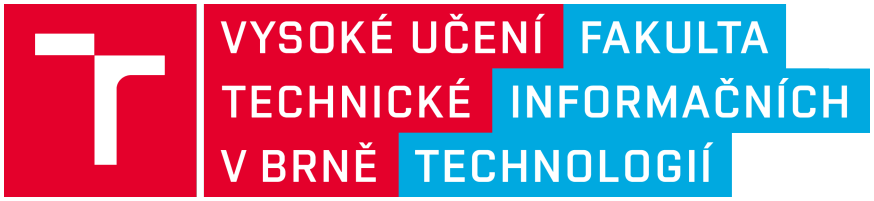


Soft Computing – Projekt

Demonštračná aplikácia algoritmu Fuzzy k-means

30. novembra 2024

Marián Tarageľ (xtarag01)

1 Úvod

Táto dokumentácia stručne predstaví riešené zadanie, implementačné riešenie a popíše ovládanie aplikácie implementovanej v projekte do predmetu SFC (Soft Computing).

2 Riešený problém

Cieľom projektu je implementovať demonštračnú aplikáciu zhlukovacieho algoritmu Fuzzy k-menas. Tento algoritmus určuje príslušnosť jednotlivých trénovacích dát do zhlukov. Jedná sa o algoritmus umelej inteligencie, konkrétne ide o učenie bez učiteľa (unsupervided learning). Táto skupina algoritmov pracuje s dátami, ktoré nie sú anotované a v prípade klasifikačnej úlohy sa ich snaží zaradiť k správnemu klastru. Medzi najpopulárnejšie algoritmy z tejto skupiny patrí algoritmus k-means. Tento projekt implementuje jeho fuzzy variantu založenú na fuzzy množinách, ktorá všeobecne konverguje k lepším riešeniam. Nasledujúci opis implementovaných algoritmov vychádza zo zdrojov [2], [3] a [4].

2.1 K-means clustering

K-menas clustering je najznámejší zhlukovací algoritmus, ktorý iteratívne prepočítava stredy jednotlivých zhlukov. Je založený na vzdialenosti trénovacích bodov od stredov tiež prototypov. Na výsledok algoritmu má vplyv použitá metrika na určenie vzdialenosti - Euklidovská, Hammingova, atď. Jeho nevýhodou oproti fuzzy variante je, že záleží na prvotnej estimácii stredov inak môžu byť výsledky nekorektné. Tento problém do veľkej miery rieši zavedenie fuzzy množín. Objektívnou fukciou pre K-means je

$$\min_{\mathbf{c}_j} \sum_{j=1}^{c} \sum_{\mathbf{x}_i \in \pi_j} \|\mathbf{x}_i - \mathbf{c}_j\|^2, \tag{1}$$

kde $\mathbf{c}_j$ sú súradnice centroidov j-tého zhluku a $\mathbf{x}_i$ sú súradnice i-tého bodu. Ekvivalentne k funkcii 1 je možné napísať funkciu 2,

$$\min_{h_{ij}, \mathbf{c}_j} \sum_{i=1}^{n} \sum_{j=1}^{c} h_{ij} \|\mathbf{x}_i - \mathbf{c}_j\|^2, \quad h_{ij} \in \{0, 1\} \tag{2}$$

kde h_{ij} vyjadruje ostrú príslušnosť do množiny, buď do nej prvok patrí $h_{ij} = 1$ alebo do množiny nepatrí $h_{ij} = 0$. Algoritmus 1 popisuje spôsob zhlukovania K-means.

Algoritmus 1: K-MEANS

1: Inicializuj stredy $\mathbf{c}_j$ (napr. na náhodne vybrané a rôzne vektory z $\mathbf{x}_i$)
2: **do**
3: $\quad \mathbf{c}_{j\_old} = \mathbf{c}_j$
4: $\quad$ Každý vektor $\mathbf{x}_i$ priraď k najbližšiemu stredu $\mathbf{c}_j$ tak, že $\|\mathbf{x}_i - \mathbf{c}_j\| \leq \|\mathbf{x}_i - \mathbf{c}_l\|, \quad l \in \langle 1, c \rangle$
5: $\quad$ Prepočítaj polohy stredov podľa vzťahu $\mathbf{c}_j = \frac{\sum_{\mathbf{x}_i \in C_j} \mathbf{x}_i}{n_j}$
6: **while** $|\mathbf{c}_{j\_old} - \mathbf{c}_j| < \epsilon$;

