Universidade Federal de Minas Gerais

Sistemas Nebulosos

Relatório TP-1

Vítor Gabriel Reis Caitité 2016111849

Belo Horizonte
2019

Introdução

O objetivo deste trabalho é entender melhor o funcionamento dos algoritmos de agrupamento K-Means e principalmente Fuzzy C-Means (FCM). Primeiramente é necessário entender bem o conceito de clustering.

O clustering pode ser considerado o mais importante problema de aprendizado não supervisionado. Assim, como qualquer outro problema desse tipo, trata-se de encontrar uma estrutura em uma coleção de dados não rotulados. Uma definição simples de agrupamento poderia ser “o processo de organizar objetos em grupos cujos membros são similares de alguma forma” (como pode ser visto na fig. 1 abaixo).

Um cluster é, portanto, uma coleção de objetos que são “semelhantes” entre eles e são “diferentes” para os objetos pertencentes a outros clusters.

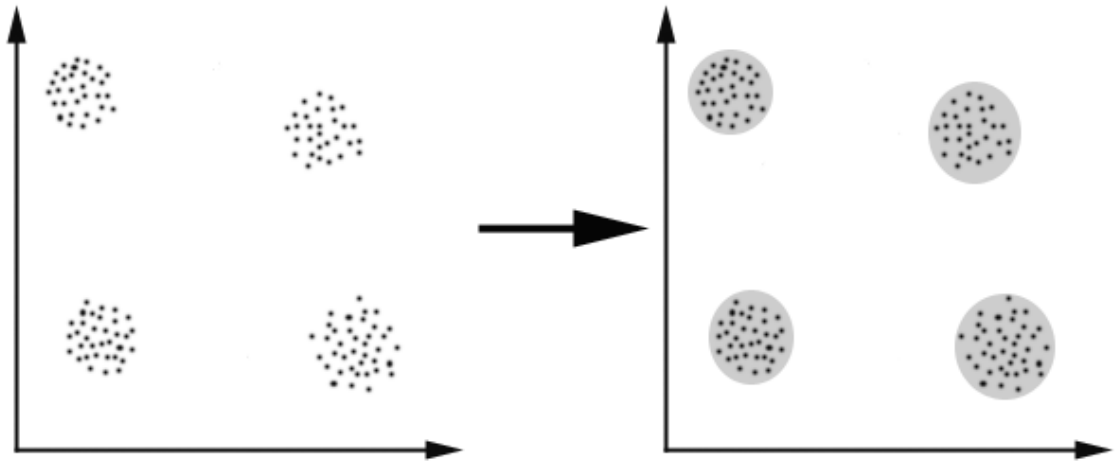


Fig. 1 - Ilustração do resultado do processo de clustering.

Na figura acima, identificamos facilmente os 4 clusters nos quais os dados podem ser divididos; o critério de similaridade é a distância: dois ou mais objetos pertencem ao mesmo cluster se estiverem “próximos” de acordo com uma determinada distância (neste caso, a distância geométrica). Isso é chamado de cluster baseado em distância.

Outro tipo de agrupamento é o agrupamento conceitual: dois ou mais objetos pertencem ao mesmo agrupamento se este definir um conceito comum a todos os objetos. Em outras palavras, os objetos são agrupados de acordo com sua adequação a conceitos descritivos, não de acordo com medidas simples de similaridade.

Assim, o objetivo do clustering é determinar o agrupamento intrínseco em um conjunto de dados não rotulados. Mas como decidir o que constitui um bom agrupamento? Pode ser demonstrado que não existe um critério “melhor” absoluto que

seja independente do objetivo final do agrupamento. Consequentemente, é o usuário que deve fornecer esse critério, de tal forma que o resultado do agrupamento atenda às suas necessidades.

Por exemplo, poderíamos estar interessados em encontrar representantes para grupos homogêneos (redução de dados), em encontrar “clusters naturais” e descrever suas propriedades desconhecidas (tipos de dados “naturais”), em encontrar agrupamentos úteis e adequados (classes de dados “úteis”) ou na localização de objetos de dados incomuns (detecção de outliers).

Algoritmo K-Means

K-Means é um dos mais simples algoritmos de aprendizado não supervisionados que resolvem o problema de clustering. O procedimento segue uma maneira simples e fácil de classificar um determinado conjunto de dados por meio de um determinado número de clusters (suponha k clusters) fixados a priori. A ideia principal é definir k centroides, um para cada cluster. Esses centroides devem ser colocados de uma maneira esperta, porque diferentes localizações causam diferentes resultados. Então, a melhor escolha é colocá-los o máximo possível longe um do outro. O próximo passo é pegar cada ponto pertencente a um determinado conjunto de dados e associá-lo ao centroide mais próximo. Quando nenhum ponto está pendente, o primeiro passo é concluído e um agrupamento inicial é feito. Neste ponto, precisamos recalcular k novos centroides como baricentros dos clusters resultantes da etapa anterior.

Depois que tivermos esses k novos centroides, uma nova ligação deve ser feita entre os mesmos pontos de conjunto de dados e o novo centroide mais próximo. Um loop foi gerado. Como resultado deste loop, podemos notar que os k centroides mudam sua localização passo a passo até que nenhuma outra mudança seja feita. Em outras palavras, os centroides não se movem mais.

Finalmente, este algoritmo visa minimizar uma *função objetivo*, neste caso, uma função de erro quadrático. A função objetivo

$$J = \sum_{j=1}^{k} \sum_{i=1}^{n} \left\| x_i^{(j)} - c_j \right\|^2 ,$$

onde $\left\|x_i^{(j)} - c_j\right\|^2$ é uma medida de distância escolhida entre um ponto de dados $x_i^{(j)}$ e o centro do cluster c_j, é um indicador da distância dos *n* pontos de dados de seus respectivos centros de cluster.

