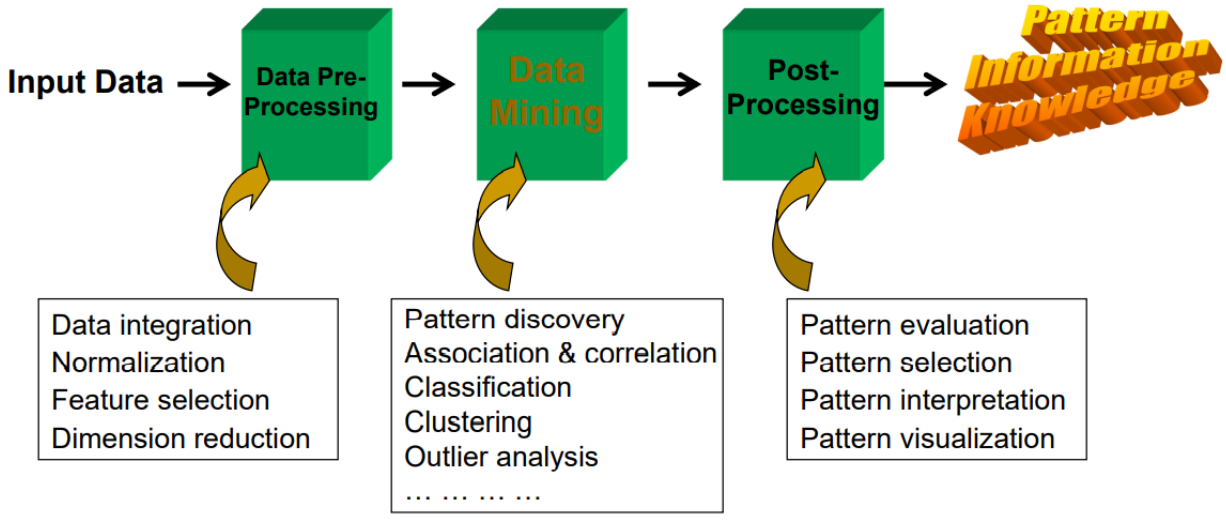
BÁO CÁO MÔN HỌC
Học máy và khai phá dữ liệu

Khai phá dữ liệu và phân cụm

Trong kỷ nguyên mới, khai phá dữ liệu đang trở thành từ khóa rất hot. Thực vậy, khai phá dữ liệu đã và đang được đẩy mạnh nghiên cứu và ứng dụng nhiều trong các lĩnh vực khác nhau và mang lại những cải thiện rất đáng kể. Những lĩnh vực có ứng dụng của khai phá dữ liệu bao gồm: thiên văn học, tin sinh học, thương mại điện tử, marketing, quản lý quan hệ khách hàng, y tế…. Đặc biệt, khai phá dữ liệu được xem là phương pháp mà đơn vị Able Danger của Quân đội Hoa Kỳ sử dụng để xác định lực lượng khủng bố đứng đầu cuộc tấn công ngày 11 tháng 9 nhằm vào nước này.

Khai phá dữ liệu là quá trình tính toán để tìm ra các mẫu, tìm ra thông tin đáng quý và tri thức từ một bộ dữ liệu để sử dụng cho quá trình đánh giá, phân tích sau này. Nó là một bước của quá trình khai phá tri thức



Các phương pháp khai phá dữ liệu có thể chia làm các phương pháp chính sau:

- Phân loại (Classification): Là phương pháp dự báo, cho phép phân loại một đối tượng vào một hoặc một số lớp cho trước.
- Hồi qui (Regression): Khám phá chức năng học dự đoán, ánh xạ một mục dữ liệu thành biến dự đoán giá trị thực.
- Phân cụm (Clustering): Một nhiệm vụ mô tả phổ biến trong đó người ta tìm cách xác định một tập hợp hữu hạn các cụm để mô tả dữ liệu.
- Tổng hợp (Summarization): Một nhiệm vụ mô tả bổ sung liên quan đến phương pháp cho việc tìm kiếm một mô tả nhỏ gọn cho một bộ (hoặc tập hợp con) của dữ liệu.
- Mô hình ràng buộc (Dependency modeling): Tìm mô hình cục bộ mô tả các phụ thuộc đáng kể giữa các biến hoặc giữa các giá trị của một tính năng trong tập dữ liệu hoặc trong một phần của tập dữ liệu.