2.2 Fuzzy k-means

Oproti klasickému K-means, kde môžu dáta patriť iba do jedného zhluku, Fuzzy k-means udáva stupeň príslušnosti do každého zhluku. Každý zhluk je vlastná fuzzy množina do ktorej jednotlive dáta patria

rôznou mierou. Fuzzy množiny nám dávajú viac informácii v porovnaní s klasickými ostrými množinami. Objektívnou funkciou Fuzzy k-means je funkcia 3, ktorú sa snaží minimalizovať. Jej podoba je nasledovná

$$\min_{h_{ij},\mathbf{c}_j} \sum_{i=1}^{n}\sum_{j=1}^{c} h_{ij}^{q}\|\mathbf{x}_i - \mathbf{c}_j\|^2, \tag{3}$$

$$\sum_{j=1}^{c} h_{ij} = 1, h_{ij} \in [0,1],$$

kde h_{ij} je stupeň príslušnosti bodu i do j-tej fuzzy množiny. Hyper parameter q je váhový koeficient taký, že $q \in (1,\infty)$. Obvykle $q = 2$. Tento parameter udáva ako veľmi „fuzzy“ budú jednotlivé množiny. Súradnice bodu sú značené ako $\mathbf{x}_i \in \mathbb{R}^d$. Potom $\mathbf{c}_j \in \mathbb{R}^d$ sú súradnice stredov zhlukov / fuzzy množín. Okrem toho sa v algoritme využívajú rovnice 4 a 5, kde prvá rovnica slúži na prepočítanie nových stupňov príslušnosti bodov do jednotlivých zhlukov a pomocou druhej môžeme spočítať nové stredy fuzzy množín. Pri zväčšení počtu iterácii algoritmu Fuzzy k-means konverguje k lokálnemu minimu alebo sedlovému bodu. Algoritmus 2 popisuje spôsob zhlukovania Fuzzy k-means.

$$h_{ij} = \frac{1}{\sum_{l=1}^{c}\left(\frac{\|\mathbf{x}_i-\mathbf{c}_j\|}{\|\mathbf{x}_i-\mathbf{c}_l\|}\right)^{\frac{2}{q-1}}} \tag{4}$$

$$\mathbf{c}_j = \frac{\sum_{i=1}^{n} h_{ij}^{q}\mathbf{x}_i}{\sum_{i=1}^{n} h_{ij}^{q}} \tag{5}$$

Algoritmus 2: Fuzzy k-means

1: Náhodne inicializuj stupne príslušnosti h_{ij}
2: **do**
3: $\quad\mathbf{c}_{j\_old} = \mathbf{c}_j$
4: $\quad$Prepočítaj polohy stredov podľa vzťahu $\mathbf{c}_j = \frac{\sum_{i=1}^{n} h_{ij}^{q}\mathbf{x}_i}{\sum_{i=1}^{n} h_{ij}^{q}}$
5: $\quad$Prepočítaj stupne príslušnosti bodov do zhlukov podľa $h_{ij} = \frac{1}{\sum_{l=1}^{c}\left(\frac{\|\mathbf{x}_i-\mathbf{c}_j\|}{\|\mathbf{x}_i-\mathbf{c}_l\|}\right)^{\frac{2}{q-1}}}$
6: **while** $|\mathbf{c}_{j\_old} - \mathbf{c}_{j\_new}| < \epsilon$;

3 Implementačné riešenie

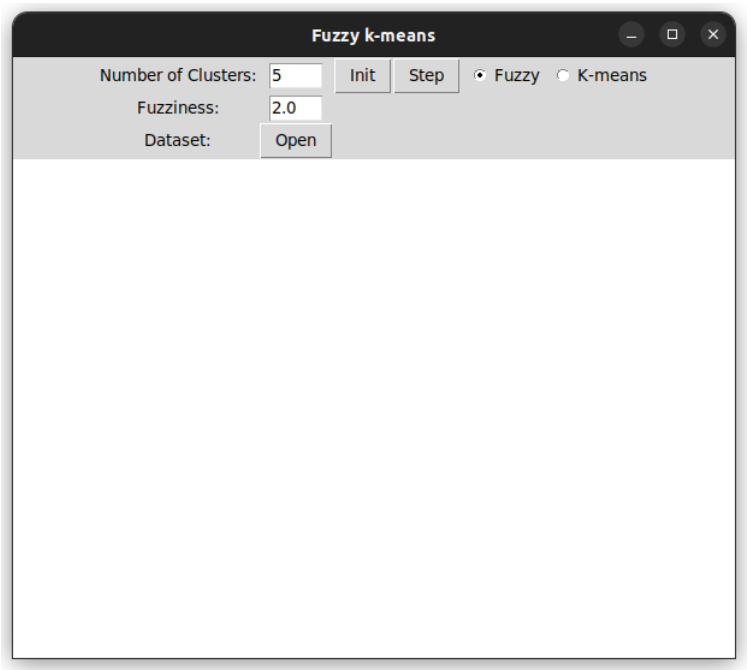
Implementačné sa jedná o demonštračnú aplikáciu napísanú v programovacom jazyku Python. Aplikácia bola rozdelená do modulov podľa návrhového vzoru MVC, kde Model (`model.py`) oddeľuje dáta, View (`view.py`) rieši vykresľovanie dát a Controller (`controller.py`) riadi celú aplikáciu. Hlavným skriptom je `main.py`, ktorý prepája celú aplikáciu. Aplikácia závisí na dvoch externých knižniciach. Prvou je Tkinter, čo je knižnica pre tvorbu GUI v programovacom jazyku Python. Druhou je knižnica Pandas, ktorá umožňuje načítať dáta zo súborov.