O algoritmo é composto pelas seguintes etapas:

1. *Coloque K pontos no espaço representado pelos objetos que estão sendo agrupados. Esses pontos representam os centroides iniciais do grupo.*
2. *Atribua cada objeto ao grupo que tem os centroides mais próximo.*
3. *Quando todos os objetos tiverem sido atribuídos, recalcule as posições dos centroides K.*
4. *Quando todos os objetos tiverem sido atribuídos, recalcule as posições dos centroides K.*
5. *Repita os passos 2 e 3 até que os centroides não se movam mais. Isso produz uma separação dos objetos em grupos a partir dos quais a métrica a ser minimizada pode ser calculada.*

Embora possa ser provado que o procedimento sempre terminará, o algoritmo k-Means não encontra necessariamente a configuração mais ideal, correspondendo ao mínimo da função objetivo global. O algoritmo também é significativamente sensível aos centros iniciais de clusters selecionados aleatoriamente. O algoritmo k-Means pode ser executado várias vezes para reduzir esse efeito.

O K-means é um algoritmo simples que foi adaptado para muitos domínios problemáticos. Será explicado agora uma extensão para trabalhar com domínios Fuzzy.

Algoritmo C-Means

Fuzzy C-means (FCM) é um método de clustering que permite que uma parte dos dados pertença a dois ou mais clusters. Este método é frequentemente usado no reconhecimento de padrões. Baseia-se na minimização da seguinte função objetiva:

$$J_m = \sum_{i=1}^{N}\sum_{j=1}^{C} u_{ij}^{m}\left\|x_i - c_j\right\|^2, \quad 1 \le m < \infty$$

onde *m* é qualquer número real maior que 1, u_{ij} é o grau de associação de x_i no cluster *j* , x_i é o *i* de dados medidos d - dimensionais, c_j é o centro de dimensão d do

cluster, e || * || é qualquer norma que expresse a semelhança entre quaisquer dados medidos e o centro.

O particionamento difuso é realizado por meio de uma otimização iterativa da função de objetivo mostrada acima, com a atualização da associação u_{ij} e os centros de clusters c_j por:

$$u_{ij} = \frac{1}{\sum_{k=1}^{C} \left(\frac{\|x_i - c_j\|}{\|x_i - c_k\|} \right)^{\frac{2}{m-1}}}, \qquad c_j = \frac{\sum_{i=1}^{N} u_{ij}^m \cdot x_i}{\sum_{i=1}^{N} u_{ij}^m}$$

Esta iteração irá parar quando $\max_{ij} \left\{ \left| u_{ij}^{(k+1)} - u_{ij}^{(k)} \right| \right\} < \varepsilon$, onde ε é um critério de terminação entre 0 e 1, enquanto *k* são os passos de iteração. Este procedimento converge para um mínimo local ou um ponto de sela de J_m. O algoritmo é composto pelas seguintes etapas:

1. *Inicializar* $U = [u_{ij}]$ *matriz,* $U^{(0)}$
2. *Calcule os vetores centrais* $C^{(k)} = [c_j]$ *com* $U^{(k)}$

$$c_j = \frac{\sum_{i=1}^{N} u_{ij}^m \cdot x_i}{\sum_{i=1}^{N} u_{ij}^m}$$

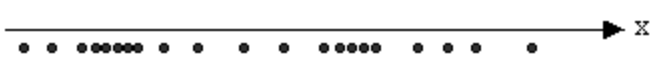
3. *Atualize* $U^{(k)}$ *,*$U^{(k+1)}$

$$u_{ij} = \frac{1}{\sum_{k=1}^{C} \left(\frac{\|x_i - c_j\|}{\|x_i - c_k\|} \right)^{\frac{2}{m-1}}}$$

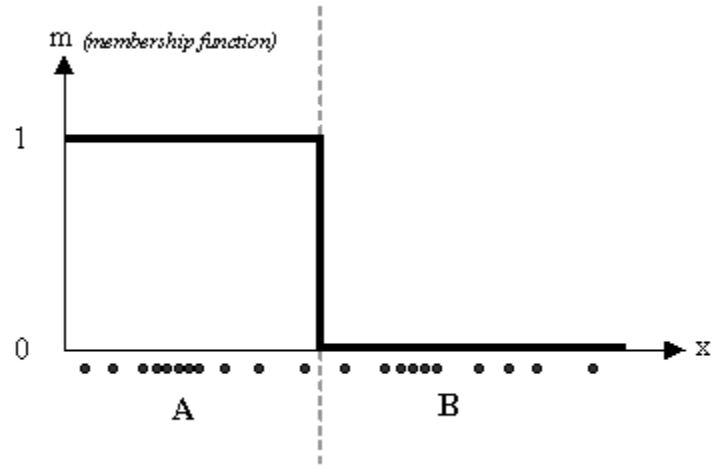
4. *Se* $\| U^{(k+1)} - U^{(k)} \| < \varepsilon$ *então STOP; caso contrário, retorne ao passo 2.*

Como já foi dito, os dados são vinculados a cada cluster por meio de uma função de associação, que representa o comportamento difuso desse algoritmo. Para fazer isso, simplesmente temos que construir uma matriz apropriada denominada U, cujos fatores são números entre 0 e 1, e representar o grau de associação entre dados e centros de clusters.

