ELE075 - Trabalho Computacional I: Fuzzy C-Means

Rúbia Reis Guerra
2013031143

07/10/2018

1 Introdução

A análise de agrupamentos consiste em um conjunto de ferramentas de análise de dados exploratória e visa ordenar diferentes objetos em grupos, de forma que o grau de associação entre dois objetos seja máximo, se eles pertencerem ao mesmo grupo, e mínimo, caso contrário. Assim, a análise de agrupamentos pode ser usada para descobrir estruturas e relacionamentos em conjuntos dados, baseando-se nos valores coletados de cada atributo de uma observação. Neste trabalho, foi estudada a implementação do Fuzzy C-Means: uma forma de agrupamento em que cada ponto do conjunto de dados pode pertencer a mais de um agrupamento.

2 Solução

Na primeira parte do estudo, foi realizada a comparação do Fuzzy C-Means (FCM) com o algoritmo K-Means. As funções de custo para cada abordagem foi descrita abaixo:

- Função de custo para o FCM:

$$J_{FCM} = \sum_{i=1}^{n}\sum_{j=1}^{k} \omega_{ij}^{m} \|\mathbf{x}_i - \mathbf{c}_j\|^2$$

- Função de custo (J) para o K-Means, dado Si = nº de observações no cluster j:

$$J_{KM} = \sum_{i=1}^{n}\sum_{j=1}^{2} \frac{1}{S_j} \|\mathbf{x}_i - \mathbf{c}_j\|^2$$

As funções de custo foram utilizadas, na implementação da solução, para a atualização da melhor configuração de centroides encontrada (menor custo).

3 Experimentos

Para a etapa de validação do FCM, foram realizados experimentos executando os algoritmos *Fuzzy C-Means* e *k-Means* 150 vezes, inicializando-se os centroides aleatoriamente e mantendo os mesmos valores iniciais para as duas abordagens.

Para a determinação de centroides com valores considerados adequados, executou-se os algoritmos k-Means e FCM com os critérios de parada descritos abaixo, armazenando, em seguida, os valores mínimos encontrados para as respectivas funções de custo. Foi contado, então, o número de vezes que o algoritmo atinge soluções adequadas até a iteração de convergência.

As implementações foram feitas na linguagem Julia (versão 0.7). A máquina utilizada para rodar os experimentos possui processador Intel Core i7 a 2,9 GHz, com 16GB de memória e sistema operacional macOS High Sierra.

3.1 Validação do FCM

Para a validação do FCM, foram definidos os seguintes critérios:

- Critério de parada **max_iter**: número de iterações máximo = 1000;
- Critério de parada **tol_iter**: para que seja considerada a convergência do algoritmo, deve-se obter um mínimo (sequencial) de iterações sem alteração dos clusters assinalados a cada observação. Para este experimento, definiu-se 10 iterações;
- Critério de aceitação **tol**: para que um conjunto de centroides seja considerado adequado, a função de custo pode variar em relação ao menor valor de custo encontrado até então em uma tolerância de, no máximo, 0.3;