Ďalšie dva skripty `kmeans.py` a `fuzzykmenas.py` sú samostatné triedy poskytujúce rozhranie a implementáciu k zhlukovacím algoritmom K-mean a Fuzzy k-means. Ich implementácia závisí iba na matematickej knižnici Numpy. Obidve poskytujú dve verejné metódy a to `init()` a `step()`, ktoré nemajú žiadne parametre. Metóda `init()` inicializuje algoritmus a `step()` postupne vykonáva jednotlivé kroky algoritmu. Konštruktor v prípade triedy `KMeans` má dva parametre a to `train_data` čo sú trénovacie dáta a `k_clusters` čo je počet zhlukov do ktorých majú byť dáta roztriedené. Konštruktor

triedy `FuzzyKMeans` je podobný, ale jemu je možné zadať ešte jeden parameter `q`, ktorý ovplyvňuje ako veľmi sú množiny „fuzzy“.

4 Ovládanie aplikácie

Ovládanie aplikácie je veľmi jednoduché a intuitívne. Po otvorení aplikácie sa zobrazí okno, kde v hornej časti je možné upravovať parametre demonštrovaných algoritmov a v dolnej časti je zobrazovaný priebeh algoritmu. Pri prvotnom otvorení na ploche ešte nie sú zobrazené trénovacie body ako je možné vidieť na obrázku 1. Tie je buď možné nahrať pomocou tlačidla `Open` alebo ak užívateľ nemá žiadne dáta stačí iba stlačiť `Init` a aplikácia si trénovacie dáta sama vygeneruje. Pri zadaní dát zo súboru sa očakáva, že každý bod bude na samostatnom riadku a súradnice x a y budú oddelené medzerou. Odovzdaný archív obsahuje pár súborov `s1.txt`, `s2.txt`, `s3.txt`, `s4.txt`, `a1.txt` a `unbalanced.txt`, ktoré boli prevzaté zo zdroja [1] ako skúšobné datasety.



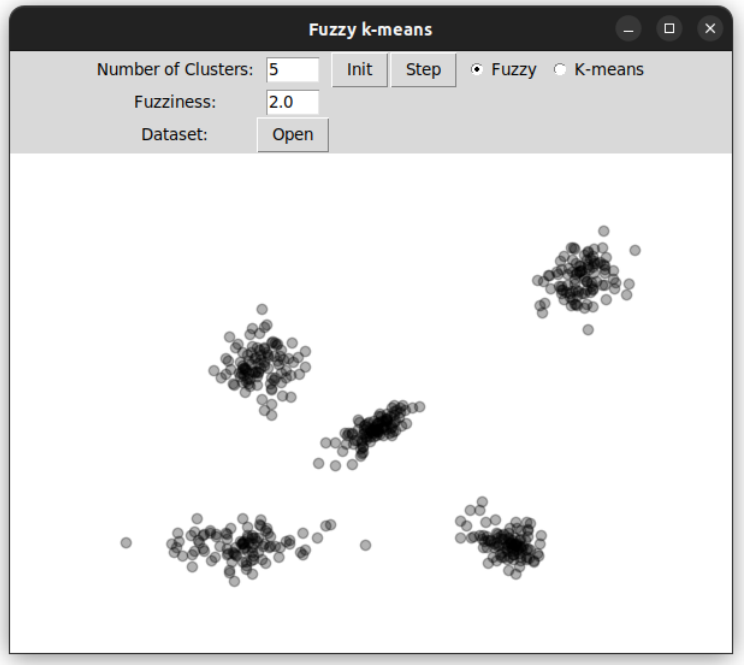
Obr. 1: Ukážka základného okna aplikácie.

Ďalej je možné v aplikácii interaktívne meniť parametre algoritmov. `Number of Clusters` je počet zhlukov do ktorých majú byť dáta kategorizované. `Fuzziness` je hyperparameter q algoritmu Fuzzy k-means, tak ako bol popísaný v sekcii 2.2. Pre potvrdenie zadanej konfigurácie je nutné stlačiť tlačidlo `Init`. Potom sa už zobrazia trénovacie dáta tak ako je možné vidieť na obrázku 2.