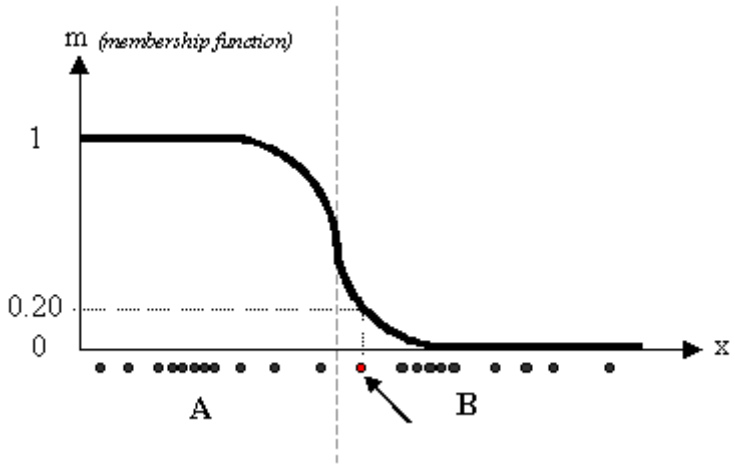
Para um melhor entendimento, podemos considerar este simples exemplo monodimensional. Dado um determinado conjunto de dados, suponha representá-lo como distribuído em um eixo. A figura abaixo mostra isso:



Olhando para a imagem, podemos identificar dois clusters nas proximidades das duas concentrações de dados. Vamos nos referir a eles usando 'A' e 'B'. Na primeira abordagem mostrada neste tutorial - o algoritmo k-means - associamos cada dado a um centroide específico; portanto, essa função de associação era assim:



Na abordagem FCM, em vez disso, o mesmo dado não pertence exclusivamente a um cluster bem definido, mas pode ser colocado de maneira intermediária. Nesse caso, a função de associação segue uma linha mais suave para indicar que cada dado pode pertencer a vários clusters com valores diferentes do coeficiente de associação.



Na figura acima, o dado mostrado como um ponto marcado vermelho pertence mais ao cluster B do que ao cluster A. O valor 0.2 de 'm' indica o grau de associação a A para tal dado. Agora, em vez de usar uma representação gráfica, introduzimos uma matriz U cujos fatores são aqueles tirados das funções de associação:

$$U_{MxC} = \begin{bmatrix} 0.8 & 0.2 \\ 0.3 & 0.7 \\ 0.6 & 0.4 \\ .. & .. \\ 0.9 & 0.1 \end{bmatrix}$$

O número de linhas e colunas depende de quantos dados e clusters estamos considerando. Mais exatamente temos C = 2 colunas (C = 2 clusters) e N linhas, onde C é o número total de clusters e N é o número total de dados. O elemento genérico é assim indicado: u_{ij} . Outras propriedades são mostradas abaixo:

$$u_{ij} \in [0,1] \qquad \forall i, j$$

$$\sum_{j=1}^{C} u_{ik} = 1 \qquad \forall i$$

$$0 < \sum_{i=1}^{N} u_{ij} < N \quad \forall N$$

Atividades

1. Fuzzy C-Means: Implemente o algoritmo de agrupamento Fuzzy C-Means (FCM). Caso seja conveniente, modifique o código do algoritmo K-Means fornecido no Moodle;

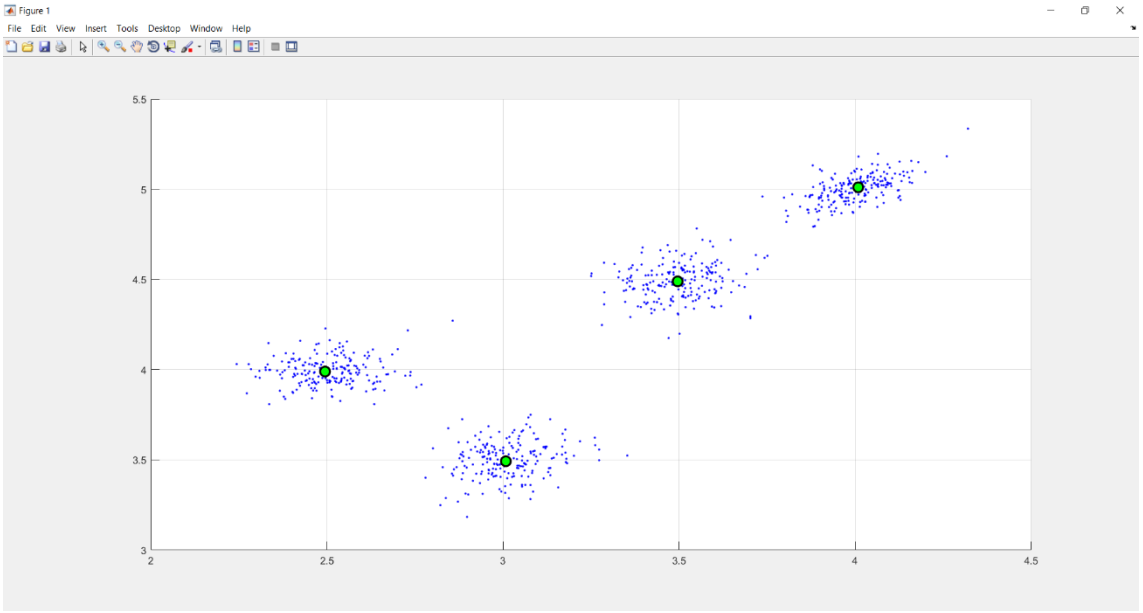
O código foi implementado de acordo com o que foi explicado acima e enviado com o nome Cmean.m.

2. Validação do FCM: Valide o algoritmo FCM com a base de dados "FCMdataset.mat". Para a validação, plote os centros dos clusters encontrados pelo algoritmo FCM sobre a base de dados fornecida. Compare os resultados obtidos pelo FCM com aqueles obtidos pelo algoritmo K-Means. A comparação deve ser em termos de: (i) número médio de iterações até a convergência e, (ii) número de vezes que o algoritmo encontra valores adequados para os centros dos clusters; Para coleta desses dados, execute os algoritmos N vezes (onde N ≥ 30) com os mesmos valores de inicialização.