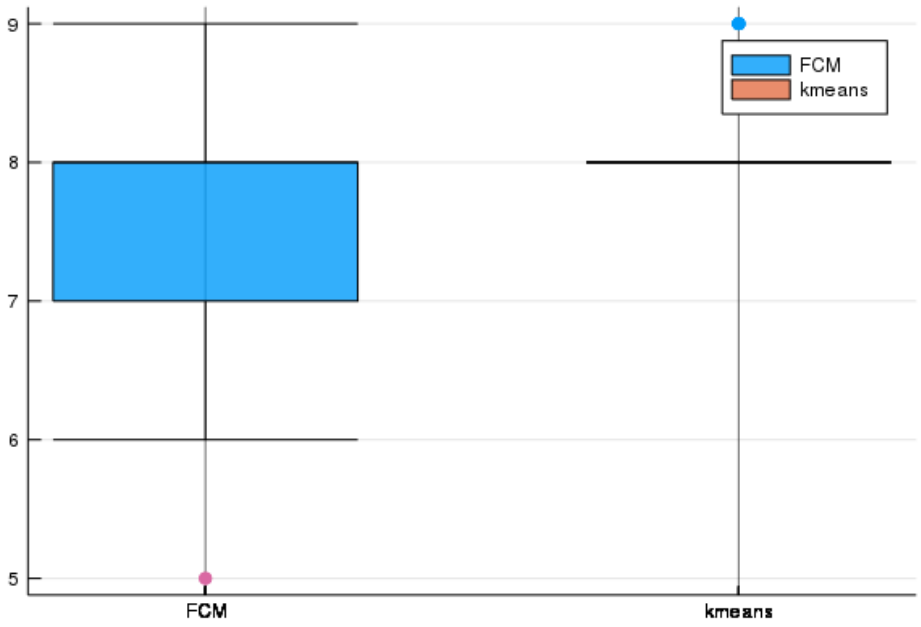


Figura 1: Número de centroides aceitáveis em 10 iterações

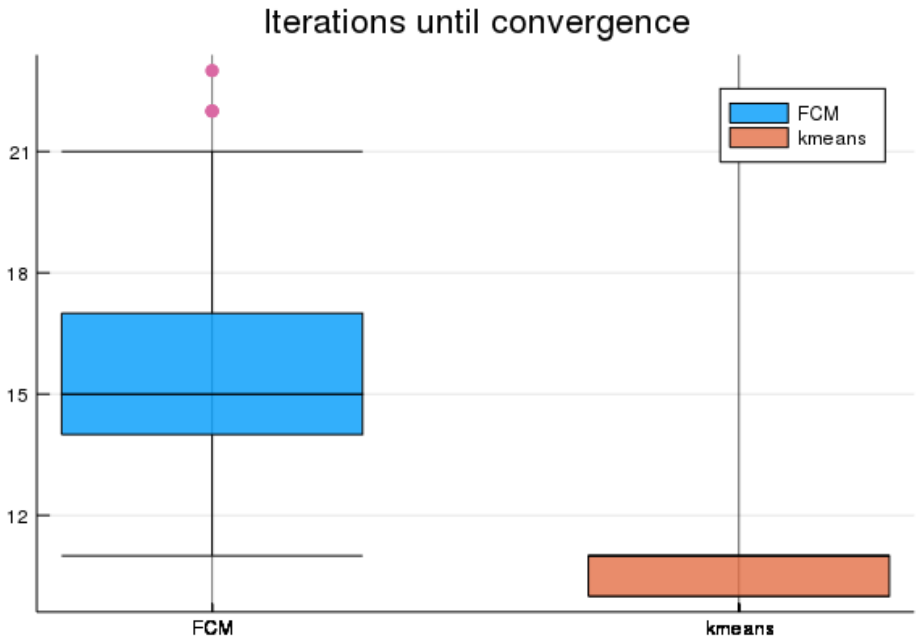


Figura 2: Número de iterações até a convergência

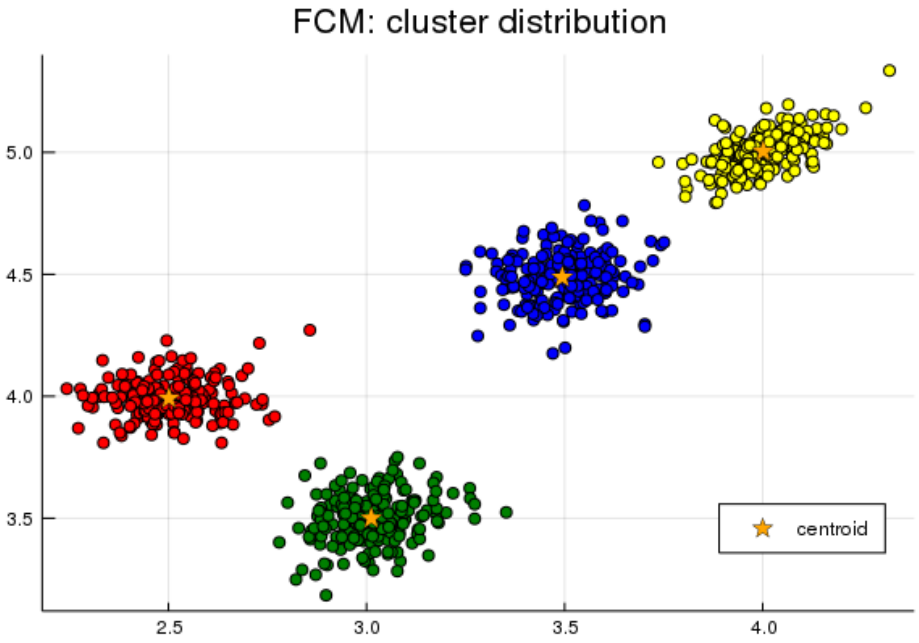


Figura 3: Agrupamentos resultantes na aplicação de *Fuzzy C-Means* nos dados sintéticos

