Shape, square

Description automatically generated**TRƯỜNG ĐẠI HỌC GIAO THÔNG VẬN TẢI**

**UNIVERSITY OF TRANSPOST AND COMMUNICATIONS**

Logo

Description automatically generated

**BÁO CÁO NGHIÊN CỨU KHOA HỌC**

***ĐỀ TÀI: K-MEANS PHÂN CỤM DỮ LIỆU VÀ ỨNG DỤNG***

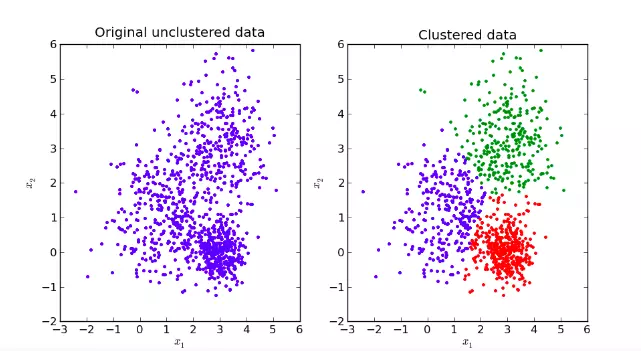
**LỜI NÓI ĐẦU**

Trong thời đại số hóa hiện nay , việc thu thập và lưu trữ dữ liệu trở nên dễ dàng hơn bao giờ hết . Tuy nhiên , để đưa ra các quyết định thông minh , chúng ta cần phải trính xuất thông tin từ các tập dữ liệu lớn và phức tạp . Phương pháp phân cụm dữ liệu là một trong những cách hiệu quả nhất để tóm tắt dữ liệu và giúp chúng ta hiểu được những mẫu trong dữ liệu . K – Means là một trong những phương pháp phân cụm dữ liệu phổ biến nhất và dễ hiểu nhất . Phương pháp này có thể giúp chúng ta phân cụm dữ liệu vào các nhóm dựa trên tương đồng giữa chúng . K – Means là một phương pháp đơn giản và nhanh chóng , phù hợp cho việc xử lí các tập dữ liệu có kích thước lớn . Trong bài báo cáo này , tôi sẽ giới thiệu về phân cụm dữ liệu và phương pháp K – Means các bước để thực hiện và ứng dụng nó vào bài toán cụ thể .

**I , BÀI TOÁN PHÂN CỤM**

**Bài toán phân cụm về cơ bản là một loại phương pháp học tập không có giám sát . Phương pháp học không có giám sát là phương pháp mà trong đó rút ra các tham chiếu từ bộ dữ liệu đầu vào không có phản hồi của các dán nhãn . Nói chung thì nó được sử dụng như một quá trình để tìm ra cấu trúc có ý nghĩa , các quá trình giải thích cơ bản , các đặc điểm tổng quát và các nhóm vốn có trong một tập hợp các ví dụ**

**Phân cụm là nhiệm vụ chia tất cả hoặc các điểm dữ liệu thành một nhóm sao cho các điểm dữ liệu trong cùng một nhóm phải giống nhau ( về các đặc điểm ) và khác các điểm dữ liệu thuộc nhóm khác ( về các đặc điểm ) . Dễ hiểu thì nó là một tập hợp các đối tượng trên cơ sở giống và khác nhau giữa chúng .**

**Ví dụ : Các dữ liệu trong biểu đồ dưới được nhóm lại với nhau vì các dữ liệu đó có thể phân loại thành một nhóm duy nhất do có chung các đặc điểm , và các cụm dữ liệu được tách riêng với nhau vì có các đặc điểm khác nhau . Chúng ta có thể phân biệt được các cụm và xác định được 3 cụm như hình bên dưới :**