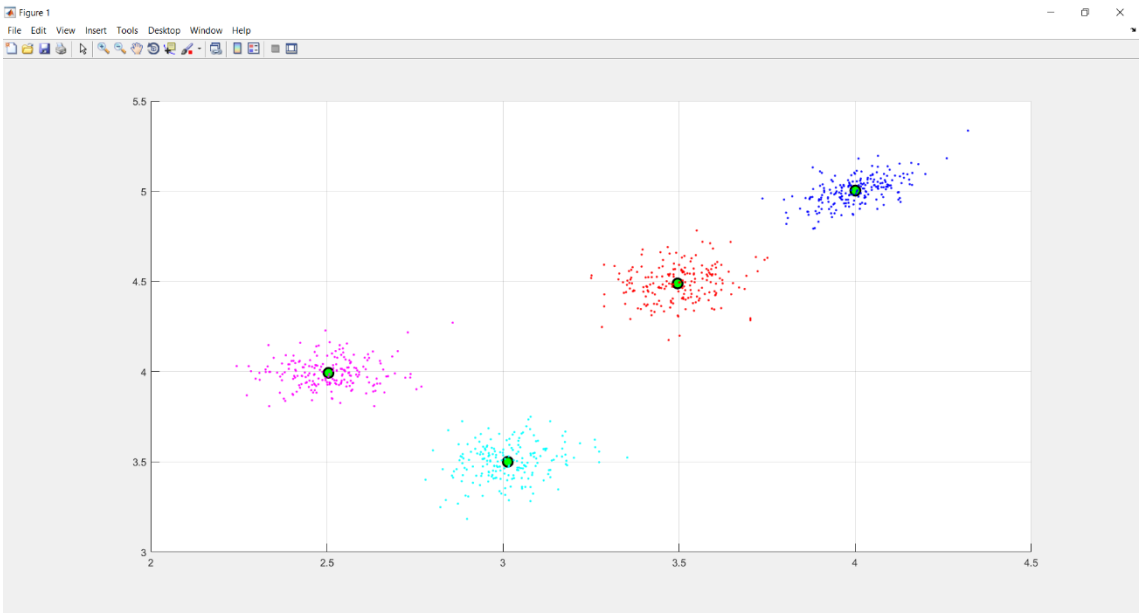
i) O número de iterações até a convergência foi
   - K-Means – 5
   - FCM - 13

ii) O número de vezes que o algoritmo encontra valores adequados para os centros dos clusters:
   - K-Means – 20
   - FCM – 30

*Pode-se observar que apesar de o K-means convergir para um resultado mais rapidamente, nem sempre ele funciona como esperado. Logo pode-se concluir que o FCM, apesar de convergir mais lentamente, é mais robusto, com menos chances de falha que o K-means. O resultado do algoritmo desenvolvido pode ser visto abaixo.



Atribuindo, por cores, os dados ao seu respectivo cluster, tem-se:



3. Segmentação de Imagens por Região: Use o algoritmo FCM para segmentar por região as imagens RGB fornecidas no diretório ImagensTeste do Moodle. Para cada imagem, escolha o número de clusters de forma empírica, com base na observação das cores das diferentes regiões. Após obter a matriz de partição U, resultado da aplicação do FCM em cada imagem, use esta matriz para colorir cada região (cluster) com a tonalidade do pixel que corresponde ao centro da região. Os pixels que apresentarem maior grau de compatibilidade (pertinência) a uma dada região devem ser coloridos com a tonalidade do pixel central daquela região.

   Arquivo CmeansImages.m

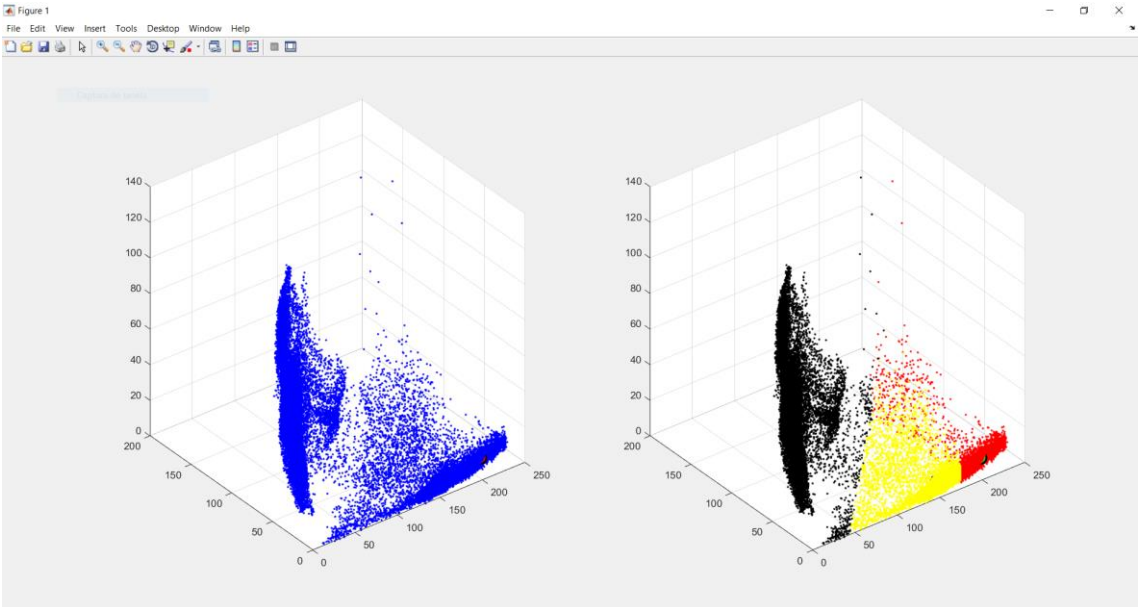
   * No algoritmo deve-se entrar com a imagem escolhida de 1 a 11, e o número de clusters.

