Mineração de Dados

Trabalho 9

Arthur do Prado Labaki - 11821BCC017

25-05, 2023

GBC212

Estudo do Conceito

K-means

O algoritmo K-means é um método de aprendizado de máquina não supervisionado usado para agrupar um conjunto de dados em clusters (grupos) com características semelhantes. Ele é amplamente utilizado em análise de dados e mineração de dados.

O objetivo principal do algoritmo K-means é particionar os dados em K clusters, onde K é um número pré-definido pelo usuário. Cada cluster é representado por um centróide, que é o ponto médio do cluster. O algoritmo K-means tenta minimizar a soma dos quadrados das distâncias entre cada ponto de dados e o centróide correspondente do cluster ao qual ele pertence.

O processo do algoritmo K-means é o seguinte:

1. Inicialização: Seleciona aleatoriamente K centróides iniciais, que são pontos no espaço de características dos dados.
2. Atribuição: Cada ponto de dados é atribuído ao centróide mais próximo com base na distância Euclidiana.
3. Atualização: Recalcula os centróides dos clusters com base nos pontos de dados atribuídos a cada cluster.
4. Repetição: Repete os passos 2 e 3 até que haja uma convergência, ou seja, até que os centróides não mudem significativamente entre as iterações.
5. Resultado: Os pontos de dados são agrupados em K clusters com base na proximidade dos centróides.

O algoritmo K-means é eficiente computacionalmente e relativamente fácil de implementar. No entanto, ele tem algumas limitações, como a necessidade de especificar o número de clusters K de antemão e sua sensibilidade à escolha inicial dos centróides. Essas limitações podem afetar a qualidade dos resultados obtidos pelo algoritmo.

Agrupamento Hierárquico

O Algoritmo Geral de Agrupamento Hierárquico Aglomerativo é um método utilizado na área de aprendizado de máquina e mineração de dados para agrupar objetos em clusters hierárquicos. Ele é usado para encontrar estruturas de agrupamento em conjuntos de dados, onde objetos similares são agrupados em clusters maiores e, em seguida, esses clusters são combinados em clusters ainda maiores, formando uma hierarquia.

O algoritmo começa considerando cada objeto como um cluster individual e, em seguida, aglomera gradualmente os clusters até que todos os objetos pertençam a um único cluster, formando assim a hierarquia. Durante o processo de aglomeração, os clusters são combinados com base em sua similaridade, que é geralmente calculada usando uma medida de distância entre os objetos.

O algoritmo segue os seguintes passos:

1. Inicialização: Cada objeto é considerado como um cluster individual.
2. Cálculo da matriz de dissimilaridade: Uma matriz de dissimilaridade é calculada, representando as distâncias ou similaridades entre os clusters.
3. Fusão dos clusters mais similares: Os dois clusters mais similares são fundidos em um único cluster, reduzindo assim o número total de clusters.
4. Atualização da matriz de dissimilaridade: A matriz de dissimilaridade é atualizada para refletir a similaridade entre o novo cluster e os clusters restantes.
5. Repetição dos passos 3 e 4: Os passos 3 e 4 são repetidos até que todos os objetos pertençam a um único cluster.

Ao final do algoritmo, é criada uma estrutura hierárquica de clusters, chamada dendrograma, que pode ser visualizada para entender a relação de similaridade entre os objetos. O dendrograma permite escolher o número de clusters desejado, cortando-o em diferentes níveis da hierarquia.

É importante destacar que existem diferentes variantes e abordagens para o algoritmo de agrupamento hierárquico aglomerativo, como o uso de diferentes medidas de distância, métodos de ligação entre clusters (por exemplo, ligação simples, ligação completa, ligação média) e critérios de parada. Cada uma dessas variações pode levar a resultados ligeiramente

diferentes e é escolhida com base nas características do conjunto de dados e nos objetivos do agrupamento.

Resolução do Exercício 1)

Considerando k = 2, encontre os grupos da base de dados a seguir partindo dos seguintes centroides: c1=(1.0, 1.0) e c2=(5.0, 7.0):

Subject	A	B
1	1.0	1.0
2	1.5	2.0
3	3.0	4.0
4	5.0	7.0
5	3.5	5.0
6	4.5	5.0
7	3.5	4.5

Figura 1: Base de dados

Como os centroides já foram escolhidos, deve ser feito o calculo das distancias de cada ponto aos centroides, para definir os clusters iniciais. Sendo subject os pontos, D(c1) a distancia do ponto ao centroide 1 e D(c2) a distancia do ponto ao centroide 2, temos:

Subject	D(c1)	D(c2)	Cluster
1	0	7.211	c1
2	1.118	6.103	c1
3	3.605	3.605	c1
4	7.211	0	c2
5	4.716	2.5	c2
6	5.315	2.061	c2
7	4.301	2.915	c2

Agora, é necessário atualizar a posição dos centroides, com base nos novos pontos em seu cluster. com isso:

Centroide 1:
Posição x: (1 + 1.5 + 3) / 3 = 1.833
Posição y: (1 + 2 + 4) / 3 = 2.333