**Tại sao phải phân cụm dữ liệu ?**

**Phân cụm rất quan trọng vì nó xác định được nhóm giữa các dữ liệu không có dán nhãn . Nó phụ thuộc vào người dùng , tiêu chí họ có thể sử dụng để đáo ứng nhu cầu của họ cần là gì ? . Chẳng hản , chúng ta có thể quan tâm để việc tìm đại diện cho các nhóm đồng nhất để rút gọn được dữ liệu , tìm các cụm dữ liệu ban đầu và mô tả các thuộc tích chưa biết của chúng , sau đó tìm các nhóm phù hợp và hữu ích . Thuật toán này phải đưa các giả định của sự giống nhau giữa các điểm và khác nhau giữa các cụm điểm .**

**Các phương pháp phân cụm :**

* **Phương pháp phân vùng : phương pháp này phân vùng các đối tượn thành K cụm và mỗi phân vùng tạo thành một cụm . Phương pháp này được sử dụng để tối ưu hóa chức năng tương tự tiêu và mục tiêu . Ví dụ : K – MEANS , CLARANS , PAM , …**
* **Phương pháp phân cụm dựa trên mật độ : phương pháp này coi các cụm là vùng dày đặc có một số điểm tương đồng và khác biệt so với vùng dày đặc thấp hơn của không gian . Phương pháp này có độ chính xác khá tốt và khả năng hợp nhất hai cụm . Ví dụ : DBSCAN, OPTICS , …**
* **Phương pháp phân cụm dựa trên phân cấp : các cụm được hình thành trong phương pháp này tạo thành cấu trúc kiểu cây dựa trên phân cấp . Các cụm mới được hình thành bằng cách sự dụng cụm đã được hình thành trước đó . Nó được chia làm 2 loại :**
* **Tích tụ**
* **Chia rẽ**

**Ví dụ : CURE , …**

* **Phương pháp phân cụm dựa trên lưới : trong phương pháp này , không gian dữ liệu được tạo thành một số ô hữu hạn tạo thành cấu trúc giống như lưới . Tất cả các hoạt động , phân cụm được được thực hiện trên các lưới này đều nhanh và không phụ thuộc và số lượng đối tượng dữ liệu .Ví dụ : STING , CLIQUE, …**

**Ứng dụng của phân cụm :**

* **Hiểu được dữ liệu ( understanding )**
* **Tóm tắt dữ liệu ( summarization )**
* **Hỗ trợ giai đoạn tiền sử lí dữ liệu ( data preprocessing )**
* **Mô tả sự phân bố dữ liệu / đối tượng ( data distribution )**
* **Nhận dạng mẫu ( pattern recognition )**
* **Phân tích dữ liệu không gian ( spatial data analysis )**
* **Xử lý ảnh ( image processing )**
* **Phân mảnh thị tường ( market segmentation )**
* **Gom cụm dữ liệu ( document clustering )**
* **…**

**II, MỘT SỐ ĐỘ ĐO KHOẢNG CÁCH TRONG PHÂN CỤM**

**1.1. Khái niệm tương tư và phi tương tự**

**Phép đo tương tự dữ liệu là khi mà các đặc tính của dự liệu đã được xác định , ta phải tìm các cách thích hợp để xác định khoảng cách giữa các đối tượng hay là phép đo , tương tự dữ lieu . Đây là các hàm để đo sự giống nhau giữa các cặp đối tương dữ liệu , thông thường các hàm này thưởng để tính độ tương tự hoặc là để tính độ phi tương tự giữa các đối tượng dữ liệu . Gía trị của hàm tính độ do tương tự càng lớn thì sự giống nhau giữa các dối tượng càng lớn và độ đo tươn tự càng nhỏ thì sự giống nhau giữa các đối tượng càng nhỏ , còn hàm tính độ phi tương tự tỉ lệ nghịch với hàm tính độ tương tự .** **Độ tương tự hoặc phi tương tự có nhiều cách để xác định, chúng thường được đo bằng khoảng cách giữa các đối tượng. Tất cả các cách đo độ tương tự đều phụ thuộc vào kiểu thuộc tính mà con người phân tích. Ví dụ, thuộc tính hạng mục thì không sử dụng độ đo khoảng cách mà sử dụng một hướng hình học của dữ liệu.**

