BỘ GIÁO DỤC VÀ ĐÀO TẠO

TRƯỜNG ĐẠI HỌC CÔNG NGHỆ TP. HCM

NGÔ MINH TRÍ

ỨNG DỤNG GOM CỤM FUZZY C-MEANS TRONG PHÂN TÍCH DỮ LIỆU MARKETING

LUẬN VĂN THẠC SĨ

Chuyên ngành: Công Nghệ Thông Tin

Mã ngành: 60480201

\

TP. HỒ CHÍ MINH, tháng 03 năm 2016

BỘ GIÁO DỤC VÀ ĐÀO TẠO

TRƯỜNG ĐẠI HỌC CÔNG NGHỆ TP. HCM

NGÔ MINH TRÍ

ỨNG DỤNG GOM CỤM FUZZY C-MEANS TRONG PHÂN TÍCH DỮ LIỆU MARKETING

LUẬN VĂN THẠC SĨ

Chuyên ngành: Công Nghệ Thông Tin

Mã ngành: 60480201

Cán bộ hướng dẫn khoa học:

PGS TS QUẢN THÀNH THƠ

TP. HỒ CHÍ MINH, tháng 03 năm 2016.

CÔNG TRÌNH ĐƯỢC HOÀN THÀNH TẠI
TRƯỜNG ĐẠI HỌC CÔNG NGHỆ TP. HCM

Cán bộ hướng dẫn khoa học : PSG.TS Quản Thành Thơ
(Ghi rõ họ, tên, học hàm, học vị và chữ ký)

Luận văn Thạc sĩ được bảo vệ tại Trường Đại học Công nghệ TP. HCM ngày 20 tháng 03 năm 2016

Thành phần Hội đồng đánh giá Luận văn Thạc sĩ gồm:
(Ghi rõ họ, tên, học hàm, học vị của Hội đồng chấm bảo vệ Luận văn Thạc sĩ)

TT	**Họ và tên**	**Chức danh Hội đồng**
1	GS.TSKH. Hoàng Văn Kiếm	Chủ tịch
2	PGS.TS Võ Đình Bảy	Phản biện 1
3	TS. Lê Văn Quốc Anh	Phản biện 2
4	TS. Lê Tuấn Anh	Ủy viên
5	TS. Nguyễn Thị Thúy Loan	Ủy viên, Thư ký

Xác nhận của Chủ tịch Hội đồng đánh giá Luận sau khi Luận văn đã được sửa chữa (nếu có).

Chủ tịch Hội đồng đánh giá LV

TRƯỜNG ĐH CÔNG NGHỆ TP. HCM
PHÒNG QLKH – ĐTSĐH

CỘNG HÒA XÃ HỘI CHỦ NGHĨA VIỆT NAM
Độc lập – Tự do – Hạnh phúc

TP. HCM, ngày..... tháng….. năm 2015

NHIỆM VỤ LUẬN VĂN THẠC SĨ

Họ tên học viên : Ngô Minh Trí — Giới tính : Nam.

Ngày, tháng, năm sinh : 25 – 11 - 1990 — Nơi sinh : Tây Ninh.

Chuyên ngành : Công Nghệ Thông Tin — MSHV : 1341860055

I - Tên đề tài:

ỨNG DỤNG GOM CỤM FUZZY C-MEANS TRONG PHÂN TÍCH DỮ LIỆU MARKETING

II- Nhiệm vụ và nội dung:

Nghiên cứu thuật toán Fuzzy C-Means và xây dụng ứng dụng phân tích dữ liệu trong marketing.

III - Ngày giao nhiệm vụ: 03/04/2015

IV- Ngày hoàn thành nhiệm vụ: 15/12/2015

V- Cán bộ hướng dẫn: PGS.TS. Quản Thành Thơ

CÁN BỘ HƯỚNG DẪN — **KHOA QUẢN LÝ CHUYÊN NGÀNH**

(Họ tên và chữ ký) — (Họ tên và chữ ký)

PGS.TS. Quản Thành Thơ

LỜI CAM ĐOAN

Tôi xin cam đoan đây là công trình nghiên cứu của riêng tôi. Các số liệu, kết quả đánh giá, nhận xét và các đề xuất cải tiến mới nêu trong Luận văn là trung thực và chưa từng được ai công bố trong bất kỳ công trình nào khác.

Tôi xin cam đoan rằng mọi sự giúp đỡ cho việc thực hiện Luận văn này cũng như các trích dẫn hay tài liệu học thuật tham khảo đã được cảm ơn đến tác giả hay ghi rõ ràng nguồn gốc thông tin trích dẫn trong Luận văn.

Học viên thực hiện Luận văn

Ngô Minh Trí

LỜI CÁM ƠN

Trước hết, cho tôi được gửi lời cảm ơn đến sự hướng dẫn và giúp đỡ tận tình của PGS.TS. Quản Thành Thơ.

Xin gửi lời cảm ơn đến toàn thể quý thầy cô Khoa Công Nghệ Thông Tin, Trung Tâm Tin Học Đại Học Công Nghệ TP. HCM đã sát cánh và cung cấp cho tôi những kiến thức quí báu trong suốt thời gian học tập và nghiên cứu thực hiện luận văn.

Tôi cũng xin gởi lời cảm ơn đến gia đình, bạn bè và những người thân đã luôn quan tâm và giúp đỡ tôi trong suốt thời gian học tập và nghiên cứu hoàn thành luận văn này.

Luận văn không thể tránh khỏi những sai sót, rất mong nhận được ý kiến đóng góp của mọi người cho luận văn được hoàn thiện hơn.

Tôi xin chân thành cảm ơn.

TP. Hồ Chí Minh, tháng 12 năm 2015

NGÔ MINH TRÍ.

TÓM TẮT

Trong bối cảnh thị trường kinh tế ngày càng cạnh tranh, Marketing mang lại mối quan hệ và lợi ích cộng hưởng giữa người tiêu dùng và doanh nghiệp. Marketing truyền thống đang dần được thay thế bằng marketing hiện đại. Bên cạnh việc cải tiến, tiêu thụ sản phẩm của truyền thống, ngày nay doanh nghiệp hướng đến nhu cầu khách hàng, tập trung vào thị trường nhất định. Điều này giúp các doanh nghiệp có lợi nhuận ổn định và chiến lược kinh doanh lâu dài.

Nghiên cứu này tập trung vào việc khai thác dữ liệu lớn và đa dạng của marketing. Dùng kỹ thuật gom cụm để phân loại dữ liệu vào các cụm. Phân tích mối quan hệ giữa cụm và dữ liệu từ đó khám phá ra được tri thức mới. Phương pháp gom cụm Fuzzy C-Means được chọn do ưu điểm mềm dẻo để xác định dữ liệu có thể thuộc một cụm hoặc nhiều cụm mà phương pháp gom cụm truyền thống chưa đáp ứng được.

Đề tài "Ứng dụng gom cụm Fuzzy C-Means trong phân tích dữ liệu marketing" sẽ giúp phân tích dữ liệu phức tạp của marketing.

Kết quả thực nghiệm cho thấy từ tập dữ liệu mẫu chưa được phân loại với các thuộc tính đều có vai trò như nhau. Sau khi chương trình phân tích xử lý, bằng cách dùng vector trọng số, người dùng tuỳ theo nhu cầu có thể phân loại dữ liệu vào các cụm khác nhau. Từ đó giúp cho việc phân tích dữ liệu của họ đơn giản hơn.

ABSTRACT

Now, business environment competitive competitive. Marketing provides relationship and benefits between consumers and businesses. Marketing traditions were replaced by modern marketing. Besides improvements, consumption products of Traditional business. In today's global business economy, understand your customers and focusing on certain markets are necessary. This will help businesses have stable profits and long-term business strategy.

This study of data mining for marketing. Using clustering techniques for classifying data into clusters. Analysis of the relationship between clusters and data that found new knowledge. Method of Fuzzy C-Means clustering are selected because of their competitive flexibility to identify the data of a cluster or multiple clusters that traditional clustering methods can not do it.

Project of "Application Fuzzy C-Means clustering in marketing data analysis" will help analyze of marketing data.

Practical results present the sample dataset with the attributes with the same role.After processing program, using vector space model, depending on user can be classified data into different clusters. That works their data analysis simpler.

MỤC LỤC

DANH MỤC CÁC TỪ VIẾT TẮT

Ký hiệu, viết tắt	Ý nghĩa tiếng Việt	Ý nghĩa tiếng anh
CSDL	Cơ sở dữ liệu	Database (DB)
KTDL	Khai thác dữ liệu	Data Mining
TF	Tần số xuất hiện 1 từ trong văn bản	Term Frequency
IDF	Tần số nghịch của 1 từ trong văn bản	Inverse Document Frequency
FCM	Fuzzy C-Means	Fuzzy C-Means

DANH MỤC CÁC BẢNG

DANH MỤC CÁC HÌNH

CHƯƠNG 1 GIỚI THIỆU

1.1/ Giới thiệu đề tài:

Dữ liệu lớn ngày càng tăng khối lượng, vận tốc, và tăng về chủng loại. Đối với các tổ chức marketing, dữ liệu lớn là kết quả cơ bản trong môi trường marketing hiện đại, được sinh ra từ thế giới kỹ thuật số của chúng ta hiện nay.

Phương pháp truyền thống của việc thu thập dữ liệu khách hàng thì thường thông qua giao dịch mua bán, gặp gỡ trực tiếp. Trong khi đó, dữ liệu khách hàng ngày nay đa dạng hơn, được thu thập bằng nhiều nguồn như: dữ liệu mua hàng trực tuyến, tỷ lệ nhấp chuột, lịch sử duyệt web, phương tiện truyền thông tương tác xã hội, thiết bị di động, dữ liệu định vị địa lý… Qua đó, các tổ chức marketing nhận thức được dữ liệu của họ đang ngày một phát triển nhiều hơn.

Các dữ liệu thu thập được ngày càng trở nên đồ sộ, phức tạp dẫn đến sự thiếu hụt hoặc không chính xác của dữ liệu được thu thập. Việc kết hợp phương pháp phân cụm với lý thuyết tập mờ là một bước đi quan trọng làm tăng độ chính xác và hiệu năng của phương pháp, hỗ trợ đắc lực trong việc trích xuất các thông tin và mẫu hữu ích từ các dữ liệu.

1.1/ Tính cấp thiết của đề tài:

Ở một nước đang phát triển như Việt Nam, số lượng doanh nghiệp liên tục tăng cao tạo nên sự cạnh tranh gây gắt. Các doanh nghiệp cần phải nhanh chóng đưa ra chiến lược, giải pháp kinh doanh. Để các người quản lý có thể đưa ra những quyết định khả thi, hiệu quả thì nguồn dữ liệu đóng vai trò khá quan trọng. Khi làm chủ được dữ liệu lớn thì họ sẽ có cơ hội thành công lớn hơn.

Nhà quản trị không có nhiều thời gian cũng như điều kiện để nghiên cứu thông tin được thu thập, nên việc xây dựng ứng dụng hỗ trợ phân tích dữ liệu là cần thiết. Ví dụ: xác định các nhóm khách hàng (khách hàng tiềm năng, khách hàng giá trị, phân loại và dự đoán hành vi khách hàng,…). Bằng việc phân tích mối quan hệ giữa cụm và dữ liệu giúp tìm ra dữ liệu mà nhà quản trị mong muốn. Điều này có thể giúp công ty có chiến lược kinh doanh hiệu quả hơn. Phương pháp gom cụm Fuzzy C-Means được chọn do ưu điểm mềm dẻo để xác định dữ liệu có thể thuộc một cụm hoặc nhiều cụm mà phương pháp gom cụm truyền thống chưa

đáp ứng được. Do đó, em đã chọn đề tài "Ứng dụng gom cụm Fuzzy C-Means trong phân tích dữ liệu marketing".

1.2/ Mục tiêu của đề tài:

Đề tài "Ứng dụng gom cụm Fuzzy C-Means trong phân tích dữ liệu marketing" sẽ giúp người sử dụng phân loại theo đặc tính của dữ liệu, theo chức năng của dữ liệu, thấy trực quan vùng dữ liệu được thu thập. Việc phân loại dữ liệu sẽ đảm bảo đầy đủ thông tin, dễ thao tác với dữ liệu cho các dự án nghiên cứu được đặt ra.

1.3/ Cấu trúc luận văn:

Chương 1: Giới thiệu

Giới thiệu lý do chọn đề tài, tính cấp thiết, mục tiêu và cấu trúc của luận văn.

Chương 2: Tổng quan

Giới thiệu tổng quan về nghiên cứu marketing. Các khái niệm về thu thập dữ liệu, về phân loại dữ liệu của marketing. Trình bày khái quát về gom cụm và các khái niệm khi tiếp cận kỹ thuật gom cụm.

Chương 3: Cơ sở lý thuyết

Khảo sát sơ lược tình hình nghiên cứu hiện nay. Trình bày lý thuyết logic mờ, lý thuyết về gom cụm. Phân tích đánh giá thuật toán K-Means và thuật toán Fuzzy C-Means.

Chương 4: Chương trình Fuzzy C-Means

Trình bày chi tiết về các phương pháp đã áp dụng trong quá trình nghiên cứu và xây dựng chương trình. Mô tả kiến trúc hệ thống cũng như cách thức hoạt động của ứng dụng

Chương 5: Thực nghiệm và đánh giá

Trình bày kết quả thực nghiệm và đánh giá phương pháp. Luận văn tiến hành thực nghiệm dựa trên phân tích kết quả mà ứng dụng Fuzzy C-Means thu được.

Chương 6: Kết luận và hướng phát triển

Tổng kết nội dung chính của luận văn và trình bày hướng phát triển trong tương lai.

CHƯƠNG 2 TỔNG QUAN

2.1/ Nghiên cứu marketing:

Nghiên cứu Marketing là quá trình thu thập và phân tích có mục đích, có hệ thống những thông tin liên quan đến việc xác định hoặc đưa ra giải pháp cho các vấn đề liên quan đến lĩnh vực marketing.

