Dvoufázový způsob vytváření nekonvexních shluků využívající metody *k*-průměrů

Marta Žambochová

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

Robust 2014

19. – 24. leden, Jetřichovice

Motivace

- Potřeba metod pro analýzu dat velkých datových souborů
- Existence velkého množství variant základních metod shlukové analýzy
- Nedostatečně popsané porovnání vhodnosti použití jednotlivých metod
- Hledání dalších variant

Základní principy metody *k*-průměrů

- Optimalizační metoda
- Hledá rozklad objektů do ***k*** shluků, pro který je součet vzdáleností jednotlivých objektů od centroidu jejich shluku minimální

(tj. minimalizace funkce $Q = \sum_{x} \left\| x - c(x) \right\|^2$)

Základní algoritmus metody *k*-průměrů

- Inicializační krok:
  - Prvotní rozdělení souboru dat do ***k*** shluků
- Krok 1:
  - Výpočet centroidů všech shluků
- Krok 2:
  - Přiřazení všech objektů k centroidům
- Krok 3:
  - Pokud došlo ke změně oproti předchozí iteraci, návrat na Krok 1.

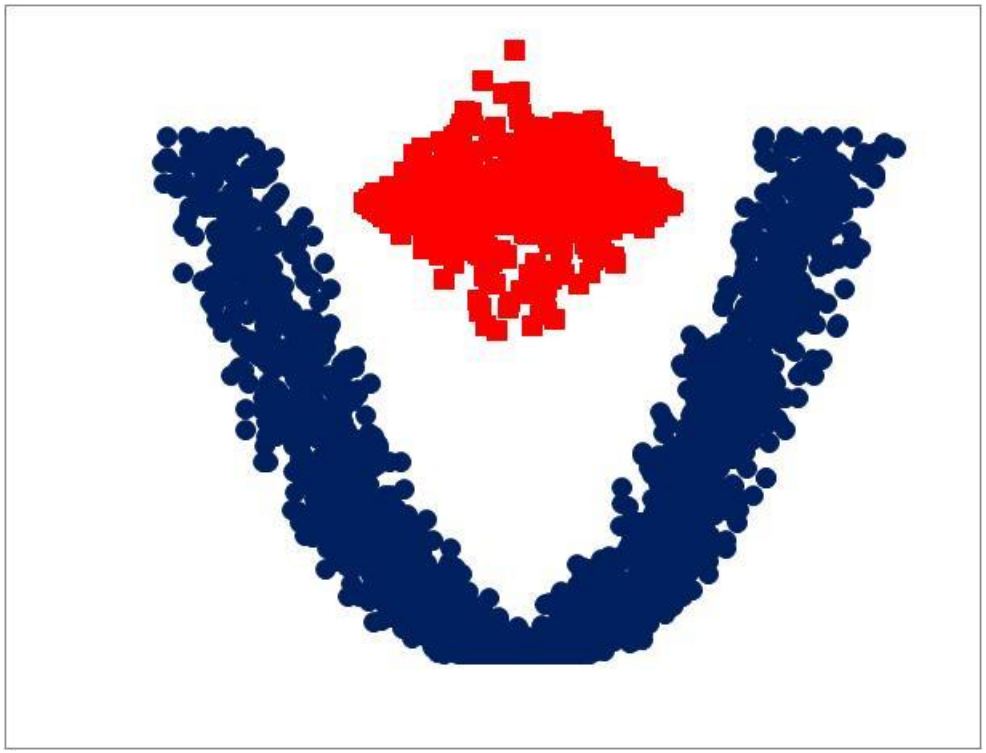
Výhody metody *k*-průměrů

- Jednoduchý princip
- Přijatelná rychlost - použitelnost pro velké soubory dat
- Relativně dobré výsledky (vzhledem k minimalizaci vnitroskupinové variability)