- Dò tìm biến đổi và độ lệch (Change and Deviation Dectection): Khám phá những thay đổi quan trọng nhất trong bộ dữ liệu.

Trong các phương pháp khai phá dữ liệu, phân cụm dữ liệu (data clustering) là quá trình tìm kiếm để phân ra các cụm dữ liệu, các mẫu dữ liệu từ tập dữ liệu lớn sao cho các đối tượng trong cùng một cụm là tương đồng, còn các đối tượng thuộc các cụm khác nhau sẽ có nét khác nhau rõ rệt.

Bài toán phân cụm kết quả học tập của sinh viên

Số học sinh, sinh viên của ngành giáo dục nói chung là một con số rất lớn. Mỗi trường, cơ sở đào tạo có số học sinh, sinh viên lớn; đặc biệt các trường đại học, như Đại học Bách Khoa Hà Nội, con số lên đến 30, 40 nghìn sinh viên. Chỉ riêng việc quản lý số sinh viên khổng lồ đã là một thách thức rất lớn huống chi việc phân tích và đánh giá, vẽ ra các chiến lược cho các sinh viên. Việc này chắc chắn không thể làm thủ công mà phải nhờ vào sự hỗ trợ của công nghệ. Công nghệ sẽ đảm nhận phân tích, tìm ra điểm sai, tìm ra lí do từ đó các nhà lãnh đạo đánh giá, đưa ra các chiến lược phù hợp, nâng cấp chất lượng cho việc giáo dục và các vấn đề xoay quanh để giảng dạy trở nên hiệu quả.

Kết quả học tập của một cá nhân phụ thuộc vào rất nhiều yếu tố. Bài toán này sẽ xem xét chủ yếu dựa trên những tác nhân chủ quan từ phía sinh viên: kết quả trung bình học kỳ, giới tính, quê quán, các đặc điểm gia đình, số lượng các học phần đăng ký cho kỳ, tỷ lệ tham dự các buổi học trên lớp…. Thực hiện việc phân cụm có thể tìm ra được những tác nhân gây ra kết quả học tập không tốt của học sinh sinh viên, sự liên kết giữa chúng từ đó tìm ra giải pháp. Kết quả học tập của học sinh được xem là tốt, không tốt sẽ dựa trên điểm trung bình của sinh viên.

Tập dữ liệu

Sử dụng https://archive.ics.uci.edu/ml/datasets/student+performance#
Chi tiết tập dữ liệu:
1 school - student's school (binary: 'GP' - Gabriel Pereira or 'MS' - Mousinho da Silveira)
2 sex - student's sex (binary: 'F' - female or 'M' - male)
3 age - student's age (numeric: from 15 to 22)
4 address - student's home address type (binary: 'U' - urban or 'R' - rural)
5 famsize - family size (binary: 'LE3' - less or equal to 3 or 'GT3' - greater than 3)
6 Pstatus - parent's cohabitation status (binary: 'T' - living together or 'A' - apart)
7 Medu - mother's education (numeric: 0 - none, 1 - primary education (4th grade), 2 as“ 5th to 9th grade, 3 as“ secondary education or 4 as“ higher education)
8 Fedu - father's education (numeric: 0 - none, 1 - primary education (4th grade), 2 as“ 5th to 9th grade, 3 as“ secondary education or 4 as“ higher education)
9 Mjob - mother's job (nominal: 'teacher', 'health' care related, civil 'services' (e.g. administrative or police), 'at_home' or 'other')
10 Fjob - father's job (nominal: 'teacher', 'health' care related, civil 'services' (e.g. administrative or police), 'at_home' or 'other')
11 reason - reason to choose this school (nominal: close to 'home', school 'reputation', 'course' preference or 'other')
12 guardian - student's guardian (nominal: 'mother', 'father' or 'other')
13 traveltime - home to school travel time (numeric: 1 - <15 min., 2 - 15 to 30 min., 3 - 30 min. to 1 hour, or 4 - >1 hour)
14 studytime - weekly study time (numeric: 1 - <2 hours, 2 - 2 to 5 hours, 3 - 5 to 10 hours, or 4 - >10 hours)
15 failures - number of past class failures (numeric: n if 1<=n<3, else 4)
16 schoolsup - extra educational support (binary: yes or no)
17 famsup - family educational support (binary: yes or no)