Nghiên cứu marketing có một vai trò rất quan trọng trong hoạt động marketing của doanh nghiệp.Nghiên cứu marketing giúp cho nhà quản trị marketing đánh giá được nhu cầu về các thông tin và cung cấp các thông tin hữu ích về các nhóm khách hàng, sự phù hợp của các biến số marketing hiện tại của doanh nghiệp cũng như các biến số môi trường không thể kiểm soát được để từ đó xây dựng và thực hiện chiến lược và các chương trình Marketing nhằm thỏa mãn những nhu cầu của khách hàng.

Nghiên cứu marketing thường được thực hiện theo một tiến trình gồm 7 bước bao gồm: (1) nhân diện vấn đề, (3) xác định mục tiêu nghiên cứu, (3) đánh giá giá trị thông tin, (4) thiết kế nghiên cứu, (5) tổ chức thu thập dữ liệu, (6) chuẩn bị, phân tích và diễn giải dữ liệu, (7) viết và trình bày báo cáo.

2.1.1/ Thu thập dữ liệu:

Thu thập dữ liệu là một bước trong quá trình nghiên cứu marketing. Xác định đầy đủ những dữ liệu nào cần thu thập và phương pháp nào được sử dụng để thu thập dữ liệu là một công việc quan trọng của nhà nghiên cứu để đảm bảo có được đầy đủ thông tin mà dự án nghiên cứu đặt ra.

Dữ liệu thu thập bao gồm nhiều loại khác nhau. Người ta có thể phân loại dữ liệu theo đặc tính của dữ liệu, theo chức năng của dữ liệu, theo địa điểm thu thập dữ liệu hoặc theo nguồn thu thập dữ liệu. Khi xác định dữ liệu cần thu thập, để đạt được mục tiêu nghiên cứu, cần phải tuân thủ theo ba yêu cầu: thông tin chứa trong dữ liệu phải phù hợp và đủ làm rõ mục tiêu nghiên cứu; dữ liệu phải xác thực trên hai phương diện độ tin cậy và giá trị và các dữ liệu phải đảm bảo thu thập nhanh với chi phí chấp nhận được.

2.1.2/ Phân loại dữ liệu:

Dữ liệu trong nghiên cứu marketing có thể thu thập từ nhiều nguồn khác nhau. Được phân thành 2 loại dữ liệu như sau:

2.1.2.1/ Dữ liệu thứ cấp:

Dữ liệu thứ cấp là loại dữ liệu được sưu tập sẵn, đã công bố nên dễ thu thập, ít tốn thời gian, tiền bạc trong quá trình thu thập.

Dữ liệu thứ cấp có vai trò quan trọng trong nghiên cứu marketing không chỉ vì các dữ liệu thứ cấp có thể giúp có ngay các thong tin để giải quyết nhanh chóng vấn đề trong một số trường hợp, nó còn giúp xác định hoặc làm rõ vấn đề và hình thành các giả thiết nghiên cứu, làm cơ sở để hoạch định thu thập dữ liệu sơ cấp. Tuy nhiên khi sử dụng dữ liệu thứ cấp phải đánh giá giá trị của nó theo các tiêu chuẩn như tính cụ thể, tính chính xác, tính thời sự và mục đích thu thập của dữ liệu thứ cấp đó. Có hai nguồn cung cấp dữ liệu thứ cấp là nguồn dữ liệu thứ cấp bên trong và nguồn dữ liệu thứ cấp bên ngoài doanh nghiệp. Dữ liệu nghiệp bên trong có thể là báo cáo về doanh thu bán hàng, chi phí bán hàng và các chi phí khác, hồ sơ khách hàng...Dữ liệu thứ cấp bên ngoài là các tài liệu đã được xuất bản có được từ các nghiệp đoàn, chính phủ, chính quyền địa phương, các tổ chức phi chính phủ, các hiệp hội thương mại, các tổ chức chuyên môn, các ấn phẩm thương mại, các tổ chức nghiên cứu Marketing chuyên nghiệp…

2.1.2.2/ Dữ liệu sơ cấp:

Dữ liệu sơ cấp có thể thu thập từ việc quan sát, ghi chép hoặc tiếp xúc trực tiếp với đối tượng điều tra; cũng có thể sử dụng các phương pháp thử nghiệm để thu thập dữ liệu sơ cấp.

Các dữ liệu sơ cấp được thu thập trực tiếp từ đối tượng nghiên cứu, có thể là người tiêu dùng,nhóm người tiêu dùng... Nó còn được gọi là các dữ liệu gốc, chưa được xử lý. Vì vậy, các dữ liệu sơ cấp giúp người nghiên cứu đi sâu vào đối tượng nghiên cứu, tìm hiểu động cơ của khách hàng,phát hiện các quan hệ trong đối tượng nghiên cứu. Dữ liệu sơ cấp được thu thập trực tiếp nên độ chính xác khá cao, đảm bảo tính cập nhật nhưng lại mất thời gian và tốn kém chi phí để thu thập.

Dữ liệu sơ cấp có thể được thu thập bằng các phương pháp nghiên cứu khác nhau. Mỗi phương pháp có những ưu điểm và hạn chế nhất định, do vậy phù hợp với những dự án

nghiên cứu nhất định. Các phương pháp nghiên cứu bao gồm: Nghiên cứu định tính, quan sát, phỏng vấn và thử nghiệm. Các nghiên cứu định tính bao gồm phỏng vấn nhóm, phỏng vấn chuyên sâu và kỹ thuật hiện hình. Phương pháp quan sát có thể được thực hiện bằng con người hoặc thiết bị. Các phương pháp phỏng vấn bao gồm phỏng vấn cá nhân trực tiếp, phỏng vấn nhóm cố định, phỏng vấn bằng điện thoại, phỏng vấn bằng thư tín. Phương pháp thử nghiệm có thể được thực trong phòng thí nghiệm hoặc thực hiện tại hiện trường. Khi thực hiện các cuộc thử nghiệm chúng ta có thể phải chịu sai lệch trong kết quả do các nguyên nhân: lịch sử, lỗi thời, bỏ ngang, hiệu ứng thử nghiệm, công cụ đo lường hoặc lấy mẫu. Do vậy, việc tổ chức một cuộc thử nghiệm cần phải chuẩn bị tốt, lường trước những sai lầm có thể xảy ra và có hướng khắc phục.

2.2/ Tổng quan về gom cụm:

2.2.1/ Các khái niệm:

Trong ngữ cảnh của ngành khoa học máy tính, người ta quan niệm rằng dữ liệu là các con số, ký hiệu, chữ cái, hình ảnh, âm thanh, ... mà máy tính có thể tiếp nhận và xử lý. Còn thông tin là tất cả những gì mà con người có thể cảm nhận được một cách trực tiếp thông qua các giác quan hoặc gián tiếp thông qua các phương tiện kỹ thuật như tivi, radio, cassette, máy tính, ... Khi dữ liệu được tổ chức lại có cấu trúc hơn, được xử lý và mang đến cho con người những ý nghĩa, hiểu biết nào đó nó trở thành thông tin. **Tri thức** là các thông tin tích hợp, bao gồm các sự kiện và mối quan hệ giữa chúng, đã được nhận thức, khám phá, hoặc nghiên cứu. Tri thức có thể được xem như là dữ liệu trừu tượng và tổng quát ở mức độ cao.

Khám phá tri thức là việc rút trích ra các tri thức chưa được nhận ra, tiềm ẩn trong các tập dữ liệu lớn một cách tự động [1]. Khám phá tri thức trong CSDL là một quá trình gồm một loạt các bước phân tích dữ liệu nhằm rút ra được các thông tin có ích, xác định được các giá trị, quy luật tiềm ẩn trong các khuôn mẫu hay mô hình dữ liệu.

Khai thác dữ liệu là một bước trong quá trình khám phá tri thức, gồm các thuật toán khai thác dữ liệu chuyên dùng với một số quy định về hiệu quả tính toán chấp nhận được để tìm ra các mẫu, các mô hình dữ liệu hoặc các thông tin có ích. Nói cách khác, mục tiêu của khai thác dữ liệu là rút trích ra những thông tin có giá trị tồn tại trong CSDL nhưng ẩn trong khối lượng lớn dữ liệu.

Phân cụm dữ liệu là một kỹ thuật trong khai phá dữ liệu, nhằm tìm kiếm, phát hiện các cụm, các mẫu dữ liệu tiềm ẩn, quan trọng trong tập dữ liệu lớn từ đó cung cấp thông tin, tri thức hữu ích.

Quá trình nhóm các đối tượng dữ liệu theo nguyên tắc: Các đối tượng trong cùng một nhóm thì tương đồng hơn so với các đối tượng khác nhóm. Trong máy học, phân cụm dữ liệu được xem là vấn đề học không có giám sát, vì nó phải giải quyết vấn đề tìm một cấu trúc trong tập hợp dữ liệu chưa biết trước các thông tin về lớp hay các thông tin về tập huấn luyện. Trong quá trình huấn luyện dữ liệu, phân cụm dữ liệu sẽ khởi tạo các lớp cho phân lớp bằng cách xác định các nhãn cho các nhóm dữ liệu.

2.2.2/ Một số khái niệm khi tiếp cận phân cụm dữ liệu:

2.2.2.1/ Một số phương pháp phân cụm điển hình:

Các kỹ thuật phân cụm có rất nhiều cách tiếp cận và dựa trên các thuật toán ứng dụng, nó đều hướng tới hai mục tiêu chung đó là chất lượng của các cụm khám phá được và tốc độ thực hiện của thuật toán.

Hiện nay, thuật toán gom cụm có thể được phân chia vào 4 nhóm sau :

- Gom cụm chọn lọc (K-Means, K-Medoids, ISODATA, K-Nearest Neighbors)
- Gom cụm mờ (Fuzzy C-Means, Fuzzy C-Ellipse, Fuzzy C-Mixed)
- Gom cụm phân cấp (Single-link, Complete-link)
- Gom cụm xác suất (COBWEB)

2.2.2.2/ Độ đo tương tự và phi tương tự:

Để phân cụm, người ta phải đi tìm cách thích hợp để xác định khoảng cách giữa các đối tượng, hay là phép đo tương tự dữ liệu. Đây là các hàm để đo sự giống nhau giữa các cặp đối tượng dữ liệu, thông thường các hàm này hoặc là để tính độ tương tự hoặc là tính độ phi tương tự giữa các đối tượng dữ liệu.

1. **Không gian metric:**

Tất cả các độ đo dưới đây được xác định trong không gian độ đo metric. Một không gian metric là một tập trong đó có xác định các 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 được gọi là một không gian metric nếu:

- Với mỗi cặp phần tử x, y thuộc X đều có 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 thoả mãn hệ tính chất sau :

(i)δ(x,y) > 0 nếu x ≠ y ;

(ii) δ(x, y)=0 nếu x =y;

(iii) δ(x,y) = δ(y,x) với mọi x,y;

(iv) δ(x,y) ≤.δ(x,z)+δ(z,y).

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.

2. Thuộc tính khoảng cách:

Sau khi chuẩn hoá, độ đo phi tương tự của hai đối tượng dữ liệu x, y được xác định bằng các metric khoảng cách như sau:

- Khoảng cách Minskowski: $d(x, y) = (\sum_{i=1}^{n} |x_i - y_i|^q)^{1/q}$ trong đó q là số tự nhiên dương.

- Khoảng cách Euclide : $d(x, y) = \sqrt{\sum_{i=1}^{n} |x_i - y_i|^2}$ đây là trường hợp đặc biệt của khoảng cách Minskowski trong trường hợp q=2.

- Khoảng cách Manhattan : $d(x, y) = \sum_{i=1}^{n} |x_i - y_i|$ đây là trường hợp đặc biệt của khoảng cách Minskowski trong trường hợp q=1.

3. Thuộc tính nhị phân :

- α là tổng số các thuộc tính có giá trị là 1 trong x,y.

- β là tổng số các thuộc tính có giá trị là 1 trong x và 0 trong y.

- γ là tổng số các thuộc tính có giá trị là 0 trong x và 1 trong y.

- δ là tổng số các thuộc tính có giá trị là 0 trong x và y.

- $\tau = \alpha + \gamma + \beta + \delta$

Các phép đo độ tương tương đồng đối với dữ liệu thuộc tính nhị phân được định nghĩa như sau :

Hệ số đối sánh đơn giản : $d(x,y) = \frac{\alpha + \delta}{\tau}$ ở đây cả hai đối tượng x và y có vai trò như nhau, nghĩa là chúng đối xứng và có cùng trọng số.

Hệ số Jacard : $d(x,y) = \frac{\alpha}{\alpha + \beta + \gamma}$ (bỏ qua số các đối sánh giữa 0-0). Công thức tính này được sử dụng trong trường hợp mà trọng số của các thuộc tính có giá trị 1 của đối tượng dữ liệu có cao hơn nhiều so với các thuộc tính có giá trị 0, như vậy các thuộc tính nhị phân ở đây là không đối xứng.

4. Thuộc tính định danh :

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

$$d(x,y) = \frac{p - m}{p}$$

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

5. Thuộc tính tỉ 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. Hoặc loại bỏ đơn vị đo của các thuộc tính dữ liệu bằng cách chuẩn hoá chúng, hoặc gán trọng số cho mỗi thuộc tính giá trị trung bình, độ lệch chuẩn. Với mỗi thuộc tính dữ liệu đã được gán trọng số tương ứng wi (1 <= i <= k), độ tương đồng dữ liệu được xác định như sau :

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

Nguyên lý thường được dùng để gom cụm dữ liệu là nguyên tắc cực tiểu khoảng cách (thường là khoảng cách Euclide).

2.2.3/ Các ứng dụng của phân cụm:

Sau đây là một số ứng dụng của phân cụm:

Giảm dữ liệu: Giả sử ta có tập cơ sở dữ liệu lớn. Phân cụm sẽ nhóm các đối tượng dữ liệu này thành cụm dữ liệu nhỏ hơn, dễ nhận thấy.