**Tất cả các độ đo dưới đây được xác định trong không gian metric. Bất kỳ một metric nào cũng là một độ đo, nhưng điều ngược lại không đúng. Để tránh sự nhầm lẫn, thuật ngữ độ đo ở đây đề cập đến hàm tính độ tương tự hoặc hàm tính độ phi tương tự. Một không gian metric là một tập trong đó có xác định “khoảng cách” giữa từng cặp phần tử, với những tính chất thông thường của khoảng cách hình học. Nghĩa là, một tập X (các phần tử của nó có thể là những đối tượng bất kỳ) các đối tượng dữ liệu trong CSDL đề cập ở trên được gọi là một không gian metric nếu:**

**- Với mỗi cặp phần tử x, y thuộc X đều xác định theo một quy tắc nào đó, một số thực δ(x, y) được gọi là khoảng cách giữa x và y.**

* **Quy tắc nói trên thỏa mãn hệ tính chất sau:**
  + **δ(x,y) > 0 nếu x ≠ y.**
  + **δ(x,y) = 0 nếu x = y.**
  + **δ(x,y) = δ(y,x) với mọi x,y.**
  + **δ(x,y) ≤ δ(x,z) + δ(z,y) với mọi x,y,z.**

**Hàm δ(x,y) được gọi là một metric của không gian. Các phần tử của X được gọi là các điểm của không gian này.**

**+ Thuộc tính khoảng cách**

**Một thành phần quan trọng trong thuật toán phân cụm là phép đo khoảng cách giữa hai điểm dữ liệu. Nếu thành phần của vectơ thể hiện dữ liệu thuộc trong cùng một đơn vị giống nhau thì nó tồn tại khoảng cách Euclidean có thể xác định được nhóm dữ liệu tương tự. Tuy nhiên, không phải lúc nào khoảng cách Euclidean cũng cho kết quả chính xác.**

**Tuy nhiên chú ý rằng đây không phải vấn đề đồ thị: vấn đề phát sinh từ công thức toán học được sử dụng để kết hợp khoảng cách giữa các thành phần đơn đặc tính dữ liệu vectơ vào trong một độ đo khoảng duy nhất mà có thể được sử dụng cho mục đích phân cụm: các công thức khác nhau dẫn tới những cụm khác nhau.**

**Các thuật toán cần có các phép đo khoảng cách hoặc độ tương tự giữa hai đối tượng để thực hiện phân cụm. Kiến thức miền phải được sử dụng để để trình bày rõ ràng phép đo khoảng thích hợp cho mỗi ứng dụng. Hiện nay, phép đo có nhiều mức độ khách nhau tùy theo từng trường hợp.**

* **Khoảng cách Minkowski được định nghĩa như sau:**

****

**Trong đó x, y là hai đối tượng với n là số lượng thuộc tính, và; dist là kích thước của dữ liệu.**

* **Khoảng cách Euclidean:**

****

**Là khoảng cách giữa hai đối tượng trong trường hợp đặc biệt q = 2.**

* **Khoảng cách Manhattan:**

****

**Là khoảng cách giữa hai đối tượng trong trường hợp đặc biệt q = 1.**

* **Khoảng cách Chebychev:**

****

**Trong trường hợp q = ∞, hữu ích để định nghĩa các đối tượng phi tương tự nếu chúng khác nhau chỉ trong một kích thước biến đổi.**

* **Bình phương khoảng cách Euclidean:**

****

* **Tỉ lệ khác nhau. Giả sử các biến là tuyệt đối.**

****

**Khoảng cách Euclidean được sử dụng phổ biến nhất để đo độ tương tự của khoảng cách Minkowski. Giả sử có hai trường hợp  và  có các biến liên tục x và y, lấy lần lượt các giá trị  và  tương ứng.**

