Soft Computing: Demonstrace činnosti algoritmu Fuzzy K-means

Katarína Grešová (xgreso00)

21. novembra 2019

1 Úvod

Ůlohou tohto projektu bolo vytvoriť demonštrátor činnosti algorimu Fuzzy K-means. V prvej časti tejto práce je popísaný samotný algormitmus a jeho implementácia. Druhá časť obsahuje popis a ovládanie vytvorenej aplikácie.

2 Fuzzy K-means

Algoritmus Fuzzy K-means (nazývaný taktiež Fuzzy C-means) vyvinutý J. C. Dunnom v roku 1973 [1] je rozšírením jednoduchého populárneho zhlukovacieho algoritmu K-means. Fuzzy K-means je viac štatisticky založená metóda a vytvára zhluky, kde konkrétny bod patrí do viacerých zhlukov s určitými pravdepodobnosťami.

Body na okraji zhluku majú menší stupeň príslušnosti než body v centre zhluku a tým je dobre popsíné rozloženie objektů v zhlukoch. Objekt tak môže patriť do viacerých zhlukov zároveň. Dajú se lepšie identifikovať objekty, které se nedajú priradiť do žiadného zhluku. Vytvorenie zhlukov na základe pravdepodobností príslušností objektov je taktiež jednoduchšie. V prípade, že každý objekt má pravdepodobnosť príslušnosti k nejakému zhluku rovnú jedna a k ostatným nulovú, potom je výsledkem pevné zhlukovánie. Naopak jednotlivé zhluky sú neurčiteľné, ak se stupeň príslušnosti každého objektu k ľubovolnému zhluku rovná prevrátenej hodnote počtu zhlukov. Súčet koeficientov príslušnosti jednotlivých objektov ke všetkým zhlukom je 1.

Algoritmus je založený na minimalizácii funkcie [2]:

$$J = \sum_{i=1}^{N} \sum_{j=1}^{C} m_{ij}^{q} \|\vec{x}_i - \vec{c}_j\| \tag{1}$$

Pseudokód algoritmu môže vyzerať nasledovne:

Algorithm 1: Fuzzy K-means

```
vyber počet zhlukov;
náhodne priraď každému bodu koeficient príslušnosti;
while zmena koeficientov príslušnosti je väčšia ako prah do
    spočítaj stred každého zhluku;
    pre každý bod spočítaj koeficient príslušnosti k jednotlivým zhlukom;
end
```

Kde hodnoty stredov zhlukov sa počítajú podľa vzťahu:

$$c_j = \frac{\sum_{i=1}^{N} m_{ij}^{q} . \vec{x}_i}{\sum_{i=1}^{N} m_{ij}^{q}} \tag{2}$$

A hodnoty koeficientov príslušnosti sa počítajú podľa vzťahu:

$$m_{ij} = \frac{1}{\sum_{k=1}^{C} \left(\frac{\|\vec{x}_i - \vec{c}_j\|}{\|\vec{x}_i - \vec{c}_k\|} \right)^{\frac{2}{q-1}}} \tag{3}$$

2.1 Implementácia

Samotný algoritmus Fuzzy K-means je implementovaný v balíku kmeans. Trieda FuzzyKMeans je vstupným bodom. Táto trieda je inicializovaná všetkými potrebnými parametrami pre algoritmus:

- points: body, ktoré majú byť zhlukované.

- centers/numOfCenters: počiatočné stredy zhlukov, prípadne číslo určujúce počet zhlukov, ktoré sa majú vytvoriť. Pre lepšiu prehľadnosť demonštrátora je počet zhlukov obmedzený na 10.

- fuzziness: parameter určujúci, ako veľmi „fuzzy“ majú výsledné zhluky byť. Čím vyššia hodnota, tým viac „fuzzy“ budú zhluky. Tento parameter odpovedá premennej q v predchádzajúcich vzorcoch.

- epsilon: parameter určujúci, pri akej malej zmene koeficientov príslušnosti už zastavíme výpočet a budeme ho považovať za úspešný.