   O resultado pode ser visto abaixo:

   - Photo001

     3 clusters

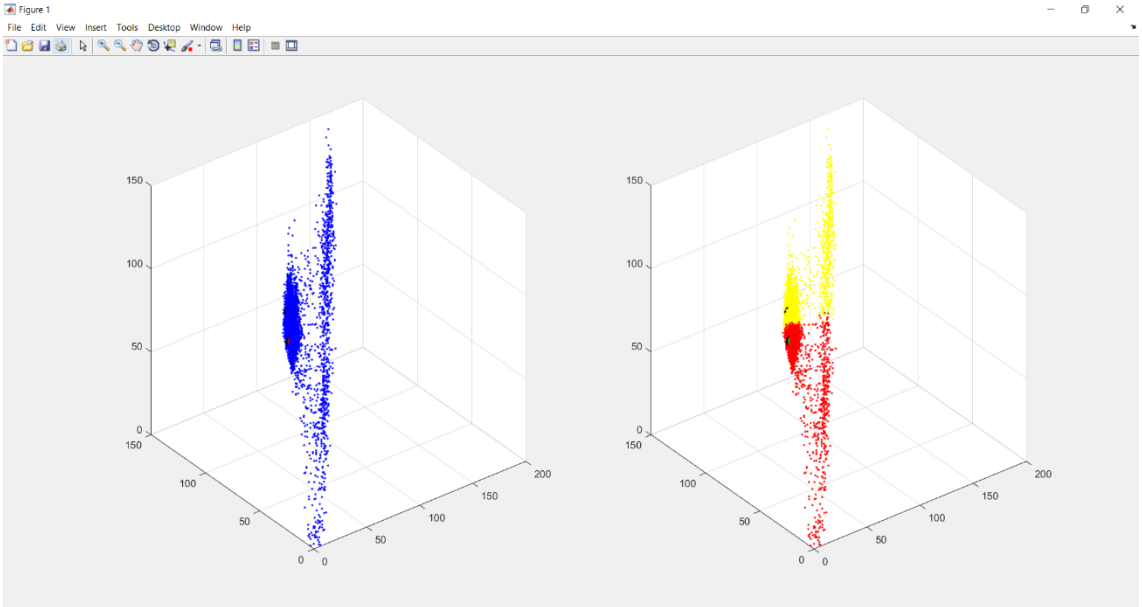
     39 iterações



- Photo002

  4 clusters

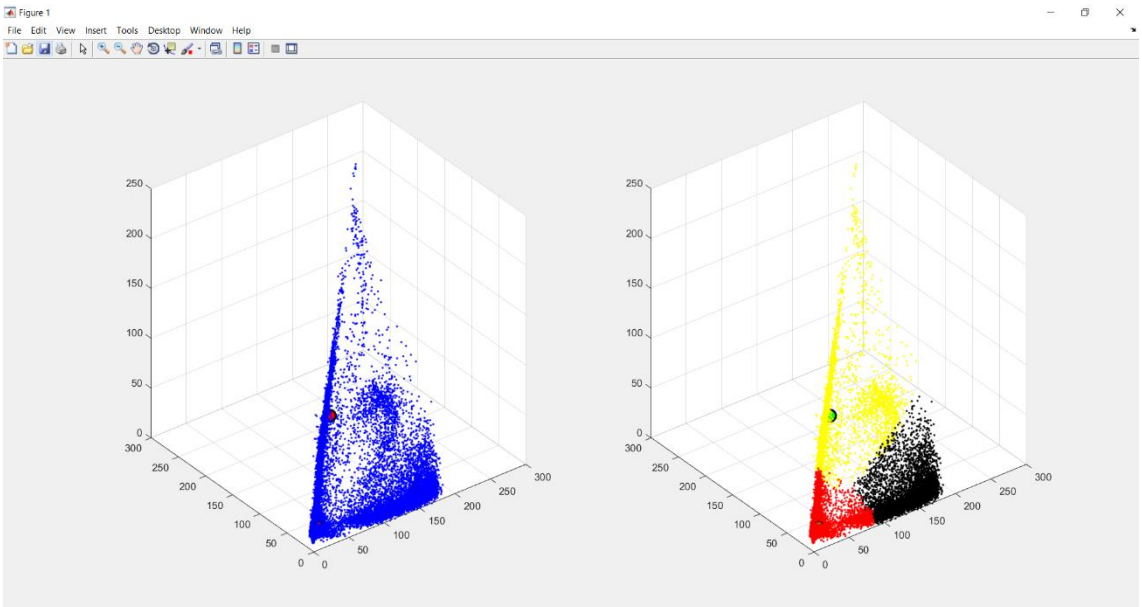
  23 iterações



- Photo003

  3 clusters

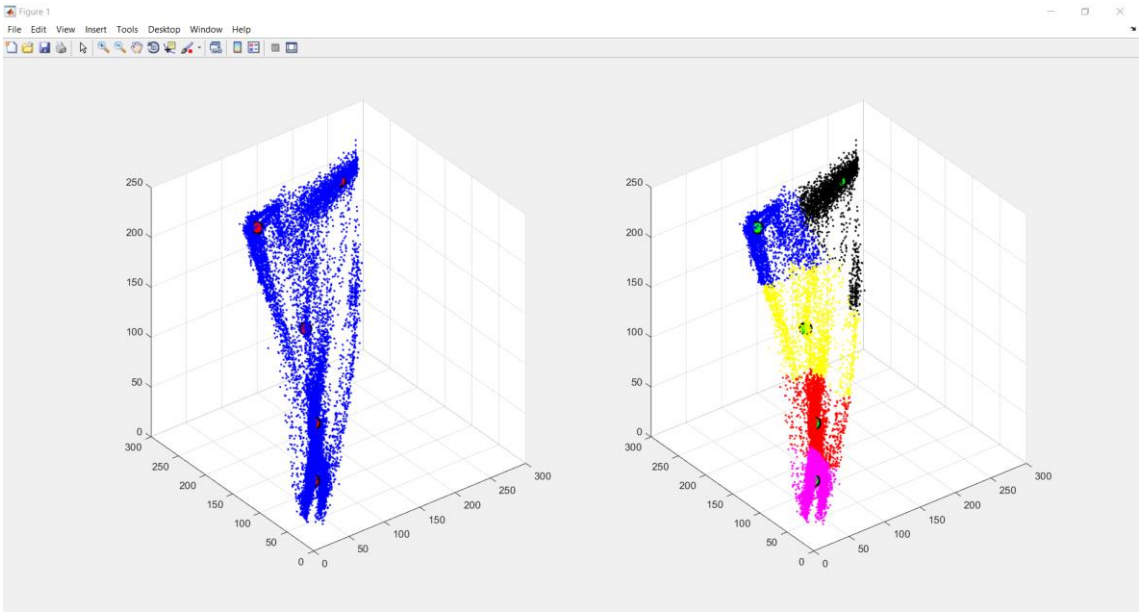
  47 iterações