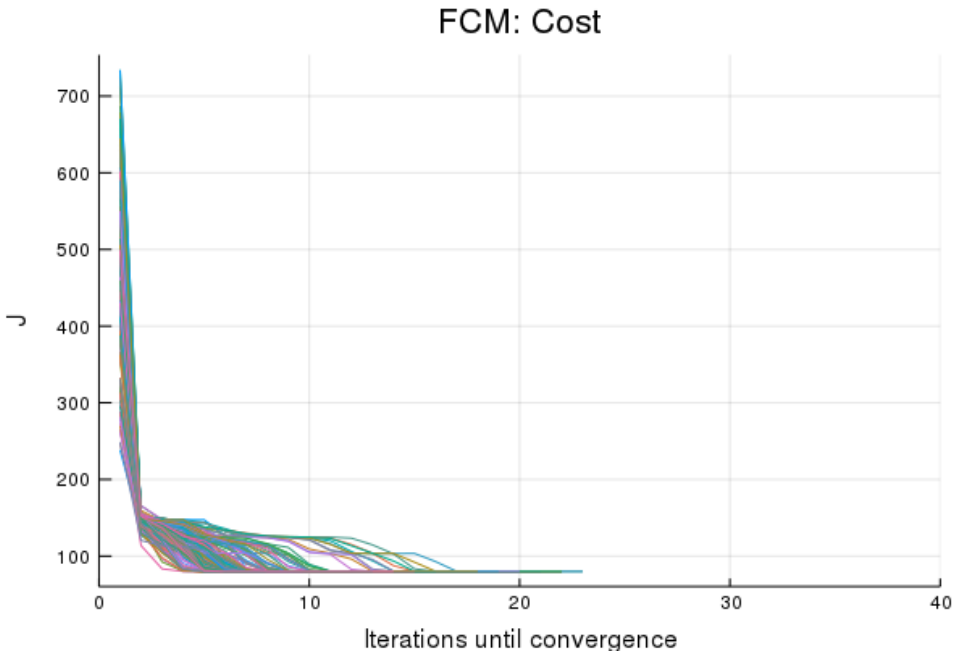


Figura 4: FCM: Função de custo x Número de iterações para convergir

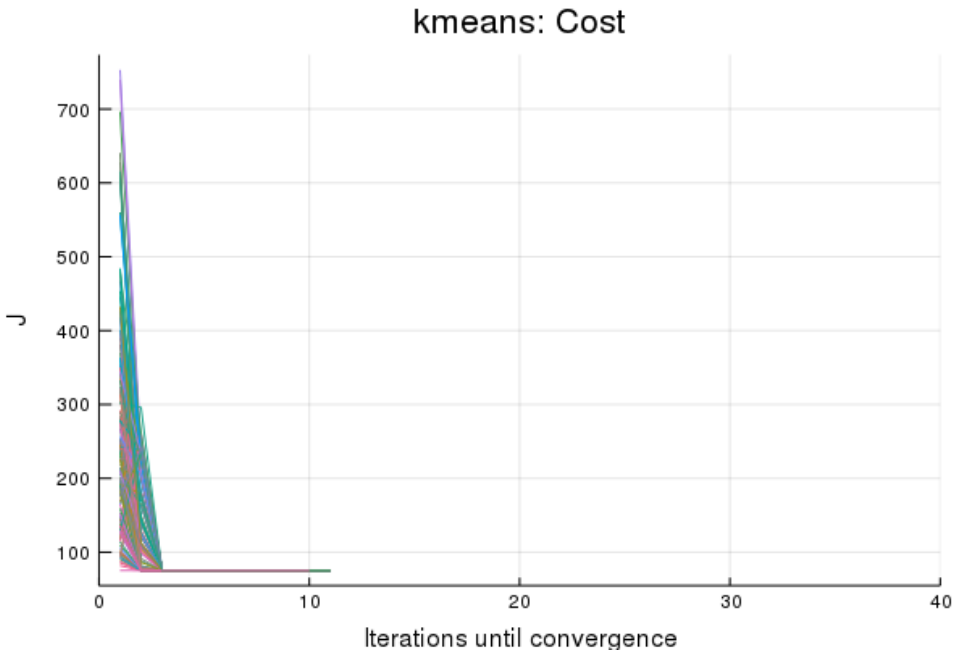


Figura 5: kmeans: Função de custo x Número de iterações para convergir

A partir dos experimentos, observou-se que ambas as implementações de *Fuzzy C-Means* e de *K-Means* encontraram um mínimo ótimo dentro do limite de iterações definido. A abordagem *fuzzy*, porém, apresentou uma convergência mais lenta, possivelmente pela maior aleatoriedade agregada no cálculo da matriz de pertinência. Observa-se, também, uma maior suavização do comportamento da função de custo de FCM, em contraste com K-Means.

3.2 Segmentação de Imagens por Região

Para a tarefa de segmentação de imagem por região, definiu-se, empiricamente, a seguinte distribuição de agrupamentos:

Arquivos	Nº de Agrupamentos
1, 2, 3, 5	8
4, 6, 7, 8	16
9, 10, 11	20

Tabela 1: Número de de agrupamentos para cada imagem

OS critérios considerados para este experimento foram:

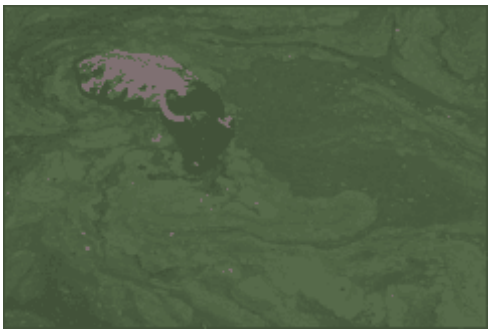