Implementácia algoritmu sa mierne odlišuje od uvedeného algoritmu a to tým, že bodom sa nepriraďuje koeficient príslušnosti náhodne. Keďže už na vstupe máme počiatočné stredy zhlukov, tak počiatočné koeficienty príslušnosti môžeme vypočítať podľa vzťahu (3).

Ďalšou triedou balíku kmeans je trieda WeightsMatrix. Táto trieda je zodpovedá za prácu s maticou koeficientov príslušnosti bodov k zhlukom.

Poslednou triedou je trieda Poin, ktorá reprezentuje bod v priestore. Pomocou tejto triedy sú reprezentované zhlukované body ako aj stredy zhlukov. Táto trieda ponúka aj metódy na výpočet Euklidovskej vzdialenosti medzi bodmi, ktorá je potrebná pri výpočte nových stredov zhlukov.

3 Manuál k aplikácii

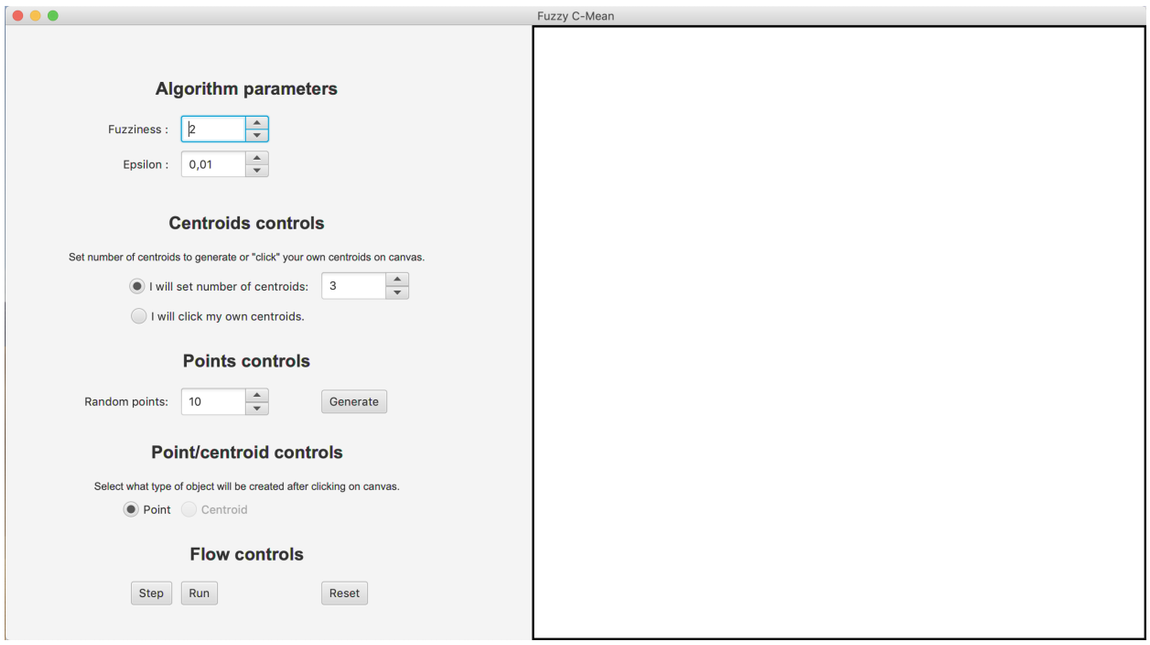
Alikácia na demonštráciu algoritmu Fuzzy K-mean bola implementovaná v jazyku Java s využitým grafickej knižnice JavaFX.

3.1 Preklad a spustenie

Aplikáciu je možné preložiť pomocou nástroja *ant*, príkazom `ant` v koreňovom adresári projektu. Tým je vytvorená zložka *build*, ktorá obsahuje preložené triedy a výsledný program v *jar* formáte. Aplikáciu je následne možné spustiť príkazom `java -jar build/dist/FuzzyKMeans.jar`.

