Fuzzy кластеровање

Маја Црномарковић 21/2017
Марко Бабић 77/2017

25. јун 2021.

Семинарски рад у оквиру курса
Рачунарска интелигенција

Садржај

1 Увод

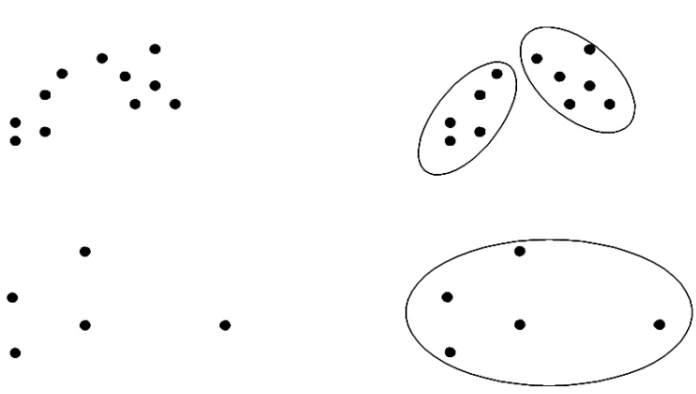
Сегментација слика је један од најпространијих начина за коректно класификовање пиксела у апликацијама заснованим на одлучивању. Сегментација је техника која партиционише слику на униформне и непреклапајуће делове, засновано на некој мери сличности. Ова техника има велики број примена у анализи слика, медицинском процесирању слика, географском информационом систему, и др. Последњих година, исказано је велико интересовање за анализу слика, и с временом се ова област све више развија. Наш рад се бави проблемом сегментације слика коришћењем технике fuzzy кластеровања. Највише пажње ће бити посвећено проблему детекције тумора на мозгу сегментовањем рендгенских снимака.

1.1 Дефиниција кластеровања

Не постоји формална дефиниција кластеровања. **Кластер анализа** је проналажење група објеката таквих да су објекти у једној групи међусобно сличнији у односу на објекте у различитим групама. **Кластеровање** се односи на поступак издвајања кластера. Проблем кластеровања се може дефинисати на следећи начин: Дат је жељени број кластера **K**, скуп података од **N** тачака и функција за мерење растојања. Потребно је пронаћи партиције скупа података тако да се минимизује вредност функције за мерење.

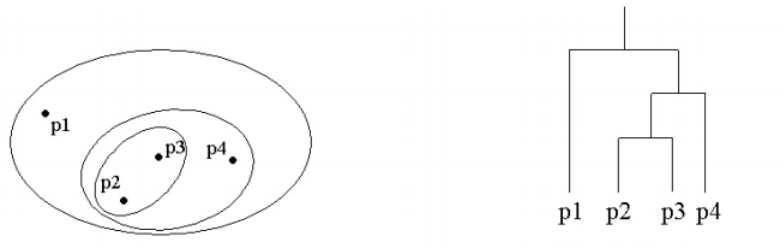
1.2 Врсте кластеровања

- **Партиционо кластеровање**: Подела скупа података у непреклапајуће подскупове (кластере) такве да је сваки податак тачно у једном подскупу.



Слика 1: Лево - почетни подаци. Десно - Партиционо кластеровање.

- **Хијерархијско кластеровање**: Скуп угњеждених кластера организован у облику хијерархијског стабла.



Слика 2: Лево - хијерархијско кластеровање. Десно - денгограм.

1.3 Различити типови кластеровања

- Ексклузивно/неексклузивно кластеровање
- Fuzzy (расплинуто/нерасплинуто) кластеровање
- Делимично/комплентно кластеровање
- Хетерогено/хомогено кластеровање

1.4 Fuzzy кластеровање

Код класичног кластеровања, подаци су подељени у одређен број дисјунктних кластера, где један елемент може припадати само једном кластеру. У fuzzy кластеровању тачка може припадати већем броју кластера са неком тежином између 0 и 1. Збир свих тежина је једнак 1. Сличне карактеристике има вероватносно кластеровање. Fuzzy кластеровање се још назива и расплинуто кластеровање.

2 Сегментација слика

Постоје многобројне методе и разноврсна литература за издвајање информација са слике и њену поделу на различите регионе. Свака од тих метода сусреће се са одређеним ограничењима која се огледају у временској сложености или тачности. Разлог за то је што не постоје јасне границе између објеката на слици. Fuzzy кластеровање показало се као веома добар начин за превазилажење овог проблема.