- Critério de parada **max_iter**: número de iterações máximo = 1000;
- Critério de parada **tol_iter**: para que seja considerada a convergência do algoritmo, deve-se obter um mínimo (sequencial) de iterações sem alteração dos clusters assinalados a cada observação. Para este experimento, definiu-se 10 iterações;

Para a atualização da matriz de pertinência e obtenção dos agrupamentos, considerou-se os valores da iteração que alcançou a menor função de custo. Os resultados podem ser observados nas figuras 6 a 8.

(a) Original 1

(b) Resultado 1, k = 8

(c) Original 2

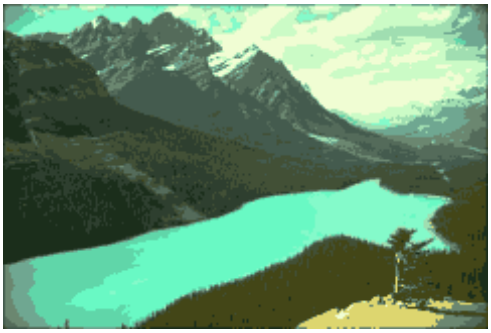


(d) Resultado 2, k = 8

(e) Original 3

(f) Resultado 3, k = 8

(g) Original 4



(h) Resultado 4, k = 16

Figura 6: Resultados do Fuzzy C-Means para segmentação de imagem

(a) Original 5

(b) Resultado 5, k = 8

(c) Original 6

(d) Resultado 6, k = 16

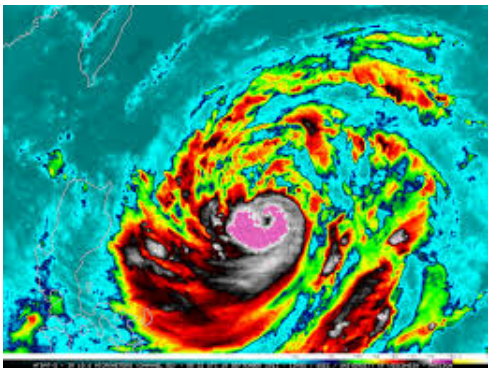
(e) Original 7

(f) Resultado 7, k = 16

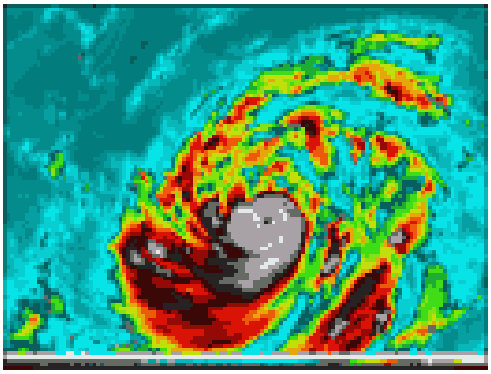
Figura 7: Resultados do Fuzzy C-Means para segmentação de imagem