Dự đoán dựa trên các cụm: Đầu tiên các đối tượng dữ liệu mang đặc điểm chung sẽ được gom nhóm vào một tập dữ liệu. Từ đó hình thành các cụm dữ liệu mang đặc điểm riêng. Sau đó, khi có một đối tượng dữ liệu mang đặc điểm chưa biết ta sẽ xác định xem nó sẽ có khả năng thuộc về cụm dữ liệu nào nhất và dự đoán được một số đặc điểm của đối tượng này nhờ các đặc trưng chung của cả cụm.

CHƯƠNG 3 CƠ SỞ LÝ THUYẾT

3.1/ Đề tài nghiên cứu thế giới:

Singh, Nigam, Pal, Mehrotra (2014) [2] đã áp dụng Fuzzy C-Means trong lĩnh vực nhận diện hình ảnh từ xa. Bài báo đã đề xuất thuật toán kết hợp giữa gom cụm Kohonen và Fuzzy Local Information C-Means (FLICM) để cải thiện hiệu năng và độ chính xác của thuật toán gom cụm. Mục tiêu của thực nghiệm là so sánh thuật toán đề xuất với hai thuật toán FCM và GIFP-FCM. Kết quả thuật toán ”Fuzzy Kohonen Local Information C-Means” (FKLICM) cho khả năng gom cụm chính xác và tối ưu hơn những phương pháp khác.

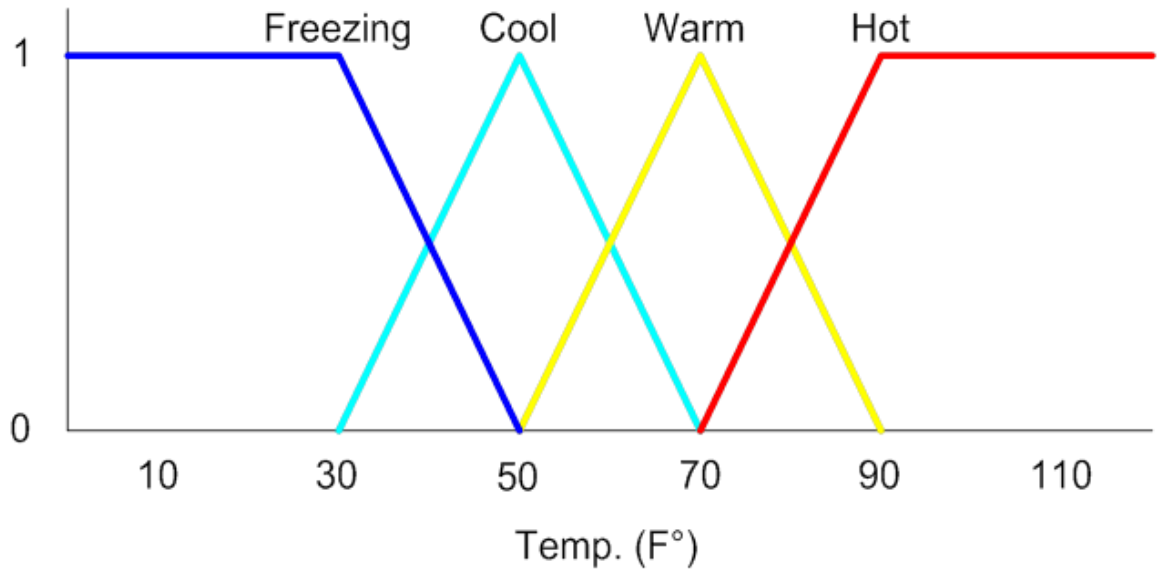
Izakian, Abraham (2011) [3] là một trong thuật toán phổ biến của kĩ thuật gom cụm bởi vì nó hiệu quả, đơn giản và dễ thực hiện. Tuy nhiên. FCM khởi tạo cảm tính. Particle swarm optimization (PSO) là công cụ tối ưu các giá trị ngẫu nhiên được dùng để giải quyết vấn đề trong tối ưu hoá. Bài báo sử dụng kết hợp phương pháp FCM và Fuzzy PSO. Kết quả thật đáng khích lệ bởi hiệu năng phương pháp đề xuất đạt được.

3.2/ Thuật toán Fuzzy C-Means:

3.2.1/ Lý thuyết fuzzy logic:

Lý thuyết fuzzy logic được Zadeh, L.A. nêu ra lần đầu tiên vào năm 1965 [4] . Lý thuyết này giải quyết các bài toán rất gần với cách tư duy của con người. Tới nay, lý thuyết logic mờ đã phát triển rất mạnh mẽ và được ứng dụng trong nhiều lĩnh vực của cuộc sống.

Theo logic truyền thống, một biểu thức logic chỉ nhận một trong hai giá trị: True hoặc False. Khác với lý thuyết logic truyền thống, một biểu thức logic mờ có thể nhận một trong vô số giá trị nằm trong khoảng số thực từ 0 đến 1. Nói cách khác, trong logic truyền thống, một sự kiện chỉ có thể hoặc là đúng (tương đương với True - 1) hoặc là sai (tương đương với False - 0) còn trong logic mờ, mức độ đúng của một sự kiện được đánh giá bằng một số thực có giá trị nằm giữa 0 và 1, tuỳ theo mức độ đúng “nhiều” hay “ít” của nó.

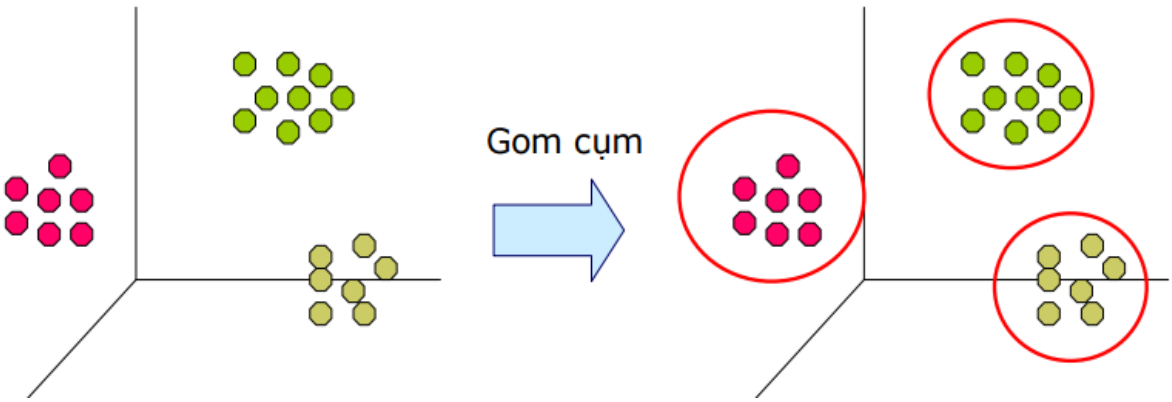


Hình 3-1 Minh hoạ biểu thức logic mờ - diễn tả mức độ nóng lạnh**.**

Giá trị của các biến trong biểu thức logic mờ không phải là các con số mà là các khái niệm, ví dụ như "nhanh", "trung bình", "chậm" hay "nóng", "vừa", "lạnh"... Chính vì vậy cách giải quyết các bài toán trong logic mờ rất gần với cách tư duy của con người

3.2.2/ Lý thuyết gom cụm (Clustering):

Gom cụm dữ liệu là phương pháp phân hoạch tập hợp dữ liệu thành nhiều tập con C sao cho mỗi tập con $c \subset C$ chứa các phần tử có những tính chất giống nhau theo tiêu chuẩn nào đó, mỗi tập con c được gọi là một cụm [5].



Hình 3-2 Minh hoạ khái niệm gom cụm.

Như vậy quá trình gom cụm là một quá trình phân các phần tử $q \in Q$ vào trong các cụm $c \subset C$.

Nguyên lý thường được dùng để gom cụm dữ liệu là nguyên tắc cực tiểu khoảng cách (thường là khoảng cách Euclide) [6].

Chúng ta sẽ tìm hiểu hai thuật toán K-Means và Fuzzy C-Means.

Trong đó với giải thuật thứ nhất (K-Means), một dữ liệu chỉ có thể thuộc duy nhất vào một phân vùng, trong khi với thuật toán thứ 2 (Fuzzy C-Means), sử dụng logic mờ trong việc phân vùng dữ liệu, mềm dẻo hơn rất nhiều, nó cho phép một lượng dữ liệu có thể thuộc vào 1 hoặc nhiều phân vùng khác nhau tùy mức độ hàm thuộc.

3.2.3/ Thuật toán K-Means:

Thuật toán phân cụm K-Means do Macqueen đề xuất trong lĩnh vực thống kê năm 1967, mục đích của thuật toán là sinh ra k cụm dữ liệu $\{C_1, C_2, \ldots, C_k\}$ từ một tập dữ liệu ban đầu gồm n đối tượng trong không gian d chiều $X_i = (x_{i1}, x_{i2}, \ldots, x_{id},)$ i= $\overline{1,n}$, sao cho hàm tiêu chuẩn $E = \sum_{i=1}^{i} \sum_{x \in C_i} D^2(x - m_i)$ đạt giá trị cực tiểu. Trong đó,m_i là trọng tâm của cụm C_i . D là khoảng cách giữa hai đối tượng.

Ưu điểm của thuật toán K-Means: Đây là một phương pháp đơn giản, hiệu quả, tự tổ chức, được sử dụng trong tiến trình khởi tạo trong nhiều thuật toán khác, hiệu xuất tương đối, thường kết thúc ở tối ưu cục bộ, có thể tìm được tối ưu toàn cục.

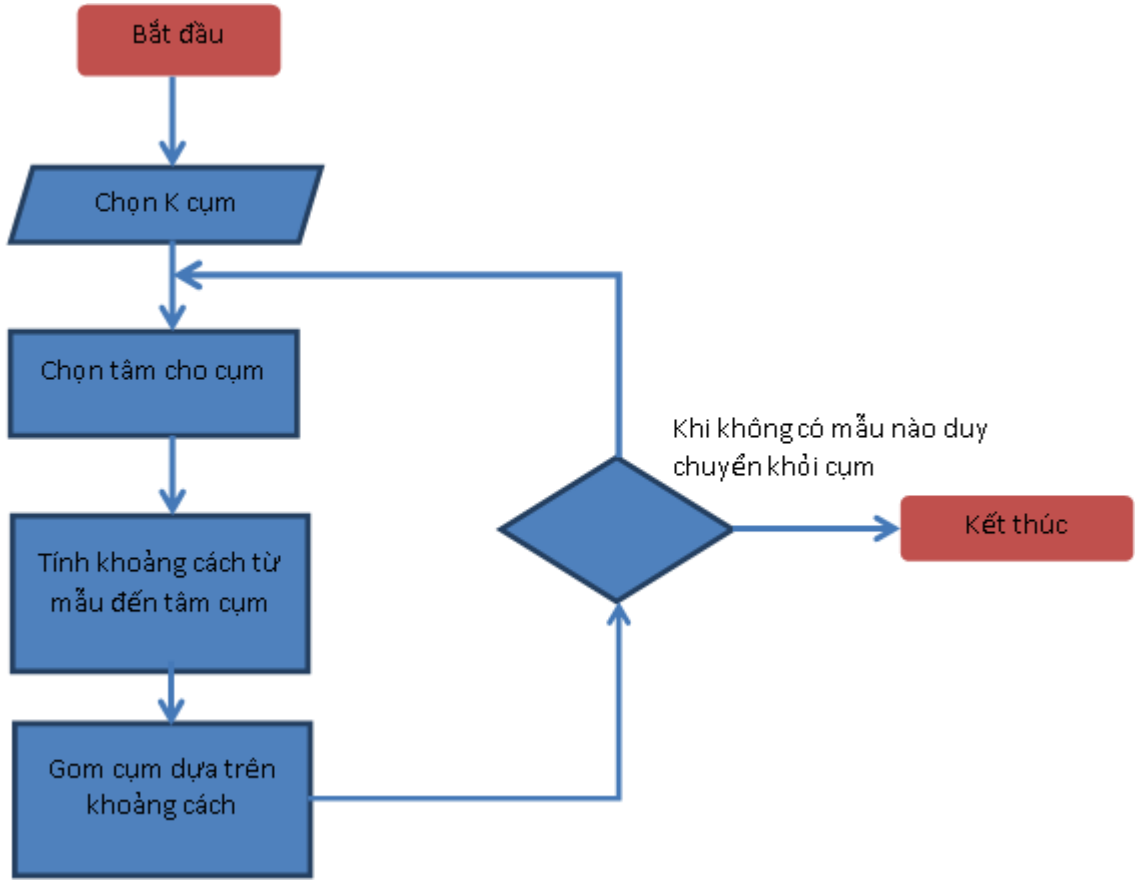
Nhược điểm của thuật toán này: Số cụm k phải được xác định trước,chỉ áp dụng được khi xác định được trị trung bình, không thể xử lý nhiễu và outliers, không thích hợp nhằm khám phá các dạng không lồi hay các cụm có kích thước khác nhau, đây là thuật toán độc lập tuyến tính.

<u>Tư tưởng của thuật toán K-means:</u>

Ý tưởng chính của thuật toán này là áp dụng nguyên lý người láng giềng gần nhất hoặc khoảng cách ngắn nhất theo định luật III Newton, nghĩa là phần tử nào gần điểm tâm của cụm c_i hơn so với các cụm c_j sẽ được gom về cụm ci.

Đầu vào của thuật toán K-Means: Số các cụm k, và CSDL có n số điểm (đối tượng) trong không gian dữ liệu.

Minh họa thuật toán K-means:



Hình 3-3 Lưu đồ của thuật toán K-means.

Các bước của thuật toán K-means:

Bước 1: Chọn ngẫu nhiên k mẫu vào k cụm. Coi tâm của cụm là chính là mẫu có trong cụm.