Centroide 2:
Posição x: (5 + 3.5 + 4.5 + 3.5) / 4 = 4.125
Posição y: (7 + 5 + 5 + 4.5) / 4 = 5.375

Com esse novos centroides, deve ser refeito o calculo das distancias de cada ponto aos centroides. Então:

Subject	D(c1)	D(c2)	Cluster
1	1.571	5.376	c1
2	0.470	4.275	c1
3	2.034	1.776	c2
4	5.640	1.845	c2
5	3.145	0.728	c2
6	3.771	0.530	c2
7	2.734	1.075	c2

Como ocorreu uma alteração de cluster, é necessário refazer o calculo das centroides, então:

Centroide 1:
Posição x: (1 + 1.5) / 2 = 1.25
Posição y: (1 + 2) / 2 = 1.5

Centroide 2:
Posição x: (3 + 5 + 3.5 + 4.5 + 3.5) / 5 = 3.9
Posição y: (4 + 7 + 5 + 5 + 4.5) / 5 = 5.1

Refazendo os cálculos das distancias, temos:

Como não houve alteração nos agrupamentos, o algoritmo K-means termina.

Subject	D(c1)	D(c2)	Cluster
1	0.559	5.021	c1
2	0.559	3.920	c1
3	3.051	1.421	c2
4	6.656	2.195	c2
5	4.160	0.412	c2
6	4.776	0.608	c2
7	3.75	0.721	c2

Resolução do Exercício 2)

Dada a mesma base de dados, utilize o método de agrupamento hierárquico Min, Max e Média, encontre os dendogramas.

Subject	A	B
1	1.0	1.0
2	1.5	2.0
3	3.0	4.0
4	5.0	7.0
5	3.5	5.0
6	4.5	5.0
7	3.5	4.5

Figura 2: Base de dados

Nesse algoritmo, devemos considerar cada ponto um cluster de começo. Apos isso, devemos calcular a matriz de dissimilaridade usando a distancia euclidiana, temos:

Como a matriz é quadrada e simétrica, foi optado por fazer somente metade dela, para não complicar. Agora é selecionado o elemento que tem menor dissimilaridade (maior similaridade) para se unirem em um cluster. Esses elementos são os pontos 5 e 7 (D = 0.5).

	D1	D2	D3	D4	D5	D6	D7
D1	0	-	-	-	-	-	-
D2	1.118	0	-	-	-	-	-
D3	3.605	2.5	0	-	-	-	-
D4	7.211	6.103	3.605	0	-	-	-
D5	4.716	3.605	1.118	2.5	0	-	-
D6	5.315	4.242	1.802	2.061	2.915	0	-
D7	4.301	3.201	0.707	2.915	0.5	1.118	0

Agora é realizado a atualização da dissimilaridade, porem existem diferentes formas, as 3 mais famosas são:

- Min = Escolher a menor distancia gerada por um ponto e esse novo cluster;
- Max = Escolher a maior distancia gerada por um ponto e esse novo cluster;
- Média = Fazer a média como a distancia gerada por um ponto e esse novo cluster;

Para o exemplo, vamos escolher fazer o Min (Single Link). Então para recalcularmos a matriz, iremos fazer:
D(1, (5,7)) = min(D(1,5), D(1,7)) = min (4.716, 4.301) = 4.301
D(2, (5,7)) = min(D(2,5), D(2,7)) = min (3.605, 3.201) = 3.201
D(3, (5,7)) = min(D(3,5), D(3,7)) = min (1.118, 0.707) = 0.707
D(4, (5,7)) = min(D(3,5), D(4,7)) = min (2.5, 2.915) = 2.5
D(6, (5,7)) = min(D(6,5), D(6,7)) = min (2.915, 1.118) = 1.118

Remontando a tabela das distancias temos:

Com isso vamos ficar refazendo as contas e as tabelas até encontrarmos um único cluster (D(1,2,3,4,5,6,7)).

	D1	D2	D3	D4	D6	D(5,7)
D1	0	-	-	-	-	-
D2	1.118	0	-	-	-	-
D3	3.605	2.5	0	-	-	-
D4	7.211	6.103	3.605	0	-	-
D6	5.315	4.242	1.802	2.061	0	-
D(5,7)	4.301	3.201	0.707	2.5	1.118	0

Valor escolhido: 0.707 (3, (5,7)):
D(1, (3,5,7)) = min(D(1,3), D(1,5), D(1,7)) = min (3.605, 4.716, 4.301) = 3.605
D(2, (3,5,7)) = min(D(2,3), D(2,5), D(2,7)) = min (2.5, 3.605, 3.201) = 2.5
D(4, (3,5,7)) = min(D(4,3), D(3,5), D(4,7)) = min (3.605, 2.5, 2.915) = 2.5
D(6, (3,5,7)) = min(D(6,3), D(6,5), D(6,7)) = min (1.802, 2.915, 1.118) = 1.118

