**Thuật toán k-means clustering**

1. **Giới thiệu**

Thuật toán K-means clustering (phân cụm K-means) nó thuộc lớp phương pháp Học không giám sát (**Unsupervised Learning)** trong học máy (**Machine Learning)**. Có rất nhiều định nghĩa khác nhau về kỹ thuật này, nhưng về bản chất ta có thể hiểu phân cụm là các qui trình tìm cách nhóm các đối tượng đã cho vào các cụm *(clusters),* sao cho các đối tượng trong cùng 1 cụm tương tự *(similar)* nhau và các đối tượng khác cụm thì không tương tự *(Dissimilar)* nhau.

Mục đích của phân cụm là tìm ra bản chất bên trong các nhóm của dữ liệu. Các thuật toán phân cụm (Clustering Algorithms) đều sinh ra các cụm (clusters). Tuy nhiên, không có tiêu chí nào là được xem là tốt nhất để đánh giá hiệu quả của của phân tích phân cụm, điều này phụ thuộc vào mục đích của phân cụm như: data reduction, “natural clusters”, “useful” clusters, outlier detection

1. ***Thuật toán***

K-Means là thuật toán rất quan trọng và được sử dụng phổ biến trong kỹ thuật phân cụm. Tư tưởng chính của thuật toán K-Means là tìm cách phân nhóm các đối tượng (objects) đã cho vào K cụm (K là số các cụm được xác đinh trước, K nguyên dương) sao cho tổng bình phương khoảng cách giữa các đối tượng đến tâm nhóm (centroid ) là nhỏ nhất.

* ***Thuật toán K-Means thực hiện qua các bước chính sau:***

1.    Chọn ngẫu nhiên K tâm (centroid) cho K cụm (cluster). Mỗi cụm được đại diện bằng các tâm của cụm.

2.    Tính khoảng cách giữa các đối tượng (objects) đến K tâm (thường dùng khoảng cách Euclidean)

3.    Nhóm các đối tượng vào nhóm gần nhất

4.    Xác định lại tâm mới cho các nhóm

5.    Thực hiện lại bước 2 cho đến khi không có sự thay đổi nhóm nào của các đối tượng

* ***Ví dụ minh họa thuật toán K-Mean:***

Giả sử ta có 4 loại thuốc A,B,C,D, mỗi loại thuộc được biểu diễn bởi 2 đặc trưng X và Y như sau. Mục đích của ta là nhóm các thuốc đã cho vào 2 nhóm (K=2) dựa vào các đặc trưng của chúng.

|  |  |  |
| --- | --- | --- |
| Đối tượng | Đặc tính 1 (X): chỉ số trọng lượng | Đặc tính 2 (Y): pH |
| Dược phẩm A | 1 | 1 |
| Dược phẩm B | 2 | 1 |
| Dược phẩm C | 4 | 3 |
| Dược phẩm D | 5 | 4 |

***Bước 1:*** Khởi tạo tâm (centroid) cho 2 nhóm. Giả sử ta chọn A là tâm của nhóm thứ nhất (tọa độ tâm nhóm thứ nhất c1(1,1)) và B là tâm của nhóm thứ 2 (tạo độ tâm nhóm thứ hai c2 (2,1)).



Y: Chỉ số pH

X: Chỉ số trọng lượng

***Bước 2:*** Tính khoảng cách từ các đối tượng đến tâm của các nhóm (Khoảng cách Euclidean)

$$D^{0}=\left[\begin{matrix}0&1\\1&0\end{matrix}\begin{matrix}3.61&5\\2.83&4.24\end{matrix}\right]\begin{matrix}c\_{1}=\left(1,1\right)&nhóm-1\\c\_{2}=\left(1,1\right)&nhóm-2\end{matrix}$$

 A B C D

$$\left[\begin{matrix}1&2\\1&1\end{matrix}\begin{matrix}4& 5\\3& 4\end{matrix}\right]\begin{matrix}X\\Y\end{matrix}$$

Mỗi cột trong ma trận khoảng cách (D) là một đối tượng (cột thứ nhất tương ứng với đối tượng A, cột thứ 2 tương ứng với đối tượng B,…). Hàng thứ nhất trong ma trận khoảng cách biểu diễn khoảng cách giữa các đối tượng đến tâm của nhóm thứ nhất (c1) và hàng thứ 2 trong ma trận khoảng cách biểu diễn khoảng cách của các đối tượng đến tâm của nhóm thứ 2 (c2).

Ví dụ, khoảng cách từ loại thuốc C=(4,3) đến tâm c1(1,1) là 3.61 và đến tâm c2(2,1) là 2.83 được tính như sau:

$$c\_{1}=\left(1,1\right)\sqrt{(4-1)^{2}+(3-1)^{2}}=3.61$$

$$c\_{2}=\left(2,1\right)\sqrt{(4-2)^{2}+(3-1)^{2}}=2.83$$

***Bước 3:*** Nhóm các đối tượng vào nhóm gần nhất

$$G^{0}=\left[\begin{matrix}1& 0\\0& 1\end{matrix}\begin{matrix} 0& 0\\ 1& 1\end{matrix}\right]\begin{matrix}nhóm-1\\nhóm-2 \end{matrix}$$

 A B C D

Ta thấy rằng nhóm 1 sau vòng lặp thứ nhất gồm có 1 đối tượng A và nhóm 2 gồm các đối tượng còn lại B,C,D.