**Tuy nhiên không có nguyên tắc tổng quát để chọn phép đo áp dụng cho bất cứ bài toán nào. Một cách đơn giản để đo độ tương tự giữa các nhóm trong khung tương tự bằng cách thay thế nhóm cho thuộc tính thứ i của đối tượng đo chẳng hạn như khoảng cách Euclidean, khoảng cách Manhattan, hoặc bình phương Mahalanobis. Ví dụ, giả sử rằng nhóm A có vectơ trung bình  và nhóm B có vectơ trung bình  , thì cách đo bằng khoảng cách Euclidean giữa hai nhóm có thể được định nghĩa là:**

****

**Cách tiếp cận khác để khoảng cách giữa phần tử gần nhất hoặc phần tử xa nhất. Cách tiếp này sử dụng các thuật toán phân cụm phân cấp chẳng hạn như liên kết đơn và liên kết đầy đủ. Vấn đề chính với hai cách tiếp cận này giống nhau là không cảm nhận được mâu thuẫn định lượng và không tính toán cho các yếu tố của các phần tử trong một nhóm.**

**Cách tiếp cận khác, là trung bình nhóm, có thể sử dụng phép đo tương tự giữa các nhóm. Cách tiếp cận này, sự giống nhau giữa các nhóm được đo bằng cách lấy giá trị trung bình cả tất cả các phép đo giữa các đối tượng cho từng cặp đối tượng trong các nhóm khác nhau. Ví dụ, trung bình phi tương tự giữa nhóm A và B có thể được định nghĩa là:**

****

**trong đó, n là tổng số các đối tượng cùng cặp, ,  và lần lượt là số các đối tượng trong đối tượng và ,  là phi tương tự của một cặp đối tượng  và , , . Hàm phi tương tự có thể dễ dàng chuyển đổi sang hàm tương tự bằng cách thay đổi cho nhau.**

**+ Thuộc tính nhị phân**

**Tất cả các phép đo được định nghĩa ở trên là đa số thích hợp cho các biến liên tục. Cho các biến danh nghĩa, “phép đo khoảng cách” là 0 nếu các trường hợp có cùng giá trị danh nghĩa, và 1 nếu các trường hợp có các giá trị danh nghĩa khác nhau, hoặc với độ đo tương tự 1 (nếu các trường hợp có cùng giá trị danh nghĩa) và 0 (nếu không giống nhau).**

**Do đó nếu xem xét p biến định danh, có thể đánh giá độ tương tự của các trường hợp bằng số các biến mà có giá trị giống nhau. Nói chung định nghĩa với một biến nhị phân mới từ mỗi biến danh nghĩa, bằng việc nhóm các nhãn danh nghĩa thành hai lớp, một nhãn là 1, nhãn khác là 0. Xây dựng và xem xét bảng ngẫu nhiên các sự kiện có thể xảy ra và định nghĩa các thuộc tính của đối tượng x, y bằng các biến số nhị phân 0 và 1.**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | | **Y** | |  |
| **1** | **0** |
| **X** | **1** | **a** | **b** | **a+b** |
| **0** | **c** | **d** | **c+d** |
|  | | **a+c** | **b+d** | **p=a+b+c+d** |

###### 

###### **Bảng tham số**

**Trong đó:**

* + **a là tổng số các thuộc tính có giá trị 1 trong hai đối tượng x, y.**
  + **b là tổng số các thuộc tính có giá trị 1 trong x và giá trị 0 trong y.**
  + **c là tổng số các thuộc tính có giá trị 0 trong x và giá trị 1 trong y.**
  + **d là tổng số các thuộc tính có giá trị 0 trong hai đối tượng x, y.**
  + **p là tổng tất cả các thuộc tính của hai đối tượng x, y.**

**Các phép đo độ tương tự của các trường hợp với dữ liệu thuộc tính nhị phân được thực hiện bằng các cách sau:**

* **Hệ số đối sánh đơn giản:**

****

**cả hai đối tượng có vai trò như nhau, nghĩa là chúng đối xứng và có cùng trọng số.**

* **Hệ số Jaccard:**

****

**Tham số này bỏ qua số các đối sánh 0-0.**

**Công thức này sử dụng trong trường hợp mà trọng số của các thuộc tính có giá trị 1 của đối tượng dữ liệu cao hơn nhiều so với các thuộc tính có giá trị 0. Như vậy thuộc tính nhị phân ở đây là không đối xứng.**

