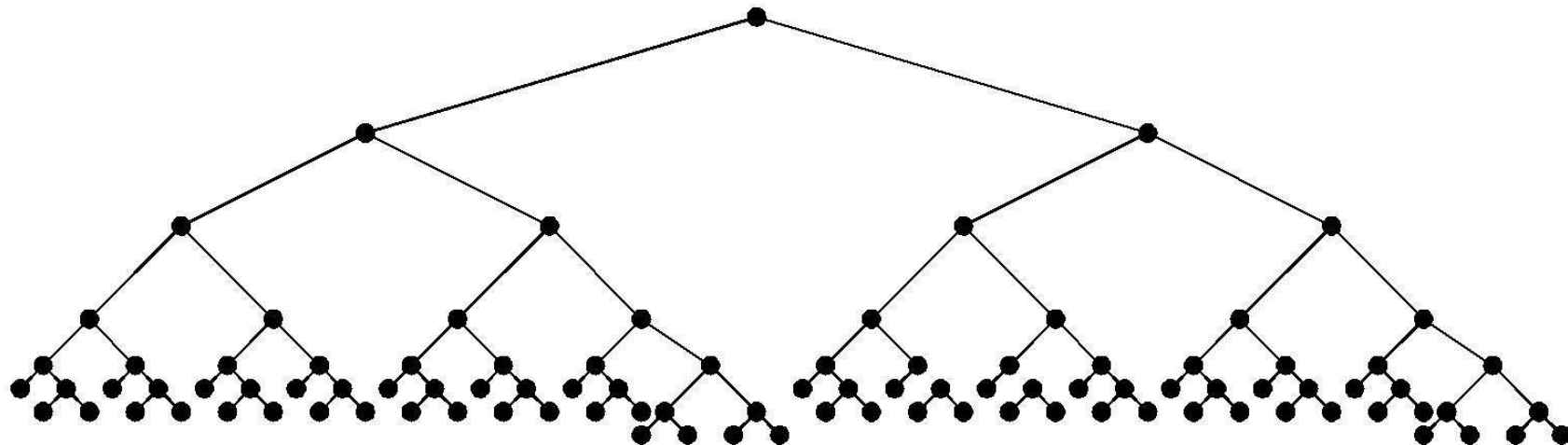
Modifikovaný filtrovací algoritmus

- dělení hyperkvádrů na úrovni průměru místo na úrovni mediánu
- stromy nejsou tak dokonale vyvážené
- nevyváženost je zapříčiněna známou vlastností aritmetického průměru, který je silně ovlivňován odlehlými hodnotami
- v části oddělené průměrem, která obsahuje odlehlé hodnoty, je umístěn mnohem menší počet objektů než v části druhé
- relativně dobře se daří odhalit odlehlé objekty