Bước 2: Tính khoảng cách giữa các mẫu còn lại đến k tâm

Bước 3: Gán các mẫu vào cụm sao cho khoảng cách từ mẫu đến tâm cụm là nhỏ nhất

Bước 4: Nếu các cụm không có sự thay đổi nào sau khi thực hiện bước 3 thì chuyển sang bước 5, ngược lại thì quay lại bước 2.

Bước 5: thuật toán kết thúc.

3.2.4/ Thuật toán Fuzzy C-Means:

Từ những năm 1920, Lukasiewicz đã nghiên cứu cách diễn đạt toán học khái niệm mờ. Năm 1965, Lofti Zadeh đã phát triển lý thuyết khả năng và đề xuất hệ thống logic mờ (fuzzy logic). Kỹ thuật này gom cụm một tập n vectơ đối tượng dữ liệu $X = \{x_1, x_2, \ldots, x_n\} \subset$

R^S thành c các nhóm mờ dựa trên tính toán tối thiểu hóa hàm mục tiêu để đo chất lượng của gom cụm và tìm trung tâm cụm trong mỗi nhóm, sao cho chi phí hàm độ đo độ phi tương tự là nhỏ nhất [7].

3.2.4.1/ Ý tưởng của thuật toán Fuzzy C-Means:

Mỗi phần tử q ∈ V ban đầu được gán cho một tập trọng số W_{qk}, trong đó W_{qk} cho biết khả năng q thuộc về cụm k, Σ(k=1,K) W_{qk} =1. Có nhiều cách tính trọng số Wqk khác nhau, trong đó W_{qk} =1/D_{qk} thường được sử dụng nhất (D_{qk} là khoảng cách từ q đến trọng tâm của cụm k) [8].

Trong quá trình gom cụm trọng số này có thể được cập nhật ở mỗi bước lặp khi trọng tâm của cụm bị thay đổi.

Sau khi kết thúc quá trình gom cụm, một cụm không có mẫn nào sẽ bị loại, do đó số cụm tìm được thường không biết trước.

3.2.4.2/ Thuật toán Fuzzy C-Means kết hợp với vector trọng số:

Các bước của thuật toán Fuzzy C-Means kết hợp véctơ trọng số:

Bước 1: Giả sử không gian dữ liệu gồm n điểm x_i x_i, i = 1..n cần phân hoạch thành c cụm ($2 \le c < n$).

Bước 2: Chọn tham số mờ hóa m > 1[1].

Bước 3: Chọn vectơ trọng số W có k thành phần, k là số thuộc tính của x_i sao cho :

$$\sum_{l=1}^{k} w_l = 1.$$

Bước 4: Khởi tạo ma trận thành viên U (c x n) với $0 \le \mu_{ji} \le 1$ sao cho:

$$\sum_{j=1}^{c} \mu_{ji} = 1, i = 1..n.$$

Bước 5: Tính trọng tâm C_j của cụm j (j = 1…c) gồm k thành phần, mỗi thành phần của nó được tính như sau:

$$c_{jl} = \frac{\sum_{i=1}^{n} \mu_{ji}^{(m)} x_{il} w_l}{\sum_{i=1}^{n} \mu_{ji}^{(m)}}, j = 1..c, l = 1..k.$$

Bước 6: Cập nhật ma trận khoảng cách D (c x n) theo độ đo khoảng cách đã chọn d_{ji} là khoảng cách từ x_i đến C_j

$$d_{ji} = \sqrt{\sum_{l=1}^{k} (x_{il} w_l - c_{jl})^2}, j = 1..c, i1..n$$

Đối với m $\to 1^+$ thì thuật toán Fuzzy C-Means trở thành thuật toán rõ. Chưa có quy tắc nào nhằm lựa chọn tham số m đảm bảo việc phân cụm hiệu quả, thông thường chọn m = 2.

Bước 7: Cập nhật ma trận thành viên U

Nếu $d_{ji} > 0$ thì $\mu_{ji} = \left[\sum_{k=1}^{c} (\frac{d_{ji}}{d_{ki}})^{\frac{2}{m-1}} \right]^{-1}$

Ngược lại nếu $d_{ji} = 0$ thì x_i trùng với trọng tâm C_j của cụm j, $\mu_{ji} = 1$.

Bước 8: Nếu sự thay đổi của ma trận U là đủ nhỏ so với bước kế trước thì chuyển đến bước 9. Ngược lại thì lặp lại từ bước 5.

Để xác định là U thay đổi nhỏ thì có thể dùng một độ đo khoảng cách ma trận nào đó như sai số trung bình, sai số lớn nhất… Ở đây chúng tôi dùng:

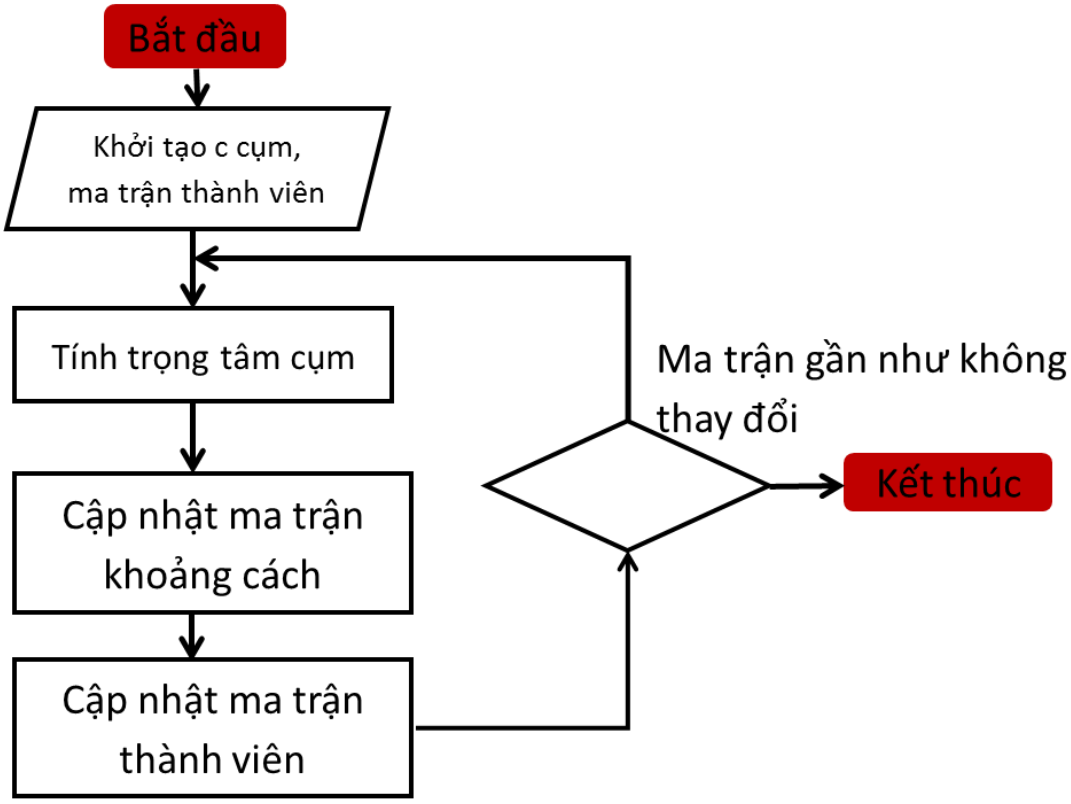
$$\max_{i,j} \left| \mu_{ji}^{(m)} - \mu_{ji}^{(m-1)} \right| < epsilon.$$

Với nghĩa $\mu_{ji}^{(m)}$ là μ_{ji} tại bước lặp thứ m.

Bước 9: Dựa trên ma trận *U*, sắp xếp các điểm dữ liệu x_i, cùng độ thuộc lớn nhất của nó vào các cụm theo qui tắc xét độ thuộc của điểm dữ liệu đó với từng cụm, điểm dữ liệu sẽ thuộc vào cụm nào có độ thuộc lớn nhất, nếu có từ hai độ thuộc lớn nhất bằng nhau trở lên thì chọn một trong số các cụm đó để đưa vào. Thuật toán kết thúc [9].

3.2.4.3/ Minh họa thuật toán Fuzzy C-Means:

Lưu đồ của thuật toán Fuzzy C-Means:



Hình 3-4 Lưu đồ của thuật toán Fuzzy C-Means.

Ví dụ minh họa của thuật toán Fuzzy C-Means:

Input: Bảng dữ liệu ban đầu có 6 điểm:

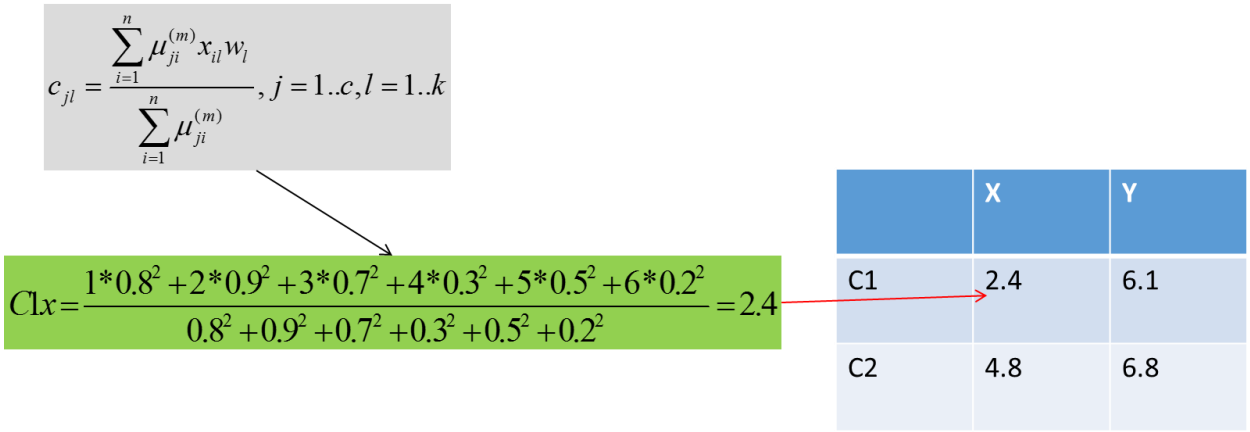
X	Y
1	6
2	5
3	8
4	4
5	7
6	9

Khởi tạo ma trận
thành viên ngẫu nhiên,
chọn số cụm = 2

C1	C2
0.8	0.2
0.9	0.1
0.7	0.3
0.3	0.7
0.5	0.5
0.2	0.8

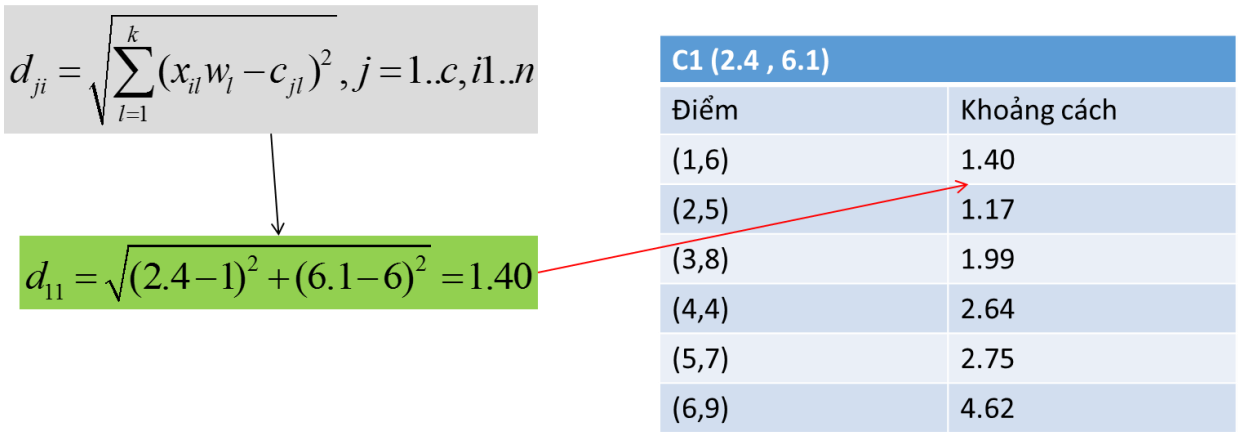
Hình 3-5 Minh họa thuật toán Fuzzy C-Means.

Tính trọng tâm của cụm:



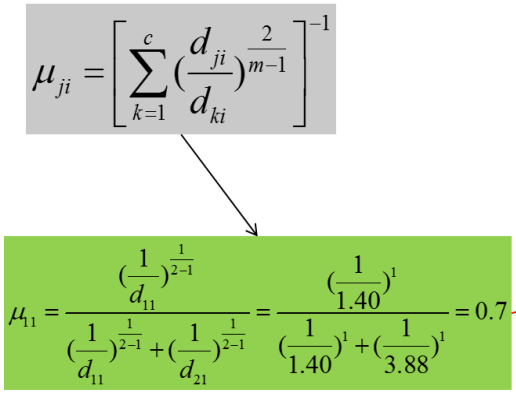
Hình 3-6 Minh họa thuật toán Fuzzy C-Means.

Tính ma trận khoảng cách:



Hình 3-7 Minh họa thuật toán Fuzzy C-Means.

Cập nhật ma trận thành viên:



$$\mu_{ji} = \left[\sum_{k=1}^{c} \left(\frac{d_{ji}}{d_{ki}} \right)^{\frac{2}{m-1}} \right]^{-1}$$

$$\mu_{11} = \frac{\left(\frac{1}{d_{11}}\right)^{\frac{1}{2-1}}}{\left(\frac{1}{d_{11}}\right)^{\frac{1}{2-1}} + \left(\frac{1}{d_{21}}\right)^{\frac{1}{2-1}}} = \frac{\left(\frac{1}{1.40}\right)^{1}}{\left(\frac{1}{1.40}\right)^{1} + \left(\frac{1}{3.88}\right)^{1}} = 0.7$$

X	y	c1	c2
1	6	0.7	0.3
2	5	0.6	0.4
3	8	0.5	0.5
4	4	0.5	0.5
5	7	0.1	0.9
6	9	0.3	0.7

Hình 3-8 Minh họa thuật toán Fuzzy C-Means.

