ĐẠI HỌC QUỐC GIA HÀ NỘI
TRƯỜNG ĐẠI HỌC CÔNG NGHỆ

VŨ THỊ THU HƯƠNG

ỨNG DỤNG PHƯƠNG PHÁP PHÂN CỤM MỜ CHO BÀI TOÁN PHÂN TÍCH THÔNG TIN RỦI RO QUẢN LÝ THUẾ DOANH NGHIỆP

LUẬN VĂN THẠC SĨ QUẢN LÝ HỆ THỐNG THÔNG TIN

Hà Nội – 2017

ĐẠI HỌC QUỐC GIA HÀ NỘI

TRƯỜNG ĐẠI HỌC CÔNG NGHỆ

VŨ THỊ THU HƯƠNG

ỨNG DỤNG PHƯƠNG PHÁP PHÂN CỤM MỜ CHO BÀI TOÁN PHÂN TÍCH THÔNG TIN RỦI RO QUẢN LÝ THUẾ DOANH NGHIỆP

Ngành: Công nghệ thông tin

Chuyên ngành: Quản lý Hệ thống thông tin

Mã số:

LUẬN VĂN THẠC SĨ QUẢN LÝ HỆ THỐNG THÔNG TIN

NGƯỜI HƯỚNG DẪN KHOA HỌC: PGS. TS. Nguyễn Đình Hóa

Hà Nội – 2017

LỜI CAM ĐOAN

Tôi xin cam đoan luận văn là công trình nghiên cứu của riêng cá nhân tôi, không sao chép của ai do tôi tự nghiên cứu, đọc, dịch tài liệu, tổng hợp và thực hiện. Nội dung lý thuyết trong trong luận văn tôi có sử dụng một số tài liệu tham khảo như đã trình bày trong phần tài liệu tham khảo. Các số liệu, chương trình phần mềm và những kết quả trong luận văn là trung thực và chưa được công bố trong bất kỳ một công trình nào khác.

Hà Nội, tháng 10 năm 2017

Học viên thực hiện

Vũ Thị Thu Hương

LỜI CẢM ƠN

Lời đầu tiên, em xin gửi lời biết ơn sâu sắc đến PGS.TS. Nguyễn Đình Hóa, TS. Lê Hoàng Sơn người đã tạo điều kiện thuận lợi, tận tình hướng dẫn, chỉ bảo, giúp đỡ em trong suốt quá trình làm luận văn.

Em cũng xin gửi lời cảm ơn đến các thầy cô giáo trường Đại học Công nghệ - Đại học Quốc Gia Hà Nội, các thầy cô khoa Công nghệ thông tin đã truyền đạt những kiến thức và giúp đỡ em trong suốt quá trình học của mình.

Và cuối cùng em xin gửi lời cảm ơn tới các đồng nghiệp, gia đình và bạn bè, những người đã luôn ủng hộ, động viên và tạo mọi điều kiện giúp đỡ để em có được kết quả như ngày hôm nay.

Hà Nội, tháng 10 năm 2017

Học viên

Vũ Thị Thu Hương

MỤC LỤC

DANH MỤC CÁC KÝ HIỆU VÀ CÁC TỪ VIẾT TẮT

Viết tắt	Thuật ngữ (Anh/Việt)	Giải thích
FCM	Fuzzy C-Mean	Một thuật toán phân cụm mờ
GTGT	Giá trị gia tăng	Tờ khai thuế giá trị gia tăng
MST	Mã số thuế	Mã số thuế doanh nghiệp

DANH MỤC HÌNH MINH HOẠ VÀ BẢNG BIỂU

MỞ ĐẦU

Công tác thanh, kiểm tra thuế là một trong những nhiệm vụ trọng tâm nhằm ngăn ngừa, phát hiện và xử lý kịp thời những vi phạm về thuế. Thực hiện tốt công tác thanh, kiểm tra thuế sẽ góp phần tăng nguồn thu cho ngân sách, tạo sự bình đẳng và công bằng xã hội về nghĩa vụ thuế của đối tượng nộp thuế. Hiện nay nhu cầu tin học hóa các quy trình nghiệp vụ của ngành Thuế nói chung và hiện đại hoá công tác thanh, kiểm tra thuế nói riêng, góp phần nâng cao hiệu quả công tác quản lý thuế ngày càng cao. Với tính chất đa dạng và phức tạp của dữ liệu trong kho dữ liệu Người nộp thuế, cần thiết phải có hướng nghiên cứu và cách tổ chức các kho dữ liệu để trích xuất thông tin phù hợp. Khai phá dữ liệu là một trong những hướng nghiên cứu phổ biến hiện nay, và phân cụm là công cụ hữu hiệu trong các bài toán khai phá dữ liệu, phân tích thông tin [3].

Mục tiêu của phân cụm là chia nhỏ các đối tượng vào các cụm sao cho các đối tượng cùng cụm là tương đồng với nhau nhất. Phân cụm có nhiều ứng dụng trong thương mại, giúp các nhà cung cấp biết được nhóm khách hàng quan trọng có các đặc trưng tương đồng nhau và đặc tả họ từ các mẫu trong cơ sở dữ liệu khách hàng. Phân cụm mờ là phương pháp phân cụm dữ liệu mở rộng trong đó mỗi điểm dữ liệu có thể thuộc về hai hay nhiều cụm với các giá trị hàm thuộc tương ứng. Năm 1969, Ruspini [17] đã giới thiệu khái niệm phân hoạch mờ để mô tả cấu trúc của một cụm mờ. Năm 1973, Dunn [18] đã mở rộng phương pháp phân cụm và đã phát triển thuật toán phân cụm mờ. Ý tưởng của thuật toán là xây dựng một phương pháp phân cụm mờ dựa trên tối thiểu hóa hàm mục tiêu. Sau đó, Bezdek [16] đã cải tiến và tổng quát hóa hàm mục tiêu mờ bằng cách thêm trọng số mũ. Cho đến nay, có rất nhiều biến thể của phân cụm mờ được ứng dụng trong các bài toán khác nhau [16].

Mục tiêu của đề tài là ứng dụng thuật toán phân cụm mờ trong phân tích thông tin rủi ro quản lý thuế doanh nghiệp. Một cơ sở dữ liệu mẫu về thông tin tờ khai thuế, báo cáo tài chính doanh nghiệp, mức độ rủi ro của 644 doanh nghiệp được sử dụng để làm đầu vào cho hệ thống phân tích rủi ro sử dụng phương pháp phân cụm mờ. Hệ thống phân tích sẽ được triển khai xây dựng và thử nghiệm kiểm chứng.

Các phần chính trong luận văn:

Chương 1: Tổng quan về phân cụm dữ liệu

Chương này giới thiệu tổng quan về khai phá dữ liệu, các giai đoạn của khai phá dữ liệu, tổng quan về phân cụm dữ liệu, các mục tiêu, một số yêu cầu của phân cụm dữ liệu và một số kỹ thuật tiếp cận trong phân cụm dữ liệu.

Chương 2: Giới thiệu bài toán phân cụm mờ và các phương pháp xác định số cụm trong gom cụm dữ liệu

Chương này đề cập đến thuật toán phân cụm mờ Fuzzy C-Mean (FCM) và các phương pháp xác định số cụm trong gom cụm dữ liệu.

Chương 3: Ứng dụng phương pháp phân cụm mờ cho bài toán phân tích thông tin quản lý rủi ro thuế doanh nghiệp

Chương này đề cập đến bài toán phân cụm doanh nghiệp dựa trên tập dữ liệu mẫu về thông tin tờ khai thuế, báo cáo tài chính doanh nghiệp của 644 doanh nghiệp. Và đưa ra kết quả khoanh vùng, lựa chọn các nhóm doanh nghiệp, các mức rủi ro quản lý thuế.

CHƯƠNG 1: TỔNG QUAN VỀ PHÂN CỤM DỮ LIỆU

1.1.Giới thiệu về khai phá dữ liệu

1.1.1. Khai phá dữ liệu là gì?

Cũng giống như khai thác tài nguyên khoáng sản, đào vàng, kim cương. Khai phá dữ liệu là quá trình khám phá tri thức có ích từ lượng dữ liệu lớn [25]. Việc khai phá dữ liệu có thể được tiến hành trên một lượng lớn dữ liệu có trong cơ sở dữ liệu, các kho dữ liệu hoặc trong các loại lưu trữ thông tin khác. Những công cụ khai phá dữ liệu có thể phát hiện những xu hướng trong tương lai, các tri thức mà khai phá dữ liệu mang lại có thể ra quyết định kịp thời [13]. Ở đây chúng ta có thể coi khai phá dữ liệu là cốt lõi của quá trình phát hiện tri thức. Quá trình phát hiện tri thức gồm các bước [14]:

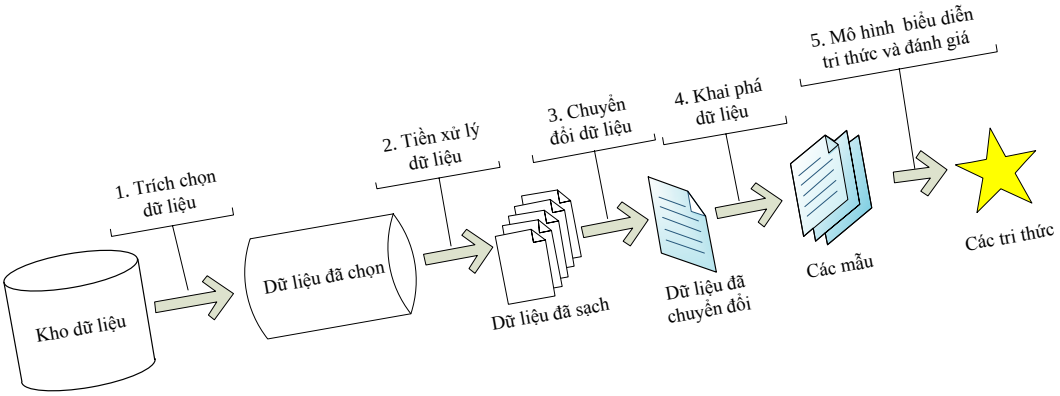
Bước 1: Trích chọn dữ liệu: Là bước chọn ra những tập dữ liệu phù hợp, cần được khai phá từ các tập dữ liệu lớn [14]

Bước 2: Tiền xử lý dữ liệu: Là bước làm sạch dữ liệu như xử lý dữ liệu không đầy đủ, dữ liệu nhiễu, dữ liệu không nhất quán, v.v [14]

Bước 3: Chuyển đổi dữ liệu: Là bước chuẩn hóa và làm mịn dữ liệu để đưa dữ liệu về dạng phù hợp, thuận lợi nhất cho quá trình khai phá dữ liệu [14]

Bước 4: Khai phá dữ liệu: Đây là bước quan trọng và tốn nhiều thời gian nhất của quá trình khám phá tri thức, sử dụng các giải thuật để đưa ra những mô hình dữ liệu [14]

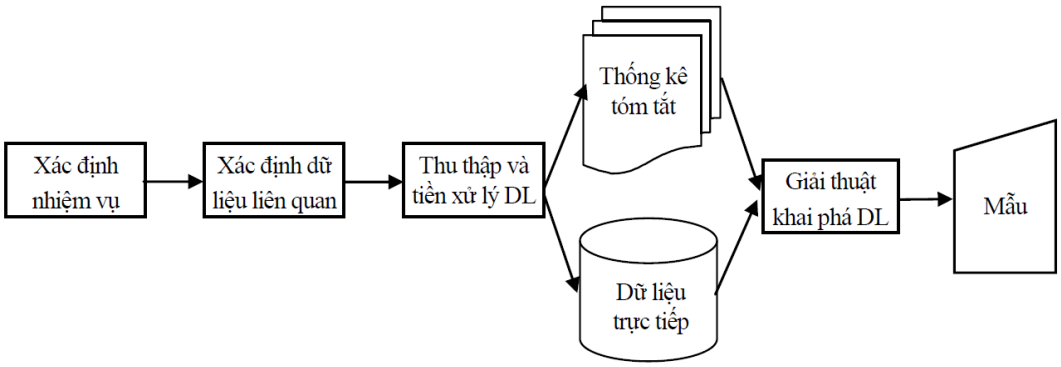
Bước 5: Mô hình biểu diễn tri thức và đánh giá: Dùng các kỹ thuật hiển thị dữ liệu để trình bày các mẫu thông tin (tri thức) và mối liên hệ đặc biệt trong dữ liệu đã được khai thác biểu diễn theo dạng gần gũi với người sử dụng như đồ thị, cây, bảng biểu, luật, v.v. Đồng thời bước này cũng đánh giá những tri thức khám phá được theo những tiêu chí nhất định [14], xác định xem liệu mô hình dữ liệu mà mình vừa tìm được có chứa thông tin hữu ích hay không, tri thức trong đó có đúng hay không?



Hình 1.1. Quá trình phát hiện tri thức [27]

1.1.2. Các giai đoạn của quá trình khai phá dữ liệu

Các giải thuật khai phá dữ liệu thường được miêu tả như những chương trình hoạt động trực tiếp trên tệp dữ liệu. Quá trình khai phá dữ liệu được thể hiện bởi mô hình sau:



Hình 1.2. Quá trình khai phá dữ liệu [15]

- Xác định nhiệm vụ: Xác định chính xác vấn đề cần giải quyết [15]
- Xác định dữ liệu liên quan: xác định các dữ liệu liên quan dùng để xây dựng giải pháp [15]
- Thu thập và tiền xử lý dữ liệu: Thu thập các dữ liệu có liên quan và xử lý chúng thành dạng sao cho giải thuật khai phá dữ liệu có thể hiểu được [15]
- Giải thuật khai phá dữ liệu: chọn thuật toán khai phá dữ liệu thích hợp và thực hiện việc khai phá dữ liệu nhằm tìm được các mẫu có ý nghĩa, các mẫu này được biểu diễn dưới dạng tương ứng với ý nghĩa của nó [15]

1.2.Tổng quan về phân cụm dữ liệu

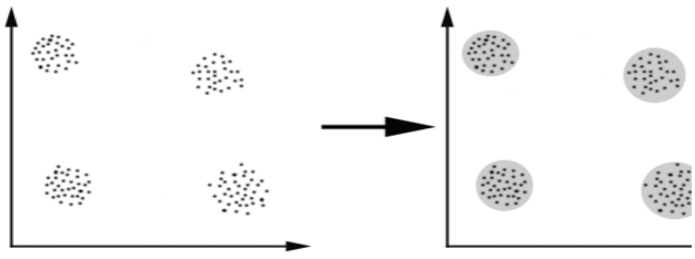
1.2.1. Khái niệm phân cụm dữ liệu

Phân cụm có ý nghĩa rất quan trọng trong hoạt động của con người. Ngay từ lúc bé, con người đã học cách làm thế nào để phân biệt giữa mèo và chó, giữa động vật và thực vật và liên tục đưa vào sơ đồ phân loại trong tiềm thức của mình. Phân cụm được sử dụng rộng rãi trong nhiều ứng dụng, bao gồm nhận dạng mẫu, phân tích dữ liệu, xử lý ảnh, nghiên cứu thị trường, v.v. Với tư cách là một chức năng khai phá dữ liệu, phân cụm có thể được sử dụng như một công cụ độc lập chuẩn để quan sát đặc trưng của mỗi cụm thu được bên trong sự phân bố của dữ liệu và tập trung vào một tập riêng biệt của các cụm để giúp cho việc phân tích đạt kết quả.