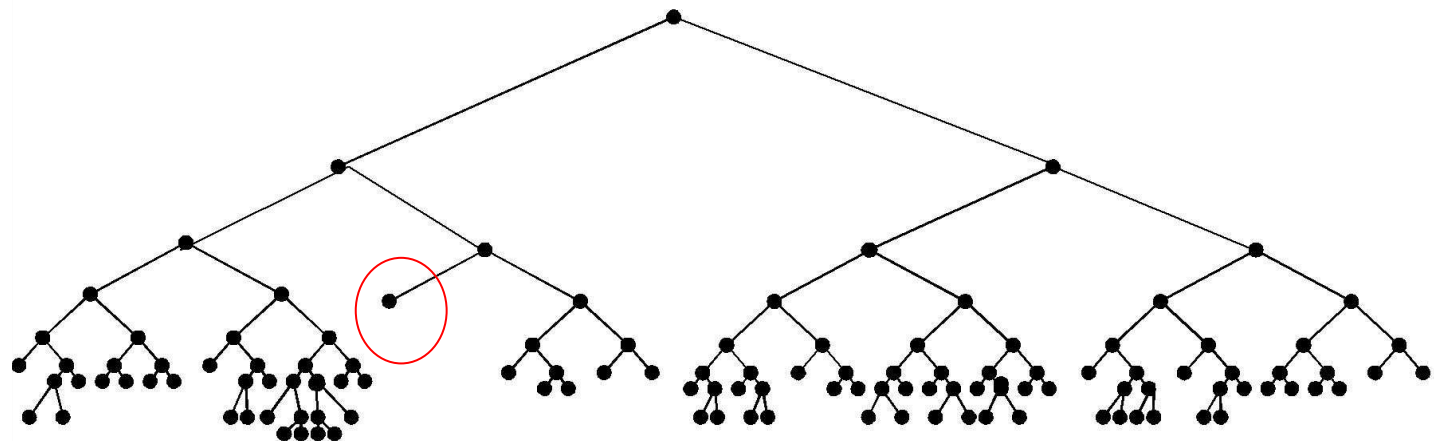
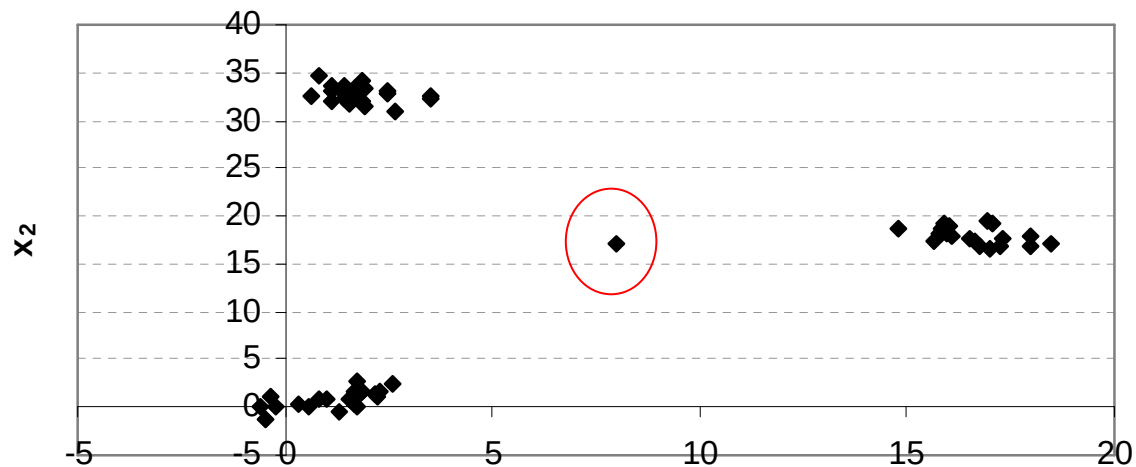
Mrkd-strom pro data s odlehlým objektem I.



Klasický mrkd-strom



Mrkd-strom pro data s odlehlým objektem II.