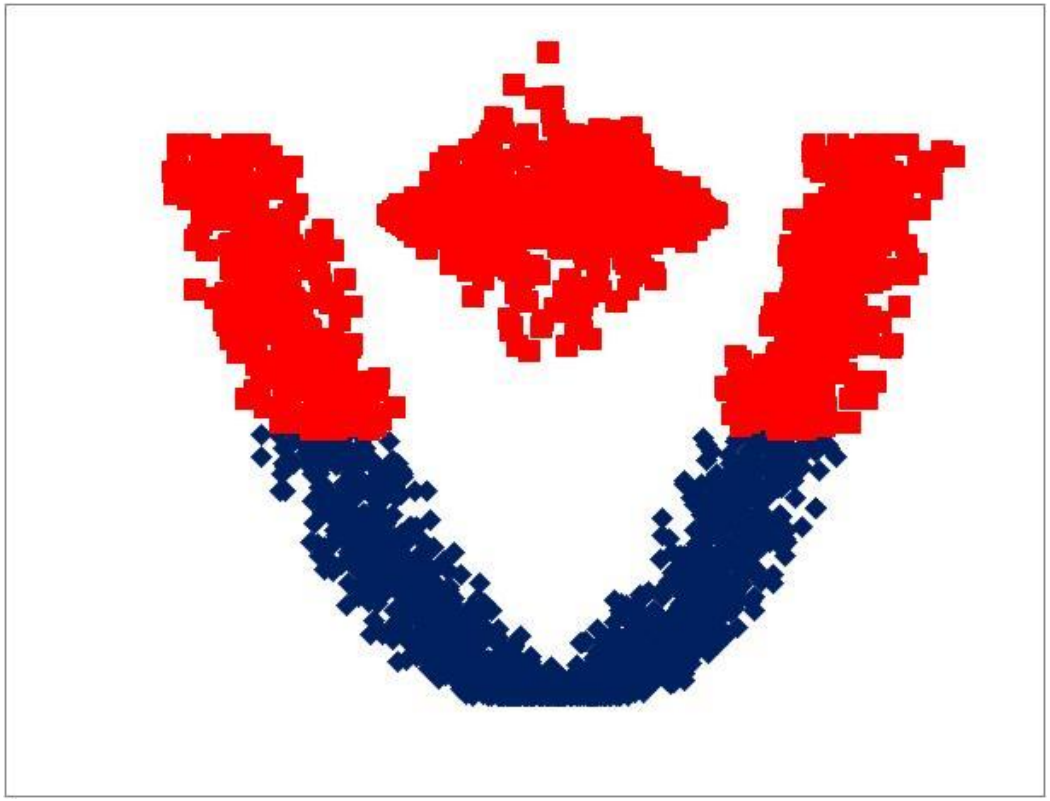
Nevýhody metody *k*-průměrů

- Nutno zadat požadovaný počet shluků
- **Hledá pouze konvexní shluky sférického tvaru**
- Hledá pouze lokální minimum
- Pro obzvlášť velké soubory velká časová náročnost
- Silný vliv inicializačního rozdělení

Ukázka I.