Phân cụm dữ liệu là một kĩ thuật trong khai phá dữ liệu nhằm tìm kiếm, phát hiện các cụm, các mẫu dữ liệu tự nhiên tiềm ẩn và quan trọng trong tập dữ liệu lớn để từ đó cung cấp thông tin, tri thức cho việc ra quyết định [10].

Phân cụm dữ liệu là sự phân chia một cơ sở dữ liệu lớn ban đầu thành các nhóm dữ liệu trong đó các đối tượng cùng nhóm tương tự như nhau. Trong mỗi nhóm, một số chi tiết có thể không quan tâm đến để đổi lấy dữ liệu đơn giản hóa. Hay ta có thể hiểu "Phân cụm dữ liệu là quá trình tổ chức các đối tượng thành từng nhóm mà các đối tượng ở mỗi nhóm đều tương tự nhau theo một tính chất nào đó, những đối tượng không tương tự tính chất sẽ ở nhóm khác" [11].

Chúng ta có thể thấy điều này với một ví dụ đơn giản như sau:

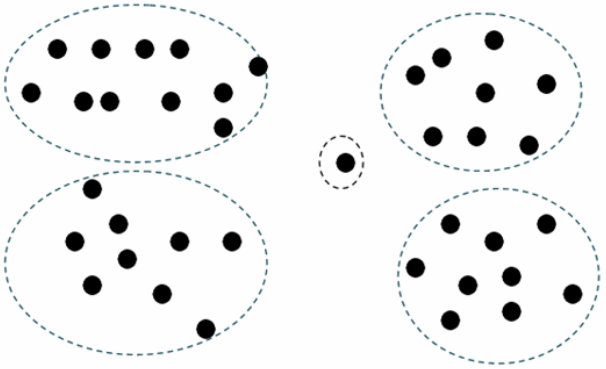


Hình 1.3. Ví dụ về phân cụm dữ liệu [22]

Trong trường hợp này, chúng ta dễ dàng xác định dữ liệu được chia thành 4 cụm dựa vào các dữ liệu đã cho, các tiêu chí tương tự để phân cụm trong trường hợp này là khoảng cách: hai hoặc nhiều đối tượng thuộc nhóm được gom lại theo một khoảng cách nhất định.

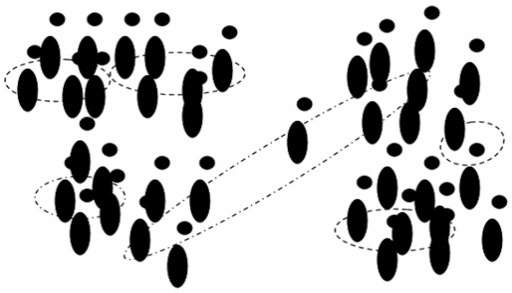
1.2.2. Các mục tiêu của phân cụm dữ liệu

Mục tiêu của phân cụm dữ liệu là chia nhỏ các đối tượng vào các cụm sao cho các đối tượng cùng cụm là tương đồng với nhau.



Hình 1.4. Ví dụ phân cụm các ngôi nhà dựa trên khoảng cách [12]

Một vấn đề thường gặp trong phân cụm là hầu hết các dữ liệu cần cho phân cụm đều có chứa dữ liệu nhiễu do quá trình thu thập thiếu chính xác hoặc thiếu đầy đủ, vì vậy cần phải xây dựng chiến lược cho bước tiền xử lí dữ liệu nhằm khắc phục hoặc loại bỏ nhiễu trước khi chuyển sang giai đoạn phân tích cụm dữ liệu. Nhiễu ở đây được hiểu là các đối tượng dữ liệu không chính xác, không tường minh hoặc là các đối tượng dữ liệu khuyết thiếu thông tin về một số thuộc tính, v.v. Một trong các kỹ thuật xử lí nhiễu phổ biến là việc thay thế giá trị các thuộc tính của đối tượng nhiễu bằng giá trị thuộc tính tương ứng. Ngoài ra, dò tìm đối tượng ngoại lai cũng là một trong những hướng nghiên cứu quan trọng trong phân cụm, chức năng của nó là xác định một nhóm nhỏ các đối tượng dữ liệu khác thường so với các dữ liệu trong cơ sở dữ liệu, tức là các đối tượng dữ liệu không tuân theo các hành vi hoặc mô hình dữ liệu nhằm tránh sự ảnh hưởng của chúng tới quá trình và kết quả của phân cụm [12].



Hình 1.5. Ví dụ phân cụm các ngôi nhà dựa trên kích cỡ [12]

Theo các nghiên cứu đến thời điểm hiện nay thì chưa có một phương pháp phân cụm tổng quát nào có thể giải quyết trọn vẹn cho tất cả các dạng

cấu trúc dữ liệu. Hơn nữa, đối với các phương pháp phân cụm cần có cách thức biểu diễn cấu trúc của dữ liệu, với mỗi cách thức biểu diễn khác nhau sẽ có tương ứng một thuật toán phân cụm phù hợp [15]. Vì vậy phân cụm dữ liệu vẫn đang là một vấn đề khó và mở, vì phải giải quyết nhiều vấn đề cơ bản một cách trọn vẹn và phù hợp với nhiều dạng dữ liệu khác nhau, đặc biệt là đối với dữ liệu hỗn hợp đang ngày càng tăng trong các hệ quản trị dữ liệu và đây cũng là một trong những thách thức lớn trong lĩnh vực khai phá dữ liệu [15].

Tóm lại, phân cụm dữ liệu cần phải giải quyết các vấn đề cơ bản như sau [4]:

- Biểu diễn dữ liệu
- Xây dựng hàm tính độ tương tự
- Xây dựng các tiêu chuẩn phân cụm
- Xây dựng mô hình cho cấu trúc cụm dữ liệu
- Xây dựng thuật toán phân cụm và xác lập các điều kiện khởi tạo
- Xây dựng các thủ tục biểu diễn và đánh giá kết quả phân cụm

1.2.3. Một số ứng dụng của phân cụm dữ liệu

Một số ứng dụng của phân cụm dữ liệu cụ thể như sau:

- Thương mại: Phân loại nhóm khách hàng, dữ liệu khách hàng
- Sinh học: Phân loại các gen với các chức năng tương đồng
- Thư viện: Phân loại các cụm sách có nội dung và ý nghĩa tương đồng nhau
- Y học: Chuẩn đoán triệu chứng, phương pháp trong điều trị y học
- Tài chính và thị trường chứng khoán: dùng để phân tích tình hình tài chính, phân tích đầu tư, phân tích cổ phiếu.
- Khai thác dữ liệu web.
- Trong công nghiệp viễn thông: Phân tích nhu cầu và phân tích các mẫu gian lận và xác định các mẫu khác thường.

1.2.4. Các yêu cầu của phân cụm dữ liệu

Theo Hoàng Thị Giao Lan và Trần Tuấn Tài [15], thuật toán phân cụm dữ liệu cần phải:

- Có khả năng mở rộng
- Có khả năng thích nghi với các kiểu dữ liệu khác nhau: kiểu số, kiểu nhị phân, dữ liệu định dạng, hạng mục, hỗn hợp, v.v
- Khám phá các cụm với hình dạng bất kỳ, do hầu hết các cơ sở dữ liệu có chứa nhiều cụm dữ liệu với các hình thù khác nhau như hình lõm, hình cầu, hình que, v.v
- Tối thiểu lượng tri thức cần cho xác định các tham số đầu vào
- Ít nhạy cảm với thứ tự của dữ liệu vào: cùng một tập dữ liệu, khi đưa vào xử lý cho thuật toán phân cụm dữ liệu với các thứ tự đầu vào của dữ liệu ở các lần thực hiện khác nhau thì không ảnh hưởng đến kết quả phân cụm
- Khả năng thích nghi với dữ liệu nhiễu cao: dữ liệu nhiễu là dữ liệu lỗi, không đầy đủ, dữ liệu rác
- Khả năng thích nghi với dữ liệu đa chiều: Thuật toán có khả năng áp dụng hiệu quả cho dữ liệu có số chiều khác nhau
- Dễ hiểu, dễ cài đặt và sử dụng

1.3.Một số kỹ thuật tiếp cận trong phân cụm dữ liệu

1.3.1. Phương pháp phân cụm phân hoạch

Với một tập dữ liệu gồm n phần tử và k ($k \leq n$) là số cụm được tạo thành. Một thuật toán phân hoạch tổ chức các phần tử dữ liệu vào k phân vùng, mỗi phân vùng thể hiện một cụm dữ liệu và thỏa mãn: mỗi cụm phải chứa ít nhất một phần tử dữ liệu và mỗi phần tử dữ liệu chỉ thuộc vào một cụm. Để đưa ra được k phân mảnh, một phương pháp phân mảnh tạo ra một phân mảnh khởi tạo, sau đó sử dụng kỹ thuật lặp để cải thiện phân mảnh bằng cách di chuyển các phần tử dữ liệu từ cụm này sang cụm khác. Tiêu chuẩn tổng quát của quá trình phân mảnh tốt là các phần tử thuộc cùng một cụm thì "gần gũi" hoặc có liên quan đến nhau, các phần tử khác cụm thì "xa nhau" hoặc rất khác nhau. Có nhiều tiêu chuẩn khác nhau để đánh giá chất lượng của các phân mảnh [8].

Tuy nhiên, phương pháp này không thể xử lí các cụm có hình dạng kỳ dị hoặc các cụm có mật độ các điểm dầy đặc. Các thuật toán phân hoạch dữ liệu có

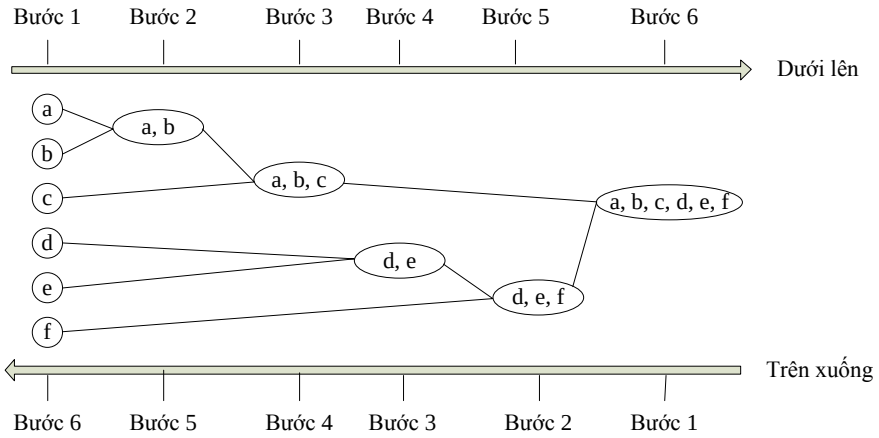
độ phức tạp rất lớn khi xác định nghiệm tối ưu toàn cục cho vấn đề phân cụm dữ liệu, do nó phải tìm kiếm tất cả các cách phân hoạch có thể được. Chính vì vậy, trên thực tế thường đi tìm giải pháp tối ưu cục bộ cho vấn đề này bằng cách sử dụng một hàm tiêu chuẩn để đánh giá chất lượng của cụm cũng như để hướng dẫn cho quá trình tìm kiếm phân hoạch dữ liệu. Như vậy, ý tưởng chính của thuật toán phân cụm phân hoạch tối ưu cục bộ là sử dụng chiến lược ăn tham để tìm kiếm lời giải [4].

1.3.2. Phương pháp phân cụm phân cấp

Quá trình thực hiện phân cụm theo phương pháp này được mô tả bởi một đồ thị có cấu trúc cây, vì vậy nó còn được gọi là phương pháp phân cụm cây. Trong đó, tập dữ liệu được sắp xếp thành một cấu trúc có dạng hình cây gọi là cây phân cụm [2]. Có hai cách tiếp cận phổ biến của kỹ thuật này đó là: hòa nhập nhóm (hay trộn các cụm), thường được gọi là tiếp cận dưới lên và phân chia nhóm (hay phân tách các cụm), thường được gọi là tiếp cận trên xuống.

Quá trình thực hiện thuật toán được biểu diễn thành cây và quyết định phân dữ liệu thành bao nhiêu cụm sẽ do người dùng quyết định. Người dùng cũng dựa trên cây này để nhận được kết quả phân cụm.

Ví dụ về phương pháp phân cụm phân cấp xem tại hình 1.6 dưới đây.



Hình 1.6. Ví dụ phương pháp phân cụm phân cấp

- *Phương pháp "dưới lên"*: Phương pháp này bắt đầu với mỗi đối tượng được khởi tạo tương ứng với các cụm riêng biệt, sau đó tiến hành nhóm các đối tượng theo một độ đo tương tự (như khoảng cách giữa hai trung tâm của hai nhóm), quá trình này được thực hiện cho đến khi tất cả các nhóm được hòa nhập vào một nhóm (mức cao nhất của cây phân cấp)

hoặc cho đến khi các điều kiện kết thúc thỏa mãn. Như vậy, cách tiếp cận này sử dụng chiến lược ăn tham trong quá trình phân cụm.

Cụ thể, phương pháp phân cụm phân cấp *dưới lên* bao gồm các bước sau [2]:

- Khởi tạo mỗi phần tử là một cụm: $c_i = \{x_i\}$, c = n

  Trong đó: c là số cụm, c_i biểu diễn cụm thứ i

  x là phần tử của cụm

  n là số phần tử của tập dữ liệu

- Khi c # 1 thực hiện lặp:
  - Chọn hai cụm gần nhất c_i và c_j theo quy tắc đã chọn
  - Trộn c_i và c_j thành $c_{ij} = c_i \cup c_j$
  - c ← c-1

Ví dụ trong hình 1.6: quá trình thực hiện phương pháp *dưới lên* cụ thể như sau:

- Bước 1: Khởi tạo mỗi phần tử a, b, c, d, e, f là một cụm. Như vậy có 6 cụm ban đầu là {a}, {b}, {c}, {d}, {e}, {f}
- Bước 2: Gộp cụm {a}, {b} thành cụm {a, b}. Các cụm thu được là: {a, b}, {c}, {d}, {e}, {f}
- Bước 3: Gộp cụm {a, b} và cụm {c} thành cụm {a, b, c}. Các cụm thu được là: {a, b, c}, {d}, {e}, {f}
- Bước 4: Gộp cụm {d} và cụm {e} thành cụm {d, e}. Các cụm thu được là: {a, b, c}, {d, e}, {f}
- Bước 5: Gộp cụm {d, e} và cụm {f} thành cụm {d, e, f}. Các cụm thu được là: {a, b, c}, {d, e, f}
- Bước 6: Gộp cụm {d, e, f} và cụm {a, b, c} thành cụm {a, b, c, d, e, f}. Cụm thu được là: {a, b, c, d, e, f}

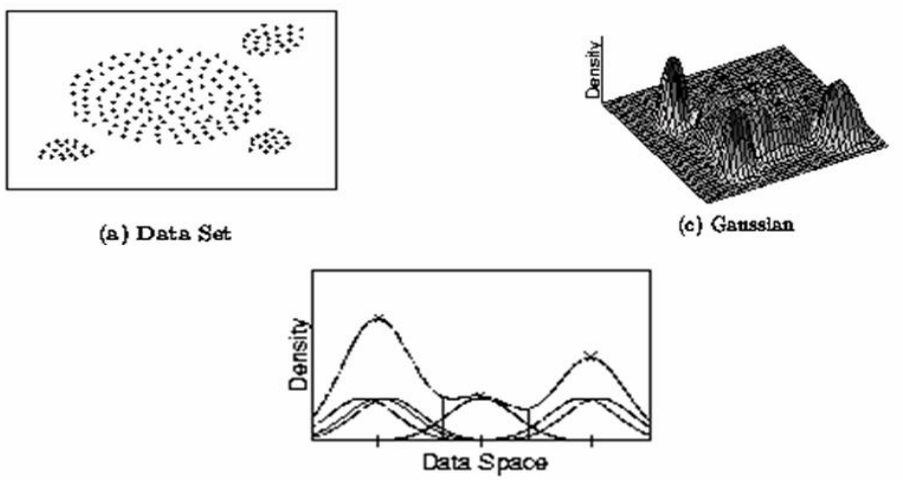
- *Phương pháp "trên xuống":* Bắt đầu với trạng thái là tất cả các đối tượng được xếp trong cùng một cụm. Mỗi vòng lặp thành công, một cụm được tách thành các cụm nhỏ hơn theo giá trị của một phép đo độ tương tự nào đó cho đến khi mỗi đối tượng là một cụm, hoặc cho đến khi điều kiện

dừng thỏa mãn. Cách tiếp cận này sử dụng chiến lược chia để trị trong quá trình phân cụm.