- Photo004

  5 clusters

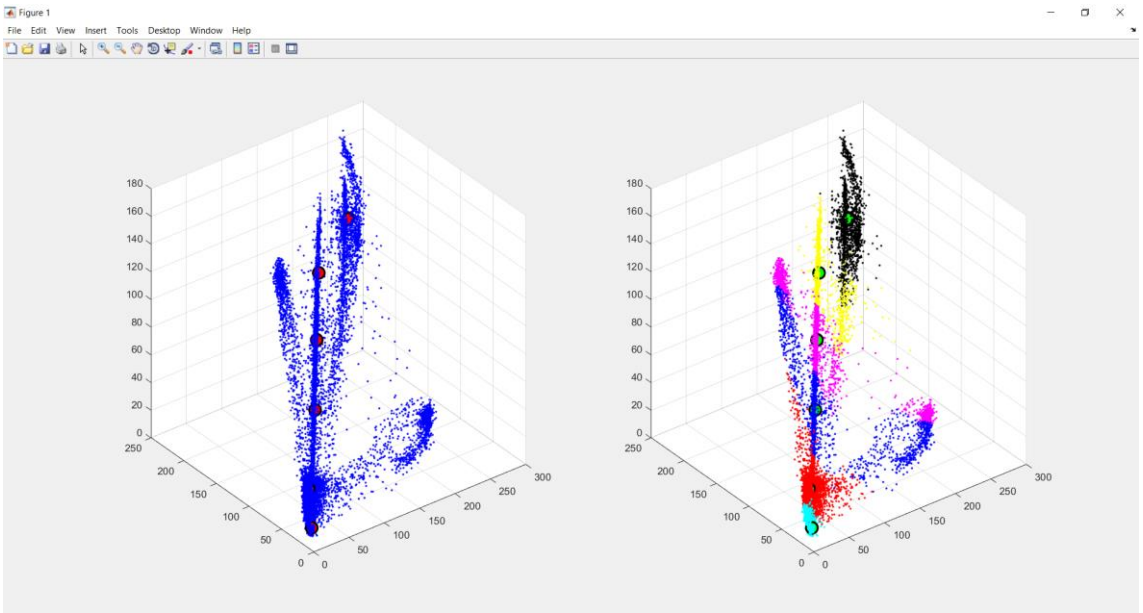
137 iterações



- <u>Photo005</u>

  6 clusters

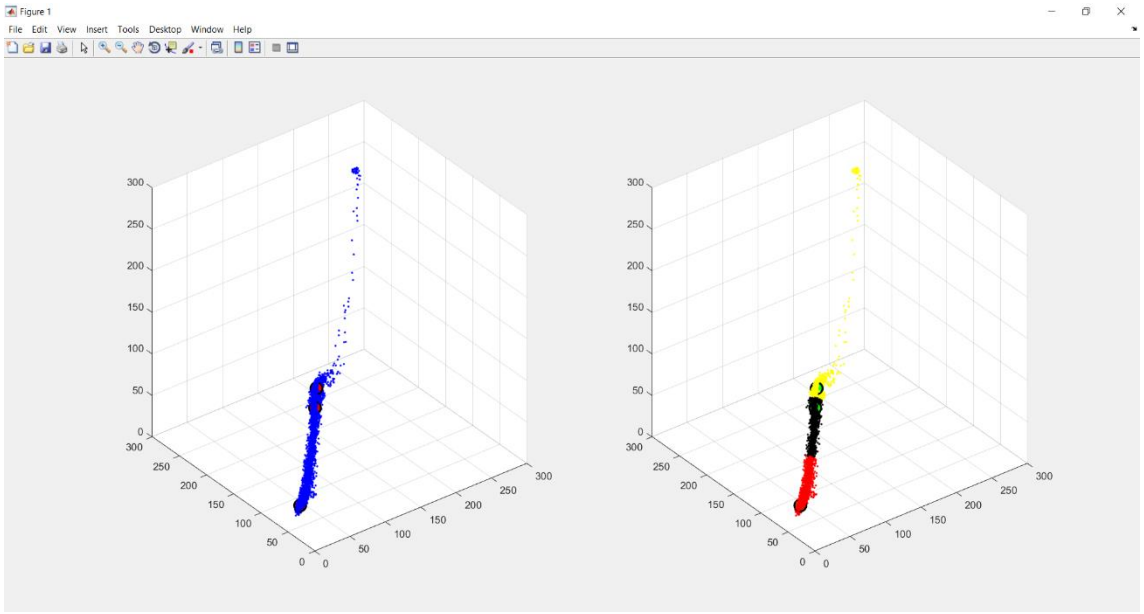
  110 iterações



- Photo006

  3 clusters

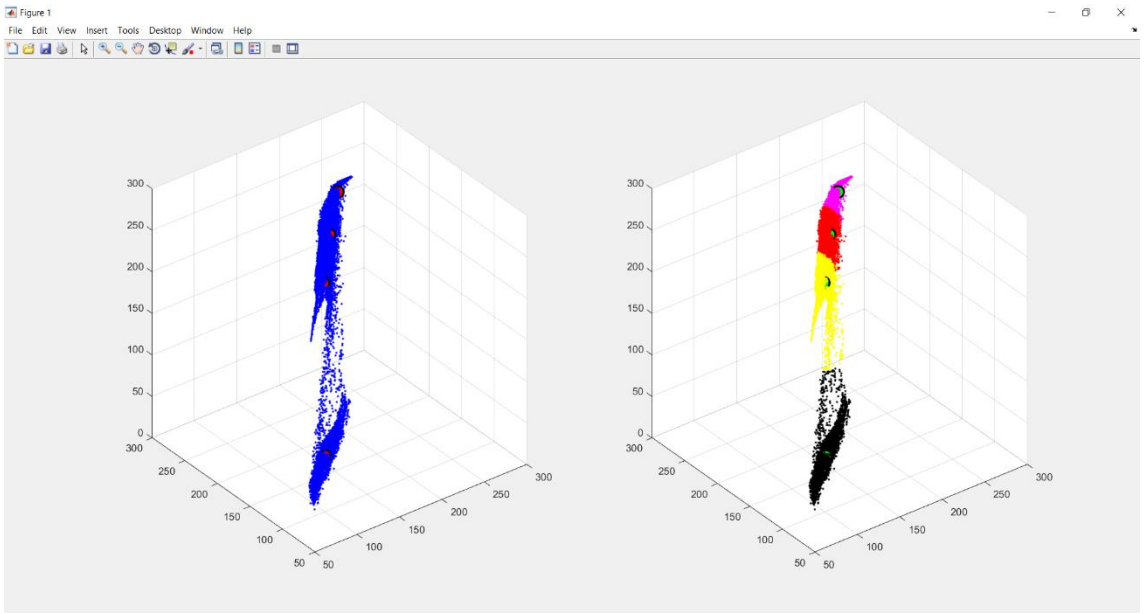
  59 iterações



- Photo007

  4 clusters