3.2 Ovládanie

Grafické užívateľské rozhranie aplikácie (viz. obr. 1) je rozdelené na dve hlavné časti: nastavenie parametrov vľavo a plátno na vizualizáciu vpravo.



Obr. 1: Grafické rozhranie aplikácie

Ľavá časť s nastavením parametrov je rozdelená do niekoľkých sekcií:

- Algorithm parameters: táto sekcia ponúka prvky na zadanie parametrov fuzziness a epsilon. Predvolená hodnota fuzziness je nastavená nahodnotu 2, čo sa udáva ako vhodná hodnota, ak nevieme viac o zhlukovaných dátach.

- Centroids controls: demonštrátor ponúka vlastné zadanie súradníc stredov počiatočných zhlukov a taktiež ponúka možnosť zadať počet požadovaných zhlukov a počiatočné stredy budú

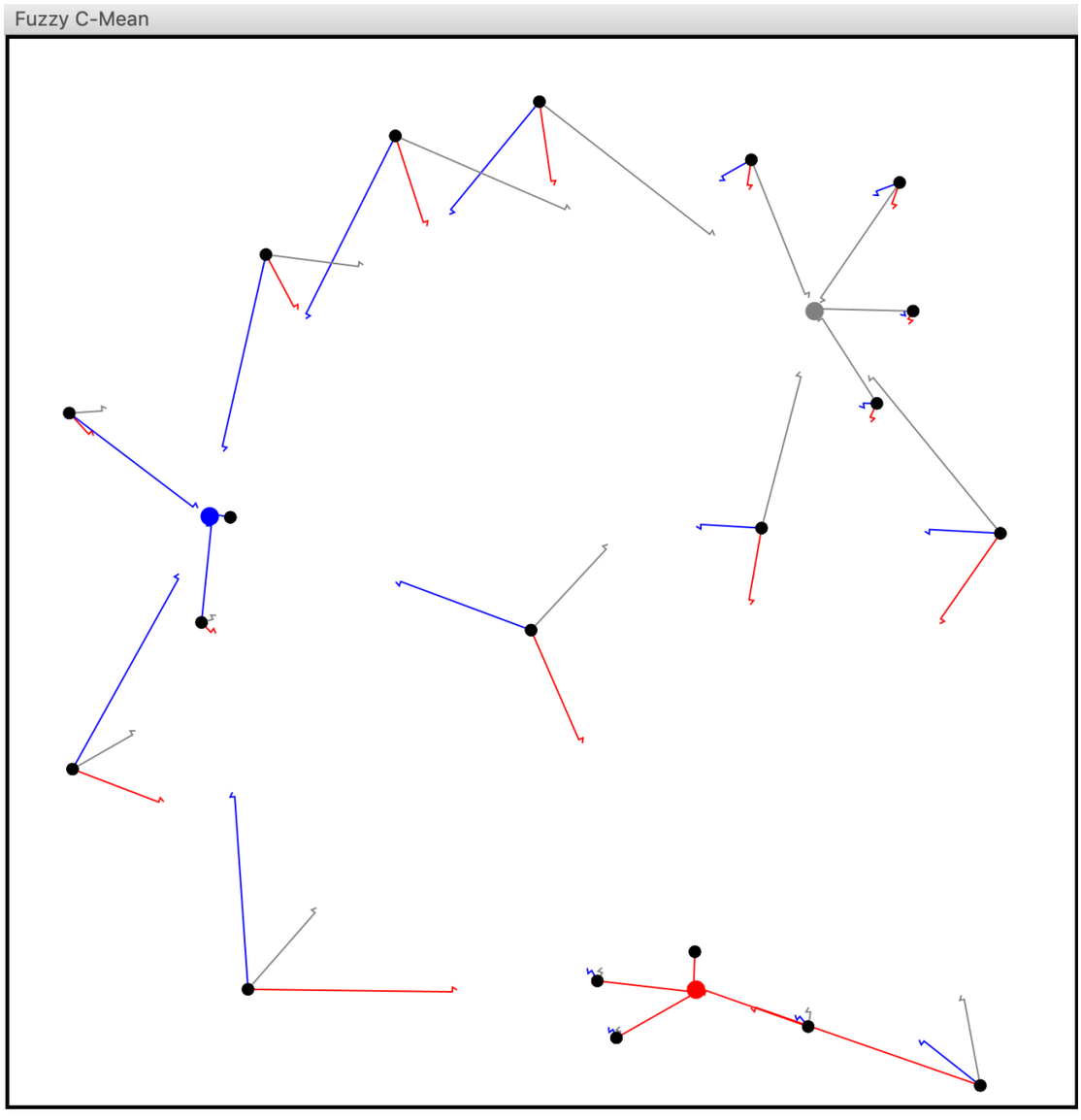
vygenerované náhodne. Výber z týchto dvoch možností sa deje práve v tejto sekcii. Predvolená je možnosť s počtom zhlukov, pretože je to pre užívateľa menej pracné.