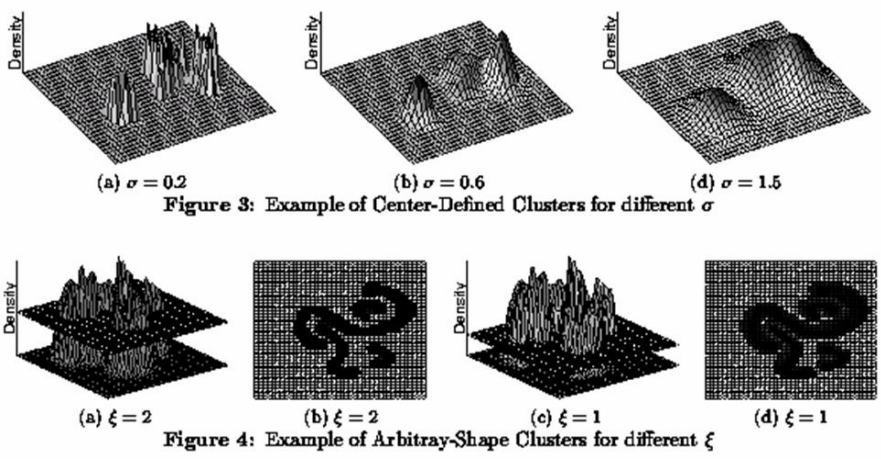
Phương pháp *trên xuống* thực hiện theo quy trình ngược với phương pháp *dưới lên*. Phương pháp này phức tạp và lâu hơn phương pháp dưới lên, thường chỉ được áp dụng khi người ta có thêm thông tin về phân bố cụm để có phương pháp tách phù hợp.

1.3.3. Phương pháp tiếp cận dựa trên mật độ

Kỹ thuật này nhóm các đối tượng dữ liệu dựa trên hàm mật độ xác định, mật độ là số các đối tượng lân cận của một đối tượng dữ liệu theo một nghĩa nào đó. Trong cách tiếp cận này, khi một dữ liệu đã xác định thì nó tiếp tục được phát triển thêm các đối tượng dữ liệu mới miễn là số các đối tượng lân cận này phải lớn hơn một ngưỡng đã được xác định trước. Phương pháp phân cụm dựa trên mật độ của các đối tượng để xác định các cụm dữ liệu có thể phát hiện ra các cụm dữ liệu với hình thù bất kỳ [4]. Kỹ thuật này có thể khắc phục được các phần tử ngoại lai hoặc giá trị nhiễu rất tốt, tuy nhiên việc xác định các tham số mật độ của thuật toán là rất khó khăn, trong khi các tham số này lại có tác động rất lớn đến kết quả phân cụm.



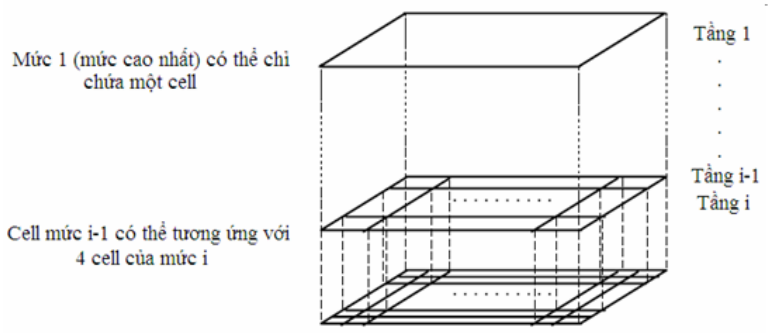
Hình 1.7. Ví dụ về phân cụm theo mật độ (1) [19]



Hình 1.8. Ví dụ về phân cụm theo mật độ (2) [19]

1.3.4. Phương pháp phân cụm dựa trên lưới

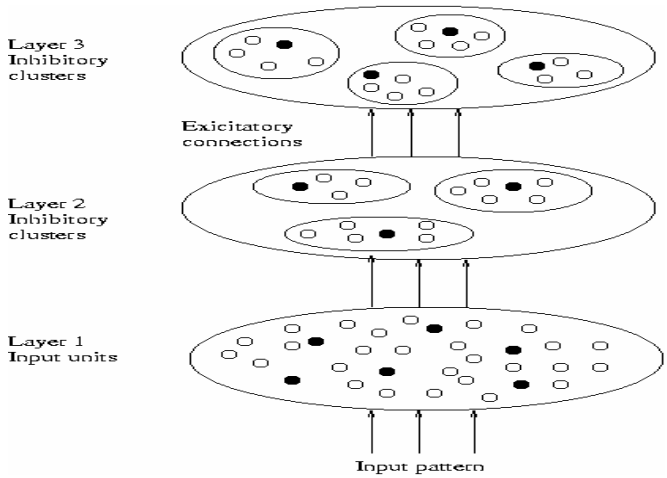
Kỹ thuật phân cụm dựa trên lưới thích hợp với dữ liệu nhiều chiều, dựa trên cấu trúc dữ liệu lưới để phân cụm, phương pháp này chủ yếu tập trung áp dụng cho lớp dữ liệu không gian. Mục tiêu của phương pháp này là lượng hóa dữ liệu thành các ô tạo thành cấu trúc dữ liệu lưới. Sau đó, các thao tác phân cụm chỉ cần làm việc với các đối tượng trong từng ô trên lưới chứ không phải các đối tượng dữ liệu. Cách tiếp cận dựa trên lưới này không di chuyển các đối tượng trong các ô mà xây dựng nhiều mức phân cấp của nhóm các đối tượng trong một ô. Phương pháp này gần giống với phương pháp phân cụm phân cấp nhưng chúng không trộn các ô, đồng thời giải quyết khắc phục yêu cầu đối với dữ liệu nhiều chiều mà phương pháp phân phân cụm dựa trên mật độ không giải quyết được. ưu điểm của phương pháp phân cụm dựa trên lưới là thời gian xử lí nhanh và độc lập với số đối tượng dữ liệu trong tập dữ liệu ban đầu, thay vào đó là chúng phụ thuộc vào số ô trong mỗi chiều của không gian lưới. [15]



Hình 1.9. Cấu trúc phân cụm dựa trên lưới [19]

1.3.5. Phương pháp phân cụm dựa trên mô hình

Phương pháp phân cụm dựa trên mô hình cố gắng để tối ưu hóa sự phù hợp giữa dữ liệu cho trước và một số mô hình toán học. Những phương pháp này thường được dựa trên giả định rằng các dữ liệu được tạo ra bởi sự hòa nhập của các phân bố xác suất cơ bản. [8]



Hình 1.10. Ví dụ về phân cụm dựa trên mô hình [19]

Phương pháp phân cụm dựa trên mô hình cố gắng khớp giữa các dữ liệu với mô hình toán học, nó dựa trên giả định rằng dữ liệu được tạo ra bằng hỗn hợp phân phối xác suất cơ bản. Các thuật toán phân cụm dựa trên mô hình có hai cách tiếp cận chính: mô hình thống kê và mạng nơron. Phương pháp này gần giống với phương pháp phân cụm dựa trên mật độ, vì chúng phát triển các cụm riêng biệt nhằm cải tiến các mô hình đã được xác định trước đó, nhưng đôi khi nó không bắt đầu với một số cụm cố định và không sử dụng cùng một khái niệm mật độ cho các cụm. [15]

CHƯƠNG 2: GIỚI THIỆU BÀI TOÁN PHÂN CỤM MỜ VÀ CÁC PHƯƠNG PHÁP XÁC ĐỊNH SỐ CỤM TRONG GOM CỤM DỮ LIỆU

2.1.Bài toán phân cụm mờ

2.1.1. Giới thiệu về phân cụm mờ

Phân cụm mờ là phương pháp phân cụm dữ liệu mở rộng trong đó mỗi điểm dữ liệu có thể thuộc về hai hay nhiều cụm thông qua giá trị hàm thuộc.

Nhiều vấn đề đã dẫn đến bài toán phân cụm mờ và ứng dụng được nói nhiều trong bài toán phân cụm mờ là: nhận dạng ảnh, xử lý thông tin, phân loại khách hàng trong ngân hàng, v.v.

Ưu điểm của phân cụm mờ so với phân cụm rõ được thể hiện trong thực tế khi mà không thể chỉ ra ranh giới rõ ràng giữa các cụm. Phân cụm rõ bắt buộc các điểm chỉ được phép thuộc vào duy nhất một cụm. Còn phân cụm mờ cho phép các điểm dữ liệu linh hoạt hơn, một điểm dữ liệu có thể thuộc vào nhiều cụm và ta đưa ra khái niệm độ thuộc để chỉ mức độ liên quan của điểm dữ liệu vào cụm mà nó thuộc. Giá trị độ thuộc nằm trong khoảng (0,1), trường hợp điểm dữ liệu không thuộc một cụm nào hay chỉ thuộc vào duy nhất một cụm là rất hiếm.

2.1.2. Thuật toán Fuzzy C-Mean (FCM)

2.1.2.1. Hàm mục tiêu

Kỹ thuật này phân hoạch một tập n vectơ đối tượng dữ liệu $X = \{x_1,\ldots,x_n\} \subset R^s$ thành c các nhóm mờ dựa trên tính toán tối thiểu hóa hàm mục tiêu để đo chất lượng của phân hoạch và tìm trung tâm cụm trong mỗi nhóm, sao cho chi phí hàm đo độ phi tương tự là nhỏ nhất. Một phân hoạch mờ vectơ điểm dữ liệu $X = \{x_1,\ldots,x_n\} \subset R^s$ là đặc trưng đầu vào được biểu diễn bởi ma trận $U =[u_{ik}]$ sao cho điểm dữ liệu đã cho chỉ có thể thuộc về một số nhóm với bậc được xác định bởi mức độ thuộc giữa [0,1]. Như vậy, ma trận U được sử dụng để mô tả cấu trúc cụm của X bằng cách giải thích u_{ik} như bậc thành viên x_k với cụm i. [4,8]

Cho $U = (u_1, u_2, \ldots u_c)$ là phân hoạch mờ gồm c cụm. Mã trận U_{cxn} như sau: [4, 8]

$$U_{cxn} = \begin{bmatrix} u_{11} u_{12} \dots u_{1n} \\ u_{21} u_{22} \dots u_{2n} \\ \dots \dots \dots \dots \dots \dots \\ u_{c1} u_{c2} \dots u_{cn} \end{bmatrix}$$

Dunn định nghĩa hàm liên tục mờ như sau: $u_{ik} d^2 (x_k, v_i)$

$$J_m(U,V) = \sum_{k=1}^{n} \sum_{i=1}^{c} u_{ik} d^2 (x_k, v_i)$$

Bezdek khái quát hóa hàm mục tiêu mờ bằng cách đưa ra trọng số mũ m>1 là bất kỳ số thực nào như sau:

$$J_m(U,V) = \sum_{k=1}^{n} \sum_{i=1}^{c} u_{ik} d^2 (x_k, v_i),\ 1 \leq m \leq \infty \qquad (1)$$

Trong đó:

$X = [x_1, \dots, x_n] \subset R^s$ là n đối tượng dữ liệu trong không gian R^s.

$m \in [1, +\infty]$ là tham số mờ.

$v_i \in R_s$ là trung tâm cụm thứ i.

$d(x_k, v_i) = d_{ik}$ là khuôn mẫu để đo khoảng cách giữa dữ liệu x_k với trung tâm cụm thứ i.

$u_{ik} \in [0,1]$ là bậc của phần tử dữ liệu x_k thuộc về cụm thứ i.

$V = [v_{ij}] = [v_l, \dots, v_c] \in R_{sxc}$ là ma trận biểu diễn các giá trị đối tượng tâm của cụm.

$U = [u_{ik}]$ là ma tra phân hoạch mờ ngẫu nhiên của X trong C phần.

Một trong các nhân tố chính ảnh hưởng tới quyết định phân cụm hợp lý các điểm là vấn đề chọn phép đo độ phi tương tự. Thực vậy, tính toán bậc thành viên u_{ik} phụ thuộc vào định nghĩa của phép đo khoảng cách d_{ik} mà là tích vô hướng trên R^s . Bình phương khoảng cách giữa vectơ mẫu x_k và trung tâm vị trí của cụm thứ i được định nghĩa như sau: [4, 8]

$$d(x_k, v_i) = \|x_k - v_i\| = \sqrt{(x_k - v_i)^T A (x_k - v_i)}$$

$$d^2(x_k, v_i) = \|x_k - v_i\|^2 = (x_k - v_i)^T A (x_k - v_i)$$

Trong đó:

A là ma trận hữu hạn dương đối xứng (p x p) bất kỳ.

$\|x_k - v_i\|^2$ biểu diễn độ lệch của dữ liệu x_k với v_i , $d(x_k, v_i)$ là tích vô hướng trên R^s.

Bậc của thành viên của x_k với cụm i thỏa mãn ràng buộc sau:

$$\begin{cases} 0 \le i_k \le 1, & 1 \le i \le c, 1 \le k \le n \\ 0 < \sum_{k-1}^{n} u_{ik} < n & 1 \le i \le c \\ \sum_{i-1}^{c} u_{ik} = 1 & 1 \le k \le n \end{cases} \quad (2)$$

Để thuận tiện, coi mảng đối tượng dữ liệu $\{x_1,....,x_n\}$ là các cột trong ma trận đối tượng dữ liệu $X = [x_{jk}] = [x_1,\ldots,x_n] \in R^{sxc}$. Ma trận phân hoạch U là một công cụ tiện lợi để mô tả cấu trúc cụm trong dữ liệu $\{x_1,\ldots x_n\}$. Định nghĩa tập tất cả các ma trận thực không suy biến cấp c x n thực hiện phân hoạch mờ n đối tượng c thành cụm dữ liệu trong không gian R^{cxn} là:

$$M_{fcn} = \{U \in R^{cxn} | \forall i,k: u_{ik} \in [0,1]; \sum_{i=1}^{c} u_{ik} = 1 < n\} \quad (3)$$

R^{cxn} là không gian của tất cả các ma trận thực cấp c x n

Thông thường người ta gọi bài toán phân cụm mờ là bài toán tìm các độ thuộc u_{ij} nhằm tối thiểu hàm mục tiêu ở trên J_m(U,V.