Thuật toán kết thúc khi ma trận thành viên gần như không thay đổi.

3.2.4.4/ Nhận xét thuật toán FCM (Fuzzy C-Means):

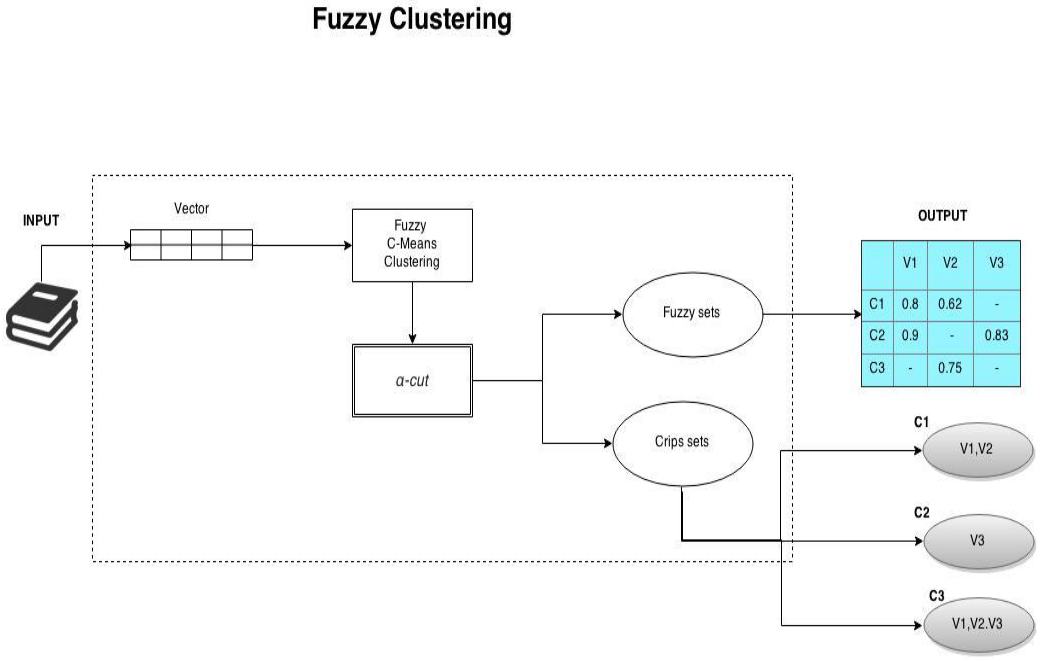
Ưu điểm của thuật toán Fuzzy C-Means: đây là phương pháp dễ thực hiện, có khả năng tìm được tối ưu toàn cục, hiệu năng tốt tương đương K Means. Thường được dùng với các vấn đề nhận dạng trong không gian đa chiều.

Nhược điểm của thuật toán Fuzzy C-Means: có công thức tính toán phức tạp, tốc độ hội tụ tuỳ thuộc vào trạng thái ban đầu của ma trận thành viên U và tham số mờ hoá m.

CHƯƠNG 4 HỆ THỐNG PHÂN TÍCH DỮ LIỆU DỰA TRÊN FCM

4.1/ Sơ đồ tổng thể hệ thống:

Chương trình Fuzzy C-Means bao gồm các khối cơ bản sau:



Hình 4-1 Mô tả hoạt động chương trình Fuzzy C-Means.

Data: dùng bộ dữ liệu (dataset) mẫu theo chuẩn UCI để làm cơ sở dữ liệu.

Biến đầu vào:

1- Age: số tuổi.

2- Job: nghề nghiệp (bao gồm: 'admin.','blue-collar','entrepreneur','housemaid','management','retired','self-employed','services','student','technician','unemployed','unknown').

3- Marital: tình trạng hôn nhân (bao gồm: 'divorced','married','single','unknown'; note: 'divorced' means divorced or widowed).

4- Education: trình độ (bao gồm: 'basic.4y','basic.6y','basic.9y','high.school','illiterate','professional.course','university.degree','unknown')

5- Default: thẻ tín dụng (bao gồm: 'no','yes','unknown').

6- Housing: nhà cửa (bao gồm: 'no','yes','unknown').

7- Loan: vay mượn (bao gồm: 'no','yes','unknown').

8- Contact: liên lạc (bao gồm: 'cellular','telephone').

9- Poutcome: kết quả chiến dịch tiếp thị trước đây (bao gồm: 'failure','nonexistent','success').

Các dữ liệu nhập vào thuộc chương trình tiếp thị trực tiếp trên điện thoại. Mục tiêu của gom cụm là để dự đoán khách hàng có hay không đồng ý đăng ký một khoản tiền gửi kỳ hạn vào ngân hàng.

10- Y: Có đồng ý cho khách hàng gửi tiền ('yes','no').

Bộ dữ liệu được sử dụng sẽ được biểu diễn thành vector không gian với mục đích tính toán và xử lý.

Bảng 4-1 Bảng bộ dữ liệu mẫu.

ID	CONTENT
1	44 "technician" "single" "secondary" "no" "yes" "no" "unknown" 5 "may" 151 1 -1 0 "unknown" "no"
2	33 "entrepreneur" "married" "secondary" "no" "yes" "yes" "unknown" 5 "may" 76 1 -1 0 "unknown" "no"
3	47 "blue-collar" "married" "unknown" "no" "yes" "no" "unknown" 5 "may" 92 1 -1 0 "unknown" "no"
4	33 "unknown" "single" "unknown" "no" "no" "no" "unknown" 5 "may" 198 1 -1 0 "unknown" "no"
5	35 "management" "married" "tertiary" "no" "yes" "no" "unknown" 5 "may" 139 1 -1 0 "unknown" "no"
6	28 "management" "single" "tertiary" "no" "yes" "yes" "unknown" 5 "may" 217 1 -1 0 "unknown" "no"
7	42 "entrepreneur" "divorced" "tertiary" "yes" "yes" "no" "unknown" 5 "may" 380 1 -1 0 "unknown" "no"
8	58 "retired" "married" "primary" "no" "yes" "no" "unknown" 5 "may" 50 1 -1 0 "unknown" "no"
9	43 "technician" "single" "secondary" "no" "yes" "no" "unknown" 5 "may" 55 1 -1 0 "unknown" "no"
10	41 "admin." "divorced" "secondary" "no" "yes" "no" "unknown" 5 "may" 222 1 -1 0 "unknown" "no"

4.2/ Mô hình không gian vector:

Mô hình không gian vector sẽ làm nhiệm vụ đưa tất cả các văn bản trong tập văn bản được mô tả bởi một tập các từ khoá hay còn gọi là các từ chỉ mục sau khi đã loại bỏ các từ ít có ý nghĩa. Các từ chỉ mục này cũng chính là các từ chứa nội dung chính của tập văn bản. Mỗi từ chỉ mục này được gán một trọng số, trọng số của một từ chỉ mục nói lên sự liên quan của nó đến nội dung của một văn bản.

Token hóa là bước xử lý để lọc ra từ khoá đại diện cho văn bản. Để tránh việc xử lý các từ vô dụng, chúng ta sẽ áp dụng một danh sách dừng cho tập các tài liệu trong dataset. Danh sách dừng là tập các từ được cho rằng không liên quan đến nội dung của tài liệu. Ví dụ "a", "the", "of", "for", "with"...là các từ dừng, mặc dù chúng có thể xuất hiện rất thường xuyên trong tài liệu. Ngoài ra, ta có thể thấy rằng một nhóm các từ có thể chia sẻ chung một từ gốc. Do vậy bước tiếp theo chúng ta sẽ định ra các nhóm từ mà trong đó các từ chỉ có sự khác biệt nhỏ về cú pháp. Ví dụ, nhóm các từ "drug", "drugged", và "drugs" sẽ cùng chia sẻ chung một từ gốc là "drug".

Cách tính TF-IDF:

Tần số xuất hiện của 1 từ trong 1 văn bản. Cách tính:

$$tf(t,d) = \frac{f(t,d)}{\max\{f(w,d) : w \in d\}}$$

Thương của số lần xuất hiện 1 từ trong văn bản và số lần xuất hiện nhiều nhất của một từ bất kỳ trong văn bản đó. (giá trị sẽ thuộc khoảng [0, 1])

f(t,d) - số lần xuất hiện từ t trong văn bản d.

max{f(w,d):w∈d} - số lần xuất hiện nhiều nhất của một từ bất kỳ trong văn bản.

Tần số nghịch của 1 từ trong tập văn bản.Tính IDF để giảm giá trị của những từ phổ biến. Mỗi từ chỉ có 1 giá trị IDF duy nhất trong tập văn bản.

$$idf(t,D) = \log \frac{|D|}{|\{d \in D : t \in d\}|}$$

|D| : - tổng số văn bản trong tập D

$|\{d \in D : t \in d\}|$: - số văn bản chứa từ nhất định, với điều kiện t xuất hiện (i.e., tf(t,d) $\neq$ 0). Nếu từ đó không xuất hiện ở bất cứ 1 văn bản nào trong tập thì mẫu số sẽ bằng 0 => phép chia cho không không hợp lệ, vì thế người ta thường thay bằng mẫu thức 1 + $|\{d \in D : t \in d\}|$.

Cơ số logarit trong công thức này không thay đổi giá trị của 1 từ mà chỉ thu hẹp khoảng giá trị của từ đó. Vì thay đổi cơ số sẽ dẫn đến việc giá trị của các từ thay đổi bởi 1 số nhất định và tỷ lệ giữa các trọng lượng với nhau sẽ không thay đổi. (nói cách khác, thay đổi cơ số sẽ không ảnh hưởng đến tỷ lệ giữa các giá trị IDF). Tuy nhiên việc thay đổi khoảng giá trị sẽ giúp tỷ lệ giữa IDF và TF tương đồng để dùng cho công thức TF-IDF như bên dưới.

Trong mô hình vector không gian, TF và IDF sẽ được kết hợp với nhau, được gọi là độ đo TF-IDF:

$$\text{TF-IDF}(t,d,D) = \text{TF}(t,d) \times \text{IDF}(t,D)$$

Những từ có giá trị TF-IDF cao là những từ xuất hiện nhiều trong văn bản này, và xuất hiện ít trong các văn bản khác. Việc này giúp lọc ra những từ phổ biến và giữ lại những từ có giá trị cao (từ khoá của văn bản đó).

V1	(0.27 ; 0.17 ; 0.14 ; 0.06 ; 0 ; 0.01 ; 0 ; 0 ; 0 ; 0.42 ; 0 ; 0 ; 0)
V2	(0.27 ; 0.3 ; 0.05 ; 0.06 ; 0 ; 0.01 ; 0 ; 0 ; 0 ; 0.42 ; 0 ; 0 ; 0)
V3	(0.42 ; 0.21 ; 0.05 ; 0 ; 0 ; 0.01 ; 0 ; 0 ; 0.35 ; 0 ; 0 ; 0 ; 0)
V4	(0.27 ; 0 ; 0.14 ; 0 ; 0 ; 0 ; 0.42 ; 0 ; 0 ; 0 ; 0 ; 0 ; 0)
V5	(0.3 ; 0.18 ; 0.05 ; 0.18 ; 0 ; 0.01 ; 0 ; 0 ; 0 ; 0.42 ; 0 ; 0 ; 0)
V6	(0.35 ; 0.18 ; 0.14 ; 0.18 ; 0 ; 0.01 ; 0 ; 0 ; 0 ; 0.42 ; 0 ; 0 ; 0)
V7	(0.27 ; 0.3 ; 0.23 ; 0.18 ; 0.01 ; 0 ; 0 ; 0 ; 0 ; 0.42 ; 0 ; 0 ; 0)
V8	(0.27 ; 0.23 ; 0.05 ; 0.21 ; 0 ; 0.01 ; 0 ; 0 ; 0 ; 0.35 ; 0 ; 0 ; 0)
V9	(0.42 ; 0.17 ; 0.14 ; 0.06 ; 0 ; 0.01 ; 0 ; 0 ; 0 ; 0.27 ; 0 ; 0 ; 0)
V10	(0.42 ; 0.2 ; 0.23 ; 0.06 ; 0 ; 0.01 ; 0 ; 0 ; 0 ; 0.42 ; 0 ; 0 ; 0)

Hình 4-2 Dòng dữ liệu vector trọng số.

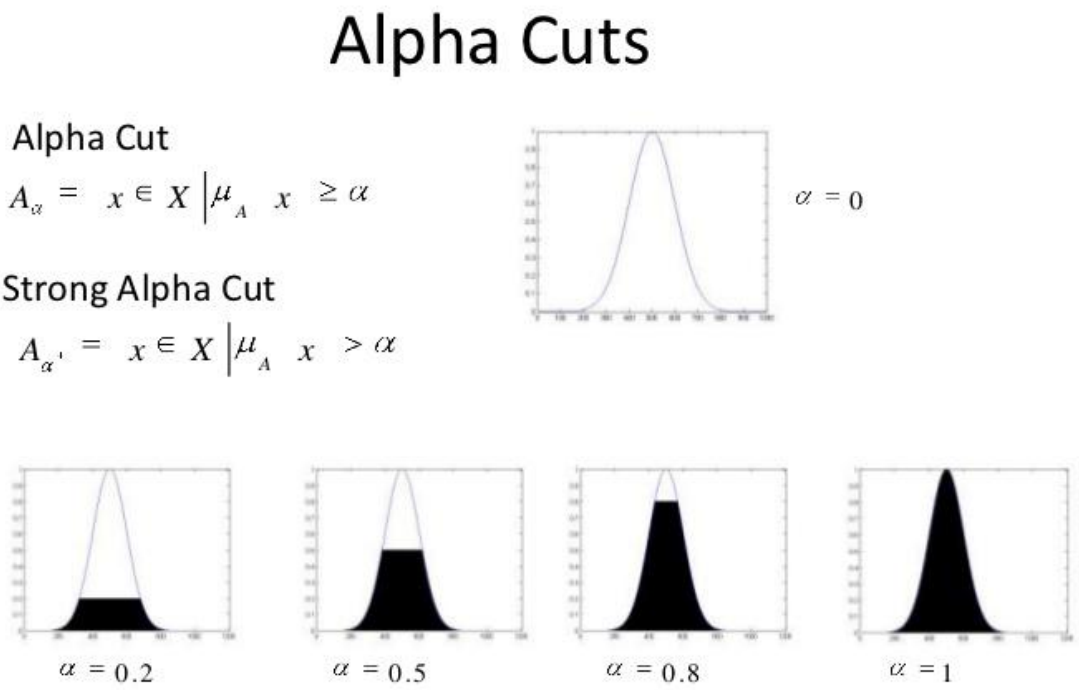
4.3/ Alpha-Cut sets:

Tập A_α được gọi là tập alpha-cut khi hàm thành viên của A thì không nhỏ hơn α ,được định nghĩa như sau:

$$A_\alpha = \{x \in X \mid \mu_A(x) \geq \alpha\}$$

Tập A$_\alpha$ ' được gọi là tập alpha-cut mạnh, được định nghĩa như sau:

$$A_\alpha' = \{x \in X \mid \mu_A(x) > \alpha\}$$



Hình 4-3 Minh hoạ alpha-cut.