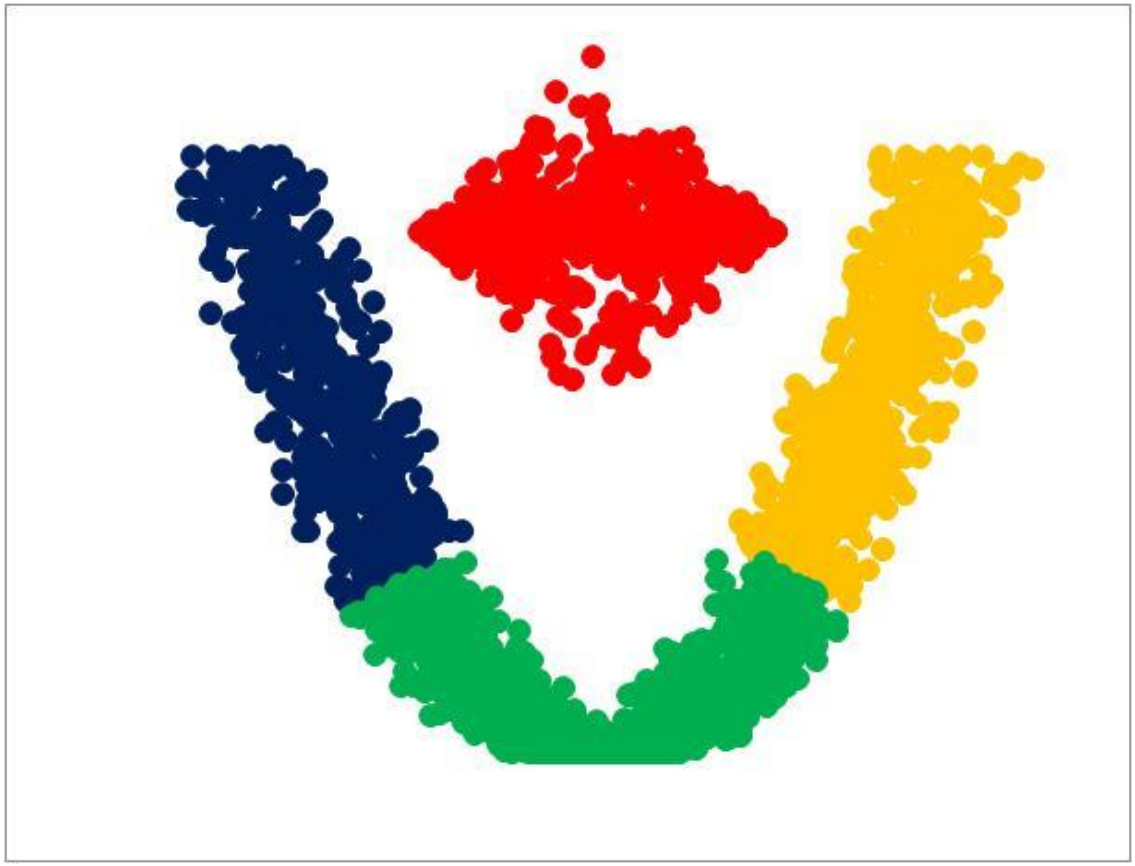


Metoda *k*-průměrů
2 shluky

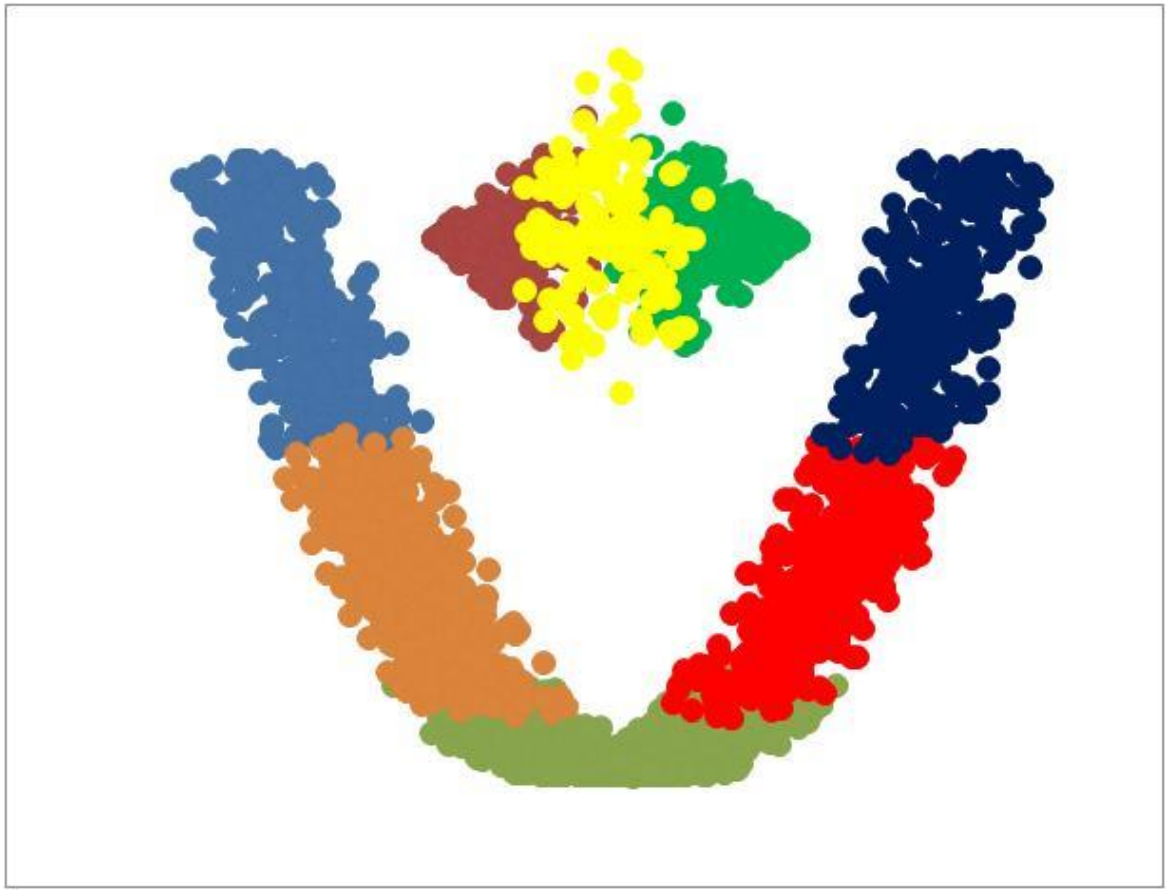