Định lý 1: *Nếu m và c là các tham số cố định và I_k là một tập được định nghĩa như sau*: [4, 8]

Với mọi số k thỏa mãn $1 \le k \le n$: $I_k = \{i | 1 \le i \le c, d_{ik} = 0\}$ (4)

thì hàm mục tiêu J_m(U,V) đạt giá trị tối thiểu:

$$\min\{J_m(U,V) = \sum_{k=1}^{n} \sum_{i=1}^{c} (u_{ik})^m d^2(x_k, v_i)\}$$

khi và chỉ khi:

$$\forall 1 \le i \le c, 1 \le k \le n : u_{ik} = \begin{cases} \dfrac{1}{\sum_{j=1}^{c} (\frac{d_{ik}}{d_{jk}})^{\frac{2}{m-1}}} & I_k = \emptyset \\ \begin{cases} 0, & i \notin I_k, I_k \ne \emptyset \\ \sum_{i \in I_k} u_{ik} = 1, & i \in I_k, I_k \ne \emptyset \end{cases} \end{cases} \quad (5)$$

$$v_i = \frac{\sum_{k=1}^{n} (u_{ik})^m x_k}{\sum_{k=1}^{n} (u_{ik})^m}, \quad 1 \le i \le c \quad (6)$$

Định lý đã được Bezdek chứng minh (nếu $m \ge 1$, $d^2_{ik} > 0$, $1 \le i \le c$) là đúng đắn.

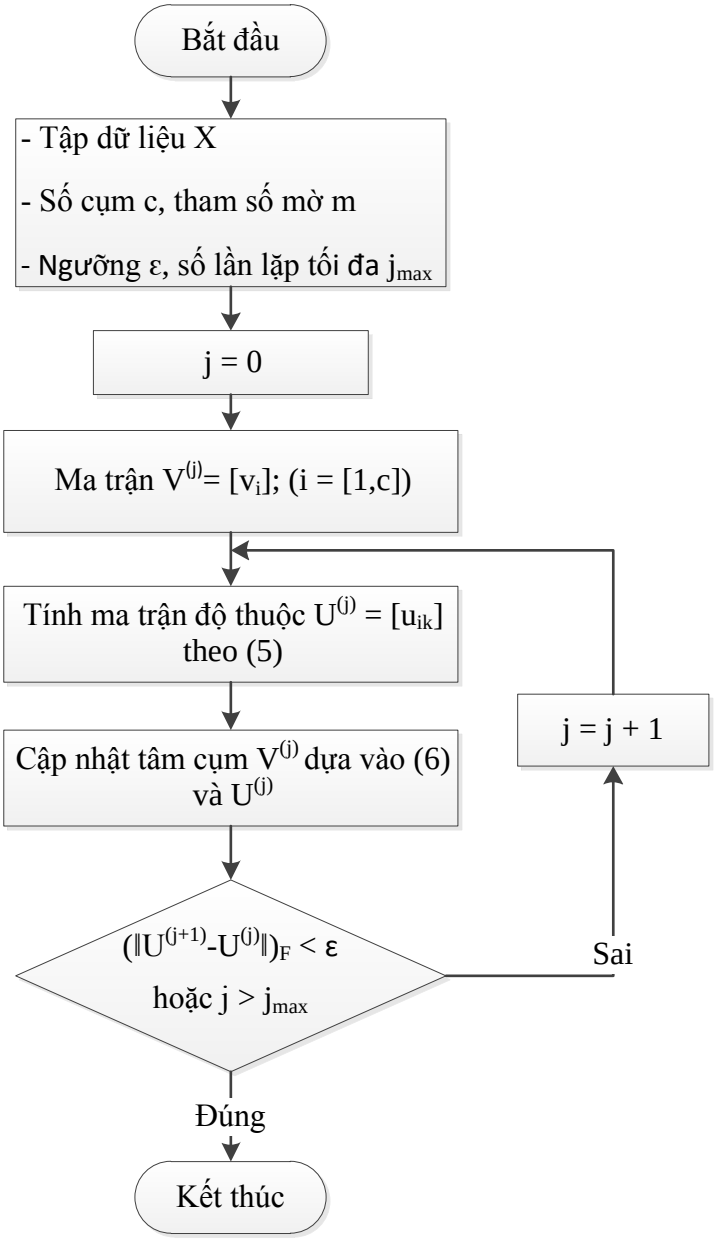
Như vậy, một phân hoạch được gọi là tối ưu thì hàm mục tiêu phải đạt giá trị tối thiểu, điều này tương đương với hai điều kiện (5) và (6) phải thỏa mãn. Từ

đó, tiến hành xây dựng thuật toán FCM như sau:

2.1.2.2. Thuật toán FCM

Thuật toán FCM cung cấp một quá trình lặp qua lại giữa phương trình (5) và (6) để tối ưu (*xấp xỉ cực tiểu*) hàm mục tiêu dựa trên đo đạc độ tương tự có trọng số giữa x_k và trung tâm cụm v_i, sau mỗi vòng lặp, thuật toán tính toán và cập nhật các phần tử u_{jk} trong ma trận phân hoạch U. Phép lặp sẽ dừng khi $\max_{ij}\{\|u_{ij}^{(k+1)}-u_{ij}^{k}\|\} \leq \varepsilon$, trong đó ε là chuẩn kết thúc giữa 0 và 1, trong khi k là các bước lặp. Thủ tục này hội tụ tới cực tiểu cục bộ hay điểm yên ngựa của $I_m(u,V)$. Thuật toán FCM tính toán ma trận phân hoạch U và kích thước của các cụm để thu được các mô hình mờ từ ma trận này [4, 8]. Các bước thực hiện cửa thuật toán FCM như sau:

THUẬT TOÁN FCM



Hình 2.1. Thuật toán FCM

Trong đó $\|*\|_F$ là chuẩn Frobenious được định nghĩa như sau:

$$\|U\|_F^2 = \sum_i \sum_k u_{ik}^2$$

và tham số ε được cho trước.

Việc chọn các tham số cụm rất ảnh hưởng đến kết quả phân cụm.

Đối với $m \to 1^+$ thì thuật toán FCM trở thành thuật toán rõ.

Đối với $m \to \infty$ thì thuật toán FCM trở thành thuật toán phân cụm mờ với: $u_{ik} = \frac{1}{c}$. Chưa có quy tắc nào nhằm lựa chọn tham số m đảm bảo cho việc phân cụm hiệu quả, nhưng thông thường chọn m = 2.

2.1.2.3. Đánh giá

Thuật toán FCM đã được áp dụng thành công trong giải quyết một số lớn các bài toàn phân cụm dữ liệu như trong nhận dạng mẫu, xử lý ảnh, y học, …

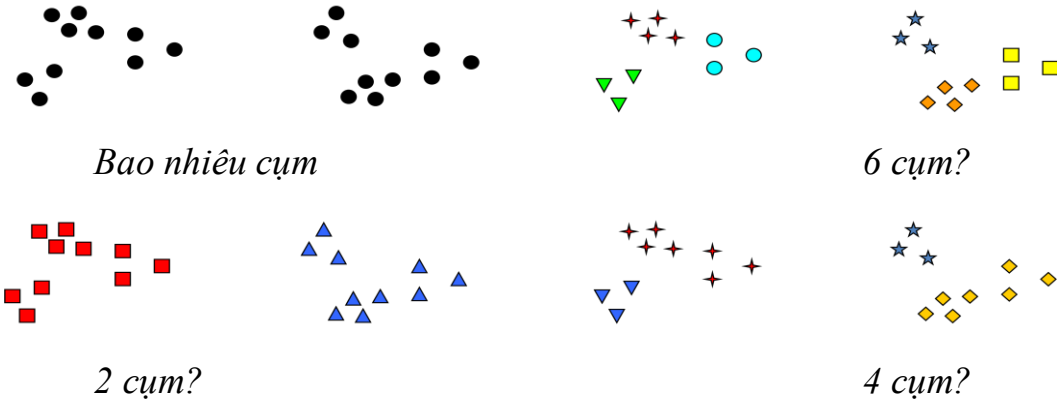
Tuy nhiên, nhược điểm lớn nhất của thuật toán FCM là nhạy cảm với các nhiễu và phần tử ngoại lai trong dữ liệu, nghĩa là các trung tâm cụm có thể nằm xa so với trung tâm thực của cụm. Do đó các cụm dữ liệu được khám phá có thể nằm rất lệch so với các cụm trong thực tế. Việc khử nhiễu và phần tử ngoại lai là một vấn đề cần được giải quyết.

Tóm lại, phân cụm mờ là một sự mở rộng của phân cụm dữ liệu bằng cách thêm vào yếu tố quan hệ giữa các phần tử và các cụm dữ liệu thông qua các trọng số trong ma trận U. Bằng cách này, chúng ta có thể khám phá ra các cụm dữ liệu phức tạp theo cách mềm dẻo từ một cụm dữ liệu đã cho. Thuật toán phân cụm mờ là một cách thức mở rộng cho các thuật toán phân cụm rõ nhằm khám phá ra các cụm dữ liệu chồng lên nhau.

2.2.Các phương pháp xác định số cụm trong gom cụm dữ liệu

Trong các thuật toán phân cụm mờ thường yêu cầu người dùng xác định trước số cụm. Số cụm là một tham số đầu vào quan trọng và ảnh hưởng nhiều tới kết quả của quá trình phân cụm, ứng với số lượng cụm khác nhau sẽ cho ra các kết quả phân cụm khác nhau, thật khó khăn để quyết định kết quả phân cụm nào là tốt nhất.

Quá trình phân cụm dữ liệu nhằm xác định các nhóm đối tượng dữ liệu tương tự, từ đó khảo sát các cụm sẽ giúp khái quát, nhanh chóng rút ra các đặc điểm của khối dữ liệu lớn. Tuy nhiên, trong hầu hết các thuật toán phân cụm, tham số số cụm không được biết trước và thuật toán thường yêu cầu người dùng phải xác định trước số lượng các cụm, ứng với mỗi số lượng cụm khác nhau sẽ cho ra các kết quả phân cụm khác nhau [2].



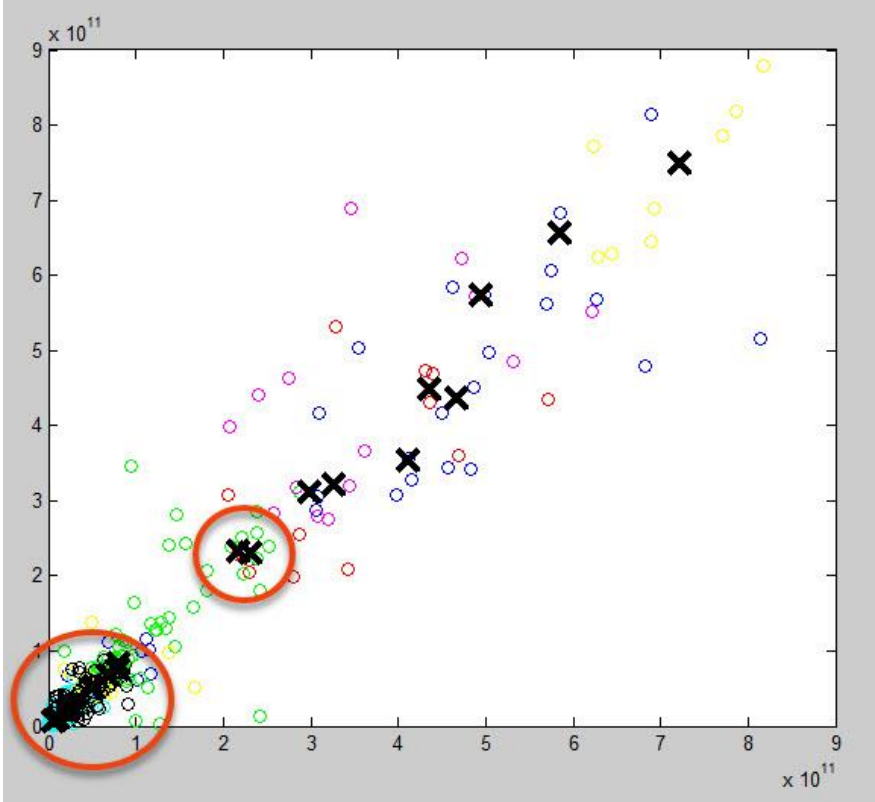
Hình 2.2. Phân cụm tập dữ liệu với số lượng cụm khác nhau [2]

2.2.1. Xác định số cụm dựa trên phương pháp truyền thống

Xác định số cụm k dựa trên phương pháp truyền thống là $\sqrt{\frac{n}{2}}$ với bộ dữ liệu có n đối tượng. Phương pháp này được thực hiện nhanh chóng nhưng độ chính xác không cao.

Ví dụ với tập dữ liệu mẫu về thông tin tờ khai thuế, báo cáo tài chính doanh nghiệp của hơn 644 doanh nghiệp, ta có:

- Số lượng đối tượng: n = 644
- Dựa trên phương pháp truyền thống, xác định số cụm k = $\sqrt{\frac{644}{2}} \approx 18$ cụm
- Hình 2.6 dưới đây biểu diễn kết quả phân cụm dữ liệu trên với số lượng cụm k = 18. Ta có thể nhìn thấy các cụm nằm gần gốc tọa độ (vùng khoanh tròn) có độ chồng nhau rất cao, các cụm phân tách không rõ ràng.



Hình 2.3. Minh họa cho phương pháp xác định số cụm dựa trên phương pháp truyền thống

2.2.2. Xác định số cụm bằng phương pháp Eblow

Xác định số cụm K dựa trên phương pháp Elbow. Phương pháp này thực hiện việc xác định số cụm dựa trên độ chính xác của việc thử các giá trị K khác nhau.

Trước tiên, tính tổng bình phương khoảng cách từ các đối tượng thuộc cụm đến tâm cụm với một số giá trị của cụm K. SSE được định nghĩa là tổng của khoảng cách bình phương giữa mỗi điểm của cụm và tâm cụm đó:

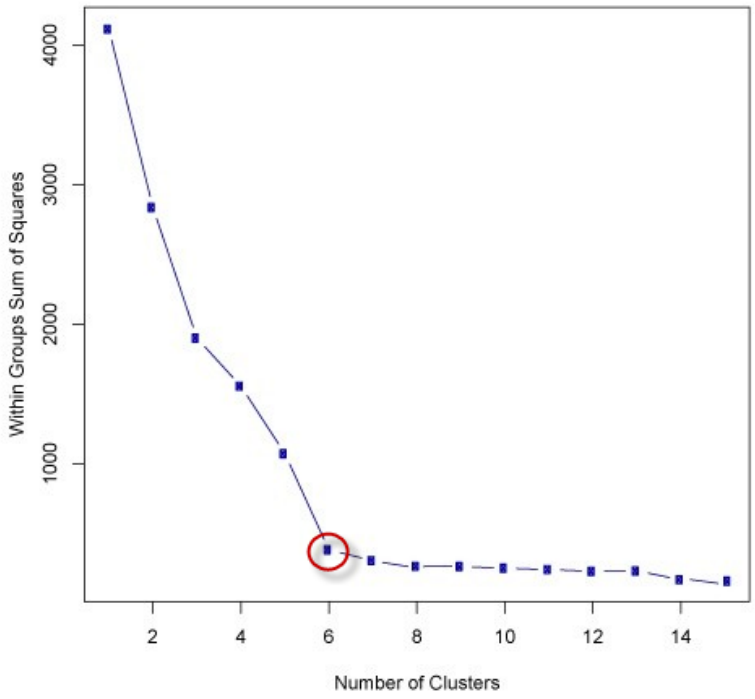
$$SSE = \sum_{i=1}^{k} \sum_{x \in c_i} d(x, c_i)^2$$

Trong đó:

- K: số cụm
- x: đối tượng thuộc cụm i
- c_i: tâm cụm i

Ý tưởng của phương pháp Elbow là chọn k mà tại đó SSE giảm đột ngột. Điều này tạo ra một hiệu ứng, tạm gọi là hiệu ứng "khuỷu tay", như trong hình

2.7 dưới đây: Trong trường hợp này, k = 6 là giá trị mà phương pháp Elbow đã chọn.