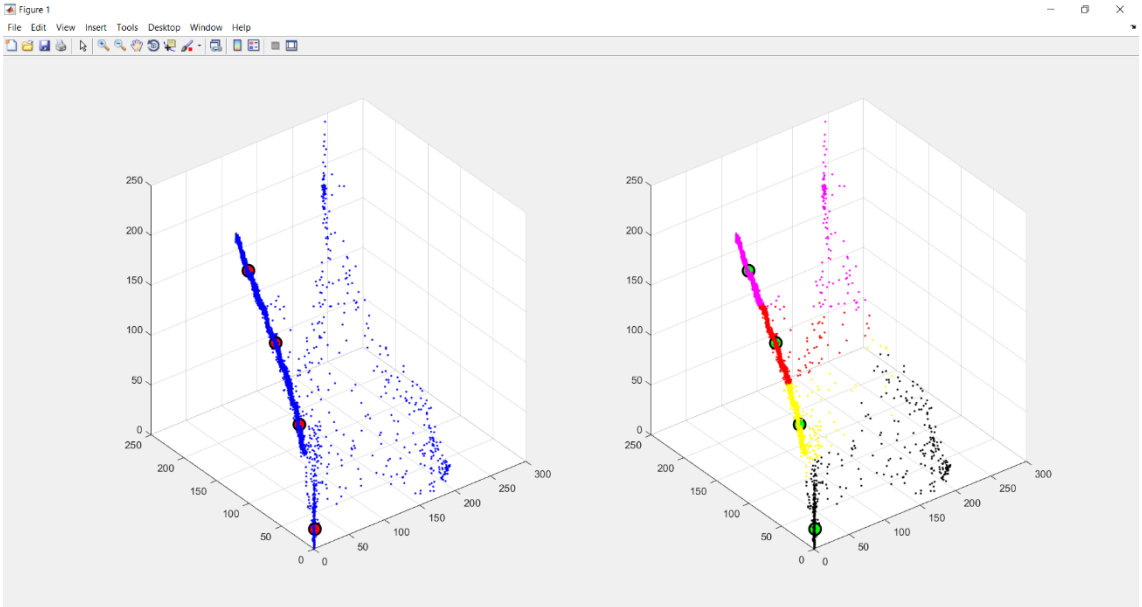
  56 iterações



- Photo008

  4 clusters

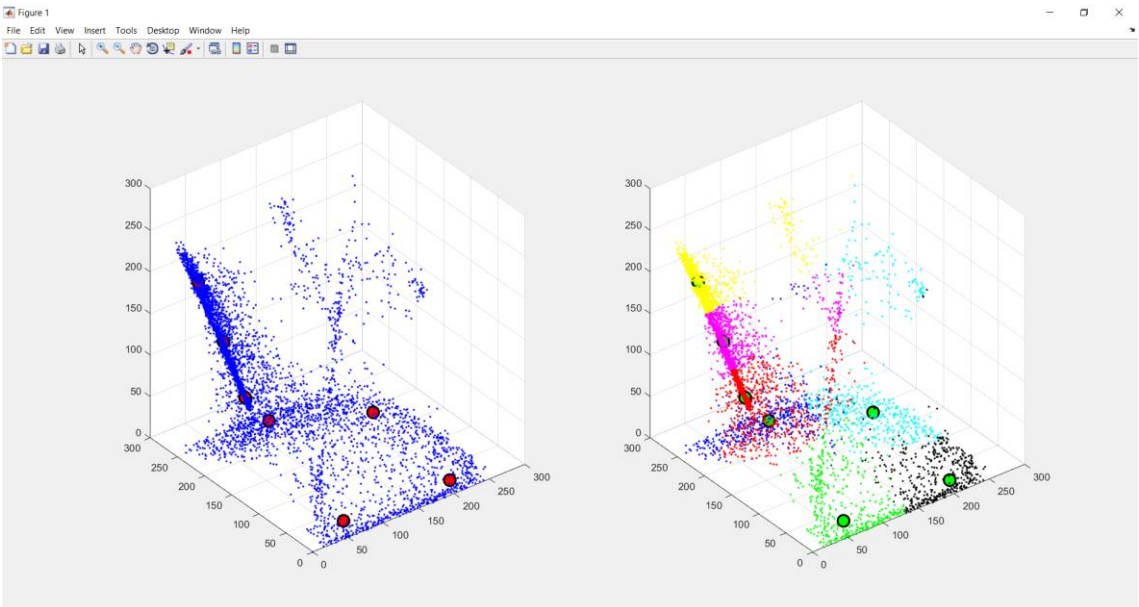
  128 iterações



- Photo009

  7 clusters

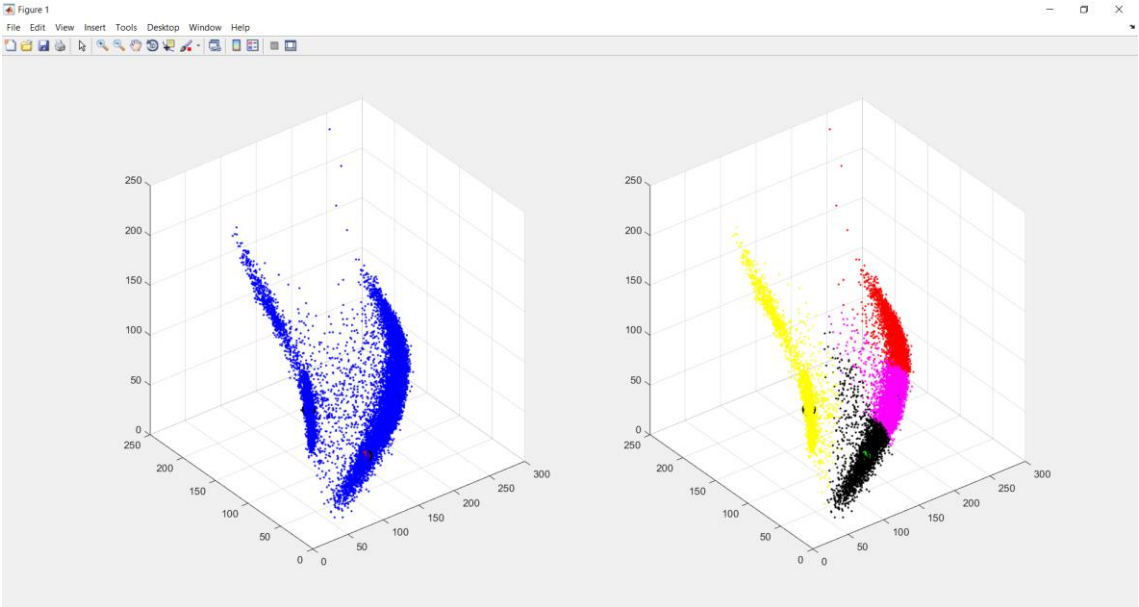
  141 iterações



- Photo010

  4 clusters

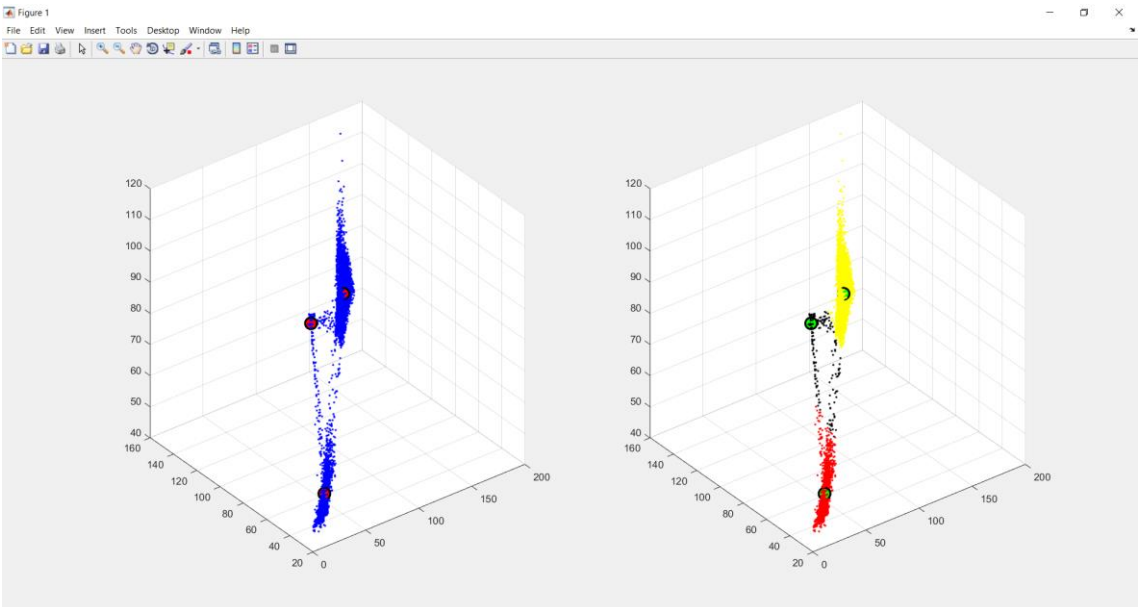
  45 iterações



- Photo011

  3 clusters

  28 iterações



* Para facilitar todas as imagens foram reduzidas em 60%.

Conclusão

Com esse trabalho foi possível entender melhor o funcionamento dos algoritmos K-means e FCM, bem como todo o processo de clustering.

Todas as atividades propostas foram realizadas com sucesso.