Metoda *k*-průměrů
3 shluky

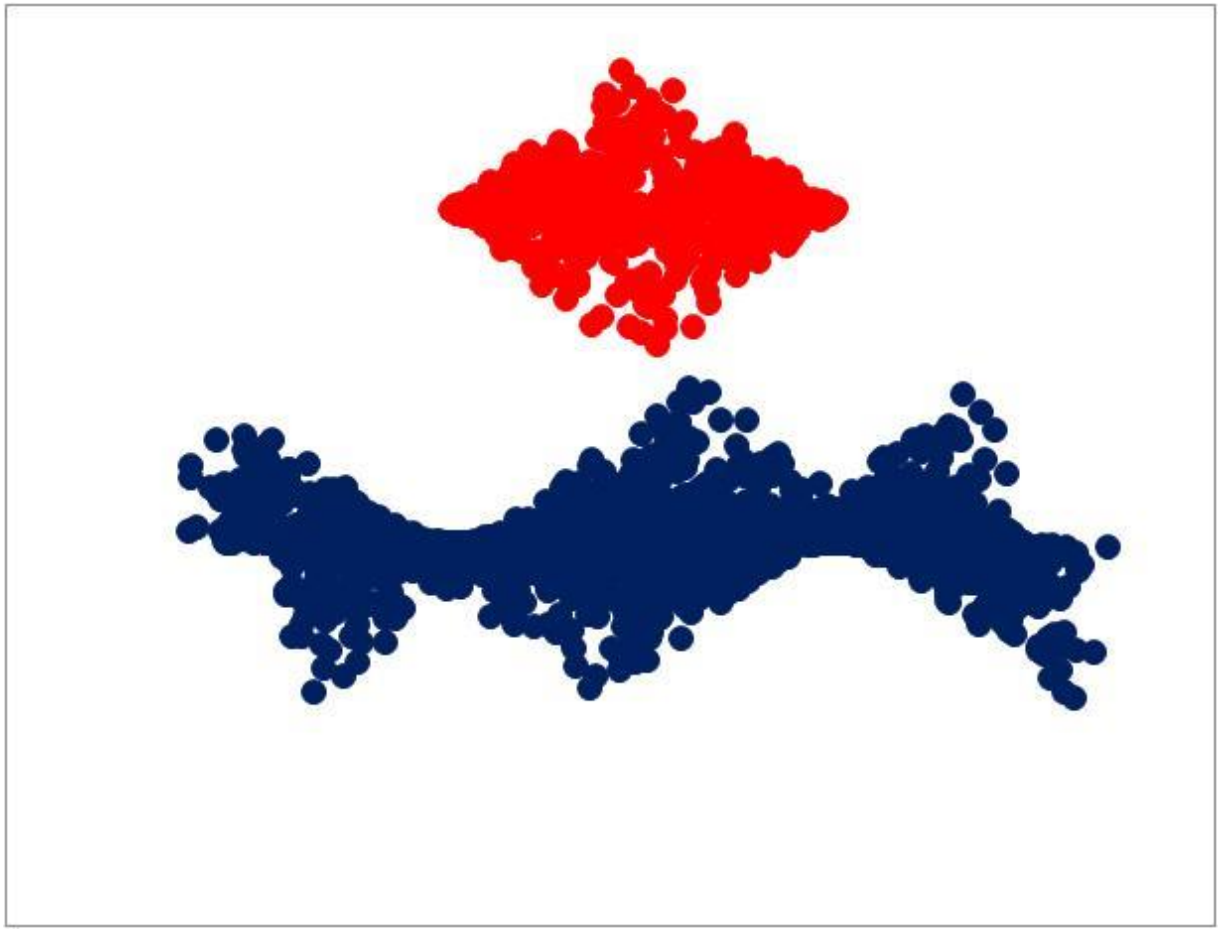
Metoda *k*-průměrů
4 shluky