(a) Original 8

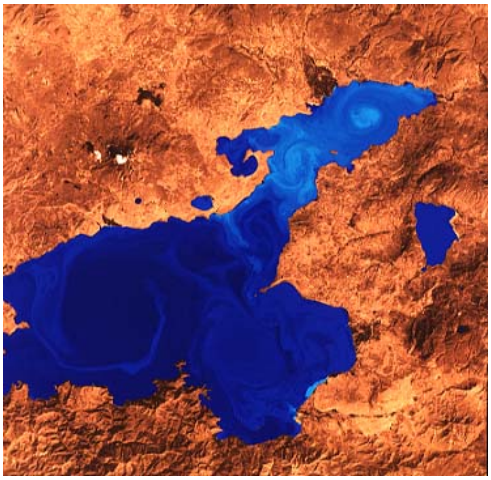
(b) Resultado 8, k = 16



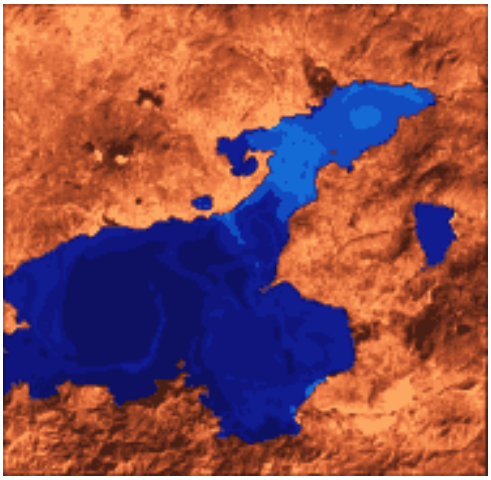
(c) Original 9



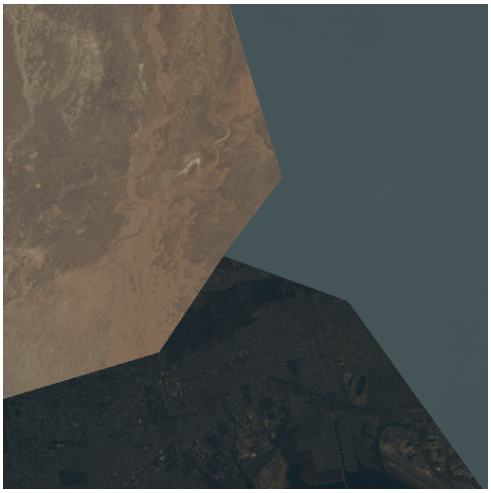
(d) Resultado 9, k = 20



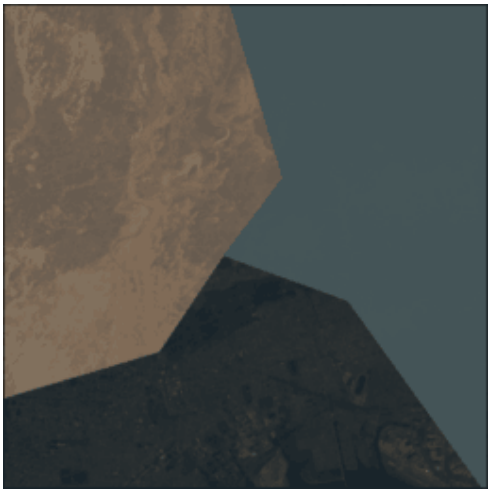
(e) Original 10



(f) Resultado 10, k = 20



(g) Original 11



(h) Resultado 11, k = 20

Figura 8: Resultados do Fuzzy C-Means para segmentação de imagem

Referências

[1] Jang, Jyh-Shing R, Chuen-Tsai Sun, and Eiji Mizutani. *Neuro-fuzzy and Soft Computing: A Computational Approach to Learning and Machine Intelligence*. Upper Saddle River, NJ: Prentice Hall, 1997.