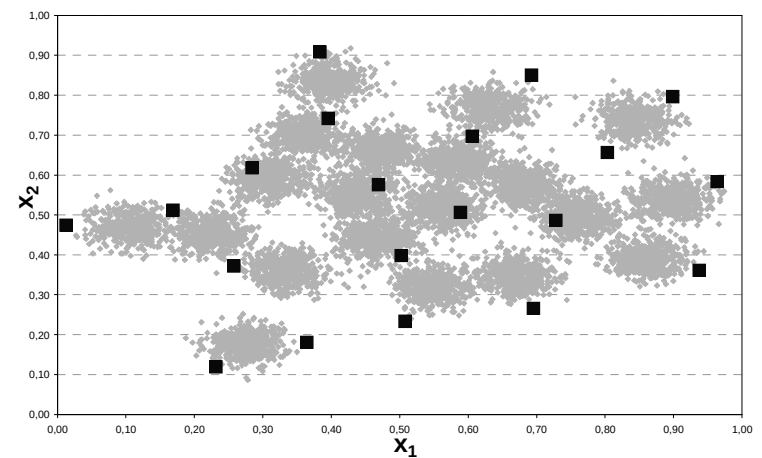
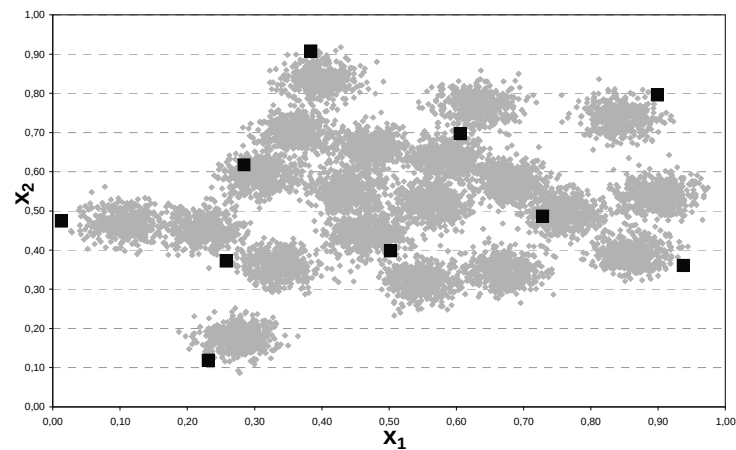
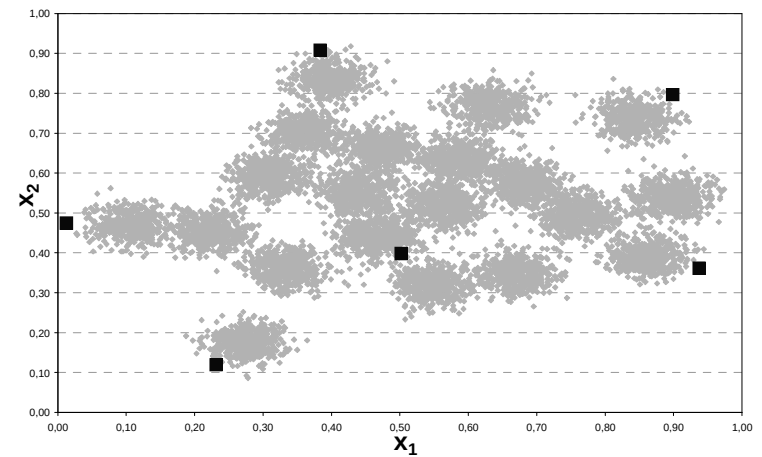
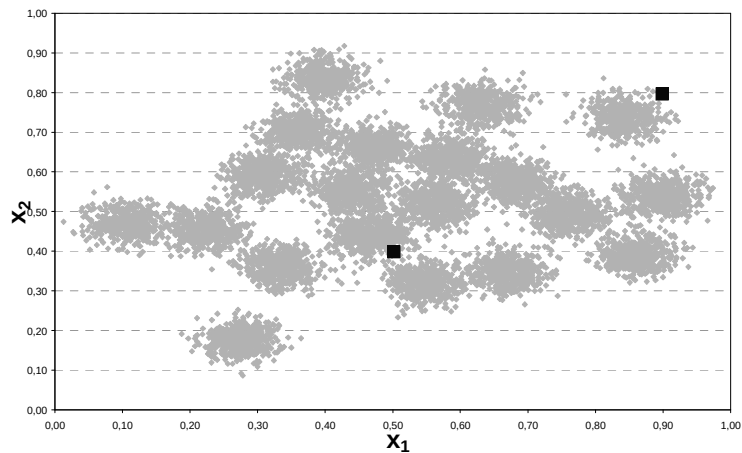
Algoritmus k -průměrů⁺⁺