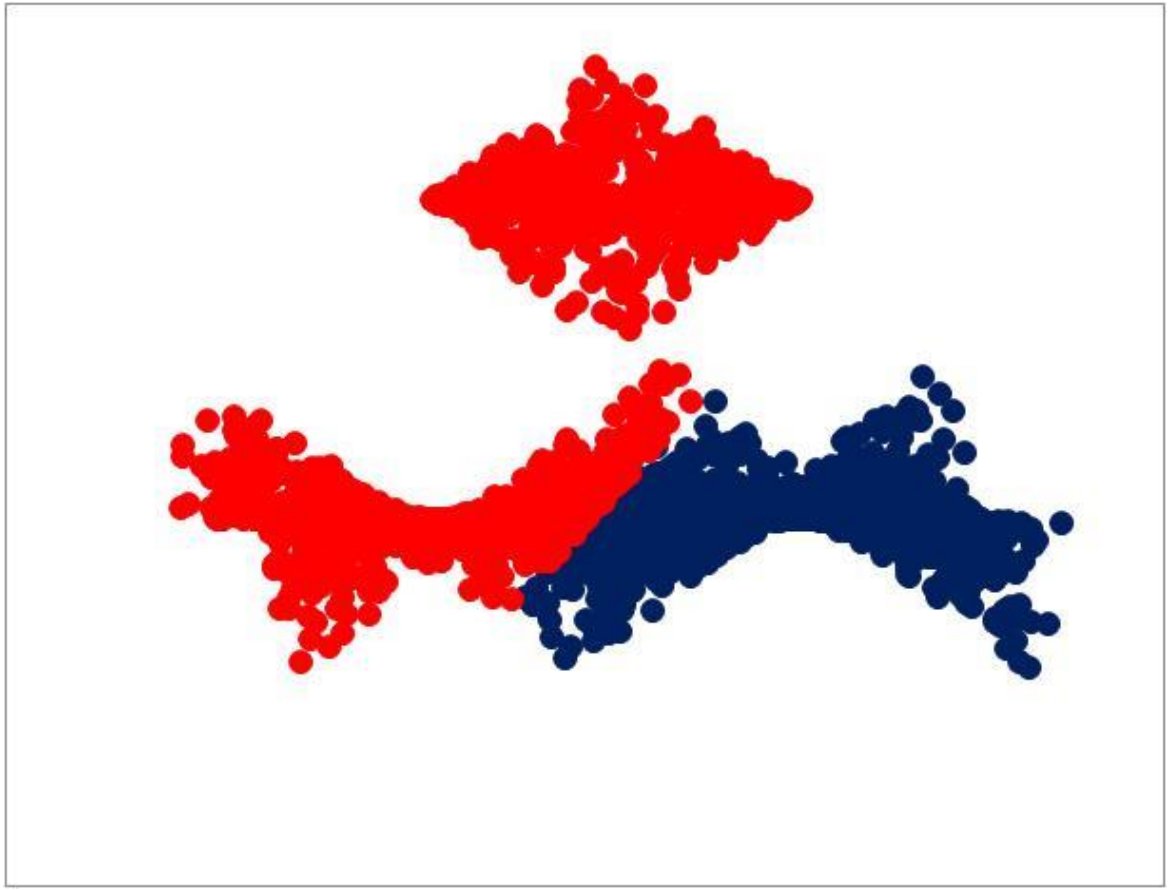
Metoda *k*-průměrů
8 shluků



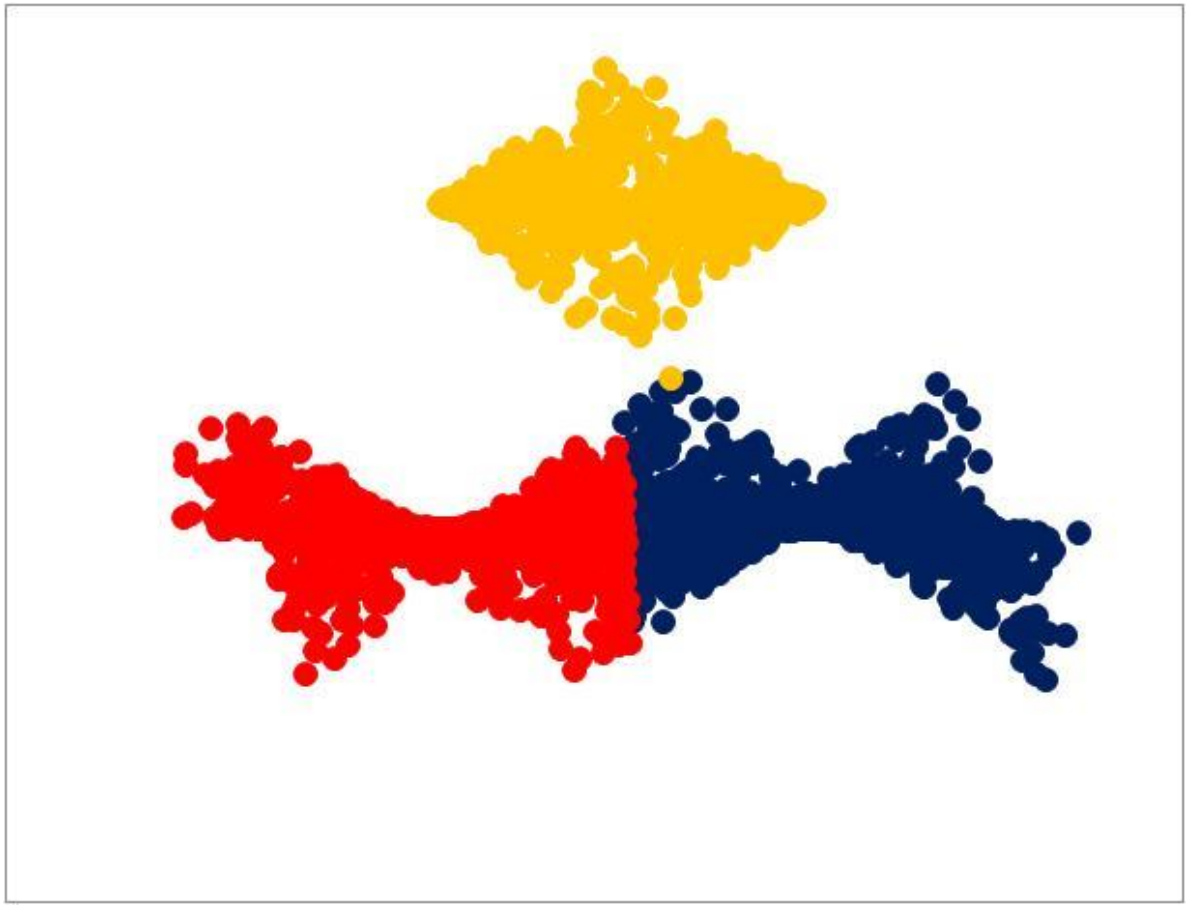
Ukázka II.