Hình 2.4. Ví dụ minh họa cách xác định số cụm bằng phương pháp Elbow

Tuy nhiên, phương pháp này vẫn còn mặt hạn chế, đó là đôi khi có nhiều hơn một điểm "khuỷu tay" hoặc không có điểm "khuỷu tay" nào cả.

2.2.3. Xác định số cụm dựa trên phương pháp phê duyệt chéo

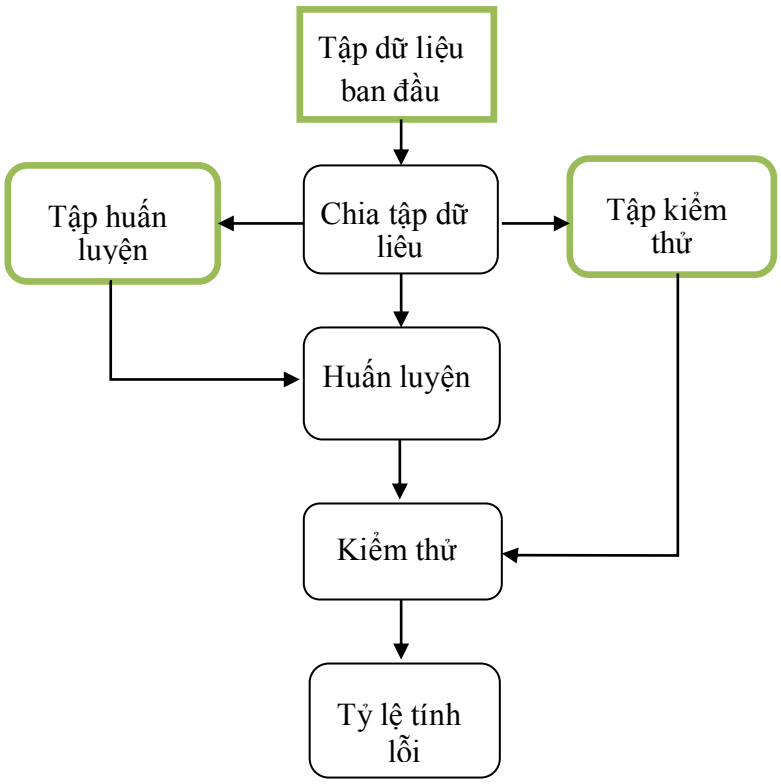
Phương pháp phê duyệt chéo (*cross validation*) chia dữ liệu thành m phần. Sử dụng m-1 phần cho mô hình gom cụm. Sử dụng phần còn lại cho việc kiểm tra chất lượng mô hình gom cụm. Kiểm tra với K>0, lặp lại m lần và tìm ra giá trị K phù hợp với dữ liệu.

Có ba phương pháp phê duyệt chéo phổ biến:

- **Từ bỏ một phần (Holdout):** Trong phương pháp từ bỏ một phần, dữ liệu được phân chia ngẫu nhiên thành 2 phần là: tập dữ liệu đào tạo và tập dữ liệu kiểm tra. Thông thường 2/3 dữ liệu cấp cho tập dữ liệu đào tạo, phần còn lại cho tập dữ liệu kiểm tra. Phương pháp này phù hợp với tập dữ liệu có kích thước lớn.
    - Toàn bộ tập dữ liệu được chia thành 2 tập con **không giao nhau:**

- Tập huấn luyện – để huấn luyện hệ thống, sử dụng cho mô hình gom cụm
- Tập kiểm thử - để kiểm tra chất lượng mô hình gom cụm
- Thường lựa chọn tập huấn luyện chiến 2/3 toàn bộ tập dữ liệu, 1/3 còn lại dùng để kiểm thử

- Các yêu cầu:
  - Bất kỳ dữ liệu nào thuộc tập kiểm thử đều không được sử dụng trong quá trình huấn luyện hệ thống
  - Bất kỳ dữ liệu nào được sử dụng trong giai đoạn huấn luyện hệ thống (thuộc tập huấn luyện) đều không được sử dụng trong giai đoạn đánh giá hệ thống.



Hình 2.5. Mô tả phương pháp từ bỏ một phần

- **Phê duyệt chéo K-nếp gấp:** Đây là nâng cấp của holdout. Toàn bộ dữ liệu được chia thành m tập con không giao nhau có kích thước xấp xỉ nhau. Thường lựa chọn m =10, hoặc 5. Phương pháp này phù hợp với tập dữ liệu vừa và nhỏ.

- Mỗi lần lặp, m-1 tập con được sử dụng cho mô hình gom cụm (tập huấn luyện), và một tập con còn lại được sử dụng để kiểm tra chất lượng mô hình gom cụm (tập kiểm thử)
- m giá trị lỗi (mỗi giá trị tương ứng với một tập con) được tính trung bình cộng để thu được giá trị lỗi tổng thể

- **Phê duyệt chéo từng phần tử** *(Leave-one-out cross validation)*: Tương tự như phê duyệt chéo K-nếp gấp nhưng tối đa hóa số tập con. Trong phương pháp này, số lượng nhóm các tập con chính bằng kích thước của tập dữ liệu (mỗi nhóm chỉ bao gồm một phần tử). Do đó phương pháp này có chi phí tính toán rất cao, chỉ phù hợp với một tập dữ liệu rất nhỏ.

2.2.4. Xác định số cụm dựa trên độ chồng và độ nén của dữ liệu

Độ nén chỉ ra mức độ tương đồng của các đối tượng dữ liệu trong một cụm và được tính toán dựa trên giá trị hàm liên thuộc của các đối tượng dữ liệu. Độ chồng nhau chỉ ra mức độ chồng nhau giữa các cụm mờ và thu được bởi tính toán tỷ lệ trùng lặp của các đối tượng dữ liệu thuộc ở hai hay nhiều cụm.

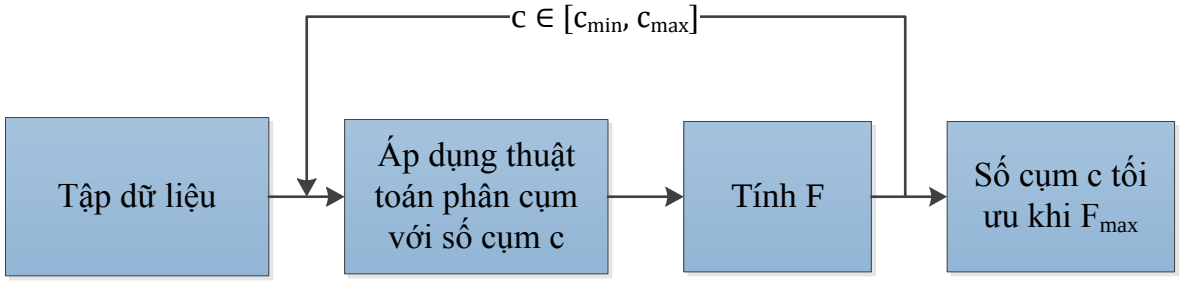
Một phân cụm tốt sẽ có sự sai khác trong mỗi cụm nhỏ (độ nén lớn) và phân tách rõ giữa các cụm (độ chồng nhau nhỏ). Do vậy, các tiêu chí được sử dụng để đánh giá chất lượng phân cụm gồm [21, 20]:

- *Độ nén:* đo mức độ tương đồng của các đối tượng dữ liệu trong một cụm. Bằng trực quan ta thấy, cụm càng tương đồng thì các điểm dữ liệu phân phối càng gần tâm cụm;
- *Độ phân tách*: đo độ tách biệt giữa các cụm. Thường được đo bằng khoảng cách giữa các cụm;
- *Độ chồng nhau*: chỉ ra mức độ chồng nhau giữa các cụm. Độ chồng nhau càng nhỏ thì các cụm càng phân tách rõ và ngược lại.

Việc ước lượng số cụm tối ưu thường được thực hiện nhờ xác định cực trị một hàm chỉ số đánh giá chất lượng phân cụm chọn trước. Quá trình đi tìm số lượng cụm tối ưu thực hiện theo lược đồ sau (được minh họa trong hình 2.6) [23, 22]:

- Thực hiện lặp thuật toán phân cụm với số cụm c lần lượt nhận giá trị trong khoảng [c_{min} , c_{max}] cho trước;

- Tính toán giá trị chỉ số đánh giá phân cụm cho mỗi kết quả phân cụm ở bước 1;
- Chọn số cụm tối ưu c ứng với kết quả phân cụm tốt nhất theo tiêu chí của chỉ số đã chọn;



Hình 2.6: Quá trình ước lượng số cụm tối ưu dựa trên độ chồng và độ nén của dữ liệu [2]

Theo [2], nếu đặt F là hiệu của hai thuộc tính độ nén và độ chồng nhau của các cụm thì bài toán trở thành bài toán đi tìm giá trị số cụm c mà tại đó hàm F đạt giá trị cực đại:

F = Compactness (c, U) – Overlap (c,U)

Trong đó:

- *Compactness (c, U)* là độ nén của các đối tượng dữ liệu trong một cụm, chỉ số sử dụng hàm đo độ nén xác định bởi [2]:

  Compactness (c, U) = $\sum_{i=1}^{c}\sum_{j=1}^{n} u_{ij}^2/u_M$,

  $u_M = max_{1\le i\le c}\{\sum_{j=1}^{n} u_{ij}^2\}$

- *Overlap (c,U)* là độ chồng của toàn bộ phân hoạch mờ, được xác định bằng tổng các giá trị độ chồng nhau giữa mỗi cặp cụm [2]:

  Overlap (c,U) = $\sum_{a=1}^{c-1}\sum_{b=a+1}^{c} O_{ab}(c,U)$

  Trong đó: O_{ab}(c,U) là độ chồng nhau giữa hai cụm C_a và C_b được tính toán từ mức độ chồng nhau O_{abj}(c,U) của mỗi đối tượng dữ liệu x_j mà nó liên thuộc đủ mạnh tới cả hai cụm mờ C_a và C_b [2].

  $O_{ab}(c,U) = \frac{1}{n}\sum_{j=1}^{n} O_{abj}(c,U)$, a,b = 1,...,c; a≠b

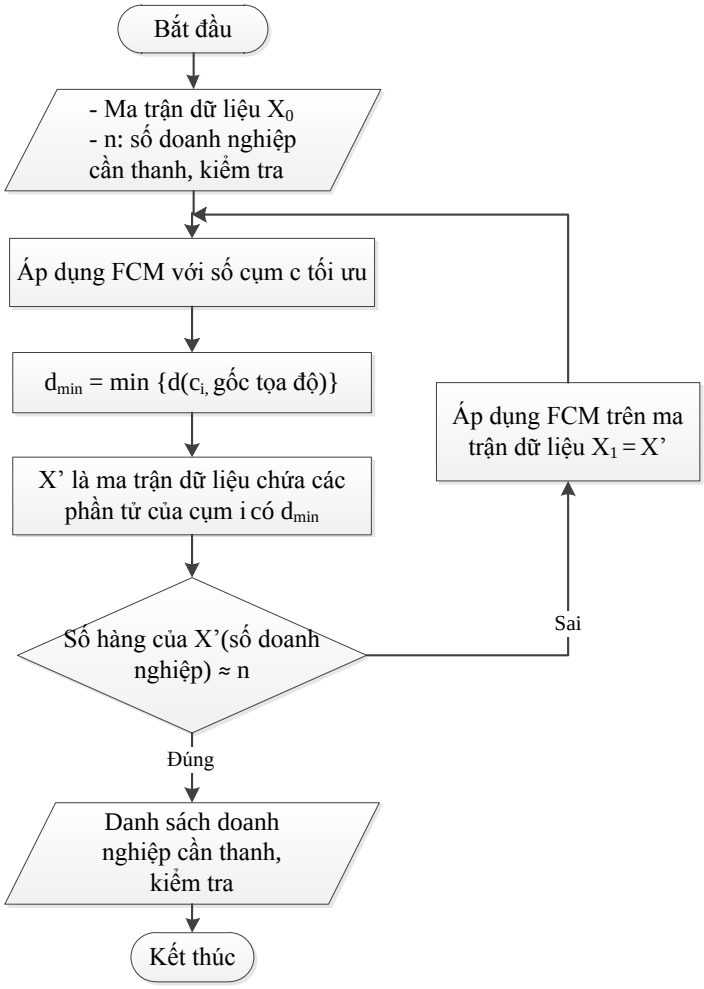
  $O_{abj}(c,U) = \begin{cases} 1 - |u_{aj} - u_{bj}| & \text{nếu } |u_{aj} - u_{bj}| \ge T_0 \text{ và } a \ne b \\ 0 & \text{ngược lại} \end{cases}$

(Giá trị T_0 nhỏ cho phép chỉ số càng hiệu lực trong trường hợp các cụm chồng nhau)

2.3. Đề xuất phương án áp dụng thuật toán FCM và phương pháp xác định số cụm vào bài toán lựa chọn nhóm doanh nghiệp rủi ro vi phạm thuế cao

Dựa vào lý thuyết bài toán phân cụm mờ và các phương pháp xác định số cụm trong gom cụm dữ liệu, luận văn đề xuất phương án áp dụng vào bài toán khoanh vùng doanh nghiệp có khả năng rủi ro vi phạm thuế cao đối với tập dữ liệu bất kỳ như hình 2.7 sau. Trong đó:

- Tập các doanh nghiệp có khả năng rủi ro vi phạm thuế cao sẽ thuộc tập dữ liệu X'.
- Tập dữ liệu đầu vào được thu thập từ các giá trị chỉ tiêu thuộc tờ khai thuế GTGT và báo cáo tài chính doanh nghiệp.



Hình 2.7. Đề xuất phương án lựa chọn nhóm doanh nghiệp rủi ro vi phạm thuế

cao

Chú giải:

- X_0 là tập dữ liệu ban đầu gồm n_1 hàng tương ứng với số doanh nghiệp và k cột tương ứng với các giá trị chỉ tiêu thuộc tờ khai khấu trừ thuế GTGT và báo cáo tài chính doanh nghiệp
- Áp dụng thuật toán FCM với tập dữ liệu đầu vào là X_0 và các tham số phù hợp. Chọn số cụm c sao cho giữa các cụm sự sai khác trong mỗi cụm nhỏ (độ nén lớn) và phân tách rõ giữa các cụm (độ chồng nhau nhỏ).
- c_i là tâm cụm thứ i.
- X' là ma trận dữ liệu của cụm thứ i, có khoảng cách giữa tâm cụm và gốc tọa độ là nhỏ nhất (d nhỏ nhất).