18 paid - extra paid classes within the course subject (Math or Portuguese) (binary: yes or no)
19 activities - extra-curricular activities (binary: yes or no)
20 nursery - attended nursery school (binary: yes or no)
21 higher - wants to take higher education (binary: yes or no)
22 internet - Internet access at home (binary: yes or no)
23 romantic - with a romantic relationship (binary: yes or no)
24 famrel - quality of family relationships (numeric: from 1 - very bad to 5 - excellent)
25 freetime - free time after school (numeric: from 1 - very low to 5 - very high)
26 goout - going out with friends (numeric: from 1 - very low to 5 - very high)
27 Dalc - workday alcohol consumption (numeric: from 1 - very low to 5 - very high)
28 Walc - weekend alcohol consumption (numeric: from 1 - very low to 5 - very high)
29 health - current health status (numeric: from 1 - very bad to 5 - very good)
30 absences - number of school absences (numeric: from 0 to 93)

these grades are related with the course subject, Math or Portuguese:
31 G1 - first period grade (numeric: from 0 to 20)
31 G2 - second period grade (numeric: from 0 to 20)
32 G3 - final grade (numeric: from 0 to 20, output target)

Đây là tập dữ liệu sử dụng trong quá trình học có giám sát, áp dụng học máy để dự đoán điểm cuối cùng (final grade) của sinh viên. Để áp dụng cho bài toán phân cụm đang xét, chúng ta bỏ thuộc tính G3 trong tập học, xét các thuộc tính khác và sử dụng G1, G2 là điểm kỳ 1 kỳ 2 để phân chia học sinh theo kết quả học tập.

Hiểu tập dữ liệu

Đề xuất bài toán

Dựa trên tập dữ liệu trên, chúng ta sẽ dùng thuật toán phân cụm để có thể phân tích 2 bài toán sau:

+ Bài toán 1: Ảnh hưởng của tình yêu với kết quả của sinh viên

+ Bài toán 2: Ảnh hưởng của giáo dục gia đình đến kết quả của sinh viên

Hướng tiếp cận bài toán

Thực hiện chia tập dữ liệu thành các phần (ví dụ: những học sinh sinh viên ở trong mối quan hệ tình cảm và không; những học sinh sinh viên được sự hỗ trợ giáo dục từ gia đình và không). Với mỗi phần của tập dữ liệu, chúng ta chạy thuật toán phân cụm để phân nhóm sinh viên theo yếu tố kết quả học tập. Từ đó chúng ta phân tích các cụm sinh viên, tìm ra vấn đề và đề ra giải pháp.

Thuật toán phân cụm

Giải quyết bài toán trên, chúng ta sử dụng 2 thuật toán phân cụm: K-means và Fuzzy C-means