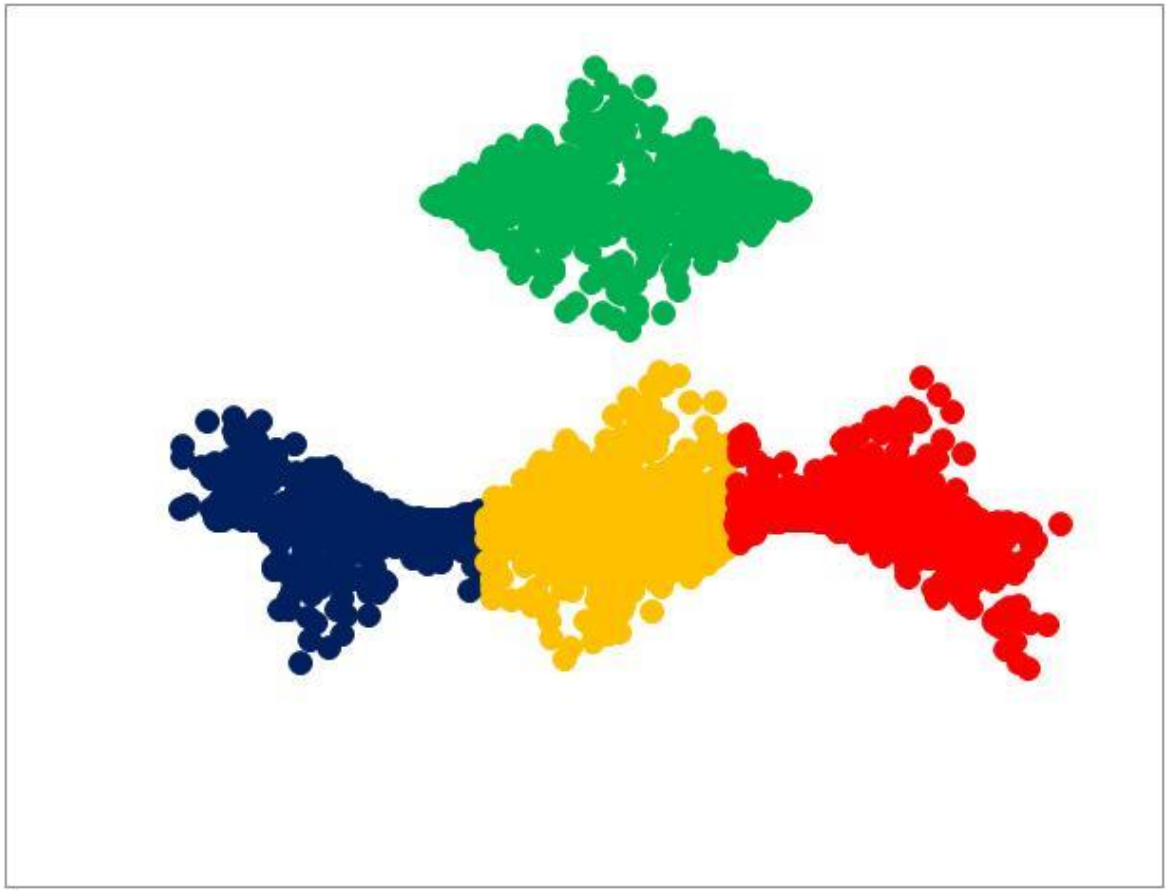
Metoda *k*-průměrů
2 shluky



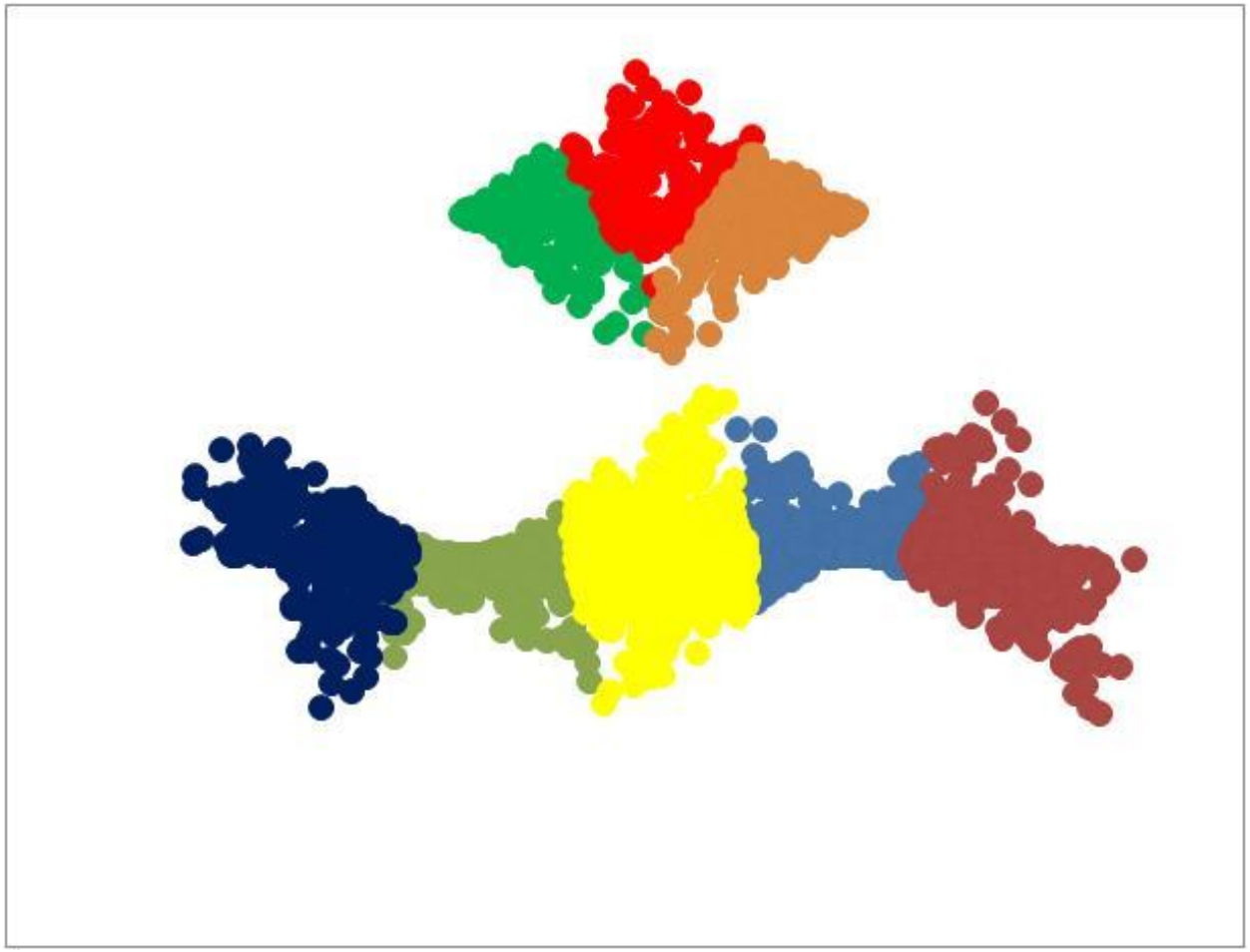
Metoda *k*-průměrů
3 shluky



Metoda *k*-průměrů
4 shluky



Metoda *k*-průměrů
8 shluků



Metody spojování shluků

- **Metoda nejbližšího souseda (Simple linkage)**
- **Metoda nejvzdálenějšího souseda (Complete linkage)**
- **Centroidní metoda (Centroid linkage)**
- **Metoda průměrné vazby (Average linkage)**
- **Mediánová metoda (Unweighted group average)**
- **Wardova metoda**

Metoda nejbližšího souseda (Simple linkage)

- Vzdálenost mezi shluky počítá tak, že vezme nejmenší ze vzdáleností každých dvou objektů z dvou různých shluků.
- Výsledné shluky nemají sférický charakter.
- Nevýhodou této metody je, že pokud existují objekty se stejnou vzdáleností od již existujících shluků, tak může dojít ke zřetězení.
- Velká časová náročnost

Metoda nejvzdálenějšího souseda (Complete linkage)

- Vzdálenost dvou shluků bere největší možnou vzdálenost ze vzdáleností každých dvou objektů z dvou různých shluků.
- Vytváří těsné shluky přibližně stejné velikosti.
- Má tendenci tvořit sférické shluky.
- Zabraňuje vzniku zřetězených shluků.
- Velká časová náročnost