  Ma trận X' gồm n_2 hàng tương ứng với số doanh nghiệp và k cột tương ứng với các giá trị chỉ tiêu thuộc tờ khai GTGT và báo cáo tài chính doanh nghiệp
- n là số doanh nghiệp cần thanh tra, kiểm tra được xác định trước.

CHƯƠNG 3: ỨNG DỤNG PHƯƠNG PHÁP PHÂN CỤM MỜ CHO BÀI TOÁN PHÂN TÍCH THÔNG TIN RỦI RO QUẢN LÝ THUẾ DOANH NGHIỆP

3.1. Mô tả bài toán

Từ trước đến nay, công tác thanh tra, kiểm tra rủi ro vi phạm thuế doanh nghiệp luôn được chú trọng để hạn chế thất thu ngân sách nhà nước, đồng thời qua đó cũng phát hiện nhiều thủ đoạn gian lận thuế của các doanh nghiệp. Ngày nay, khi tiến hành thanh tra, kiểm tra thuế, Cơ quan thuế có xu hướng chuyển đổi từ việc thanh tra, kiểm tra theo diện rộng, nặng tính cảm tính sang thanh tra, kiểm tra theo hệ thống tiêu thức lựa chọn khách quan, khoa học, đi vào chiều sâu theo mức độ vi phạm rủi ro. Sự thay đổi này, sẽ tăng tính hiệu quả của hoạt động thanh kiểm tra, mở rộng số lượng, trường hợp doanh nghiệp được cơ quan thuế giám sát việc tuân thủ nghĩa vụ thuế. Việc lựa chọn đối tượng thanh tra, kiểm tra theo xu hướng ngẫu nhiên, dàn trải và không phụ thuộc vào việc phân tích mức độ rủi ro vi phạm thuế của doanh nghiệp, sẽ được thay thế bởi một cơ chế lựa chọn đối tượng bị thanh tra, kiểm tra tập trung vào nhóm doanh nghiệp có rủi ro vi phạm thuế.

Hiện nay, Tổng cục Thuế đã ban hành bộ tiêu chí đánh giá rủi ro bao gồm 20 tiêu chí tĩnh (áp dụng thống nhất tất cả các cục thuế) và các tiêu chí động (do từng Cục thuế tự xây dựng phù hợp với thực tế địa phương theo gợi ý của Tổng cục Thuế) nhằm lựa chọn các trường hợp kiểm tra, thanh tra [25]. Tuy nhiên các tiêu chí và chỉ số đánh giá mức độ tuân thủ thuế của doanh nghiệp còn có nhiều bất cập. Các chuyên gia nghiệp vụ thuế xây dựng bộ tiêu chí phân tích rủi ro có công thức phân tích như sau:

Điểm rủi ro = α_1 * Tiêu chí 1 +...+α_n * Tiêu chí n.

Trong đó, chuyên gia nghiệp vụ xác định trọng số α cho từng tiêu chí chủ yếu dựa vào kinh nghiệm. Bản chất hệ thống là quản lý rủi ro vì các tham số α_1,... α_n do chuyên gia xác định theo kinh nghiệm mà chưa được "học tự động từ dữ liệu" [24].

Do đó, trong phạm vi luận văn này, luận văn đề xuất cách tiếp cận phân cụm dữ liệu để đưa ra phương án khoanh vùng, lựa chọn các nhóm đối tượng, doanh nghiệp có rủi ro vi phạm thuế mà không sử dụng số liệu từ kinh nghiệm

của các chuyên gia. Dựa vào các dữ liệu trên tờ khai thuế, báo cáo tài chính của các doanh nghiệp, sử dụng thuật toán phân cụm để phân loại, khoanh vùng các đối tượng, từ đó giúp tăng cường tính hiệu quả trong việc lựa chọn trường hợp thanh tra, kiểm tra.

3.2. Dữ liệu đầu vào

Dựa vào bài toán đặt ra tại mục 3.1, luận văn lựa chọn tập dữ liệu đầu vào là các giá trị chỉ tiêu trên tờ khai khấu trừ thuế GTGT và báo cáo tài chính doanh nghiệp, vì bất kỳ doanh nghiệp nào cũng phải kê khai tờ khai khấu trừ thuế GTGT và báo cáo tài chính (các loại tờ khai khác chỉ một số doanh nghiệp phải kê khai):

- Tờ khai khấu trừ thuế GTGT của doanh nghiệp giúp Nhà nước kiểm soát được hoạt động, sản xuất, nhập khẩu, kinh doanh hàng hóa nhờ kiểm soát được hệ thống hóa đơn, chứng từ, khắc phục được nhược điểm của thuế doanh thu là trốn thuế. Qua đó, còn cung cấp cho công tác nghiên cứu, thống kê, quản lý những số liệu quan trọng [30].
- Báo cáo tài chính là những báo cáo tổng hợp nhất về tình hình tài sản, vốn chủ sở hữu và nợ phải trả cũng như tình hình tài chính, kết quả kinh doanh trong kỳ của doanh nghiệp. Báo cáo tài chính có ý nghĩa quan trọng đối với công tác quản lý doanh nghiệp cũng như đối với các cơ quan chủ quản và các đối tượng quan tâm [29].

Dữ liệu đầu vào được thu thập từ dữ liệu mẫu trên thông tin tờ khai khấu trừ thuế GTGT, báo cáo tài chính doanh nghiệp và được lưu trữ trong tệp *data.csv*. Cấu trúc dữ liệu trong tệp *data.csv* bao gồm:

- 13 cột tương ứng với các giá trị chỉ tiêu thuộc tờ khai khấu trừ thuế GTGT và giá trị chỉ tiêu thuộc báo cáo tài chính của doanh nghiệp. Cụ thể:
    - 7 cột tương ứng với 7 giá trị chỉ tiêu thuộc tờ khai khấu trừ thuế GTGT trong kỳ
    - 5 cột tương ứng với giá trị 5 chỉ tiêu thuộc tờ khai khấu trừ thuế GTGT kỳ trước
    - 1 cột tương ứng với giá trị vốn đầu tư của chủ sở hữu (số cuối kỳ) trên báo cáo tài chính của doanh nghiệp
- 644 hàng tương ứng với dữ liệu trên tờ khai khấu trừ thuế GTGT và báo

cáo tài chính của 644 doanh nghiệp.

Chi tiết các chỉ tiêu thuộc tờ khai thuế khấu trừ GTGT và báo cáo tài chính doanh nghiệp trong tập dữ liệu được thể hiện tại bảng 3.1 như sau:

Bảng 3.1. Mô tả thông tin các chỉ tiêu các cột dữ liệu thuộc tập dữ liệu *data.csv*

STT	Mã chỉ tiêu	Tên chỉ tiêu	Kiểu dữ liệu	Ghi chú
Các cột chứa giá trị chỉ tiêu thuộc tờ khai thuế GTGT				
1.	#34	Tổng doanh thu của hàng hoá dịch vụ bán ra trong kỳ	Kiểu số	Lấy giá trị trên tờ khai khấu trừ thuế GTGT tại kỳ kiểm tra và kỳ liền trước đó
2.	#23	Doanh số hàng hoá dịch vụ mua vào trong kỳ	Kiểu số	
3.	#35	Tổng số thuế hàng hóa, dịch vụ bán ra trong kỳ	Kiểu số	
4.	#24	Số thuế GTGT của hàng hóa, dịch vụ mua vào	Kiểu số	
5.	#29	Doanh số hàng hóa, dịch vụ bán ra chịu thuế suất %	Kiểu số	
6.	#25	Tổng số thuế GTGT được khấu trừ kỳ này	Kiểu số	Lấy giá trị trên tờ khai GTGT tại kỳ kiểm tra
7.	#43	Thuế GTGT còn được khấu trừ chuyển kỳ sau	Kiểu số	
Cột chứa giá trị chỉ tiêu thuộc báo cáo tài chính doanh nghiệp				
8.	#411	Vốn đầu tư của chủ sở hữu	Kiểu số	Lấy giá trị số cuối kỳ

Luận văn lựa chọn lấy giá trị của một số chỉ tiêu trên tờ khai khấu trừ thuế GTGT kỳ liền trước đó, và các giá trị trên tờ khai khấu trừ thuế GTGT kỳ kiểm tra, vốn đầu từ của chủ sở hữu do các chỉ tiêu này có ý nghĩa rất quan trọng trong việc đánh giá doanh nghiệp:

- Đánh giá sự biến động của việc kê khai doanh thu và thuế GTGT

của hàng hóa, dịch vụ mua vào, bán ra

- Đánh giá và theo dõi sự biến động về thuế GTGT của hàng hoá bán ra giữa các kỳ nhằm phát hiện những bất thường có thể xảy ra
- Đánh giá và theo dõi sự biến động về thuế GTGT của hàng hoá mua vào giữa các kỳ nhằm phát hiện những bất thường có thể xảy ra
- Đánh giá và theo dõi sự biến động doanh thu hoạt động xuất khẩu, xây lắp công trình cho doanh nghiệp chế xuất, vận tải quốc tế... giữa các kỳ nhằm phát hiện những bất thường có thể xảy ra
- Đánh giá và theo dõi sự biến động về kê khai thuế GTGT đầu ra và hàng tồn kho
- Đánh giá mức độ tuân thủ kê khai thuế GTGT khi phát sinh doanh thu hàng hóa dịch vụ bán ra không chịu thuế GTGT và việc phân bổ thuế GTGT đầu vào được khấu trừ tương ứng
- Đánh giá tỷ lệ tăng doanh thu so với vốn chủ sở hữu của đơn vị
- Đánh giá mức độ tuân thủ về việc kê khai thuế GTGT đầu ra của doanh nghiệp

3.3. Lựa chọn công cụ, môi trường thực nghiệm

Với bài toán phân cụm các doanh nghiệp rủi ro quản lý thuế theo tập dữ liệu đã đặt ra ở mục 3.2, ngôn ngữ được sử dụng trong chương trình là ngôn ngữ Matlab. Ngôn ngữ lập trình này hỗ trợ trong rất nhiều ứng dụng như:

- Xây dựng chương trình giải quyết các bài toán về toán học
- Xây dựng các chương trình mô phỏng, thống kê
- Đặc biệt ngôn ngữ lập trình Matlab hỗ trợ hệ logic mờ, cung cấp các thư viện về các hàm dữ liệu logic mờ

Vì vậy, việc lựa chọn ngôn ngữ lập trình Matlab trong phần ứng dụng này

sẽ tận dụng được các thư viện sẵn có nhằm hỗ trợ quá trình xây dựng thuật toán.

3.4. Phương pháp phân cụm và lựa chọn số cụm

3.4.1. Xác định phương pháp phân cụm

- Dữ liệu của các doanh nghiệp khá tương đồng, khi phân cụm rủi ro vi phạm cho doanh nghiệp không có ranh giới rõ ràng để khẳng định một doanh nghiệp là rủi ro vi phạm cao hay không. Ranh giới đó là mờ. Ta chỉ có thể nói doanh nghiệp đó rủi ro cao ở mức độ bao nhiêu phần trăm. Do đó khi phân cụm doanh nghiệp, sẽ có nhiều đối tượng nằm trong ranh giới giữa các cụm, đối tượng có thể thuộc vào nhiều cụm.
- Khái niệm "rủi ro" về bản chất là mờ, vì:
    - Có nhiều mức độ rủi ro khác nhau: Rủi ro cao, rủi ro vừa, rủi ro thấp, hay không rủi ro
    - Có yếu tố bất định, ngẫu nhiên
    - Mức độ rủi ro được xác định tùy theo quan điểm của người đánh giá

Do đó với bài toán phân tích thông tin rủi ro quản lý thuế doanh nghiệp nên biểu diễn bằng tập mờ, sẽ cho kết quả tốt hơn, luận văn lựa chọn phương pháp phân cụm mờ để ứng dụng vào bài toán đặt ra tại mục 3.1 và tập dữ liệu đầu vào đưa ra tại mục 3.2.

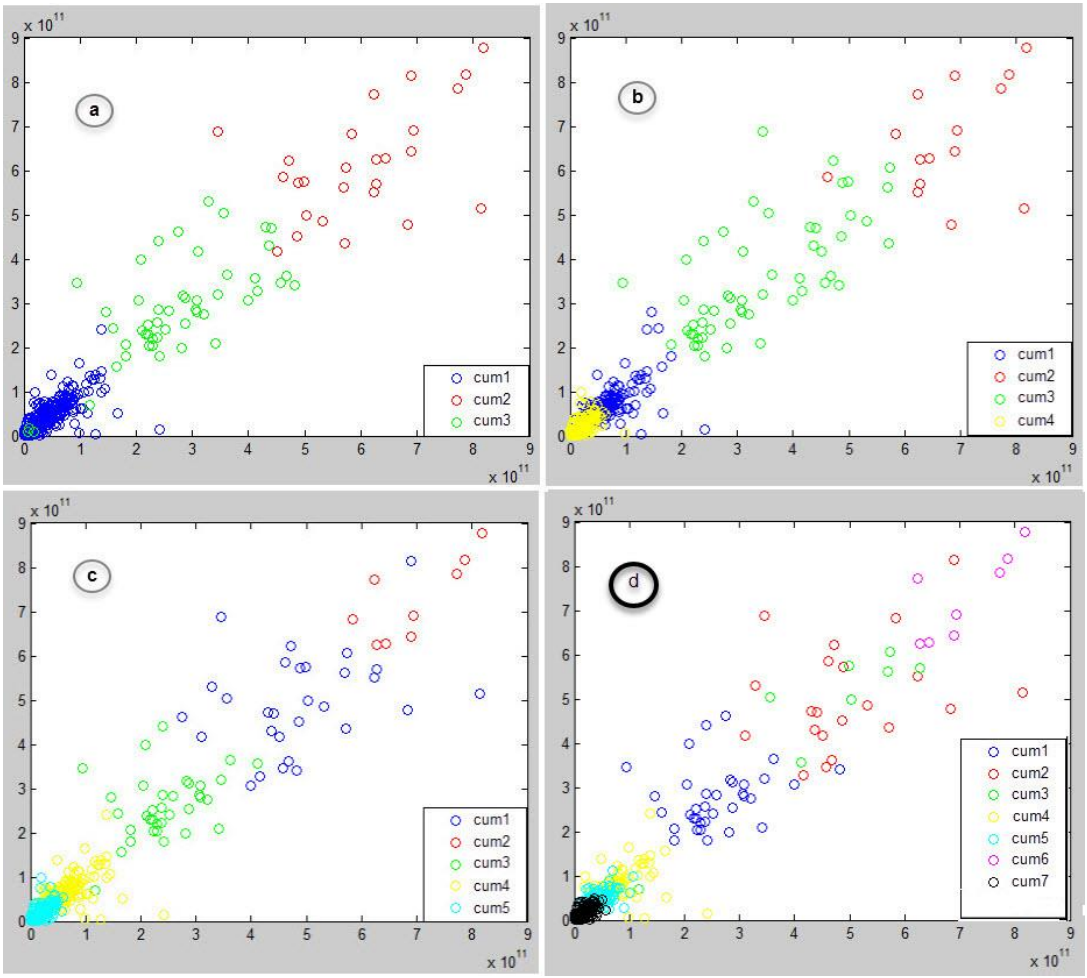
3.4.2. Lựa chọn số cụm

Quá trình phân cụm dữ liệu nhằm xác định các nhóm đối tượng dữ liệu tương tự, từ đó khảo sát các cụm sẽ giúp khái quát, nhanh chóng rút ra các đặc điểm của khối dữ liệu lớn. Tuy nhiên, trong hầu hết các thuật toán phân cụm, tham số số cụm không được biết trước và thuật toán thường yêu cầu người dùng phải xác định trước số lượng các cụm, ứng với mỗi số lượng cụm khác nhau sẽ cho ra các kết quả phân cụm khác nhau [2].