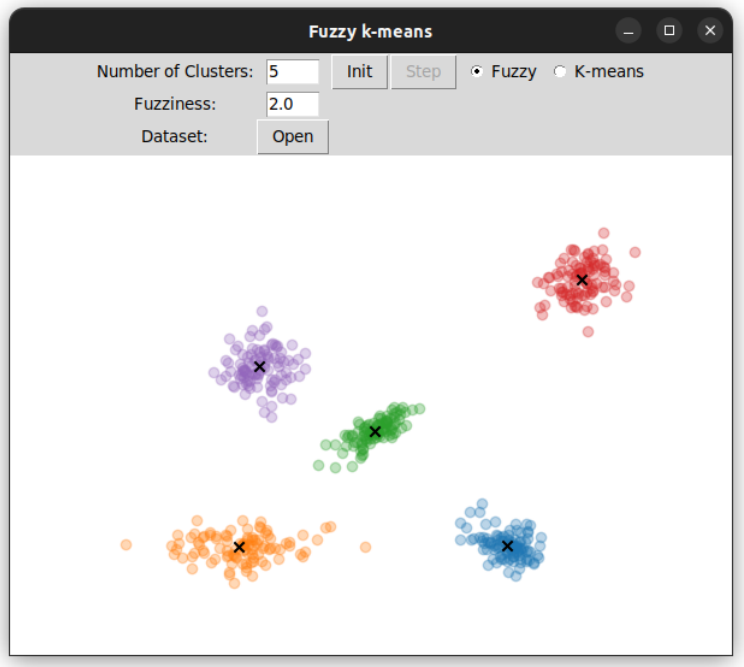
Potom je možné krokovať jednotlivá algoritmy pomocou tlačidla `Step`, ktoré vykoná ďalší krok zvoleného algoritmu. V ponuke sú dva zhlukovacie algoritmy prvý je `Fuzzy` (Fuzzy k-means) a druhý je `K-means` (K-means) medzi ktorými je možné prepínať pomocou radio tlačidla. Stlačením radio tlačidla automaticky začne algoritmus od začiatku.

Ak algoritmus skončí ako napríklad na obrázok 3, tak sa nedá pokračovať v krokovaní a tlačidlo step sa zablokuje. Do terminálu sa následne vypíšu údaje o finálnych pozíciach centroidov. Pre začatie nového behu algoritmu s rovnakými dátami sa môže použiť tlačidlo `Init`. Pre začatie nového behu s inými dátami sa stlačí tlačidlo `Open`, vyberie sa súbor z ktorého sa majú dáta nahrať a stlačí sa `Init`.

Aplikáciu je jednoducho možné spustiť pomocou skriptu `run.sh`, prípadne volaním príkazu `python3 main.py`. Pri druhej variante je však potrebné mať naištalované knižnice `numpy`, `matplotlib`, `pandas` a `tkinter`.



Obr. 2: Ukážka trénovacích dát po stlačení tlačidla `Init`.



Obr. 3: Ukážka ukončeného algoritmu Fuzzy k-means.

5 Záver

Úlohou tejto aplikácie je demonštrovať fungovanie dvoch zhlukovacích algoritmov, ktoré boli bližšie opísané v sekcii 2. Aplikácia dokáže obidva algoritmy postupne krokovať tak, aby bolo lepšie viditeľné ich správanie. Taktiež je možné interaktívne zadávať parametre a meniť ich. Výstupom sú klasifikované body do jednotlivých zhlukov.

Literatúra

[1] Fänti, P. a Sieranoja, S. *K-means properties on six clustering benchmark datasets.* 2018. Dostupné z: http://cs.uef.fi/sipu/datasets/.

[2] Janoušek, V. *Fuzzy množiny, fuzzy logika, vybrané aplikace* [SFC lecture]. 2024. Dostupné z: https://www.fit.vut.cz/study/course/SFC/private/lectures2023/2023-sfc-fuzzy.pdf. [cit. 2024-11-19]. Fakulta informačních technologií VUT v Brně.

[3] Kočí, R. *Strojové učení* [IZU lecture]. 2023. [cit. 2024-11-19]. Fakulta informačních technologií VUT v Brně.

[4] Yong, P.; Keding, C.; Feiping, N.; Bao Liang, L. a Wanzeng, K. Two-Dimensional Embedded Fuzzy Data Clustering. *IEEE Transactions on Emerging Topics in Computational Intelligence.* 1. vyd., 2023, zv. 7, č. 4, s. 1263–1275. Dostupné z: https://ieeexplore.ieee.org/document/9956777.