	D1	D2	D4	D6	D(3,5,7)
D1	0	-	-	-	-
D2	1.118	0	-	-	-
D4	7.211	6.103	0	-	-
D6	5.315	4.242	2.061	0	-
D(3,5,7)	3.605	2.5	2.5	1.118	0

Valor escolhido: 1.118 (1, 2):
D(4, (1,2)) = min(D(4,1), D(4,2)) = min (7.211, 6.103) = 6.103
D(6, (1,2)) = min(D(6,1), D(6,2)) = min (5.315, 4.242) = 4.242
D((3,5,7), (1,2)) = min(D(3,1), D(3,2), D(5,1), D(5,2), D(7,1), D(7,2)) =
= min (3.605, 2.5, 4.716, 3.605, 4.301, 3.201) = 2.5

	D4	D6	D(1,2)	D(3,5,7)
D4	0	-	-	-
D6	2.061	0	-	-
D(1,2)	6.103	4.242	0	-
D(3,5,7)	2.5	1.118	2.5	0

Valor escolhido: 1.118 (6, (3,5,7)):

D(4, (3,5,6,7)) = min(D(4,3), D(4,5), D(4,6), D(4,7)) = min (3.605, 2.5, 2.061, 2.915) = 2.061
D((1,2), (3,5,6,7)) = min(D(1,3), D(1,5), D(1,6), D(1,7), D(2,3), D(2,5), D(2,6), D(2,7)) =
= min (3.605, 4.716, 5.315, 4.301, 2.5, 3.605, 4.242, 3.201) = 2.5

	D4	D(1,2)	D(3,5,6,7)
D4	0	-	-
D(1,2)	6.103	0	-
D(3,5,6,7)	2.061	2.5	0

Valor escolhido: 2.061 (4, (3,5,7)):
D((1,2), (3,4,5,6,7)) = min(D(4,1), D(4,2), D(4,3), D(4,5), D(4,6), D(4,7), ...) =
= min (7.211, 6.103, 3.605, 2.5, 2.915, ...) = 2.5

	D(1,2)	D(3,4,5,6,7)
D4	0	-
D(3,5,6,7)	2.5	0

Com isso, ja temos o calculo completo, agora basta fazer o dendograma da base de dados encontrada, de acordo com os dados obtidos. Para facilitar a criação do dendograma, foi feito uma tabela com os dados obtidos.

Nó	Fusão	Nível
1	5 e 7	0.5
2	3 e (5,7)	0.707
3	1 e 2	1.118
4	6 e (3,5,7)	1.118
5	4 e (3,5,6,7)	2.061
6	(1,2) e (3,4,5,6,7)	2.5

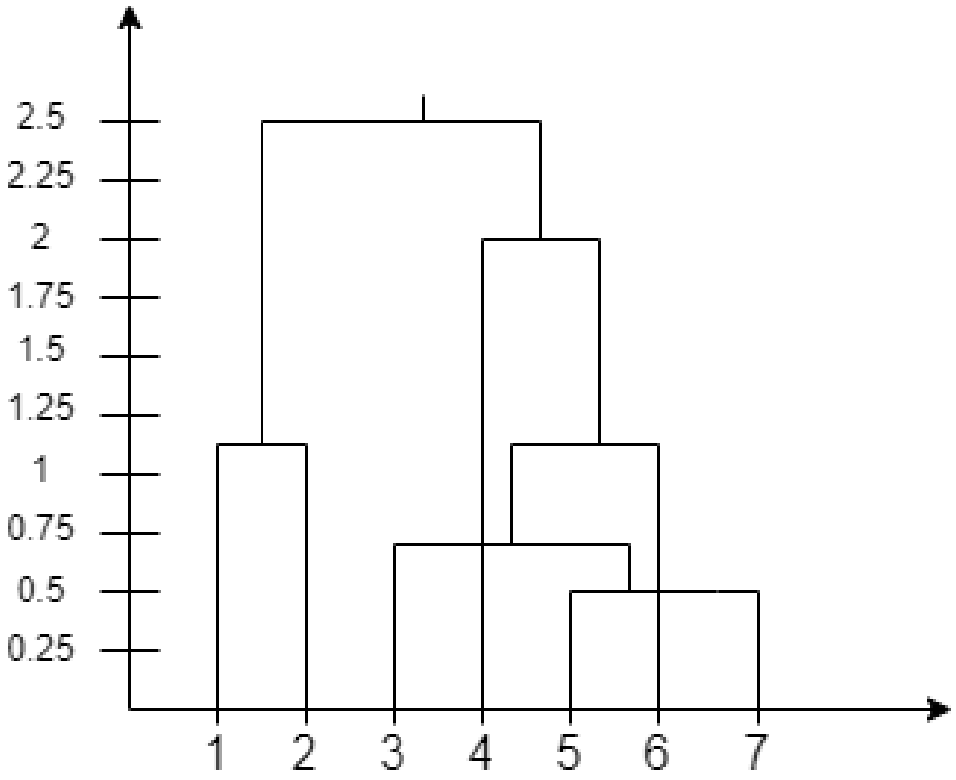

Figura 3: Dendograma Single Link