Сегментација слика коришћењем fuzzy кластеровања била је предмет великог интересовања кроз године. Неки од алгоритама који се баве овом темом су Fuzzy C-Means (FCM), Gustafson-Kesse (GK), Gaussian Mixture Decomposition (GMD), Fuzzy C-Varieties (FCV), Adaptive Fuzzy-C varieties

(AFC), Fuzzy C-Shell (FCS), Fuzzy C-Spherical Shells (FCSS), Fuzzy C-Rings, Fuzzy C-Quadric Shells (FCQS), Fuzzy C- Rectangular Shells (FCRS) и други. Међу свим горе наведеним алгоритмима, метод FCM је најприхваћенији начин за сегментацију слика јер омогућава мањи губитак информација у односу на алгоритме класичног кластеровања.

2.1 Fuzzy c-means

Fuzzy c-means је први пут представио Dunn а потом га је модификовао Bezdek.

Следи детаљан опис алгоритма:

- Улазни параметри: Подаци које желимо да кластерујемо (скуп елемената x дименизје n) и број који означава колико кластера желимо да добијемо (k).
- Излазни парамтери: Матрица припадности кластерима (димензије $n \times k$) и матрица центроида кластера(димензије $k \times d$ где је d димензија сваког појединачног елемента из скупа x).

Кораци алгоритма:

1. Насумично додељујемо вредности за све тежине у ознаци: w_{ij}, $1 \leq i \leq n, \quad 1 \leq i \leq n$ уз услове:
   (а) $\sum_{j=1}^{k} w_{ij} = 1, \quad \forall i \in 1, 2, ..., n$
   (б) $0 < \sum_{i=1}^{n} w_{ij} < n, \quad \forall j \in 1, 2, ..., k$
2. Рачунање центроида за све кластере у ознаци c_j помоћу формуле:
$$c_j = \frac{\sum_{i=1}^{n} w_{ij}^p x_i}{\sum_{i=1}^{n} w_{ij}^p} \tag{1}$$
   Напомена: ако је $p = 0$, имаћемо понашање класичног k-means алгоритма.
3. Ажурирање вредности матрице припадности користећи формулу:
$$w_{ij} = \frac{\frac{1}{dist(x_i, c_j)}^{\frac{1}{p-1}}}{\sum_{q=1}^{k} \frac{1}{dist(x_i, c_q)}^{\frac{1}{p-1}}} \tag{2}$$
4. Понављати кораке 2. и 3. док центроиди не остану исти у две итерације за редом.

Мера квалитета кластеровања је сума квадратне грешке(енг. Sum of Squared Error):

$$SSE = \sum_{j=1}^{k} \sum_{i=1}^{n} w_{ij}^p dist(x_i, c_j), \quad p \in 1, ..., \infty \tag{3}$$

2.1.1 Наша имплементација алгоритма fuzzy c-means

Решење проблема сегментације слика уз помоћ алгоритма fuzzy c-means смо имплементирали у програмском језику Python. Алгоритам је имплементиран у функцији fuzzy_c_means. Она као аргументе прима:

- *data* - низ података(то је уствари матрица која је димензије $n \times d$);
- n - цео број који означава број података које желимо да кластерујемо;
- k - цео број који означава број кластера;
- d - цео број који означава димензију појединачног податка из скупа података које желимо да кластерујемо;
- p - цео број који означава парамеар за fuzzy формулу којом одредјујемо степен припадности неког податка за сваки од кластера;
- *max_iter* - цео број који означава максимални број итерација због безбедности,