- Arthur, D., Vassilvitskii, S. (2007)
- postupný výběr inicializačních středů
 - cílem je rovnoměrné rozmístění inicializačních středů mezi daty
 - první střed vybrán náhodně ze všech objektů
 - další jsou vybrány postupně, vždy dle pravděpodobnosti

$$P = \frac{D(\mathbf{x}')^2}{\sum_x D(\mathbf{x})^2}$$

kde $\mathbf{x}'$ je zkoumaný objekt, $D(\mathbf{x})$ je nejkratší vzdálenost objektu $\mathbf{x}$ od nejbližšího středu ze všech doposud vybraných

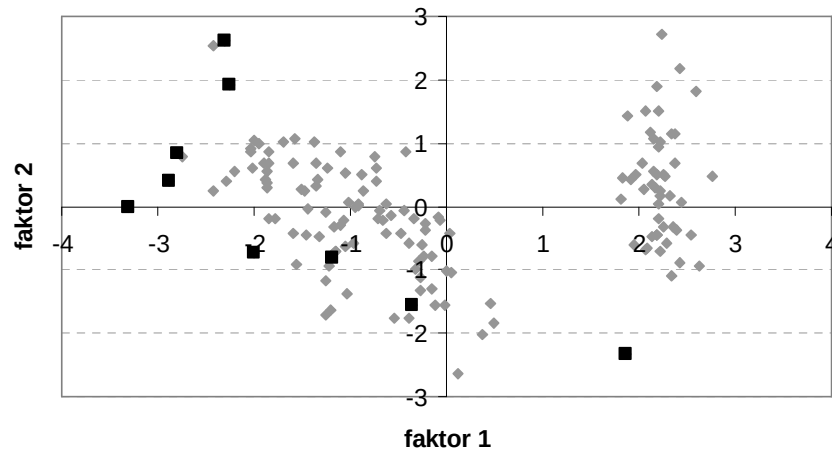
Tvorba inicializačních center



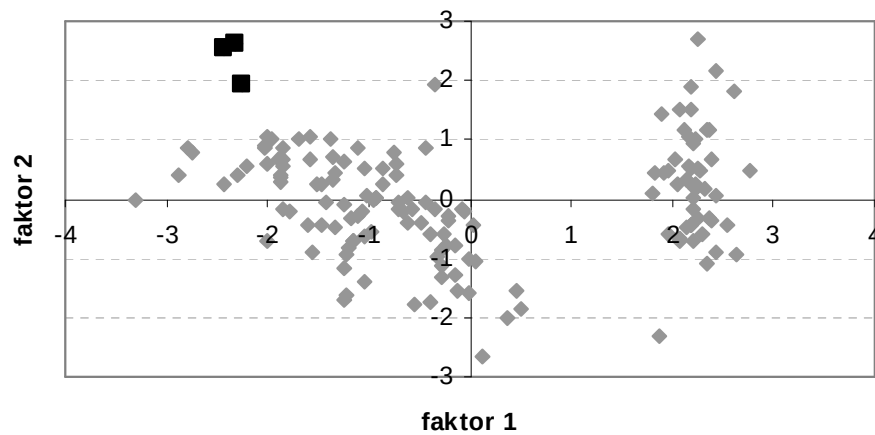
Data

- soubor IRIS
 - 150 objektů,
4 numerické proměnné, 3 shluky
- soubor VOWEL
 - 528 objektů,
10 numerických proměnných, 11 shluků
- soubor GENER
 - 1 000 000 objektů,
2 numerické proměnné, 20 shluků

