Z2 c dokumentacia

Nikolas Knapik

November 2024

1 Úvod

Táto dokumentácia analyzuje implementáciu zhlukovania v 2D priestore. Hlavným cieľom je analyzovať priestor a rozdeliť ho na k zhlukov pomocou dvoch variánt aglomeratívneho zhlukovania:

- Použitie centroidu ako stredového bodu
- Použitie medoidu ako stredového bodu

Úspešnosť klastrovania sa hodnotí na základe priemernej vzdialenosti bodov od stredu zhluku. Úspešný zhluk nesmie mať priemernú vzdialenosť väčšiu ako 500

2 Popis Algoritmu

2.1 Generovanie bodov

Pre prvých 20 bodov sa náhodne generujú X a Y súradnice v intervale [-5000,5000]. Pomocou set() sa zabezpečia unikátne súradnice v prípade duplicít.

Ďalších 20 000 bodov sa generuje od náhodne vybraného už existujúceho bodu od ktorého sa generujú nové súradnice s offsetom 100

2.2 Algomeratívne zhlukovanie

1. Inicializácia:

   Každý z 20 020 vygenerovaných bodov je na začiatku zaevidovaný ako samostatný zhluk. Vzdialenosti medzi všetkými dvojicami zhlukov sú uložené v 2D matici vzdialeností, ktorej diagonála je 0, ktorá predstavuje vzdialenosť zhluku od samého seba

2. Iterácia zhlukovania:

   - Nájde sa najbližšia dvojica zhlukov (pomocou minimálnej hodnoty v matici vzdialeností)
   - Zhluky sa zlúčia a aktualizuje sa ich stred:
     - **Centroid:** Priemerná súradnica bodov v novom zhluku.
     - **Medoid:** Bod v zhluku s najmenšou celkovou vzdialenosťou k ostatným bodom.
   - Matica vzdialeností sa aktualizuje len pre nový zhluk.

3. Ukončenie:

   - Algoritmus sa zastaví, keď priemerná vzdialenosť bodov od stredu pre všetky zhluky klesne pod 500

2.3 Vyhodnocovanie úspešnosti

Úspešnosť klastrovania sa hodnotí ako percento zhlukov, ktoré spĺňajú kritérium priemernej vzdialenosti bodov od stredu menšej ako 500.

3 Reprezentácia údajov

- **Body:** Uložené ako zoznam dvojíc (X,Y).
- **Matica vzdialeností:** 2D numpy matica s hodnotami vzdialeností medzi všetkými dvojicami bodov.
- **Zhluky:** Slovník, kde kľúč je ID zhluku a hodnota je zoznam bodov.

- **Stredy zhlukov:**
  - Centroid: Priemerná súradnica bodov v zhluku.
  - Medoid: Bod s najmenšou celkovou vzdialenosťou k ostatným bodom.

4 Optimalizácie

Počas optimalizácie som si zvýšil offset na 300 aby som lepšie videl vytvorené kódy. Funkčnosť kódu som vo väčšine prípadoch testoval na 1000 až 5000 bodoch. Importol som knižnicu time aby som vedel testovať čas, za ktorý program zbehne. Pri stále dlhých časoch aj pri menších počtoch bodov som import a implementoval niektoré funkcie z numpy na urýchlenie procesov hlavne pri vytvorení a spravovaní matice a počítanie vzdialeností. Toto mi výrazne urýchlilo bežanie programu. Takisto som rýchlosť testoval bez vizualizácie, ktorá mi už pri malých počtoch zaberala aj do dvoch minút.

4.1 Porovnanie Centroidu a Medoidu

V mojom kóde sa prepínanie používania centroidu a medoidu prepína pomocou use centroids = True alebo = False Každý spôsob má na výpočet svoju vlastnú funkciu, ktorá sa volá na základe use centroids. Kedže pracujem s veľkým množstvom dát, výhodnejšie je pre mňa použitie centroidu kedže výpočet medoidu je náročnejší na čas. Výber medoidu pre výpočet podobných dát môže byť efektívne kvôli faktu, že medoid je konkrétny bod z daného klusteru.