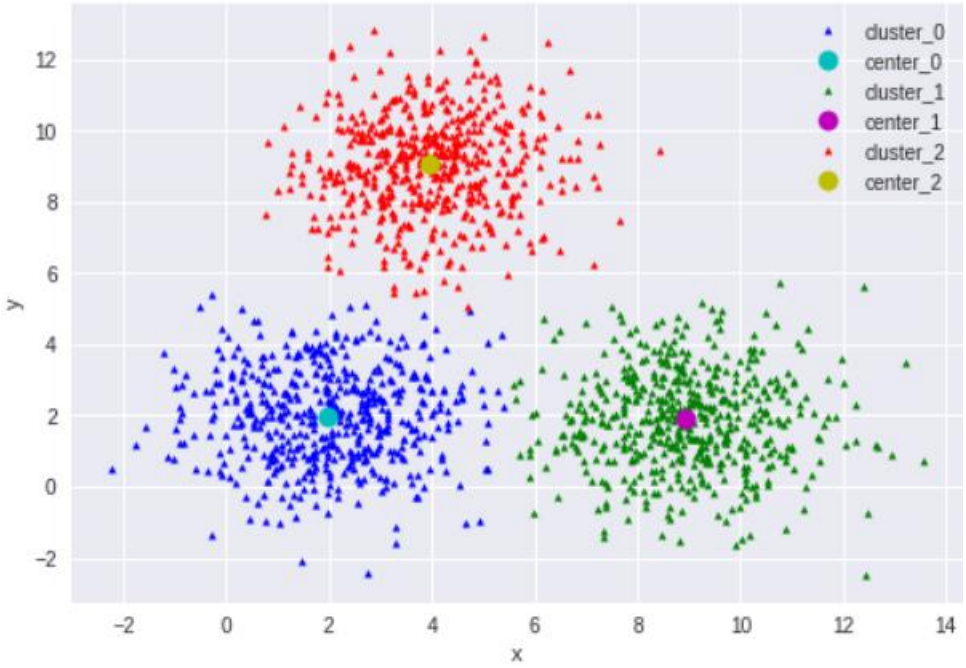
Phân cụm K-means

Đặc điểm của k-means và giải thuật

- Là phương pháp phổ biến nhất trong các phương pháp phân cụm dựa trên chia cắt
- Giải thuật k – means phân chia tập dữ liệu thành k cụm (k được xác định trước):
  + Mỗi cụm có một điểm trung tâm gọi là centroid

+ Các điểm dữ liệu thuộc cùng một cụm có sự tương tự cao với nhau, và khác biệt so với các dữ liệu thuộc cụm khác

- Đầu vào: một tập dữ liệu (các dữ liệu được biểu diễn bằng một dải các thuộc tính)
  Đầu ra: nhóm mà mỗi điểm dữ liệu thuộc vào



- Các bước chính của phân cụm k-means:
  + Xác định k là số các cụm
  + Xác định ngẫu nhiên k ví dụ học (gọi là seeds) để làm điểm trung tâm ban đầu của k cụm đó
  + Đối với mỗi ví dụ x, gán nó vào cụm có điểm trung tâm của cụm đó gần với x nhất
  + Tính toán lại điểm trung tâm của cụm dựa trên các ví dụ thuộc cụm đó tại thời điểm hiện tại
  + Dừng lại nếu điều kiện hội tụ thỏa mãn; nếu không, quay lại bước 3

Các vấn đề cần chú ý của k-means

+ "Gần" có ý nghĩa như thế nào?

Với một bộ dữ liệu, đặc biệt là bộ dữ liệu lớn, chúng ta không thể trực quan hóa chúng và xác định các điểm mình cho là "gần" bằng mắt. Sự "gần" hay tương tự giữa các điểm dữ liệu sẽ được đo bằng các hàm khoảng cách.

Các hàm khoảng cách thường dùng: Euclid, city-block, minkowski… Điển hình sử dụng nhiều nhất chính là hàm khoảng cách Euclid. Trong bài toán này chúng ta sử dụng khoảng cách Euclid.

Với hai điểm x, y trong không gian m chiều, công thức tính khoảng cách Euclid là:

$$d(x, y) = \sqrt{\sum_{i=1}^{m}(x_i - y_i)^2}$$

Khi tính bước 3, chúng ta thực hiện tính khoảng cách từ điểm dữ liệu x đến các centroid dựa trên công thức trên.

+ Điều kiện hội tụ?

Qua các vòng lặp, làm sao để biết khi nào thuật toán phân cụm k-means dừng lại?

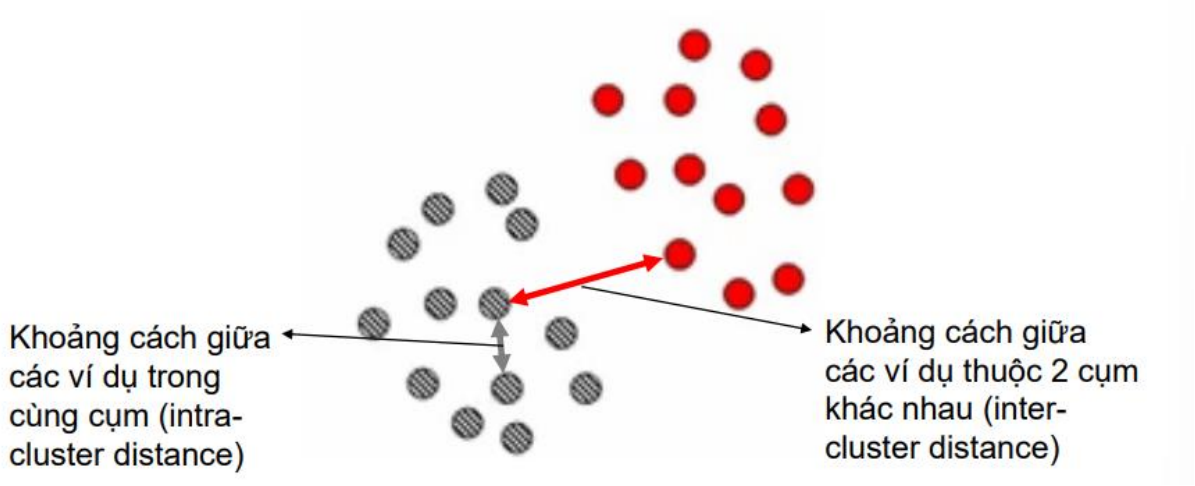
Các điều kiện hội tụ đó là: không có (hoặc rất ít) việc gán lại các ví dụ vào cụm khác; không có (hoặc rất ít) việc thay đổi các điểm trung tâm của các cụm.