4.4/ Crisp sets và Fuzzy sets:

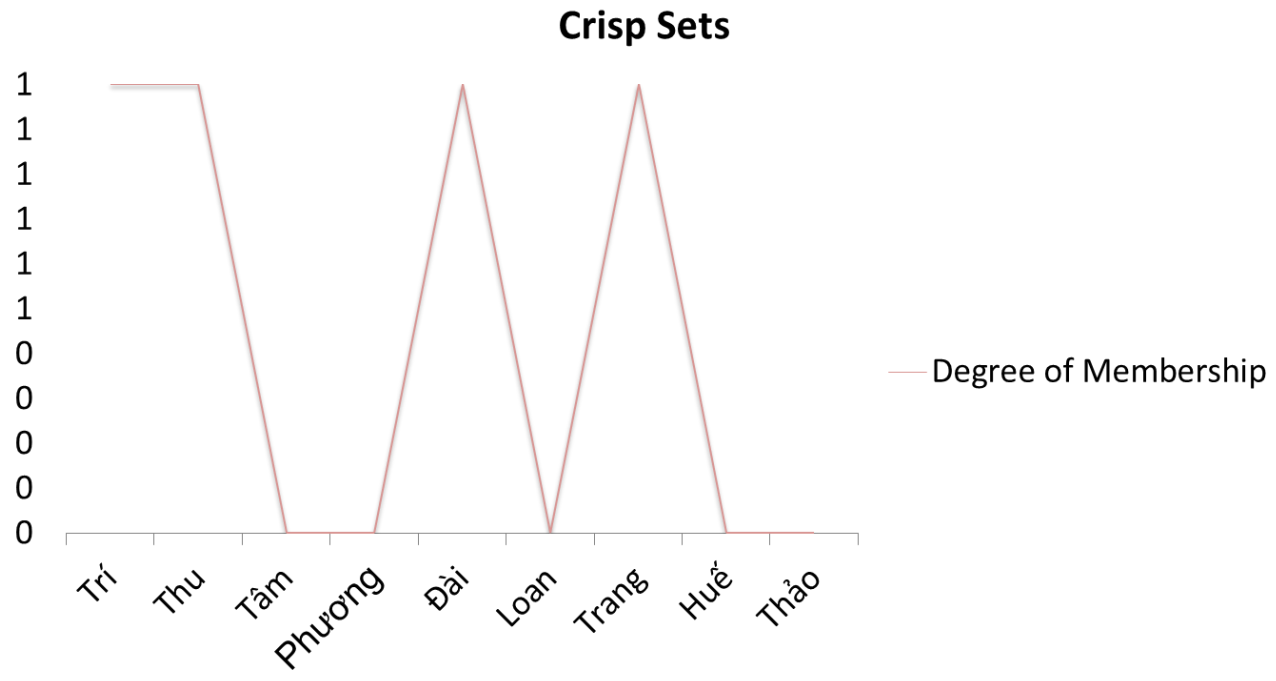
Hình 4-4 Mô tả Crisp sets và Fuzzy sets.

4.4.1/ Tập rõ (Crisp sets):

Tên	Chiều cao	Điểm thành viên (Membership)	
		Crisp	Fuzzy
Trí	200	1	1.00
Thu	166	1	0.89
Tâm	178	0	0.78
Phương	163	0	0.06
Đài	166	1	0.15
Loan	165	0	0.78
Trang	169	1	0.25
Huế	170	0	0.98
Thảo	171	0	0.00

Hình 4-5 Mô tả dữ liệu của tập rõ và tập mờ.

Trong rất nhiều ứng dụng , phương pháp đơn giản và khá phổ dụng là C-Means . Để phân biệt phương pháp này với Fuzzy C-Means chúng ta sẽ tạm gọi nó là Crisp C-Means . Tập rõ (crisp set) là tập hợp truyền thống theo quan điểm của Cantor. Gọi A là một tập hợp rõ, một phần tử x có thể có $x \in A$ hoặc $x \notin A$, Có thể sử dụng hàm χ để mô tả khái niệm thuộc về. Nếu $x \in A$, $\chi(x) = 1$, ngược lại nếu $x \notin A$, $\chi(x) = 0$. Hàm χ được gọi là hàm đặc trưng của tập hợp A.



Hình 4-6 Biểu diễn Crisp sets bằng đồ thị.

4.4.2/ Tập mờ (fuzzy sets):

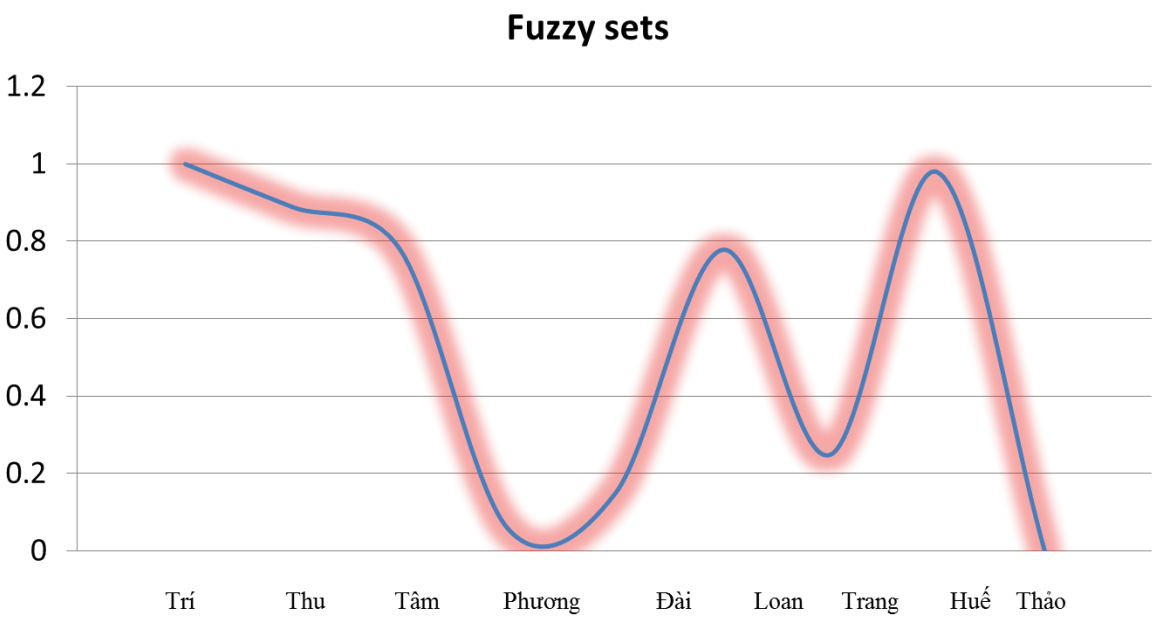
Gọi X là không gian các đối tượng và x là các phần tử tổng quát thuộc X. Khi đó, theo Zadeh (1965), một tập mờ A trong X được định nghĩa là tập các cặp như sau:

$$A = \{ x, \mu_A(x) \mid x \in X \}$$

trong đó, $\mu_A(x)$ được gọi là hàm thành viên của tập mờ A. Hàm thành viên này chỉ mức độ thuộc của x trong không gian X và có giá trị từ 0 đến 1. Hay ký hiệu khác khi X là không gian liên tục: (ký hiệu này không phải chỉ hàm tích phân mà chỉ sự hội các phần tử liên tục).

$$A = \int_X \mu_A(x) \mid x$$

Dễ dàng nhận thấy, nếu như tập mờ A chỉ toàn những hàm thành viên có giá trị 0 hoặc 1 thì A trở thành một tập rõ



Hình 4-7 Biểu diễn Fuzzy sets bằng đồ thị.

4.5/ Chương trình gom cụm Fuzzy C-Means:

Chạy từng bước của chương trình với CSDL:

ID	CONTENT
1	44 "technician" "single" "secondary" "no" "yes"
2	33 "entrepreneur" "n "secondary" "no" "ye

ID	Term	ID_DOCUMENT	FrequencyIn Document	FrequencyIn ALL_Document
1	Technician	1	1	81
2	Single	1	1	225
3	Secondary	1	3	361

$$tf(t,d)=\frac{f(t,d)}{max\{f(w,d):w\in d\}}=\frac{1}{4}=0.25$$

$$idf(t,D)=\log\frac{|D|}{|\{d\in D:t\in d\}|}=\log\frac{1533}{82}=1.271$$

TF-IDF(t,d,D)= TF(t,d) × IDF(t,D)
=0.25x1.271=0.31793

ID	Term	ID_DOCUMENT	TF_IDF
1	44	1	0.31793206930
2	technician	1	0.20785842833
3	Secondary	1	0.06596391648

Hình 4-8 Mô tả các bước tính TF-IDF.

Bảng dữ liệu **dataInput_IS** chứa dữ liệu đầu vào:

Bảng 4-2 Bảng dữ liệu dataInput_IS.

Tên thuộc tính	Kiểu dữ liệu	Chú thích
ID	Int	Khoá chính, tự động tăng dần
CONTENT	Nvarchar[MAX]	Chứa nội dung văn bản được thu thập từ phỏng vấn khách hàng, quan sát, báo cáo kinh doanh...

Bảng dữ liệu **TermFrequency** chứa tần số xuất hiện của thuật ngữ t trong tài liệu hiện tại và so với các tài liệu khác trong tập dữ liệu đầu vào.

Bảng 4-3 Bảng dữ liệu TermFrequency.

Thuộc tính	Kiểu dữ liệu	Chú thích
ID	Int	Khoá chính, tự động tăng dần
Term	Nvarchar[500]	Thuật ngữ trong văn bản ID_DOCUMENT gồm có một từ,hai từ, ba từ...
ID_DOCUMENT	Int	Chứa khoá của văn bản trong bảng dataInput_IS
FrequencyInDocument	Int	Số lần xuất hiện của thuật ngữ Term trong tài liệu ID_DOCUMENT
FrequencyInAll_Document	Int	Số lần xuất hiện của thuật ngữ Term trong tất cả các tài liệu trong tập dữ liệu đầu vào

Bảng dữ liệu **TF_IDF** chứa kết quả tính toán theo phương pháp TF_IDF cho các thuật ngữ trong văn bản trong tập dữ liệu đầu vào

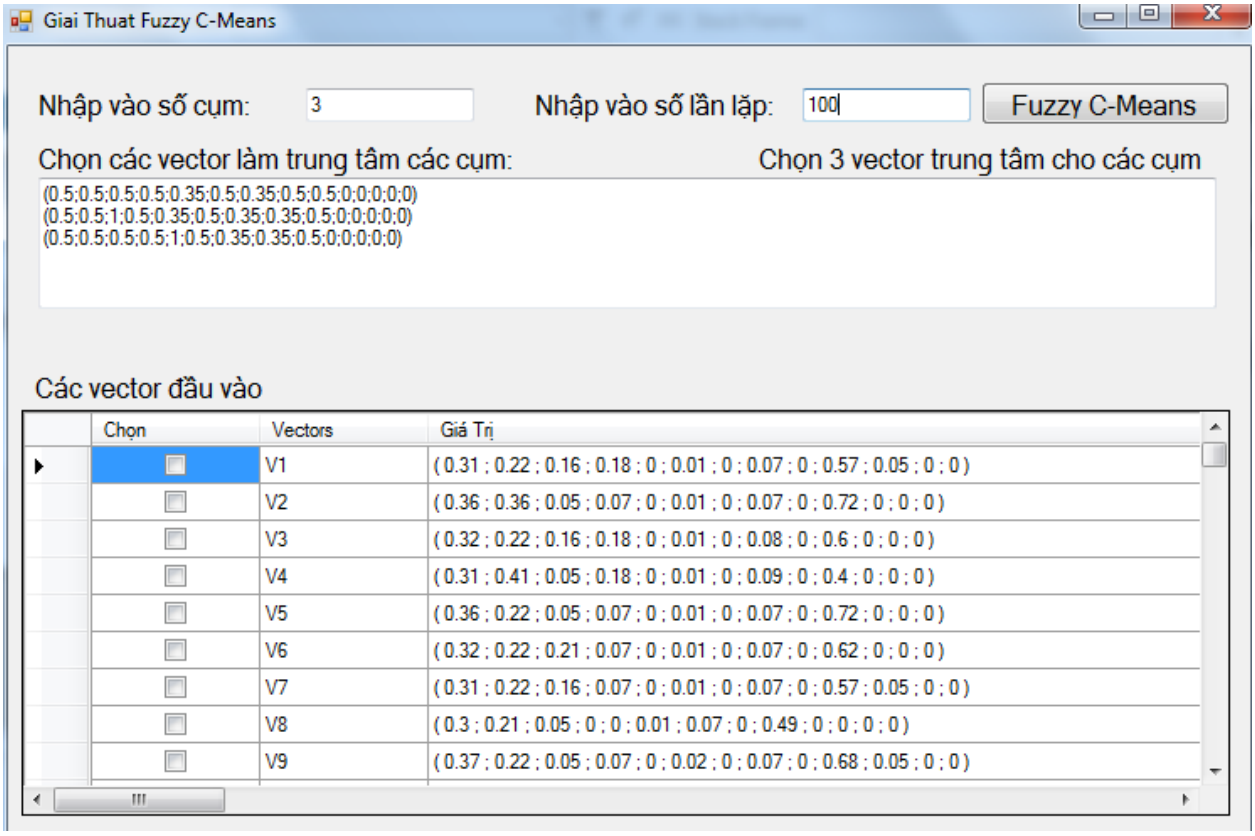
Bảng 4-4 Bảng dữ liệu TF-IDF.