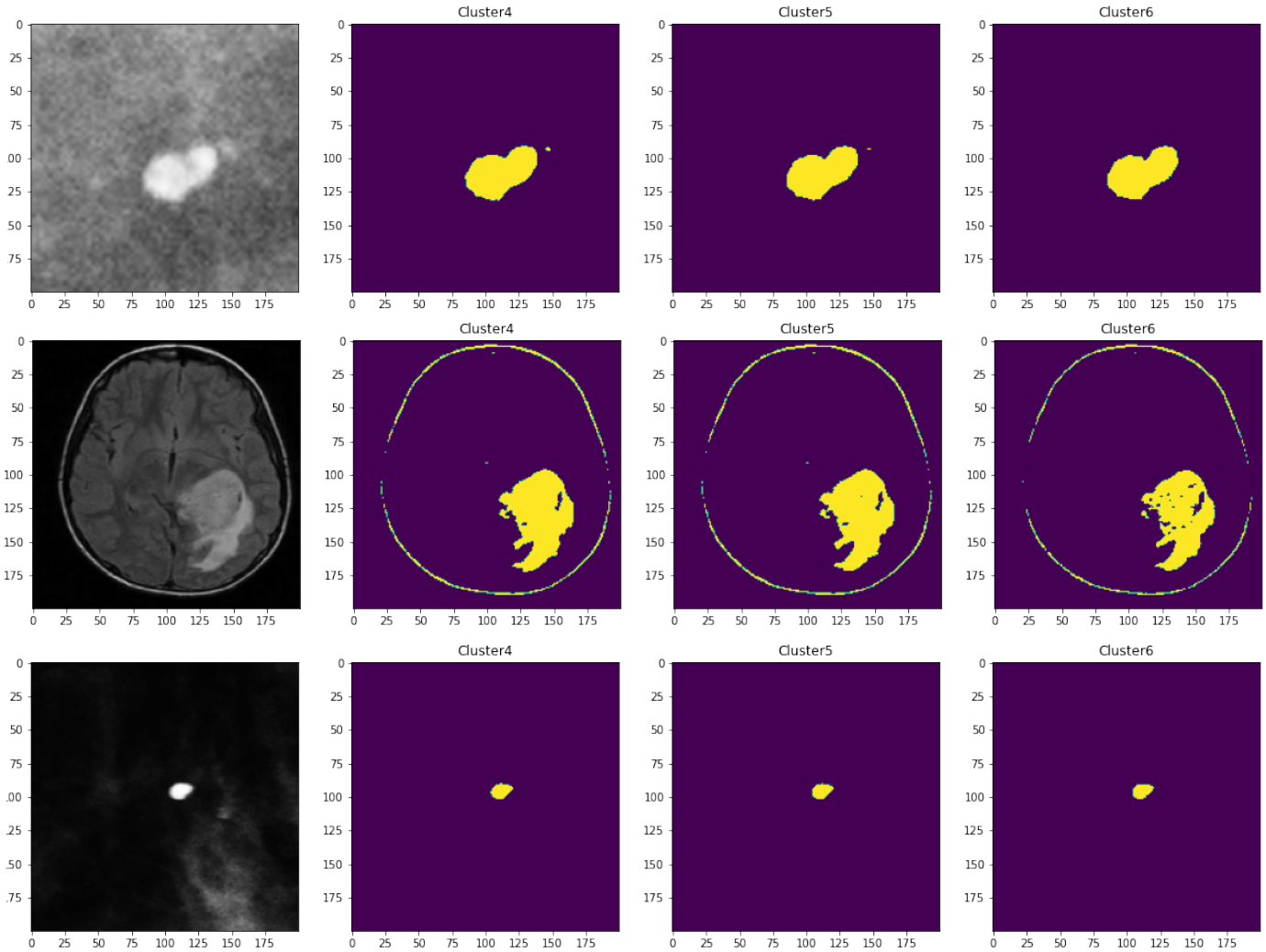
док као повратну вредност враћа:

- матрицу $n \times k$ у којој ће елемент на позицији бити w_{ij}, тј. тежина са којом i-ти елемент припада j-том кластеру;
- матрицу $k \times d$ у којој ће се чувати центроиди свих кластера.

Дакле, како је једна слика представљена као матрица пиксела где је сваки пиксел димензије 3, њу парсирамо у низ како бисмо је могли, као аргумент, проследити нашој функцији. Након што наша функција као повратну вредност врати матрицу припадности података(у нашем случају пиксела димензије 3) кластерима сваком од пиксела додељујемо вредност центроида кластера за који је вредност највећа у матрици припадности. Затим низ пиксела враћамо у матрицу полазних димензија и посматрамо је као слику. У нашем случају врло је битно да се јасно сегментује одређени детаљ на слици и због тога бинаризујемо слику, односно представљамо је са само две боје.
Python библиотеке коришћене у нашем решењу су:

- numpy
- matplotlib.pyplot
- os
- cv2
- time
- math

2.1.2 Резултати нашег алгоритма fuzzy c-means на неким ренд-генским снимцима тумора



2.2 K-means

K-меанс алгоритам је итеративни алгоритам који покушава поделити скуп података у K засебних подгрупа(кластера).