Trong bài toán này, điều kiện dừng là: không có hoặc có rất ít sự gán lại các ví dụ vào cụm khác.

+ Làm sao xác định được độ hiệu quả của k-means?

Thuật toán phân cụm nói chung là dạng học không có giám sát (learn by observations) vậy nên không có nhãn đầu ra trong tập học để xác minh tính hiệu quả của k-means. Việc đánh giá hiệu quả phân cụm là rất thách thức. Các nguyên tắc của phân cụm sẽ được sử dụng để đánh giá: Sự gắn kết các ví dụ trong cùng cụm là tối đa, và sự tương tự giữa các ví dụ thuộc 2 cụm khác nhau là tối thiểu.

Các phương pháp đo độ chính xác thường dùng là RMSSTD, R-squared, Dunn-index, Davies-Bouldin index.



+ Xác định k thế nào là tối ưu

Việc xác định k một cách không có cơ sở sẽ dẫn đến độ chính xác của thuật toán không cao, và việc rút ra tri thức từ các cụm sẽ khó khăn. Mặt khác, k là tham số phải xác định từ ngay bước đầu tiên của thuật toán vậy nên phải xác định một k đủ tốt. Các phương pháp xác định k thường dùng là elbow method và silhouette coefficient.

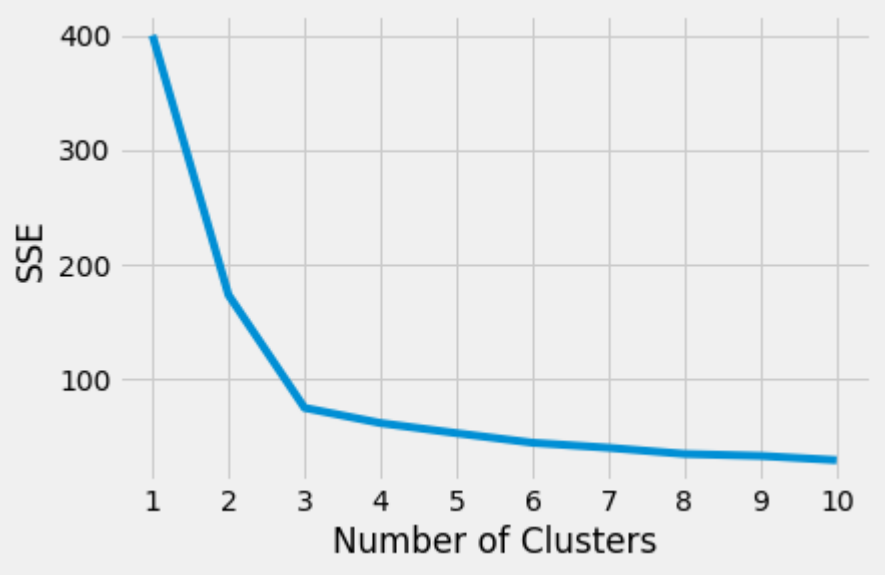
Muốn biết tập dữ liệu có được phân cụm đủ tốt hay không, chúng ta sử dụng Inertia hay SSE (Sum of squared errors). Công thức tính SSE như sau:

$$SSE = \sum_{i\equiv 1}^{n} \sum_{j=1}^{k} w^{(i_j j)}\, d\left(x^i - \mu^j\right)$$

Trong đó: n là số điểm dữ liệu, k là số cụm, x^i là ví dụ học thứ I, μ^j là centroid của cụm thứ j

$w^{(i_j j)} = 1\ nếu\ x^i$thuộc cụm j

Chúng ta nghiên cứu sử dụng elbow method để chọn k: chạy k-means với k tăng dần và ghi lại SSE với từng k. SSE sẽ giảm khi chúng ta tăng k. Tuy nhiên tồn tại một giá trị k mà tại đó SSE giảm một cách ngoạn mục, và từ giá trị k đó SSE giảm không đáng kể. Ta chọn giá trị k đó.