****

****

****

**Các giá trị được định nghĩa trong khoảng [0, 1] và có thể biến đổi sang độ đo phi tương tự bằng biểu thức: .**

**+ Thuộc tính định danh**

**Độ đo phi tương tự giữa hai đối tượng x và y được định nghĩa như sau:**

****

**trong đó, m là số thuộc tính đối sánh tương ứng trùng nhau, p là tổng số các thuộc tính.**

**+ Thuộc tính có thứ tự**

**Phép đo độ phi tương tự giữa các đối tượng dữ liệu với thuộc tính thứ tự được thực hiện như sau: Giả sử i là thuộc tính thứ tự có  giá trị (là kích thước miền giá trị):**

**- Các trạng thái  được sắp xếp thứ tự như nhau: , có thể thay thế mỗi giá trị của thuộc tính bằng giá trị cùng loại với .**

**- Mỗi thuộc tính có thứ tự có các miền giá trị khác nhau, vì vậy phải chuyển đổi chúng về cùng miền giá trị [0, 1] bằng cách thực hiện phép biến đổi sau cho mỗi thuộc tính:**

****

**- Sử dụng công thức tính độ phi tương tự của thuộc tính khoảng đối với các giá trị , đây chũng chính là độ phi tương tự của thuộc tính có thứ tự.**

**+ Thuộc tính tỷ lệ**

**Có nhiều cách khác nhau để tính độ tương tự giữa các thuộc tính tỉ lệ. Một trong những số đó là sử dụng công thức tính logarit cho mỗi thuộc tính , ví dụ ,  đóng vai trò như thuộc tính khoảng. Phép biến đổi logarit này thích hợp trong trường hợp các giá trị của thuộc tính là số mũ.**

**Trong thực tế, khi tính độ tương tự dữ liệu, chỉ xem xét một phần các thuộc tính đặc trưng đối với các kiểu dữ liệu hoặc là đánh trọng số cho tất cả các thuộc tính dữ liệu. Trong một số trường hợp, loại bỏ đơn vị đo của các thuộc tính dữ liệu bằng cách chuẩn hóa chúng, hoặc gán trọng số cho mỗi thuộc tính giá trị trung bình, độ lệch chuẩn. Các trọng số này có thể sử dụng trong các độ đo khoảng cách trên, ví dụ với mỗi thuộc tính dữ liệu đã được gán trọng số tương ứng , độ tương đồng dữ liệu được xác định như sau:**

****

**Có thể chuyển đổi giữa các mô hình cho các kiểu dữ liệu trên, ví dụ như dữ liệu kiểu hạng mục có thể chuyển đổi thành dữ liệu nhị phân hoặc ngược lại. Giải pháp này rất tốn kém về chi phí tính toán, do vậy, cần phải cân nhắc khi áp dụng cách thức này.**

**Tóm lại, tùy từng trường hợp dữ liệu cụ thể mà có thể sử dụng các mô hình tính độ tương tự khác nhau. Việc xác định độ tương đồng dữ liệu thích hợp, chính xác đảm bảo khách quan là rất quan trọng, góp phần xây dựng thuật toán PCDL có hiệu quả cao trong việc đảm bảo chất lượng cũng như chi phí tính toán.**

**III, THUẬT TOÁN K-Means**

**1.1 Giới thiệu chung về thuật toán**

**Như ta đã biết thuật toán K – Means clustering ( thuật toán phân cụm K- Means là một trong những thuật toán cơ bản nhất trong học không giám sát (unsupervised learning ) . Thuật ngữ “K-Means” được J . MacQueen giới thiệu vào năm 1967 và đã được phát triển dựa trên ý tưởng của H.Steinhaus đề xuất vào năm 1956 .Thuật toán tiêu chuẩn được đề xuất lần đầu tiên vào năm 1957 bởi Stuart Lloyd của Bell Labs như một kỹ thuật cho điều chế mã xung , mặc dù nó không được xuất bản dưới dạng một bài báo cho đến năm 1982 . Năm 1965, Edward W.Forgy đã công bố về cơ bản cùng một phương pháp , thế nên đôi khi nó được gọi là Lloyd-Forgy .**