Следи детаљан опис алгоритма:

- Улазни параметри: Подаци које желимо да кластерујемо (скуп елемената x дименизје n) и број који означава колико кластера желимо да

добијемо (k).

- Излазни парамтери: Матрица припадности кластерима (димензије $n \times 2$) и матрица центроида кластера(димензије $k \times d$ где је d димензија сваког појединачног елемента из скупа x).

  Кораци алгоритма:

  1. Наводимо жељени број кластера.
  2. Насумично додељујемо вредности за центроиде свих кластера у ознаци c_j.
  3. Рачунање центроида за све кластере као аритметичке средине свих елемената у кластеру.
  4. За сваки податак ажурирамо кластер ком он припада тако да припада кластеру чији центроид му је најближи.
  5. Понављамо кораке 2. и 3. док центроиди не остану исти у две итерације за редом.

  Мера квалитета кластеровања је сума квадратне грешке(енг. Sum of Squared Error).

2.2.1 Наша имплементација алгоритма k-means

Решење проблема сегментације слика уз помоћ алгоритма k-means смо имплементирали у програмском језику Python. Алгоритам је имплементиран у функцији k_means. Она као аргументе прима:

- $data$ - низ података(то је уствари матрица која је димензије $n \times d$);
- n - цео број који означава број података које желимо да кластерујемо;
- k - цео број који означава број кластера;
- d - цео број који означава димензију појединачног податка из скупа података које желимо да кластерујемо;
- max_iter - цео број који означава максимални број итерација због безбедности,

док као повратну вредност враћа:

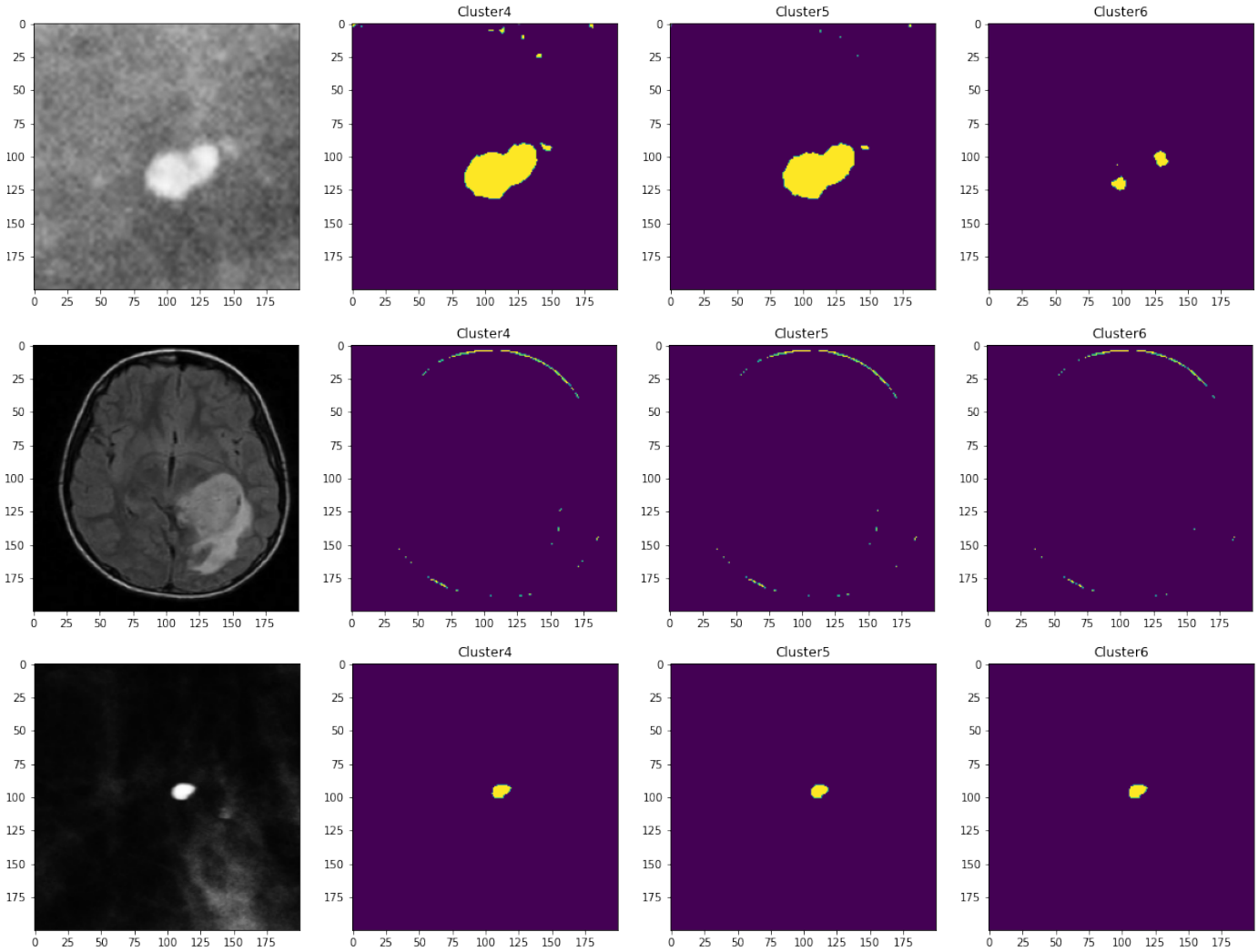
- матрицу $n \times 2$ у којој ће се уз елемент налазити ознака кластера ком припада;
- матрицу $k \times d$ у којој ће се чувати центроиди свих кластера.