Ví dụ với đồ thị này, chúng ta chọn k = 3.

Áp dụng vào bài toán thực tế

Đối với bài toán 1

Thực hiện chia tập dữ liệu thành 2 bên: một bên có trong mối quan hệ tình cảm, một bên không. Thực hiện việc hiểu cơ bản 2 tập này (% giới tính, % sống ở thành thị, nông thôn,…)

Sử dụng phương pháp elbow để quyết định số cụm tối ưu đối với mỗi phần của tập dữ liệu

Thực hiện phân cụm k – means

Rút ra kết luận dựa trên một số thuộc tính

Đối với bài toán 2

Thực hiện chia tập dữ liệu thành 2 bên: một bên được sự hỗ trợ giáo dục của bố mẹ, một bên không. Thực hiện việc hiểu cơ bản 2 tập này (% giới tính, % sống ở thành thị, nông thôn,…)

Sử dụng phương pháp elbow để quyết định số cụm tối ưu đối với mỗi phần của tập dữ liệu

Thực hiện phân cụm k – means

Rút ra kết luận dựa trên một số thuộc tính

Phân cụm Fuzzy C-means

Đặc điểm của Fuzzy C-means và thuật toán

- Trong khi k-means là một thuật toán phân cụm cứng (hard clustering), có nghĩa là mỗi điểm dữ liệu sẽ chỉ được gán cho một cụm duy nhất, thì Fuzzy C-means là một thuật toán phân cụm mềm (soft clustering) và phân cụm mờ (fuzzy clustering)

- Kết quả của Fuzzy C-means khác so với k-means: một điểm dữ liệu có thể thuộc nhiều cụm. Đầu ra của một điểm dữ liệu sau khi chạy qua thuật toán FCM sẽ là bộ xác suất.

Để dễ hiểu: giả sử có ba cụm A, B, C và một điểm dữ liệu x. Sau FCM, x được gán một bộ như sau (0,1; 0,3; 0,6) ứng với ba cụm A, B và C. Điều này có nghĩa là 10% khả năng x thuộc cụm A, 30% khả năng x thuộc cụm B và 60% khả năng x thuộc cụm C

- Thuật toán FCM rất giống với k-means. Điểm khác biệt duy nhất là ở đầu ra, khi mà k-means cho đầu ra cứng còn FCM cho đầu ra "mềm mại" hơn dựa trên một chỉ số gọi là chỉ số đo mức độ thành viên (membership grade).

Membership grade có giá trị từ 0 đến 1. Nếu nó có giá trị 0 thì điểm dữ liệu đó hoàn toàn không thuộc về cụm đang xét, nếu là 1 thì nó hoàn toàn thuộc về cụm đang xét. Có thể nhìn nhận k-means cũng sử dụng membership grade này, nhưng sẽ chỉ là 0 hoặc là 1 tương ứng với việc điểm dữ liệu thuộc 1 cụm duy nhất. FCM thì membership grade có giá trị từ 0 đến 1, chỉ ra phần trăm khả năng mà nó thuộc về cụm đang xét.

- Trong FCM, chúng ta đi cực tiểu hóa hàm sau:

$$\sum_{j=1}^{k} \sum_{x \in cj} u_{ij}^{m} \left(x_i - \mu_j\right)^2$$

Trong đó,

+ uij chính là membership grade, là độ đo xác định xem liệu điểm dữ liệu x_i có thuộc về cụm C_j hay không. Tất nhiên với mỗi điểm xác định thì tổng các giá trị membership grade của nó đối với các cụm sẽ bằng 1

+ μ_j là điểm centeroid của cụm j ; m là tham số mờ (fuzziness parameter hoặc fuzzifier).

Tham số mờ xác định độ mờ của việc phân cụm, m có giá trị thực nằm trong khoảng từ 1 đến vô cùng. Nếu m = 1, FCM trở thành phân cụm cứng như k-means.