**Trong thuật toán K – Means clustering chúng ta không biết nhãn ( label ) của từng điểm dữ liệu . Mục đích là làm thế nào để phân các dữ liệu thành các cụm ( cluster ) khác nhau sao cho dữ liệu trong cùng một cụm có cùng tính chất giống nhau . Vậy nên thuật toán K – Means clustering đã sử dụng phương pháp tạo và cập nhật trung tâm để phân nhóm các điểm dữ liệu cho trước vào các nhóm khác nhau . Đầu tiên chúng sẽ tạo ra các điểm chung ngẫu nhiên . Sau đó gán mỗi điểm trong tập dữ liệu vào trung tâm mà gần với nó nhất . Sau đó chúng sẽ cập nhật lại trung tập và tiếp tục lặp lại các bước đã kể trên . Điều kiện dừng của thuật toán K – Means clustering đó là : Khi các trung tâm không thay đổi trong 2 vòng lặp kế tiếp nhau .Thế nhưng việc đạt được 1 kết quả hoàn hoản rất là khó và tốn nhiều thời gian , vậy nên thường thì người ta sẽ cho dừng thuật toán khi đạt được 1 kết quả gần đúng và điều đó chấp nhận được .**

**Ví dụ :**

**Một nhãn hàng muốn tạo ra những chính sách ưu đãi cho những khách hàng khác nhau của họ dựa trên sự tương tác giữa mỗi khách hàng với nhãn hàng đó ( số năm là khách hàng , số tiền khách hàng đã chi , độ tuổi , giới tính , thành phố , nghề nghiệp , ... ) Gỉa sử nhãn hàng đó có rất nhiều dữ liệu của rất nhiều khách hàng những chưa biết cách chia toàn bộ khách hàng đó thành một số nhóm / cụm khác nhau . Vì thế chúng ta nên sử dụng thuật toán K – Means clustering . Sau khi đã phân ra được từng nhóm , nhân viên của nhãn hàng có thể chọn một vài khách hàng để xem khách hàng đó thuộc vào nhóm nào .**

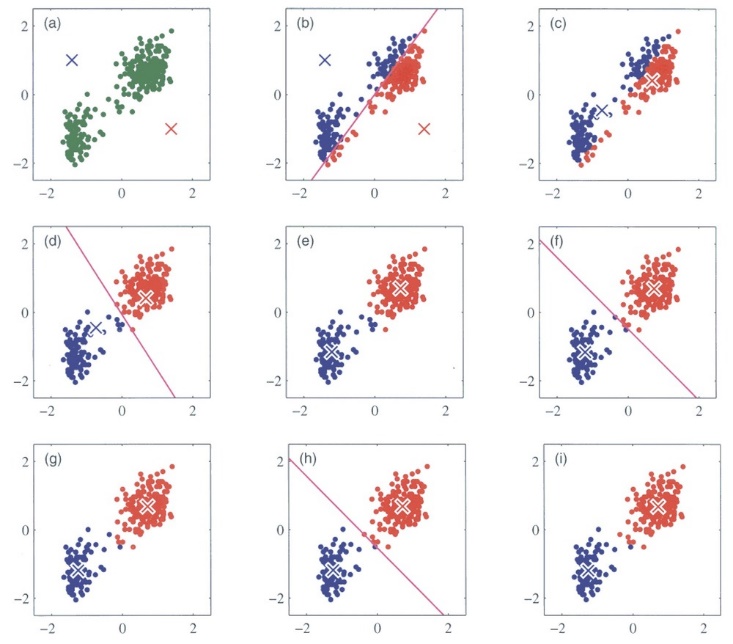
**1.2 Thuật toán K – Means .**

**Biểu diễn dữ liệu đầu vào là D = {x1,x2,x3,…,xn} với xi là vector có n chiều trong không gian Euclidean. Thuật toán phân cụm D thành K cụm dữ liệu nhỏ trong đó mỗi cụm có điểm trung tâm được gọi là centroid ( K là một hằng số cho trước ) .**