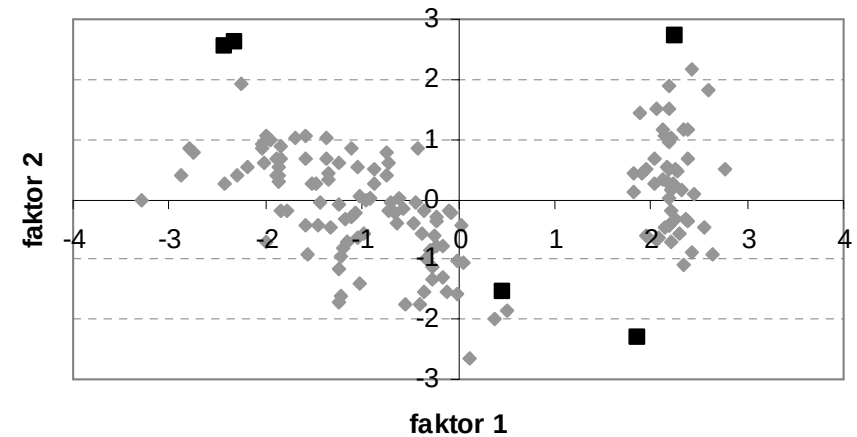
Soubor IRIS



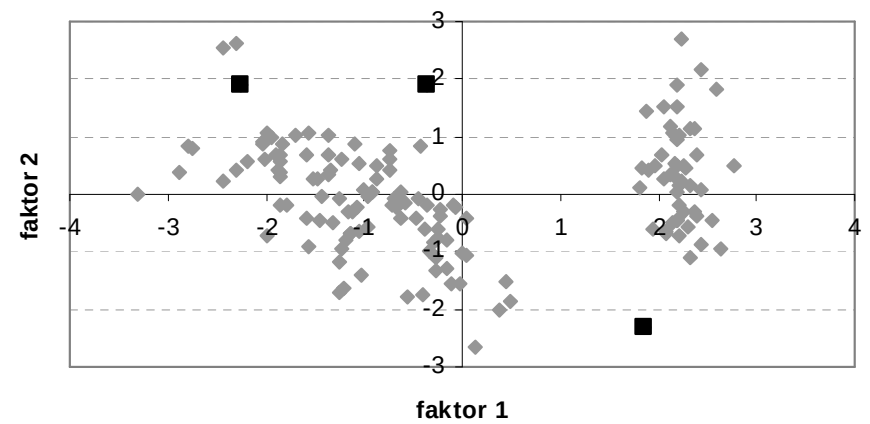
Dvoufázový algoritmus



k -průměrů⁺⁺ ($k = 10$, počet = 3)