Дакле, како је једна слика представљена као матрица пиксела где је сваки пиксел димензије 3, њу парсирамо у низ како бисмо је могли, као аргумент, проследити нашој функцији. Након што наша функција

као повратну вредност врати матрицу припадности података(у нашем случају пиксела димензије 3) кластерима сваком од пиксела додељујемо вредност центроида кластера ком припада. Затим низ пиксела враћамо у матрицу полазних димензија и посматрамо је као слику. У нашем случају врло је битно да се јасно сегментује одређени детаљ на слици и због тога бинаризујемо слику, односно представљамо је са само две боје.
Python библиотеке коришћене у нашем решењу су:

- numpy
- matplotlib.pyplot
- os
- cv2
- time
- math

2.2.2 Резултати нашег алгоритма k-means на неким рендгенским снимцима тумора



3 Закључак

Кластеровање је један од најкоришћенијих начина за сегментацију слика. Огромна предност fuzzy кластеровања у односу на остале технике сегментације слика је то што је врло отпоран на неправилне границе. Током рада користили смо више различитих слика које смо сегментовали уз испробавање различитих вредности одређених параметара и различитог броја кластера и у већини случајева, код fuzzy c-means алгоритма, најбоље се показало коришћење 5 кластера уз параметар fuzzy формуле (p) који је једнак 2, док се код k-means алгоритма не може једноставно утврдити за који број кластера даје најбоље резултате с обзиром на то да смо у неким ситуацијама при раду са експерименталним подацима добијали најбољу сегментацију за један број кластера, а онда када бисмо опет покренули алгоритам најбоља сегментација би се добила за други број кластера. Сегментација слика уз помоћ алгоритама fuzzy кластеровања је своју примену између осталог нашла у медицини где је веома битна прецизност резултата. Наравно, fuzzy кластеровање има широк спектар примене поред сегментације слика у разним другим областима неке од њих су економија, вештачка интелигенција и друге.

3.1 Поређење резултата

Поређењем резултата које враћају алгоритми fuzzy c-means и k-means можемо видети да су слике приметно боље сегментоване алгоритмом fuzzy c-means. Његова сложеност је већа од сложености алгоритма k-means па сам процес сегментације траје нешто дуже. Такође предност алгоритма fuzzy c-means је то што се у пракси показало да даје скоро потпуно исте резултате за исти број кластера на истим сликама, док код алгоритма k-means то није случај.

4 Литература

- M.S. YANG - A Survey of Fuzzy Clustering, October 1993.
- XL Xie, G Beni - A validity measure for fuzzy clustering, 1991.
- Donald E. Gustafson, William C. Kessel - Fuzzy clustering with a fuzzy covariance matrix, 1979.
- Imad Dabbura - K-means Clustering: Algorithm, Applications, Evaluation Methods, and Drawbacks, September 2018
- Мирјана Маљковић - Скалабилни кластер алгоритми, 2008.
- Ненад Митић - Предавање о кластер анализи на Математичком факултету, 2020.