- Giá trị u_{ij}^{m} được xác định bởi công thức sau:

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

- Tập hợp các giá trị u_{ij}^{m} tạo thành ma trận U

Ma trận U thay đổi qua từng vòng lặp

- Centroid của mỗi cụm được xác định bằng công thức sau:

$$C_j = \frac{\sum_{x \in C_j} u_{ij}^m x}{\sum_{x \in C_j} u_{ij}^m}$$

Trong đó, Cj là centroid của cụm j

- Thuật toán FCM diễn ra như sau:
    1. Chọn số cụm k và chọn tham số m. Khởi tạo ma trận U: $U^{(0)}$
    2. Tính toán tâm của các cụm cho mỗi bước lặp

$$C_j = \frac{\sum_{x \in C_j} u_{ij}^m x}{\sum_{x \in C_j} u_{ij}^m}$$

    3. Cập nhật lại ma trận U theo công thức

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

    4. Kiểm tra điều kiện hội tụ: sự thay đổi của ma trận U không đáng kể

    Nếu chưa hội tụ, quay lại bước 2.

Các vấn đề cần chú ý của Fuzzy C-means:

+ Chọn số cụm k như thế nào

Số cụm k sẽ được chọn như k-means. Chiến lược FCM áp dụng trong bài toán này sẽ là: với một điểm dữ liệu x, ta sẽ gán x cho cụm mà nó có khả năng lớn nhất thuộc vào, hay membership grade của nó với

cụm đó là cao nhất. Việc này cho ta một giải thuật phân cụm cứng như k-means, từ đó cách chọn số cụm k sẽ tương tự như k-means đã trình bày ở trên

+ Đo độ hiệu quả như thế nào

Như đã nói, với một điểm dữ liệu x, ta sẽ gán x cho cụm mà nó có khả năng lớn nhất thuộc vào, hay membership grade của nó với cụm đó là cao nhất. Việc này cho ta một giải thuật phân cụm cứng như k-means, từ đó cách đánh giá độ hiệu quả sẽ tương tự như k-means đã trình bày ở trên

+ Xác định giá trị của siêu tham số m như thế nào

Đây là một vấn đề rất thách thức. Theo một số nghiên cứu, giá trị m nên sử dụng thuộc khoảng [1,5;4]

+ Sự thay đổi "không đáng kể" là như thế nào?

Sự thay đổi không đáng kể của ma trận U cũng là vấn đề nan giải. Nó quyết định trực tiếp đến thời gian mà thuật toán chạy, số vòng lặp của thuật toán và độ chính xác. Một giá trị e được xác định, nếu mà ma trận U thay đổi không quá giá trị e thì thuật toán dừng. Vấn đề xảy ra, nếu e quá nhỏ thì thuật toán hội tụ rất chậm; nếu e quá lớn thì độ chính xác của thuật toán giảm.

Giải pháp đặt ra là chọn một e vừa đủ nhỏ, đồng thời giới hạn số vòng lặp tối đa cho thuật toán

Áp dụng vào bài toán thực tế

Đối với bài toán 1

Thực hiện chia tập dữ liệu thành 2 bên: một bên có trong mối quan hệ tình cảm, một bên không. Thực hiện việc hiểu cơ bản 2 tập này (% giới tính, % sống ở thành thị, nông thôn,…)

Sử dụng phương pháp elbow để quyết định số cụm tối ưu đối với mỗi phần của tập dữ liệu.

Thực hiện phân cụm Fuzzy C-means với điều kiện: với một điểm dữ liệu x, ta sẽ gán x cho cụm mà nó có khả năng lớn nhất thuộc vào, hay membership grade của nó với cụm đó là cao nhất.

Rút ra kết luận dựa trên một số thuộc tính.

Đối với bài toán 2

Thực hiện chia tập dữ liệu thành 2 bên: một bên được sự hỗ trợ giáo dục của bố mẹ, một bên không. Thực hiện việc hiểu cơ bản 2 tập này (% giới tính, % sống ở thành thị, nông thôn,…)

Sử dụng phương pháp elbow để quyết định số cụm tối ưu đối với mỗi phần của tập dữ liệu.

Thực hiện phân cụm Fuzzy C-means với điều kiện: với một điểm dữ liệu x, ta sẽ gán x cho cụm mà nó có khả năng lớn nhất thuộc vào, hay membership grade của nó với cụm đó là cao nhất.

Rút ra kết luận dựa trên một số thuộc tính.