***Bước 4:*** Tính lại tọa độ các tâm cho các nhóm mới dựa vào tọa độ của các đối tượng trong nhóm. Nhóm 1 chỉ có 1 đối tượng A nên tâm nhóm 1 vẫn không đổi, c1(1,1). Tâm nhóm 2 được tính như sau:

$$c\_{2}=\left(\frac{2+4+5}{3},\frac{1+3+4}{3}\right)=\left(\frac{11}{3}, \frac{8}{3}\right).$$



X: Chỉ số trọng lượng

Y: Chỉ số pH

***Bước 5:*** Tính lại khoảng cách từ các đối tượng đến tâm mới

$$D^{1}=\left[\begin{matrix}0&1\\3.14&2.36\end{matrix}\begin{matrix}3.61&5\\0.47&1.89\end{matrix}\right]\begin{matrix}c\_{1}=\left(1,1\right)&nhóm-1\\c\_{2}=\left(\frac{11}{3},\frac{8}{3}\right)&nhóm-2\end{matrix}$$

 A B C D

$$\left[\begin{matrix}1& 2\\1& 1\end{matrix}\begin{matrix}4& 5\\3& 4\end{matrix}\right]\begin{matrix}X\\Y\end{matrix}$$

***Bước 6:*** Nhóm các đối tượng vào nhóm

$$G^{1}=\left[\begin{matrix}1& 1\\0& 0\end{matrix}\begin{matrix} 0& 0\\ 1& 1\end{matrix}\right]\begin{matrix}nhóm-1\\nhóm-2 \end{matrix}$$

 A B C D

***Bước 7:*** Tính lại tâm cho nhóm mới

$$c\_{1}=\left(\frac{1+2}{2},\frac{1+1}{2}\right)=\left( 1\frac{1}{2}, 1 \right)$$

$$c\_{2}=\left(\frac{4+5}{2},\frac{3+4}{2}\right)=\left( 4\frac{1}{2}, 3\frac{1}{2}\right).$$



X: Chỉ số trọng lượng

Y: Chỉ số pH

***Bước 8:*** Tính lại khoảng cách từ các đối tượng đến tâm mới

$$D^{2}=\left[\begin{matrix}0.5&0.5\\4.30&3.54\end{matrix}\begin{matrix}3.20&4.61\\0.71&0.71\end{matrix}\right]\begin{matrix}c\_{1}=\left(1\frac{1}{2},1\right)&nhóm-1\\c\_{2}=\left(4\frac{1}{2},3\frac{1}{2}\right)&nhóm-2\end{matrix}$$

 A B C D

$$\left[\begin{matrix}1& 2\\1& 1\end{matrix}\begin{matrix} 4& 5\\ 3& 4\end{matrix}\right]\begin{matrix}X\\Y\end{matrix}$$

***Bước 9:*** Nhóm các đối tượng vào nhóm

$$G^{2}=\left[\begin{matrix}1& 1\\0& 0\end{matrix}\begin{matrix} 0& 0\\ 1& 1\end{matrix}\right]\begin{matrix}nhóm-1\\nhóm-2 \end{matrix}$$

 A B C D

Ta thấy G2 = G1 (Không có sự thay đổi nhóm nào của các đối tượng) nên thuật toán dừng và kết quả phân nhóm như sau:

|  |  |  |  |
| --- | --- | --- | --- |
| Đối tượng | Đặc tính 1 (X): chỉ số trọng lượng | Đặc tính 2(Y): pH | Nhóm (kết quả) |
| Dược phẩm A | 1 | 1 | 1 |
| Dược phẩm B | 2 | 1 | 1 |
| Dược phẩm C | 4 | 3 | 2 |
| Dược phẩm D | 5 | 4 | 2 |

Thuật toán K-Means có ưu điểm là đơn giản, dễ hiểu và cài đặt. Tuy nhiên, một số hạn chế của K-Means là hiệu quả của thuật toán phụ thuộc vào việc chọn số nhóm K (phải xác định trước) và chi phí cho thực hiện vòng lặp tính toán khoảng cách lớn khi số cụm K và dữ liệu phân cụm lớn.

Ta thấy G2 = G1 (Không có sự thay đổi nhóm nào của các đối tượng) nên thuật toán dừng và kết quả phân nhóm như sau:

|  |  |  |  |
| --- | --- | --- | --- |
| Đối tượng | Đặc tính 1 (X): chỉ số trọng lượng | Đặc tính 2(Y): pH | Nhóm (kết quả) |
| Dược phẩm A | 1 | 1 | 1 |
| Dược phẩm B | 2 | 1 | 1 |
| Dược phẩm C | 4 | 3 | 2 |
| Dược phẩm D | 5 | 4 | 2 |

Thuật toán K-Means có ưu điểm là đơn giản, dễ hiểu và cài đặt. Tuy nhiên, một số hạn chế của K-Means là hiệu quả của thuật toán phụ thuộc vào việc chọn số nhóm K (phải xác định trước) và chi phí cho thực hiện vòng lặp tính toán khoảng cách lớn khi số cụm K và dữ liệu phân cụm lớn.