- Points controls: na zjednodušenie prípravy demonštrácie bola pridaná táto sekcia, ktorá slúži na náhodné vygenerovanie zadaného počtu bodov na plátno.
- Point/centroid controls: v prípade, že má užívateľ záujem zhlukovať body so špecifickým rozložením, má na to možnosť. Po kliknutí na plátno sa na danom mieste zaznamená bod. Ak má úživateľ záujem ručne zvoliť aj stredy počiatočných zhlukov, potom po zvolení tejto možnosti v sekcii *Centroids controls*, zvolí v tejto sekcii možnosť *Centroids*. Následne sa po kliknutí na plátno zaznamená stred nového zhluku.
- Flow controls: táto sekcia slúži na ovládanie behu zhlukovania. Tlačítko *Step* vykoná jeden cyklus algoritmu – spočíta nové stredy zhlukov a nové koeficienty príslušnosti. Tlačítko *Run* spustí cyklus zhlukovania až kým nebude platiť ukončovacia podmienka. Po stlačení *Step* alebo *Run* je pozastavená možnosť pridávania bodov na plátno a zmena parametrov tiež nemá efekt. Tlačítko *Reset* vymaže výsledky predchádzajúceho zhlukovania, vyčistí plátno a znova povolí zadávanie bodov a parametrov.

3.3 Vizualizácia koeficientov príslušnosti

Najzložitejším problémom demonštrátoru Fuzzy K-means bolo vizualizovať koeficienty príslušnosti. Pri klasickom algoritme K-means patrí každý bod práve jednému zhluku. Vizualizácia sa dá spraviť napríklad farebným rozlíšením jednotlivých zhlukov. Pri algoritme Fuzzy K-means je to zložitejšie, pretože každý bod patrí každému zhluku s určitou pravdepodobnosťou.

Bol zvolený prístup k vizualizácii pomocou vektorov, viz Obr. 2. Každému zhluku je priradená určitá farba a z každého bodu vychádza vektor ku každému zhluku s danou farbou. Dĺžka velktoru je vypočítaná z koeficientu príslušnosti. Čím väčšia dĺžka, tým väčšia príslušnosť k danému zhluku.



Obr. 2: Vizualizácia koeficientov príslušnosti

4 Záver

V rámci tohto projektu bola implementovaná aplikácia na demonštráciu zhlukovacieho algoritmu Fuzzy K-means. Program bol implementovaný v jazyku Java s využitím knižnice JavaFX a je preložiteľný na serveri *merlin*.

Literatúra

[1] Joseph C Dunn. A fuzzy relative of the isodata process and its use in detecting compact well-separated clusters. 1973.

[2] František Zbořil. Fuzzy množiny, fuzzy logika, fuzzy inference. `https://www.fit.vutbr.cz/study/courses/SFC/private/19sfc_9.pdf`. Accessed: 2019-11-20.