Khi áp dụng thuật toán phân cụm cho từng bài toán cụ thể, việc ước lượng số cụm ảnh hưởng lớn đến chất lượng phân cụm. Một phân cụm tốt sẽ có sự sai khác trong mỗi cụm nhỏ (độ nén lớn) và phân tách rõ giữa các cụm (độ chồng nhau nhỏ). Do vậy, trong phạm vi bài toán đã nêu tại mục 3.1 và tập dữ liệu mẫu *data.csv* đặt ra tại mục 3.2, luận văn lựa chọn việc xác định số cụm dựa trên độ

chồng và độ nén của dữ liệu (phương pháp này đã được trình bày tại mục 2.2.4). Cụ thể như sau:

- Thực hiện lặp thuật toán phân cụm mờ trên tập dữ liệu *data.csv* với số cụm c nằm trong khoảng [3, 7]. Hình 3.1 dưới đây là kết quả thu nhận được:



Hình 3.1. Kết quả phân cụm dữ liệu với số cụm c = [3, 7]

(a) Tập dữ liệu gồm 3 cụm

(b) Tập dữ liệu gồm 4 cụm

(c) Tập dữ liệu gồm 5 cụm

(d) Tập dữ liệu gồm 7 cụm

- Áp dụng công thức tính độ tương đồng của các đối tượng trong một cụm, độ chồng nhau giữa các cụm và F là hiệu của hai thuộc tính độ nén và độ chồng nhau của các cụm (công thức được nêu tại mục 2.2.4), luận văn

tính độ chồng nhau của mỗi đối tượng x_j với $T_0 = 0.1$, tính hàm F tương ứng với số cụm c=[3,7], được kết quả như bảng 3.2 sau:

Bảng 3.2. Kết quả tính F với số cụm c=[3,7]

c	Compactness (c, U)	Overlap (c,U)	F
3	1,337962	0,266365	**1,071597**
4	2,000024	1,151229	0,848795
5	2,178677	1,768209	0,410468
6	2,644531	3,049731	-0,4052
7	2,845703	3,949323	-1,10362

Số cụm c là tối ưu khi hàm F đạt giá trị cực đại. Dựa vào kết quả bảng 3.2, nhận thấy: trong phạm vi bài toán đã nêu tại mục 3.1 và tập dữ liệu mẫu *data.csv* đặt ra tại mục 3.2, số cụm tối ưu là c = 3.

3.5. Kết quả thực nghiệm

Trong phần thực nghiệm, luận văn áp dụng thuật toán FCM với các tham số: tham số mờ m = 2, sai số ε = 0.01, số lần lặp tối đa là 1000, số cụm c = 3.

Môi trường lập trình là Matlab, với cấu hình máy tính: Ram 4GB, tốc độ xử lý của CPU là 2.30 GHz

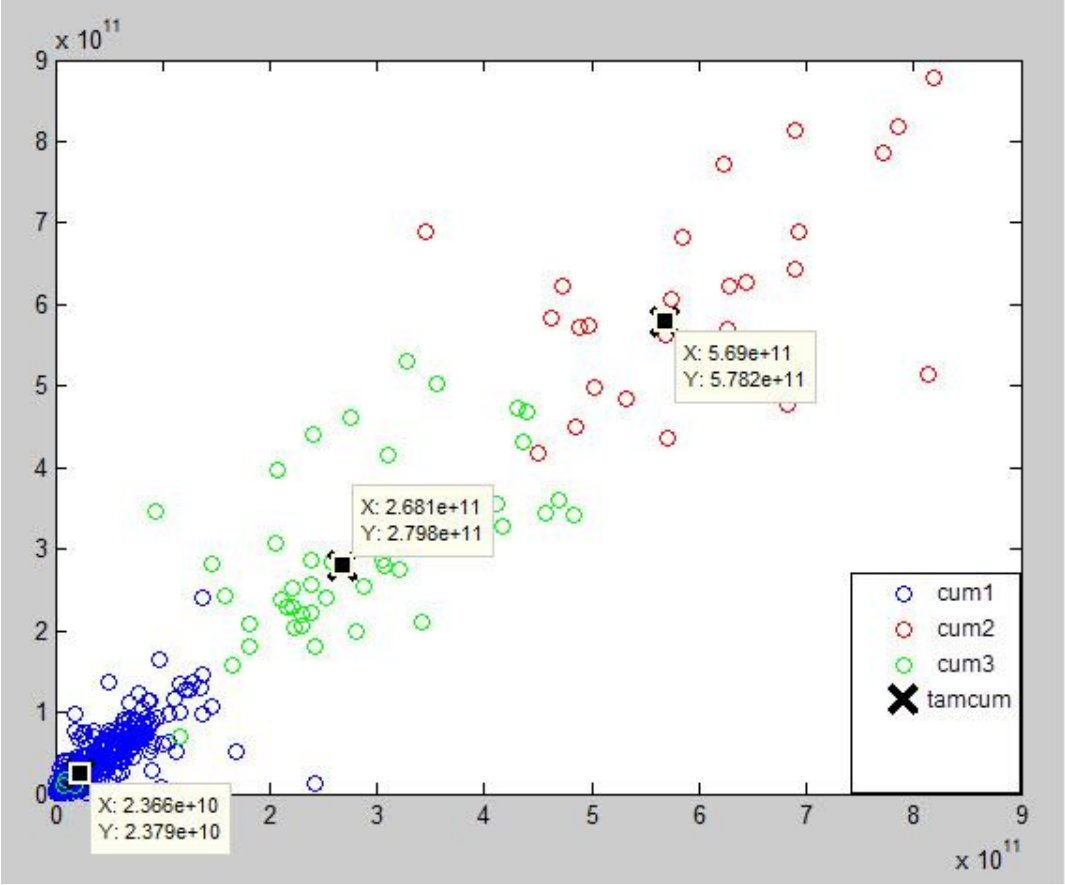
3.5.1. Kết quả phân loại doanh nghiệp

3.5.1.1. Kết quả phân cụm trên tập dữ liệu data.csv

Kết quả phân cụm doanh nghiệp rủi ro vi phạm với tập dữ liệu *data.scv* với số cụm c = 3 được thể hiện tại bảng 3.2 và hình 3.2 dưới đây:

Bảng 3.3. Kết quả phân cụm doanh nghiệp trên tập dữ liệu *data_cum.csv*

STT	Thứ tự cụm	Số doanh nghiệp thuộc cụm
1	1	568
2	2	26
3	3	50



Hình 3.2. Kết quả phân cụm dữ liệu với tập dữ liệu data.csv

3.5.1.2. So sánh kết quả phân cụm doanh nghiệp với mức rủi ro vi phạm thuế tương ứng được đánh giá từ kinh nghiệp của chuyên gia

Theo các chuyên gia nghiệp vụ thuế, doanh nghiệp rủi ro vi phạm thuế được chia làm 3 mức: mức 0, mức 1 và mức 2.

Luận văn đã tiến hành thu thập thông tin rủi ro vi phạm thuế của 644 doanh nghiệp thuộc tập dữ liệu *data.csv* (thông tin rủi ro vi phạm này được tính toán dựa trên kinh nghiệm của các chuyên gia nghiệp vụ thuế) và tiến hành so sánh với kết quả phân cụm doanh nghiệp (bảng 3.2 mục 3.5.1.1) được kết quả như bảng 3.3 sau:

Bảng 3.3. So sánh kết quả phân cụm dữ liệu *data.csv* với thông tin rủi ro vi phạm thuế

STT	Thứ tự cụm	Số doanh nghiệp thuộc cụm	Tỷ lệ dữ liệu so với mức rủi ro vi phạm
1	1	568	- Mức 0: 0% - Mức 1: 37.68% - Mức 2: 62.32%
2	2	26	- Mức 0: 80.77% - Mức 1: 19.23% - Mức 2: 0%
3	3	50	- Mức 0: 4% - Mức 1: 80% - Mức 2: 16%

Dựa vào bảng 3.3, nhận thấy các đối tượng trong cùng một nhóm có độ tương đồng nhau tương đối cao về mức rủi ro vi phạm thuế, đại đa số các doanh nghiệp trong cùng một cụm có cùng giá trị mức rủi ro, cụ thể:

- Cụm 1: 62.32% doanh nghiệp thuộc mức rủi ro 2
- Cụm 2: 80.77% doanh nghiệp thuộc mức rủi ro 0
- Cụm 3: 80% doanh nghiệp thuộc mức rủi ro 1

3.5.1.3. Xác định doanh nghiệp thuộc cụm

Tập dữ liệu ban đầu của doanh nghiệp có chứa thông tin chi tiết của doanh nghiệp (bao gồm MST, tên doanh nghiệp, địa chỉ, ...), khi trích xuất thông tin vào tập *data.csv* để thực nghiệm chỉ sử dụng các thông tin các giá trị chỉ tiêu trên tờ khai khấu trừ thuế GTGT và báo cáo tài chính doanh nghiệp. Do đó sau khi có kết quả phân cụm cho tập dữ liệu *data.csv*, luận văn tiến hành ánh xạ thông tin phân cụm trên tập *data.csv* với thông tin chi tiết ban đầu để xác định doanh nghiệp thuộc cụm.

MST	Tên Doanh nghiệp	Địa chỉ Doanh nghiệp	Kết quả phân nhóm theo FCM	Kết quả phân nhóm thực tế
2300108456	Công ty TNHH Canon Việt Nam chi nhánh Quế Võ	Xã Đông Tiến, Huyện Quế Võ, Bắc Ninh	3	3
2300220602	Công ty điện lực Bắc Ninh	Phường Vũ Ninh, Tp Bắc Ninh, Tỉnh Bắc	3	3
2300205562	Công ty Cổ phần Tập đoàn Dabaco Việt Nam	Phường Vệ An, Tp Bắc Ninh, Tỉnh Bắc	1	1
0101395280	Công ty TNHH Canon Việt Nam chi nhánh Tiên Sơn	Xã Dũng Liệt, Huyện Yên Phong, Bắc N	3	2
2300232767	Công ty TNHH Funing Precision Component	Thị trấn Hồ, Huyện Thuận Thành, Bắc N	3	3
2300236955	Công ty Cổ phần VS Industry Việt Nam	Phường Khúc Xuyên, Tp Bắc Ninh, Tỉnh	2	2
2300205883	Công ty TNHH GoerTek Vina	Thị trấn Lim, Huyện Tiên Du, Bắc Ninh	3	3
2300240140	Công ty TNHH Một thành viên Chánh Phát	Phường Vũ Ninh, Tp Bắc Ninh, Tỉnh Bắc	2	2
2300247266	Công ty cổ phần may mặc xuất khẩu Việt Nam - Ba Lan	Phường Khúc Xuyên, Tp Bắc Ninh, Tỉnh	1	1
2300657128	Công ty Trách nhiệm hữu hạn Đại Tấn Phát	Phường Khúc Xuyên, Tp Bắc Ninh, Tỉnh	1	1
2300345954	Công ty TNHH HaNoi Doosung Tech	Thị trấn Hồ, Huyện Thuận Thành, Bắc N	3	1

Hình 3.3. Xác định doanh nghiệp thuộc cụm

Lưu ý: Thông tin chi tiết các doanh nghiệp trên hình 3.3 chỉ mang tính chất tham khảo.

3.5.2. Kết luận

Dựa vào kết quả thực nghiệp, bộ dữ liệu đầu vào, nhận thấy cách chọn các tiêu chí, thuộc tính dữ liệu đầu vào và cách phân cụm của luận văn phù hợp với mục tiêu bài toán đặt ra là phân tích thông tin rủi ro quản lý thuế. Trong công tác quản lý rủi ro vi phạm thuế nên có 3 giá trị mức rủi ro.

Các chuyên gia nghiệp vụ thuế xác định 3 mức rủi ro vi phạm thuế của doanh nghiệp lần lượt là:

- Mức 0: rủi ro vi phạm thấp – không rủi ro
- Mức 1: rủi ro vi phạm vừa
- Mức 2: rủi ro vi phạm cao

Dựa vào kết quả phân cụm doanh nghiệm tập dữ liệu *data.csv*, ta thấy: các doanh nghiệp có rủi ro vi phạm cao thường tập trung tại các cụm nằm gần gốc tọa độ Oxy, các doanh nghiệp trong các phân cụm càng xa gốc tọa độ thì mức rủi ro vi phạm càng giảm (xem chi tiết kết quả phân cụm tại hình 3.2 và bảng 3.3).

Kết quả phân loại, khoanh vùng các đối tượng doanh nghiệp theo mức độ rủi ro vi phạm này sẽ giúp tăng tính hiệu quả trong việc lựa chọn, phân tích thông tin rủi ro quản lý thuế doanh nghiệp, tăng tính hiệu quả của hoạt động thanh tra kiểm tra, mở rộng số lượng, trường hợp doanh nghiệp được cơ quan thuế giám sát việc tuân thủ nghĩa vụ thuế.

3.6. Ứng dụng kết quả thực nghiệm vào bài toán khoanh vùng, lựa chọn nhóm doanh nghiệp có khả năng rủi ro vi phạm thuế cao

Dựa vào kết quả thực nghiệm (mục 3.5.2): các doanh nghiệp có rủi ro vi phạm cao thường tập trung tại cụm dữ liệu nằm gần gốc tọa độ Oxy, áp dụng phương án khoanh vùng doanh nghiệp có khả năng rủi ro vi phạm thuế cao đối với tập dữ liệu bất kỳ được đề xuất tại hình 2.7 (mục 2.3) của luận văn với các dữ liệu đầu vào như sau:

- (1): X_0 là tập dữ liệu *data.csv* (tập dữ liệu *data.csv* được mô tả tại mục 3.2)

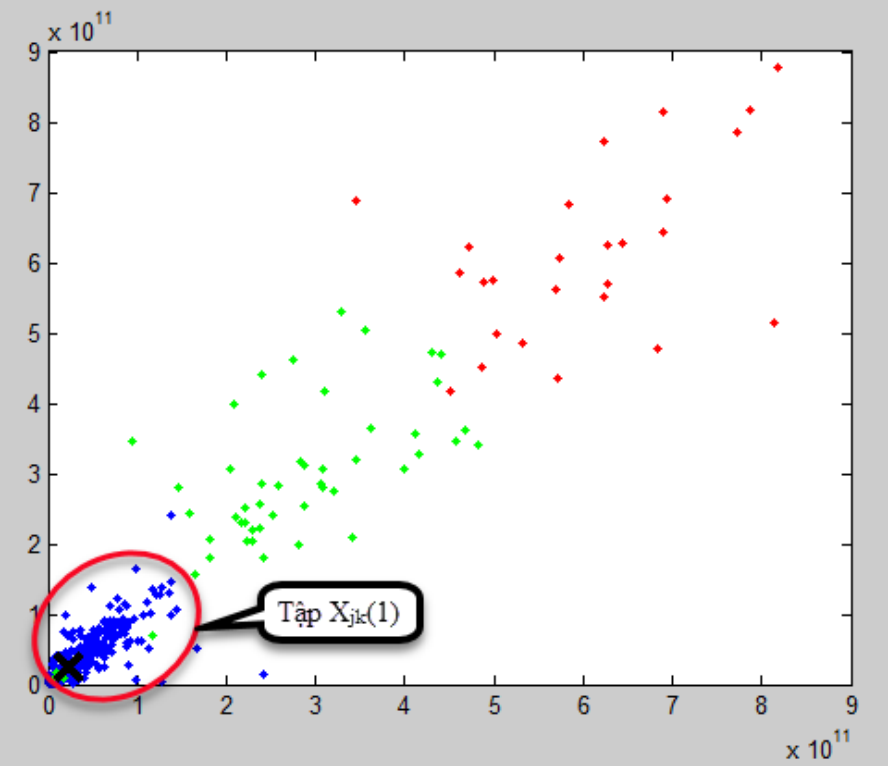
  Chọn n = [100, 200]

- (2): Áp dụng thuật toán FCM trên môi trường Matlab với các tham số: tham số mờ m = 2, sai số $\varepsilon = 0.01$, số lần lặp tối đa là 1000, số cụm c = 3

- Kết quả mong muốn: Tập dữ liệu doanh nghiệp rủi ro vi phạm X’ với n = [100, 200] và số doanh nghiệp rủi ro vi phạm cao chiếm ≥ 70% tập dữ liệu X’.