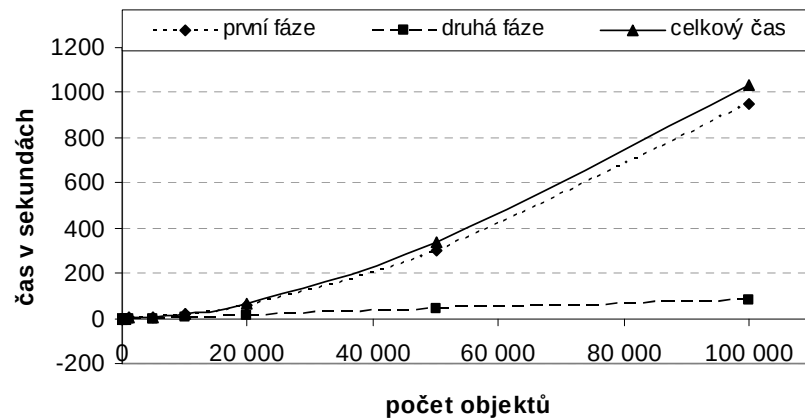


Modifikovaný filtrovací algoritmus

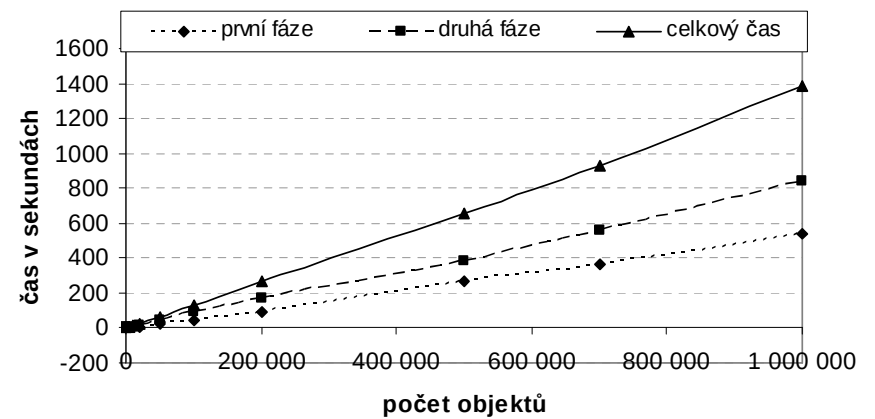


k -průměrů⁺⁺ ($k = 20$, počet = 1-2)

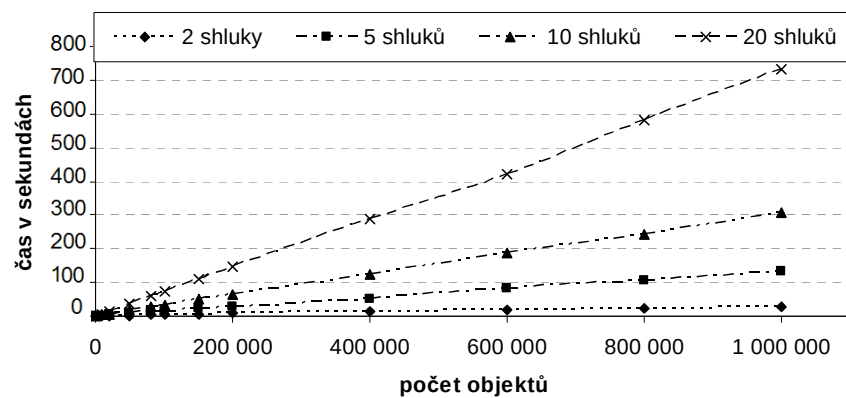
Soubor GENER



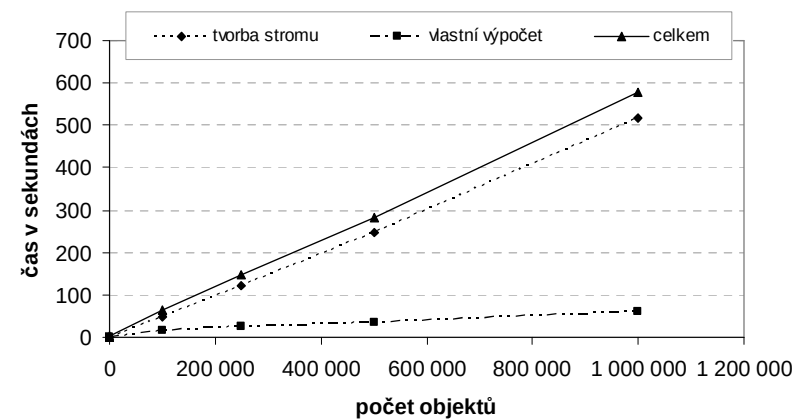
Dvoufázový algoritmus, limitní Q



Dvoufázový algoritmus, počet iterací



k -průměrů⁺⁺



Modifikovaný filtrovací algoritmus

Shrnutí

- popis a otestování několika vybraných variant metody k -průměrů
- výsledky dvoufázového algoritmu a algoritmu k -průměrů⁺⁺ jsou závislé na nastavení vstupních parametrů programu
- výsledky varianty filtrovacího algoritmu nejsou ovlivněny žádným uživatelským nastavením
- nejrychlejší modifikovaný filtrovací algoritmus
- existuje ještě řada dalších variant, které by bylo dobré prozkoumat
- další výzkum bude zaměřen na testování jednotlivých algoritmů na speciálně vygenerovaných velkých souborech dat s cílem dokonalejšího porovnání variant