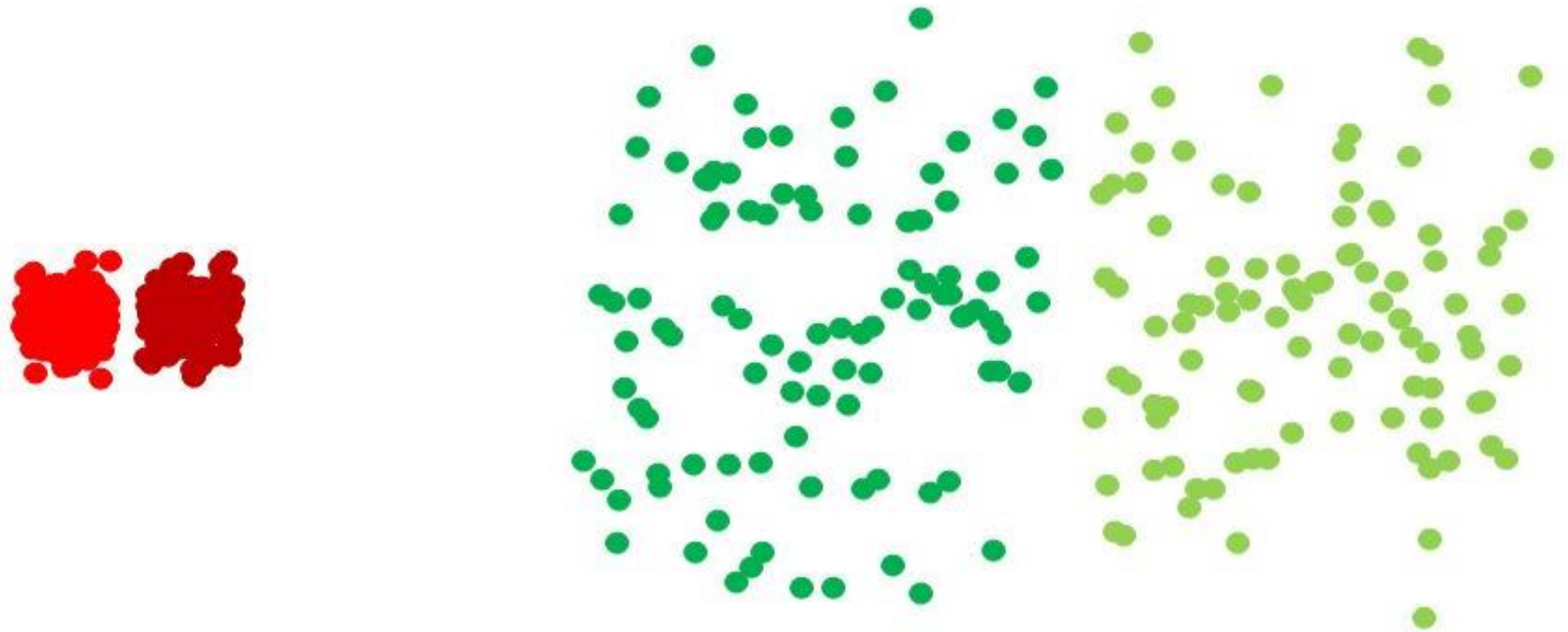
Centroidní metoda (Centroid linkage)

- Pro spočítání nepodobnosti objektů se využívá euklidovská metrika, v které se změří vzdálenosti těžišť shluků.
- Menší časová náročnost

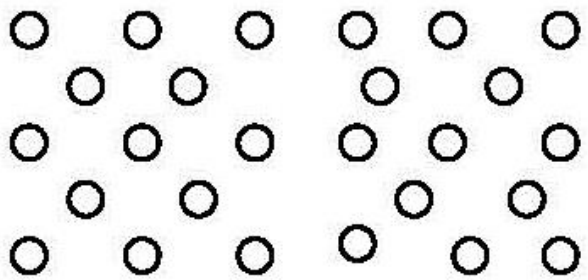
Wardova metoda

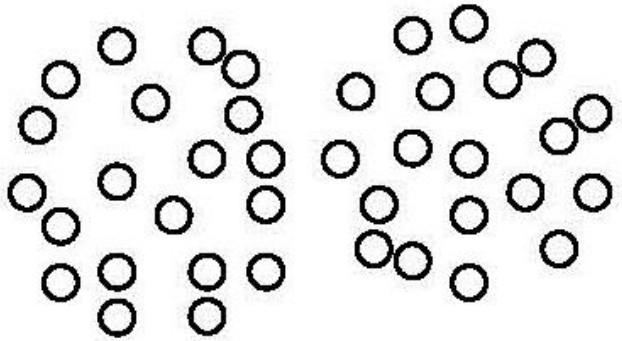
- Minimalizuje přírůstek celkového vnitroskupinového součtu čtverců.
- Vytváří shluky srovnatelné velikosti.
- Tvoří sférické shluky.
- Velmi vysoká časová náročnost

Problém slučování I.



Problém slučování II.





Algoritmus CHAMELEON

- Ve svém průběhu využívá slučování shluků na základě dynamického programování.
- Bere v úvahu dvě charakteristiky
  - Relative Inter-Connectivity
  - Relative Closeness
- Využívá uživatelsky nastavitelných prahových konstant T_{RI} a T_{RC}, které se v průběhu programu dynamicky mění, a to v případě, že existuje více možností na sloučení, nebo naopak žádná.
- Složitost algoritmu $O(nm + n\log n + m^2 \log m)$, přičemž $O(n\log n)$ se týká první části zpracování

Shrnutí

- Práce obsahuje popis varianty metody *k*-průměrů pro hledání shluků obecně nesférického tvaru
- Složitost (by měla být) lineární
- Je ještě nutno technicky vylepšit
- Je nutno ověřit doporučený počet shluků v 1. fázi
- Je nutno ověřit pro větší různorodost vzhledu shluků, pro různé typy rozložení dat

Děkuji za pozornost