Thuộc tính	Kiểu dữ liệu	Chú thích
ID	Int	Khoá chính, tự động tăng dần
Term	Nvarchar[500]	Thuật ngữ cần tính theo phương pháp tf-idf trong tài liệu ID_DOCUMENT
ID_DOCUMENT	Int	Tài liệu chứa thuật ngữ Term trong tập dữ liệu đầu vào
TF_IDF	Float	Giá trị của thuật ngữ được tính theo phương pháp tf-idf

Chương trình sẽ tính toán xuất kết quả đầu ra là tập fuzzy sets hoặc tập crisp sets tuỳ theo kết quả mong muốn.

CHƯƠNG 5 THỰC NGHIỆM – ĐÁNH GIÁ KẾT QUẢ

Chương trình Fuzzy C-Means được xây dựng bằng ngôn ngữ C Sharp trên Visual Studio 2012 với giao diện như sau:



Hình 5-1 Chương trình Fuzzy C-Means.

5.1/ Thực nghiệm:

Demo với chương trình Fuzzy C-Means:

Đầu tiên, chương trình sẽ chạy với CSDL đã được import vào SQL Server Management Studio 2008. Chương trình sẽ sử phân tách chuỗi dữ liệu đã nhập vào thành các Term (một từ, hai từ, ba từ…). Kế tiếp FrequencyInDocument, FrequencyInAll_Document sẽ được tính toán.

Tiếp theo, sử dụng công thức TF-IDF, ta tính được vector trọng số của mỗi văn bản. Chương trình sau khi thực thi xong giải thuật Fuzzy C-Means sẽ tự động kết xuất ra một tập tin excel (.xls) theo định dạng

	A	B	C	D
1	Clusters	C1	C2	C3
2	V1	0.46	0.07	0.47
3	V2	0.45	0.11	0.44
4	V3	0.46	0.06	0.48
5	V4	0.4	0.19	0.41
6	V5	0.47	0.08	0.45
7	V6	0.46	0.07	0.46
8	V7	0.47	0.06	0.47
9	V8	0.12	0.76	0.12
10	V9	0.49	0.06	0.45

Hình 5-2 Kết quả chương trình Fuzzy C-Means trên excel.

Với V1,V2,V3… là các vector (tương ứng với các tài liệu trong tập dữ liệu đầu vào). C1,C2,C3… là các cụm.

Tiếp theo, chương trình trải qua α-cut xuất kết quả đầu ra là tập fuzzy sets hoặc tập crisp sets tuỳ theo kết quả mong muốn.

Thực nghiệm 1:

Giả sử người dùng mong muốn thu tập fuzzy sets

Bắt đầu thuật toán với các tham số: tham số mờ hoá m=2, tiêu chuẩn hội tụ (epsilon)= 0.01, và chọn 1533 khách hàng (n=1533) có 13 thuộc tính (k=13), được phân vào 3 cụm (c=3). Khi đó kết quả gom cụm như sau:

Clusters	C1	C2	C3
V1	0.46	0.47	0.07
V2	0.45	0.44	0.11
V3	0.47	0.47	0.06
V4	0.4	0.41	0.19
V5	0.47	0.45	0.08
V6	0.46	0.46	0.07
V7	0.46	0.48	0.06
V8	0.12	0.12	0.76
V9	0.48	0.45	0.06
V10	0.48	0.47	0.05
V11	0.44	0.44	0.12
V12	0.43	0.42	0.14
V13	0.47	0.47	0.06
V14	0.46	0.44	0.11
V15	0.47	0.45	0.08
V16	0.45	0.49	0.06
V17	0.46	0.48	0.06
V18	0.19	0.19	0.62
V19	0.39	0.4	0.2
V20	0.43	0.43	0.15

Hình 5-3 Kết quả với số cụm bằng 3.

Cũng với các tham số như trên, khi chúng ta phân thành 6 cụm thì kết quả thu được như sau:

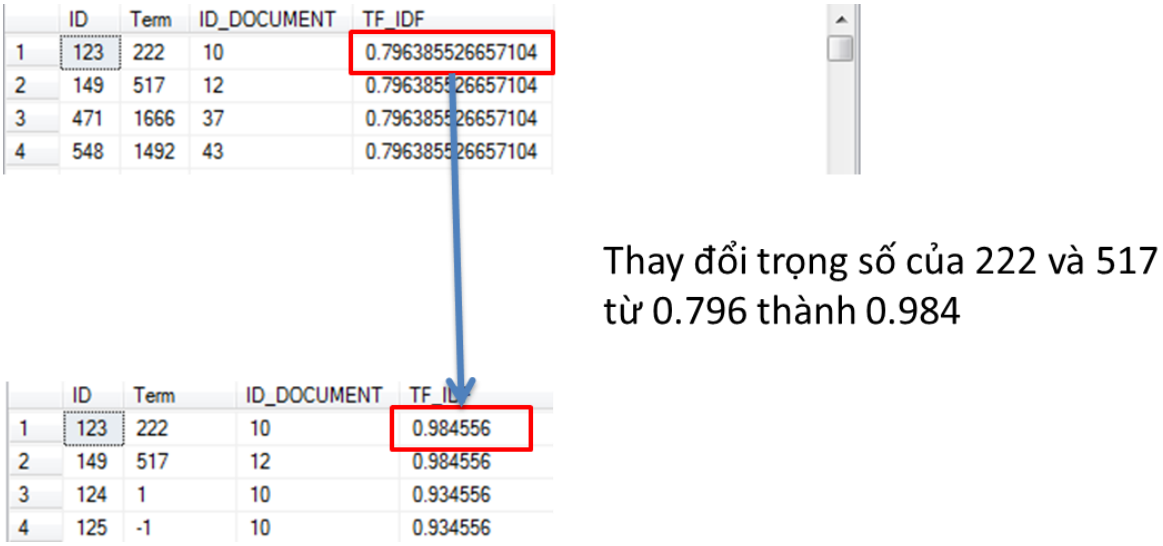
	A	B	C	D	E	F	G
1	Clusters	C1	C2	C3	C4	C5	C6
2	V1	0.19	0.2	0.03	0.2	0.19	0.2
3	V2	0.19	0.19	0.05	0.19	0.19	0.19
4	V3	0.2	0.19	0.02	0.19	0.2	0.19
5	V4	0.18	0.19	0.07	0.19	0.18	0.18
6	V5	0.2	0.19	0.03	0.19	0.2	0.2
7	V6	0.2	0.19	0.03	0.19	0.2	0.19
8	V7	0.19	0.2	0.02	0.2	0.19	0.2
9	V8	0.09	0.09	0.53	0.09	0.09	0.09
10	V9	0.2	0.19	0.03	0.19	0.2	0.2
11	V10	0.2	0.19	0.02	0.19	0.2	0.21

Hình 5-4 Kết quả với số cụm bằng 6.

Thực nghiệm cho thấy, khi chúng ta phân nhiều cụm hơn so với số cụm hiện tại thì các đối tượng cùng một cụm mà có độ thuộc càng xa nhau thì thường sẽ tách ra các cụm khác nhau.

Thực nghiệm 2:

Ở thực nghiệm 1 do các trọng số cho các thuộc tính đều bằng nhau nên ta chưa thấy được sự đặc trưng của cụm. Trong thực nghiệm này, điều chỉnh độ thuộc lên cao để thu kết quả tốt hơn.



Hình 5-5 Mô tả điều chỉnh trọng số TF-IDF.

Kết quả thu được :

	A	B	C	D	E	F	G
1	**Clusters**	**C1**	**C2**	**C3**	**C4**	**C5**	**C6**
2	V1	0.22	0.03	0.11	0.22	0.21	0.22
3	V2	0.21	0.05	0.12	0.21	0.19	0.21
4	V3	0.22	0.03	0.1	0.22	0.2	0.22
5	V4	0.18	0.08	0.21	0.18	0.19	0.18
6	V5	0.23	0.04	0.1	0.23	0.18	0.23
7	V6	0.22	0.03	0.11	0.22	0.2	0.22
8	V7	0.22	0.03	0.09	0.22	0.21	0.22
9	V8	0.09	0.54	0.11	0.09	0.09	0.09
10	V9	0.23	0.03	0.09	0.23	0.18	0.23
11	V10	0.24	0.02	0.07	0.24	0.18	0.24
12	V11	0.2	0.05	0.15	0.2	0.2	0.2
13	V12	0.2	0.07	0.14	0.2	0.19	0.2
14	V13	0.23	0.03	0.09	0.23	0.2	0.23
15	V14	0.22	0.05	0.12	0.22	0.18	0.22
16	V15	0.22	0.04	0.11	0.22	0.19	0.22
17	V16	0.22	0.03	0.1	0.22	0.22	0.22
18	V17	0.22	0.02	0.09	0.22	0.22	0.22
19	V18	0.12	0.35	0.16	0.12	0.13	0.12
20	V19	0.17	0.08	0.22	0.17	0.19	0.17

Hình 5-6 Kết quả sau khi điều chỉnh trọng số.

Qua bước phân tích này, rõ ràng độ thuộc 0.984 thể hiện đặc trưng hơn hẳn độ thuộc 0.769.

Thực nghiệm 3:

Giả sử khách hàng mong muốn thu được tập crisp sets

Bắt đầu thuật toán với các tham số: tham số mờ hoá m=2, tiêu chuẩn hội tụ (epsilon)= 0.33, và chọn 1533 khách hàng (n=1533) có 13 thuộc tính (k=13), được phân vào 3 cụm (c=3). Khi đó kết quả gom cụm như sau:

	A	B	C	D
1	**Clusters**	**C1**	**C2**	**C3**
2	V1	0.36	0	0
3	V2	0.35	0	0
4	V3	0.36	0	0
5	V4	0	0	0
6	V5	0.35	0	0
7	V6	0.35	0	0
8	V7	0.36	0	0
9	V8	0	0.34	0
10	V9	0.35	0	0
11	V10	0.36	0	0
12	V11	0.34	0	0
13	V12	0.34	0	0
14	V13	0.36	0	0
15	V14	0.35	0	0

Hình 5-7 Kết quả dữ liệu crisp sets.

Dễ dàng nhận thấy dữ liệu thu được đã trở thành tập rõ nếu hàm thành viên là 0 và 1.

Thực nghiệm 4:

Thực nghiệm này sẽ diễn tả quá trình alpha-cut dữ liệu. Với tham số α lần lượt là α=0.01, α=0.02, α=0.03, α=0.04, α=0.05, α=0.06. Để từ đó làm nổi rõ các dữ liệu với hàm thành viên nào sẽ quyết định đặc trưng của cụm.

Đầu tiên, chúng ta sẽ tính khoảng cách giữa các cụm.

Clusters	C1	C2	C3
V1	0.1	0.44	0.46
V2	0.12	0.44	0.43
V3	0.61	0.19	0.19
V4	0.37	0.31	0.32
V5	0.1	0.46	0.44
V6	0.1	0.44	0.46
V7	0.16	0.41	0.43
V8	0.13	0.44	0.43
V9	0.2	0.4	0.4
V10	0.17	0.41	0.43

Tính khoảng cách giữa các cụm

$$C12 = |C1 - C2|$$

$$C23 = |C2 - C3|$$

$$C13 = |C1 - C3|$$

0.01	c12	c23	c13
	0.34	0.02	0.36
	0.32	0.01	0.31
	0.42	0	0.42
	0.06	0.01	0.05
	0.36	0.02	0.34
	0.34	0.02	0.36
	0.25	0.02	0.27

Hình 5-8 Khoảng cách giữa các cụm với α=0.01.

Làm tương ứng với các tham số α khác, ta được bảng sau:

0.01	c12	c23	c13		0.02	c12	c23	c13		0.05	c12	c23	c13
	0.34	0.02	0.36			0.34	0.02	0.36			0.35	0	0.35
	0.32	0.01	0.31			0.32	0.01	0.31			0.32	0.02	0.3
	0.42	0	0.42			0.42	0	0.42			0.4	0	0.4
	0.06	0.01	0.05			0.06	0.01	0.05			0.06	0.01	0.05
	0.36	0.02	0.34			0.36	0.02	0.34			0.37	0.04	0.33
	0.34	0.02	0.36			0.34	0.02	0.36			0.34	0	0.34
	0.25	0.02	0.27			0.25	0.02	0.27			0.25	0.01	0.26

0.03	c12	c23	c13		0.04	c12	c23	c13		0.06	c12	c23	c13
	0.35	0	0.35			0.35	0	0.35			0.35	0	0.35
	0.32	0.02	0.3			0.32	0.02	0.3			0.32	0.02	0.3
	0.4	0	0.4			0.4	0	0.4			0.4	0	0.4
	0.06	0.01	0.05			0.06	0.01	0.05			0.06	0.01	0.05
	0.37	0.04	0.33			0.37	0.04	0.33			0.37	0.04	0.33
	0.34	0	0.34			0.34	0	0.34			0.34	0	0.34
	0.25	0.01	0.26			0.25	0.01	0.26			0.25	0.01	0.26

Hình 5-9 Khoảng cách giữa các cụm của từng α.