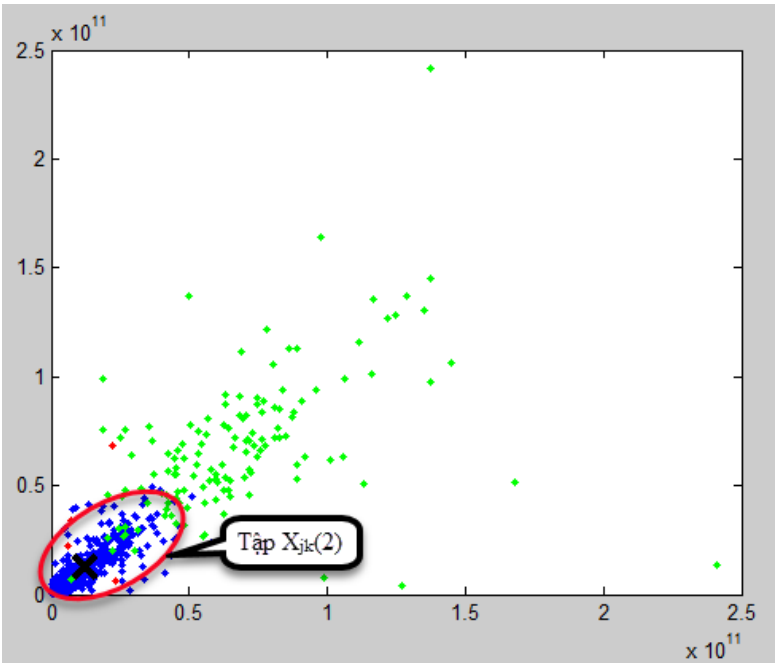
➔ Kết quả thực nghiệm:

- Áp dụng quy trình hình 2.7 lần 1: X’(1) chứa 568 doanh nghiệp và được mô phỏng trong hình 3.4 (các đối tượng thuộc tệp có dạng chấm màu xanh dương)



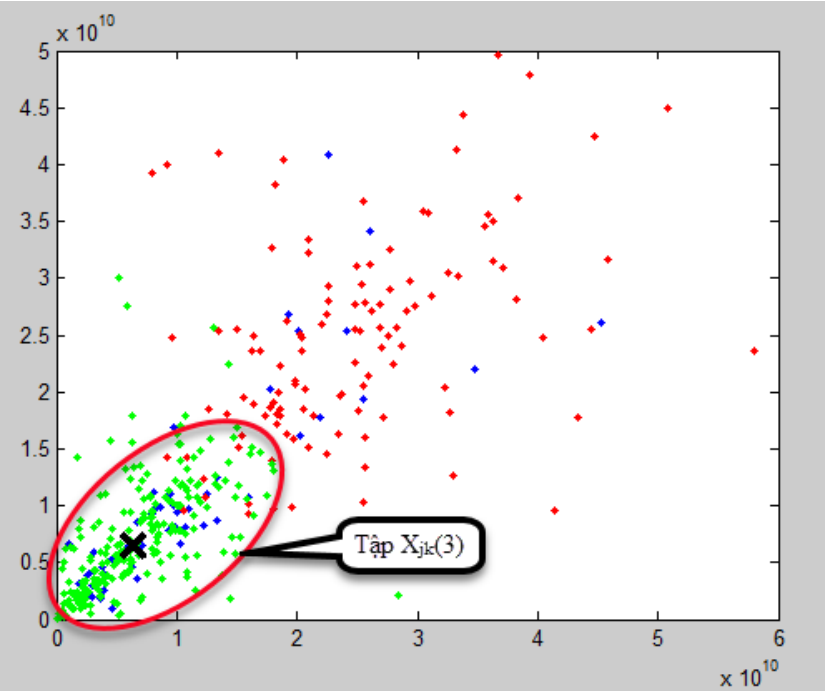
Hình 3.4. Mô phỏng tập dữ liệu X'(1)

- Áp dụng quy trình hình 2.7 lần 2 (X_1 = X'(1)): X'(2) chứa 425 doanh nghiệp và được mô phỏng trong hình 3.5 (các đối tượng thuộc tệp có dạng chấm màu xanh dương)



Hình 3.5. Mô phỏng tập dữ liệu X'(2)

- Áp dụng quy trình hình 2.7 lần 3 (X’ = X’(2)): X’(3) chứa 255 doanh nghiệp và được mô phỏng trong hình 3.6 (các đối tượng thuộc tệp có dạng chấm màu xanh lá)



Hình 3.6. Mô phỏng tập dữ liệu X’(3)

- Tương tự, áp dụng quy trình hình 2.7 lần 4 (X_1 = X’(3)), áp dụng thuật toán FCM với số cụm c =2 (do lúc này số dữ liệu thuộc tập X_1 chỉ còn 255 doanh nghiệp, nên luận văn lựa chọn chia làm 2 cụm).

  Kết quả thu được: X’(4) chứa 146 nghiệp, thỏa mãn j = [100, 200]

- Tính tỷ lệ doanh nghiệp rủi ro vi phạm cao trong tập dữ liệu nhận được bằng cách ánh xạ tương ứng MST doanh nghiệp với tập dữ liệu *data.csv* ban đầu để lấy ra mức rủi ro. Ta được kết quả như sau:

  Tập X’(4) có chứa: 71.233% (104/146) doanh nghiệp rủi ro vi phạm cao và 28.767 % (42/146) doanh nghiệp rủi ro vi phạm vừa. Thỏa mãn kết quả mong muốn.

KẾT LUẬN VÀ HƯỚNG PHÁT TRIỂN

KẾT LUẬN

Ngày nay, khai phá dữ liệu đang là lĩnh vực thời sự của ngành công nghệ thông tin thế giới nói chung và Việt Nam nói riêng. Khai phá dữ liệu đang được ứng dụng rất rộng rãi trong nhiều lĩnh vực của đời sống. Một trong những bài toán quan trọng trong lĩnh vực khai phá dữ liệu là bài toán phân cụm dữ liệu. Phân cụm dữ liệu, nói một cách khái quát là việc tự động sinh ra các cụm dựa vào sự tương tự của các đối tượng dữ liệu. Trong các kỹ thuật phân cụm dữ liệu, kỹ thuật phân cụm dữ liệu theo hướng tiếp cận mờ là một lĩnh vực nghiên cứu rộng lớn và đầy triển vọng. Với đề tài ***"Ứng dụng phương pháp phân cụm mờ cho bài toán phân tích thông tin rủi ro quản lý thuế"***, luận văn đã tập trung tìm hiểu, nghiên cứu và đạt được một số kết quả sau đây:

- Nắm bắt các khái niệm liên quan đến khai phá dữ liệu, phân cụm dữ liệu
- Phân tích một số phương pháp phân cụm dữ liệu như: phương pháp phân cụm phân hoạch, phương pháp phân cụm phân cấp, phương pháp tiếp cận dựa trên mật độ, phương pháp phân cụm dựa trên lưới và phương pháp phân cụm dựa trên mô hình.
- Tìm hiểu được một số phương pháp xác định số cụm trong gom cụm dữ liệu dựa trên phương pháp truyền thống, phương pháp Eblow, phương pháp phê duyệt chéo và phương pháp xác định số cụm dựa trên độ chồng, độ nén của dữ liệu.
- Tìm hiểu về thuật toán phân cụm mờ FCM, cài đặt thuật toán trên môi trường Matlab và thử nghiệm phân cụm các doanh nghiệp rủi ro vi phạm thuộc bộ dữ liệu mẫu về thông tin tờ khai thuế, báo cáo tài chính doanh nghiệp của 644 doanh nghiệp.
- Phân loại, khoanh vùng các đối tượng doanh nghiệp theo mức độ rủi ro vi phạm giúp tăng tính hiệu quả trong việc lựa chọn, phân tích thông tin rủi ro quản lý thuế doanh nghiệp, tăng tính hiệu quả của hoạt động thanh tra kiểm tra, mở rộng số lượng, trường hợp doanh nghiệp được cơ quan thuế giám sát việc tuân thủ nghĩa vụ thuế.

Tuy nhiên bên cạnh những kết quả đã đạt được em tự thấy luận văn còn nhiều hạn chế như về mặt trình bày những vấn đề đã hiểu, chương trình thử

nghiệm chỉ dừng ở một thuật toán phân cụm, dữ liệu đầu vào còn nhiều hạn chế. Thời gian nghiên cứu và trình độ của bản thân có hạn nên không thể tránh hỏi những thiếu sót, rất mong nhận được những ý kiến đóng góp từ quý thầy cô, anh chị và các bạn.

HƯỚNG PHÁT TRIỂN

Trên cơ sở những nghiên cứu và tìm hiểu trong luận văn, trong thời gian tới em định hướng sẽ tiếp tục nghiên cứu, mở rộng đề tài bằng cách nghiên cứu các kỹ thuật khai phá dữ liệu khác. Nghiên cứu thêm một số kỹ thuật phân cụm và đặc biệt là phân cụm mờ ứng dụng vào một số bài toán thực tế.

TÀI LIỆU THAM KHẢO

Tiếng Việt

1. An Hồng Sơn (2008), *Nghiên cứu một số phương pháp phân cụm mờ và ứng dụng*, Đại học Thái Nguyên.
2. Nguyễn Trung Đức (2013), *Tiếp cận mờ trong phân cụm dữ liệu*, Trường Đại học Công nghệ - Đại học Quốc gia Hà Nội.
3. Đặng Tiến Dũng (2003), *Tìm hiểu khái niệm quản lý và quản lý thuế*, Tạp chí thuế nhà nước.
4. Lê Tuấn Tú (2011), *Nghiên cứu xây dựng luật mờ từ dữ liệu theo phân cụm* – ĐH Công nghệ thông tin và Truyền thông.
5. Phạm Thị Thu (2007), *Thuật toán phân cụm dữ liệu mờ*, Trường Đại học dân lập Hải Phòng.
6. Nguyễn Trung Sơn (2009), *Phương pháp phân cụm và ứng dụng*, luận văn thạc sĩ Khoa học máy tính.
7. Trần Nguyên Hương (2009), *Một số thuật toán phân cụm cơ bản trong Data mining*
8. Trần Thị Yến (2012), *Phân cụm dữ liệu trừ mờ và ứng dụng*, luận văn thạc sĩ Công nghệ thông tin.
9. Vũ Hải Thuyết (2012), *Nghiên cứu một số giải thuật trong phân cụm dữ liệu,* luận văn thạc sĩ chuyên ngành Truyền dữ liệu và mạng máy tính.
10. Vũ Minh Đông (2010), *Một số phương pháp phân cụm dữ liệu*, Đại học dân lập Hải Phòng.
11. Nguyễn Hoàng Tú Anh (2009), Giáo trình *Khai thác dữ liệu và ứng dụng*, Đại học KHTN Tp Hồ Chí Minh.
12. Nguyễn Thế Đạt (2017), *Nghiên cứu mô hình phân cụm có thứ bậc các đồ thị dữ liệu*, Đại học Công nghệ thông tin và Truyền thông.
13. Hoàng Thị Minh Châu (2010), *Các giải pháp cải tiến của thuật toán FCM và CFCM nhằm tăng tốc độ tính toán*, luận văn thạc sĩ.
14. Hoàng Văn Dũng (2007), *Khai phá dữ liệu web bằng kỹ thuật phân*

cụm, luận văn thạc sĩ khoa học.

15.Hoàng Thị Lan Giao, Trần Tuấn Tài (2011), *Ứng dụng phân cụm dữ liệu trong việc phân tích, đánh giá kết quả học tập của học sinh*.

Tiếng Anh

16.Bezdek, J. C., Ehrlich, R., & Full, W. (1984), *FCM: The fuzzy c-means clustering algorithm*, Computers & Geosciences, 10(2-3), 191-203.

17.Ruspini E.H. (1969), *A new approach to clustering, Information and. Control*.

18.Dunn J.C. (1973), *A fuzzy relative of the ISODATA process and its use in detecting compact Well-Separated clusters*, Journal of Cybernetics.

19.Jiawei Han and Micheline Kamber (2007), *Data Mining Concepts and Techniques*, Chapter 1 & Chapter 8 (Intelligent Database Systems Research Lab School of Computing Science Simon Fraser University, Canada).

20.W. Wang, Y. Zhang (2007), *On fuzzy cluster validity indices, ScienceDirect*, vol. 158, pp. 2095-2117.

21.K.R. Zalik (2010), *Cluster validity index for estimation of fuzzy clusters of different sizes and densities*, Pattern Recognition. 43, pp. 3374-3390.

22.Q. Zhao (2012), *Cluster validity in clustering methods*, Publications of the University of Eastern Finland.

23.D.W. Kim, K.H. Lee, D. Lee (2004), *On cluster validity index for estimation of the optimal number of fuzzy clusters*, Pattern Recognition 37, pp. 2009–2025.

Một số trang web

24.http://www.taichinhdientu.vn/tap-chi-efinance/phan-tich-rui-ro-nguoi-nop-thue-mau-chot-o-con-nguoi-148789.html

25.http://vneconomy.vn/tai-chinh/quan-ly-rui-ro-trong-kiem-tra-thanh-tra-thue-la-gi-20160408110926l2.htm

26.http://gizteam.com/tong-quan-ve-khai-pha-du-lieu/

27. http://home.deib.polimi.it/matteucc/Clustering/tutorial_html/index.html

28. https://bienuit.wordpress.com/2013/09/07/quy-trinh-khai-pha-du-lieu-process-of-data-mining/

29. http://ketoanthue24h.com/bao-cao-tai-chinh-la-gi-khai-niem-y-nghia-cua-bctc/

30. http://www.tuvanluatvietnam.vn/vn/service/thue-gia-tri-gia-tang-30.html