**Các bước thực hiện trong thuật toán :**

* **Đầu vào : Tập dữ liệu D , với K là số cụm , phép đo khoảng cách giữa 2 điểm là d(x,y)**
* **Khởi tạo : Khởi tạo K điểm dữ liệu trong D làm điểm trung tâm ( centroid)**
* **Lặp lại các bước sau cho đến khi hội tụ :**
  + **Bước 1 : Với mỗi điểm dữ liệu , gán điểm đó vào cluster sao cho khoảng cách đến centroid là nhỏ nhất .**
  + **Bước 2 : Với mỗi cluster , xác định lại điểm trung tâm của tất cả các điểm dữ liệu được gán vào cluster đó**

**Ví dụ của một số bước :**



**Điều kiện để thuật toán dừng ( hội tụ ) :**

**Thuật toán được dừng khi có một trong các cách sau :**

* **Trong vòng lặp có ít các điểm dữ liệu bị thay đổi sang cluster khác .**
* **Điểm trung tâm không bị thay đổi nhiều .**
* **Giá trị hàm mất mát không thay đổi nhiều .**

**Hàm mất mát :**

**Nếu ta coi** mk **là centroid của mỗi cluster và ước lượng tất cả các điểm được phân**

**vào cluster này bởi** mk , thì một điểm dữ liệu xi được phân vào cluster K sẽ bị sai số là ( xi – mk ) . Chúng ta mong muốn sai số này có giá trị tuyệt đói nhỏ nhất nên ta sẽ tìm các để đại lượng này nhỏ nhất :

Vậy sai số cho toàn bộ dữ liệu sẽ là :

ERROR =

**Trong đó Ck là cluster thứ K ,** mk là điểm trung tâm của cluster Ck tương ứng .

Xác định điểm trung tâm của cluter

Để xác định điểm trung tâm của cluster ta sử dụng công thức như sau :

=

Trong đó Ck là cluster thứ k , mk là điểm trung tâm của cluster Ck tương ứng .

Phép đo khoảng cách

Trong thuật toán **K – Means clustering** để đánh giá mức độ giống nhau hay khác nhau giữa 2 điểm dữ liệu ta có thể sử dụng các phép đo khoảng cách khác nhau . Ngoài khoảng cách Euclidean , tùy thuộc vào từng bài toán chungstacos thể sử dụng các phương pháp đo khác nhau như : cosine , manhattan , … )

d(x,) = =

Mọi phương pháp tính khoảng cách giữa 2 vector đều có thể được sử dụng . Mỗi cách tính khoảng cách thể hiện cách nhìn nhận khác về dữ liệu

* Có vô số cách tính khoảng cách
* Cách tính khoảng cách nào là tốt nhất ? Tùy thuộc vào từng bài toán

Một số nhược điểm của thuật toán :

* Chỉ áp dụng với dữ liệu có thuộc tính số
* Khám phá các cụm có dạng hình cầu , không thích hợp với hiện tìm các cụm hình dáng không lồi hay các cụm có hình khác xa nhau
* Nhạy cảm với các phần tử ngoại lại , phần tử nhiễu , phần tử cận biên cụm ( những phần tử như vậy có thể ảnh hưởng đến giá trị trung bình đáng kể )
* Chọn tập điểm trung tâm ban đầu cũng ảnh hưởng nhiều đến chất lượng cụm sinh ra

Thuật toán **K – Means clustering** được xếp vào lớp bài toán NP , chính vì vậy để phát triển thuật toán này người ta thường kết hợp với phỏng đoán .