Kế tiếp tính trung bình của các cụm với công thức: $TB = \frac{C12 + C23 + C13}{3}$

Clusters	0.01	0.02	0.03	0.04	0.05	0.06
V1	0.24	0.24	0.233333	0.233333	0.233333	0.233333
V2	0.213333	0.213333	0.213333	0.213333	0.213333	0.213333
V3	0.28	0.28	0.266667	0.266667	0.266667	0.266667
V4	0.04	0.04	0.04	0.04	0.04	0.04
V5	0.24	0.24	0.246667	0.246667	0.246667	0.246667
V6	0.24	0.24	0.226667	0.226667	0.226667	0.226667
V7	0.18	0.18	0.173333	0.173333	0.173333	0.173333

Hình 5-10 Ma trận trung bình của các vector.

Nhận xét: khi biến α càng tăng, khoảng cách trung bình giữa các cụm càng thu hẹp lại (nhỏ dần). Điều này chứng tỏ khoảng cách giữa vector thuộc cụm và tâm của cụm tương ứng sẽ ngắn lại. Tăng dần giá trị α sẽ tìm được khoảng cách ngắn nhất. Dữ liệu sẽ được phân vào cụm chính xác hơn. Chúng ta còn thấy khi tăng α đến giá trị nào đó, khoảng cách giữa các cụm sẽ không thay đổi nhiều.

Ngoài ra, những vector có thay đổi giá trị α thì khoảng cách vẫn không thay đổi. Những vector này sẽ quyết định đặc trưng của cụm. Bảng vector trọng số và bảng dữ liệu thực tế sau sẽ giúp ta thấy rõ điều này.

	Vector trọng số	Cụm
V1	(0.27 ; 0.17 ; 0.14 ; 0.06 ; 0 ; 0.01 ; 0 ; 0 ; 0 ; 0.42 ; 0 ; 0 ; 0)	1
V2	(0.27 ; 0.3 ; 0.05 ; 0.06 ; 0 ; 0.01 ; 0 ; 0 ; 0 ; 0.42 ; 0 ; 0 ; 0)	1
V3	(0.42 ; 0.21 ; 0.05 ; 0 ; 0 ; 0.01 ; 0 ; 0 ; 0.35 ; 0 ; 0 ; 0 ; 0)	2
V4	(0.27 ; 0 ; 0.14 ; 0 ; 0 ; 0 ; 0.42 ; 0 ; 0 ; 0 ; 0 ; 0 ; 0)	2
V5	(0.3 ; 0.18 ; 0.05 ; 0.18 ; 0 ; 0.01 ; 0 ; 0 ; 0 ; 0.42 ; 0 ; 0 ; 0)	1
V6	(0.35 ; 0.18 ; 0.14 ; 0.18 ; 0 ; 0.01 ; 0 ; 0 ; 0 ; 0.42 ; 0 ; 0 ; 0)	1
V7	(0.27 ; 0.3 ; 0.23 ; 0.18 ; 0.01 ; 0 ; 0 ; 0 ; 0 ; 0.42 ; 0 ; 0 ; 0)	1
V8	(0.27 ; 0.23 ; 0.05 ; 0.21 ; 0 ; 0.01 ; 0 ; 0 ; 0 ; 0.35 ; 0 ; 0 ; 0)	1
V9	(0.42 ; 0.17 ; 0.14 ; 0.06 ; 0 ; 0.01 ; 0 ; 0 ; 0 ; 0.27 ; 0 ; 0 ; 0)	1
V10	(0.42 ; 0.2 ; 0.23 ; 0.06 ; 0 ; 0.01 ; 0 ; 0 ; 0 ; 0.42 ; 0 ; 0 ; 0)	1
V11	(0.3 ; 0.2 ; 0.14 ; 0.06 ; 0 ; 0.01 ; 0 ; 0 ; 0 ; 0.35 ; 0 ; 0 ; 0)	1
V12	(0.3 ; 0.17 ; 0.05 ; 0.06 ; 0 ; 0.01 ; 0 ; 0 ; 0 ; 0.42 ; 0 ; 0 ; 0)	1
V13	(0.27 ; 0.17 ; 0.05 ; 0 ; 0 ; 0.01 ; 0 ; 0 ; 0.42 ; 0 ; 0 ; 0 ; 0)	2
V14	(0.25 ; 0.27 ; 0.05 ; 0.06 ; 0 ; 0.01 ; 0 ; 0 ; 0 ; 0.42 ; 0 ; 0 ; 0)	1
V15	(0.27 ; 0.23 ; 0.05 ; 0.21 ; 0 ; 0.01 ; 0 ; 0 ; 0 ; 0.42 ; 0 ; 0 ; 0)	1
V16	(0.42 ; 0.2 ; 0.14 ; 0 ; 0 ; 0.01 ; 0 ; 0 ; 0.35 ; 0 ; 0 ; 0 ; 0)	2
V17	(0.25 ; 0.21 ; 0.05 ; 0.21 ; 0 ; 0.01 ; 0 ; 0 ; 0 ; 0.35 ; 0 ; 0 ; 0)	1
V18	(0.23 ; 0.23 ; 0.05 ; 0.21 ; 0 ; 0.01 ; 0 ; 0 ; 0 ; 0.42 ; 0 ; 0 ; 0)	1
V19	(0.27 ; 0.27 ; 0.05 ; 0.06 ; 0 ; 0.01 ; 0 ; 0 ; 0 ; 0.35 ; 0 ; 0 ; 0)	1
V20	(0.35 ; 0.21 ; 0.05 ; 0.06 ; 0 ; 0.01 ; 0 ; 0 ; 0 ; 0.42 ; 0 ; 0 ; 0)	1

Hình 5-11 Số liệu 20 dòng vector trọng số

	age;	job	marital	education	default	housing	loan	contact	day	month	unknown	poutcome	y	
V1	58	management	married	tertiary	no	yes	no	unknown		5 may	261	1-10 unknown	no	1
V2	44	technician	single	secondary	no	yes	no	unknown		5 may	151	1-10 unknown	no	1
V3	33	entrepreneur	married	secondary	no	yes	yes	unknown		5 may	76	1-10 unknown	no	2
V4	47	blue-collar	married	unknown	no	yes	no	unknown		5 may	92	1-10 unknown	no	2
V5	33	unknown	single	unknown	no	no	no	unknown		5 may	198	1-10 unknown	no	1
V6	35	management	married	tertiary	no	yes	no	unknown		5 may	139	1-10 unknown	no	1
V7	28	management	single	tertiary	no	yes	yes	unknown		5 may	217	1-10 unknown	no	1
V8	42	entrepreneur	divorced	tertiary	yes	yes	no	unknown		5 may	380	1-10 unknown	no	1
V9	58	retired	married	primary	no	yes	no	unknown		5 may	50	1-10 unknown	no	1
V10	43	technician	single	secondary	no	yes	no	unknown		5 may	55	1-10 unknown	no	1
V11	41	admin.	divorced	secondary	no	yes	no	unknown		5 may	222	1-10 unknown	no	1
V12	29	admin.	single	secondary	no	yes	no	unknown		5 may	137	1-10 unknown	no	1
V13	53	technician	married	secondary	no	yes	no	unknown		5 may	517	1-10 unknown	no	2
V14	58	technician	married	unknown	no	yes	no	unknown		5 may	71	1-10 unknown	no	1
V15	57	services	married	secondary	no	yes	no	unknown		5 may	174	1-10 unknown	no	1
V16	51	retired	married	primary	no	yes	no	unknown		5 may	353	1-10 unknown	no	2
V17	45	admin.	single	unknown	no	yes	no	unknown		5 may	98	1-10 unknown	no	1
V18	57	blue-collar	married	primary	no	yes	no	unknown		5 may	381	1-10 unknown	no	1
V19	60	retired	married	primary	no	yes	no	unknown		5 may	219	1-10 unknown	no	1
V20	33	services	married	secondary	no	yes	no	unknown		5 may	541	1-10 unknown	no	1

Hình 5-12 Bảng 20 dòng dữ liệu đầu vào

Giải thích:

Bảng 20 dòng dữ liệu được nhập vào sẽ có các vector trọng số tương ứng nằm trong bảng 20 dòng vector trọng số. Dữ liệu được phân thành 3 cụm tương ứng với 3 vector sau:

	Vector tâm cụm
v3	(0.42 ; 0.21 ; 0.05 ; 0 ; 0 ; 0.01 ; 0 ; 0 ; 0.35 ; 0 ; 0 ; 0 ; 0)
v20	(0.35 ; 0.21 ; 0.05 ; 0.06 ; 0 ; 0.01 ; 0 ; 0 ; 0 ; 0.42 ; 0 ; 0 ; 0)
v35	(0.25 ; 0.17 ; 0.23 ; 0.06 ; 0 ; 0.01 ; 0 ; 0 ; 0 ; 0.35 ; 0 ; 0 ; 0)

Hình 5-13 Vector trọng tâm của 3 cụm.

Ta có thể thấy cụm 2 sẽ có các đặc trưng sau:

Marital: married, Housing: yes, Loan:no, Poutcome: unknown, Education: secondary.

5.2/ Đánh giá kết quả:

Qua các thực nghiệm đã trình bày, phản ánh được đầy đủ tính chất cơ bản của phương pháp gom cụm Fuzzy C-Means. Dữ liệu đầu vào cũng đã được vào các cụm tương đối chính xác ứng với đặc trưng của cụm. Tuy vậy vẫn phải trải qua các phân tích thủ công để thu được kết quả mong muốn.

CHƯƠNG 6 KẾT LUẬN VÀ HƯỚNG PHÁT TRIỂN

6.1/ Kết luận:

Đây là chương trình phân tích dữ liệu thu thập trong marketing giúp đỡ người dùng trong việc phân loại nhóm khách hàng, tìm kiếm khách hàng tiềm năng từ đó có kế hoạch cho thị trường mục tiêu, chiến lược kinh doanh về lâu dài.

Kết quả hiện thực của các thực nghiệm cho thấy bước đầu xây dựng thành công chương trình Fuzzy C-Means bao gồm các chức năng:

- Ứng dụng Logic mờ để thu thập dữ liệu marketing.
- Có khả năng làm việc với lượng dữ liệu lớn.
- Có khả năng khám phá ra các cụm chưa được gán nhãn.
- Xử lý được các tập dữ liệu chưa biết giá trị.
- Phân loại dữ liệu, phần thành các cụm tương đồng.
- Chương trình sẽ tính toán dựa vào dữ liệu đầu vào để đưa ra kết quả mong muốn.

Fuzzy C-Means phân cụm dữ liệu khá mềm dẻo khi xét đối tượng thuộc một cụm tuỳ theo độ thuộc của đối tượng đó vào cụm. Tuy nhiên Fuzzy C-Means lại xét tất cả các thuộc tính của đối tượng đều có vai trò như nhau. Trong thực tế, dữ liệu thường phức tạp và có những thuộc tính có ý nghĩa hơn hẳn các thuộc tính khác. Vì vậy việc đưa vector trọng số để điều chỉnh ý nghĩa của các thuộc tính là cần thiết. Điều này làm cho việc gom cụm chính xác và linh hoạt hơn.Khi có vector trọng số, người dùng tuỳ theo tình huống mà điều chỉnh để việc phân tích dữ liệu đáp ứng được các yêu cầu trong thực tế.

6.2/ Hướng nghiên cứu tiếp theo:

Luận văn đã thử nghiệm thuật toán Fuzzy C-Means với cơ sở dữ liệu ngẫu nhiên, đơn giản nên đánh giá hiệu năng và độ hiệu quả của việc phân loại cụm chưa được trực quan. Trên cơ sở nghiên cứu đã được trình bày trong luận văn, tôi tiếp tục nghiên cứu các thuật toán trong gom cụm dữ liệu, cũng như sử dụng kết hợp giữa các thuật toán để cải tiến và khắc phục nhược điểm của thuật toán Fuzzy C-Means. Ngoài ra khi đưa vào nguồn dữ liệu thực tế sẽ điều chỉnh để chương trình đáp ứng được nhu cầu và xây dựng thêm tính năng mới.

TÀI LIỆU THAM KHẢO:

[3] Hesam Izakian, Ajith Abraham (2010). ”*Fuzzy C-means and fuzzy swarm for fuzzy clustering problem”.* Expert Systems with Applications Volume 38, Issue 3, March 2011, Pages 1835–1838.

[1] Jiawai Han and Micheline Kember (2011). *Data Mining Concept and Techniques* (Second Edition).

[5] John Wiley and Sons (1999). “*Fuzzy Cluster Analysis”.* ISBN: 978-0-471-98864-9.

[2] Krishna Kant Singh, M. J. Nigam, Kirat Pal, Akansha Mehrotra (2014). *” A Fuzzy Kohonen Local Information C-Means Clustering for Remote Sensing Imagery”.* IETE Technical Review, Volume 31, Issue 1, 2014, pages 75-81.

[7] Lý Thành (2008). “*Giới thiệu một số thuật toán gom cụm mờ. ứng dụng thuật toán gom cụm mờ (fuzzy clustering), mô hình xich markov để phân loại, dự báo, giải quyết các tình trạng kẹt xe”.* Đại học Công nghệ Thông tin.

[4] Lotfi A. Zadeh (1965). “Fuzzy sets”. Information and Control. 8: 338–353.

[9] Nguyễn Đình Thuân, Đoàn Huấn (2012). “*Sử dụng thuật toán gom cụm mờ khai phá cơ sở dữ liệu ERP trong doanh nghiệp dược phẩm*.” Tập san tin học quản lý, Tập 02, Số 2, 2012, 9-17p.

[6] Patrick Andre Pantel (2003). “*Clustering by Committee*”. Thesis Doctor of Philosophy, University of Alberta, 15-25p.

[8] Quan Thanh Tho, Siu Cheung Hui, A.C.M. Fong, Tru Hoang Cao (2006). *“Automatic Fuzzy Ontology Generation for Semantic Web”*.IEEE Transactions on Knowledge & Data Engineering, 2006 Vol 18, No.06 – June.