5 Vizualizácia

Všekty bodu sú vykreslené v 2D priestore aj pred aj po zhlukovaní. Po zhlukovaní sú zhluky rozdelené podľa farby a centroid/medoid každého zhluku je označný X

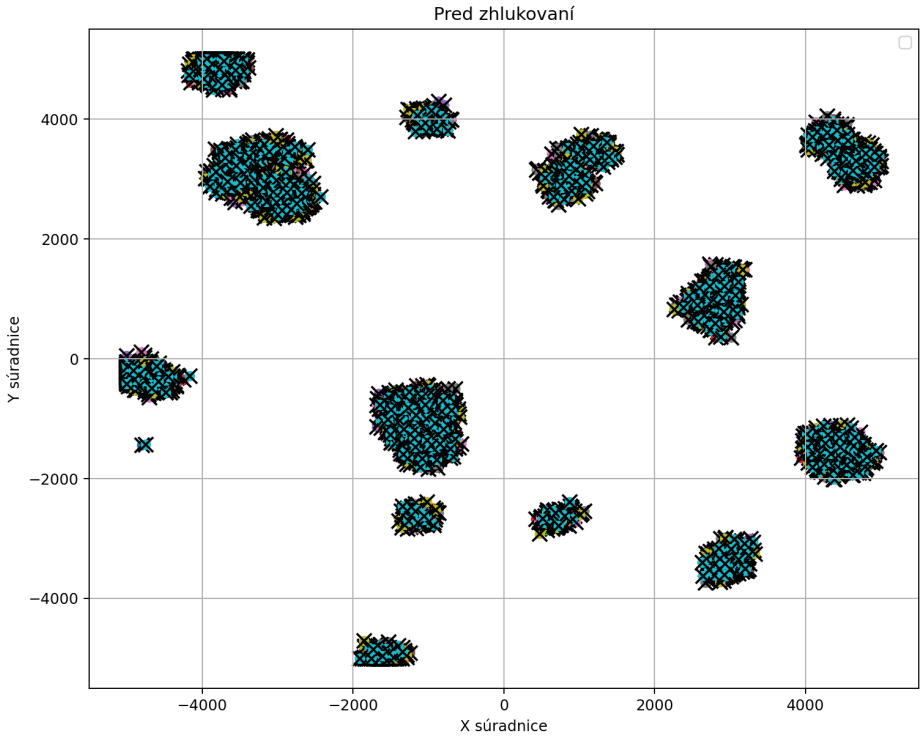

Figure 1: Priklad 20 000 bodov pred zhlukovaním

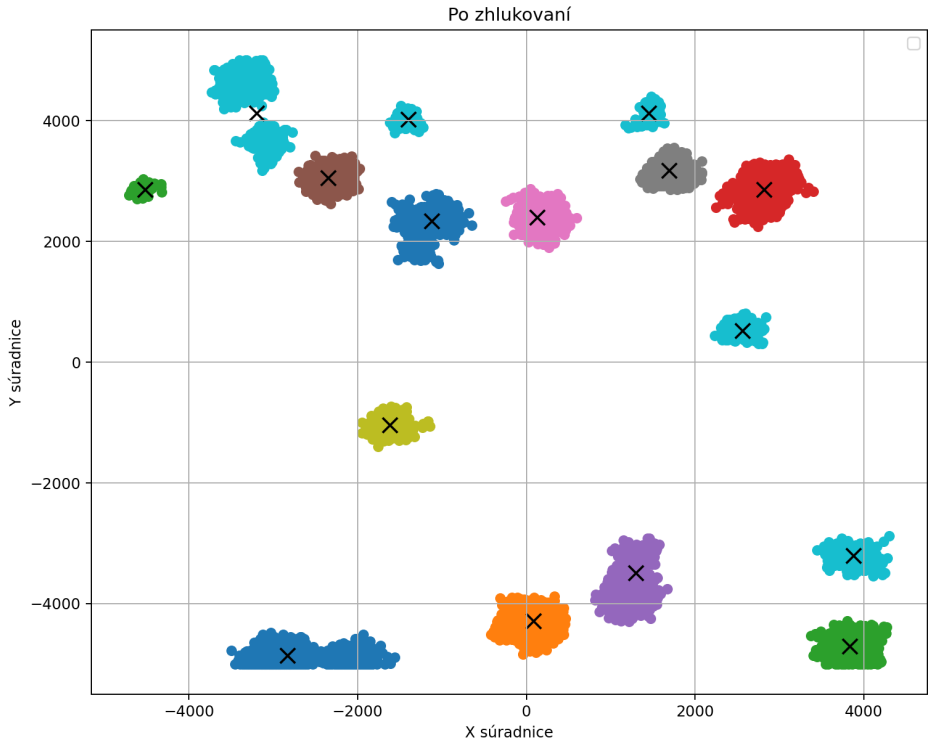

Figure 2: 20 000 bodov. Centroid ako stred

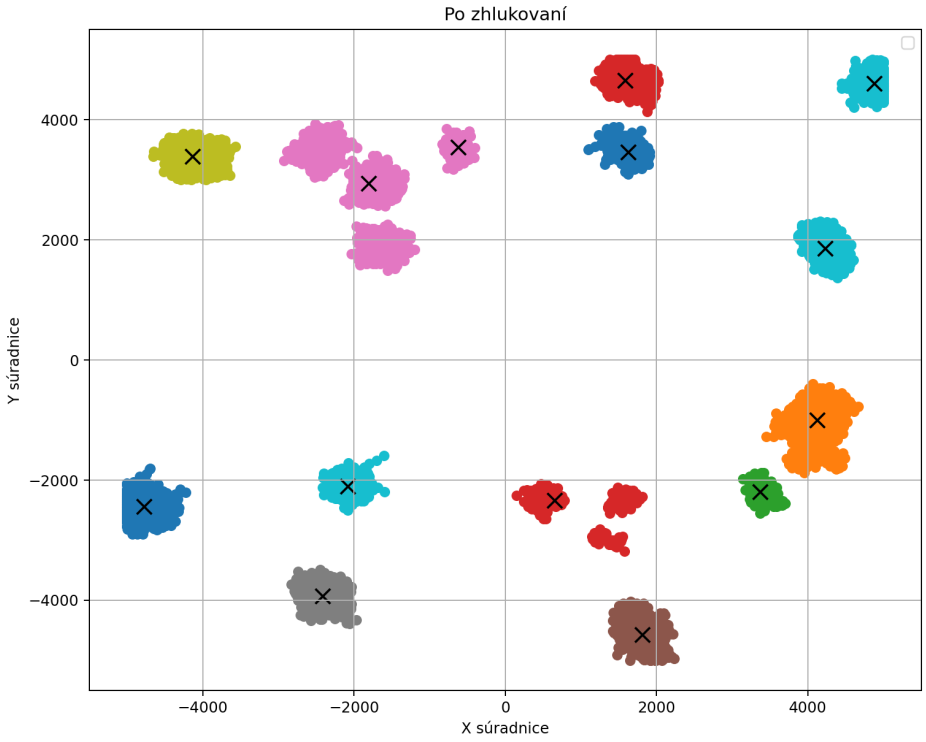

Figure 3: 20 000 bodov. Medoid ako stred

6 Čas behu kódu

Kód mi aj po optimalizácií bežal pomerne dlho. Kedže algoritmus má kubickú zložitosť, pri 20 000 bodoch mi algortimus trval až cez 3 hodiny. Na základe funkcie time som zachytil čas pre centroid 216 minút. Pre medoid mi kód pre 20 000 bodov bežal 360 minút. Taktiež som zistil že vizualizácia zaberá pomerne veľký čas hlavne pri prvom vykreslení. Algoritmus bez vizualizacie už pri 1000 vyhadzoval zrýchlenie skoro o 2 minúty.

7 Záver

Na základe môjho kódu a vykonanej optimalizácie sa centroid preukázal ako efektívnejšie predstavenie stredu zhluku. Úspešnosť je hodnotená na konci programu a je vypočítaná podľa priemernej vzdialenosti bodov od stredu zhluku. Vizualizácia poskytla jasný tvorbu zhlukov a potvrdil funkčnosť implementovaných algoritmov. Celkovo táto implementácia ukázala, že aglomeratívne zhlukovanie je robustný prístup na anlýzu dát v 2D priestore.