

ĐẠI HỌC QUỐC GIA TP. HCM
TRƯỜNG ĐẠI HỌC CÔNG NGHỆ THÔNG TIN

ĐOÀN HUẤN

**PHÁT TRIỂN MỘT SỐ PHƯƠNG PHÁP NHẬN DIỆN VÀ
PHÂN TÍCH RỦI RO THEO HƯỚNG TIẾP CẬN
GOM CỤM MỜ**

Chuyên ngành Khoa học máy tính
Mã số: 62.48.01.01

TÓM TẮT LUẬN ÁN TIẾN SĨ KHOA HỌC MÁY TÍNH

THÀNH PHỐ HỒ CHÍ MINH - 2019

Công trình được hoàn thành tại **Trường Đại học Công nghệ Thông tin,
Đại học Quốc gia TP.HCM.**

Người hướng dẫn khoa học 1: PGS. TS Nguyễn Đình Thuần
Người hướng dẫn khoa học 2: PGS. TS Hồ Bảo Quốc

Phản biện độc lập 1: PGS. TS Nguyễn Đức Dũng
Phản biện độc lập 2: PGS. TS Lê Hoàng Thái

Luận án đã được bảo vệ trước
Hội đồng chấm luận án họp tại:

Vào lúc giờ ngày tháng năm

Có thể tìm luận án tại:
Thư viện Quốc gia Việt Nam.
Thư viện Trường Đại học Công nghệ Thông tin, ĐHQG-HCM.

CHƯƠNG 1 GIỚI THIỆU

1.1 Động cơ nghiên cứu

Quản lý rủi ro là một một trong những nhiệm vụ quan trọng hàng đầu của các tổ chức, doanh nghiệp. Mục đích của việc quản lý rủi ro là nhằm phát hiện, phân tích để ngăn ngừa, kiểm soát, giảm tối thiểu các thiệt hại do các sự kiện tiêu cực gây ra và tận dụng các cơ hội do sự kiện tích cực mang lại. Quản lý rủi ro đã thu hút được nhiều sự quan tâm nghiên cứu của các nhà khoa học. Mặc dù vậy, những thành tựu đạt được của quản lý rủi ro trong doanh nghiệp nói chung vẫn còn ở mức sơ khởi và nó mở ra một lĩnh vực nghiên cứu mới. Trong thực tế có một nguồn dữ liệu lớn, phong phú sẵn có ghi nhận các hoạt động của các doanh nghiệp trong các cơ sở dữ liệu của các hệ thống phần mềm quản lý như ERP, CRM¹, ... có thể khai thác thành thông tin, tri thức nhất là việc nhận diện và phân tích các rủi ro nằm tiềm ẩn trong đó. Khoa học máy tính, cụ thể là các kỹ thuật khám phá tri thức có khả năng rất mạnh trong việc tìm kiếm thông tin, tri thức nằm tiềm ẩn trong dữ liệu. Nên có thể nói đây là khoảng trống dành cho việc phát triển, cải tiến và sử dụng các kỹ thuật khám phá tri thức xây dựng các công cụ, các phương pháp hỗ trợ cho việc nhận diện và phân tích một số loại rủi ro liên quan đến các hoạt động của doanh nghiệp. Việc phát triển, cải tiến các kỹ thuật của khám phá tri thức và áp dụng vào bài toán quản lý rủi ro mà cụ thể hơn là nhận diện và phân tích rủi ro là một thách thức đặt ra và cần thiết nhằm mục đích trả lời câu hỏi: “Trong cơ sở dữ liệu lớn, phong phú sẵn có của doanh nghiệp lưu trữ dữ liệu về các đối tượng, vậy những đối

¹ ERP: Enterprise Resource Planning, CRM: Customs Relationship Management

tượng này tiềm ẩn loại rủi ro gì và có thể nhận diện và phân tích nó không?”

Quản lý rủi ro là một bài toán có nhiều thách thức vì nó phụ thuộc vào rất nhiều yếu tố bởi tùy theo lĩnh vực mà tổ chức, doanh nghiệp hoạt động, tùy theo qui mô mà độ phức tạp cũng có thể khác nhau. Sự phức tạp và không chắc chắn trong nhiều vấn đề thực tế của bài toán đòi hỏi phải có những phương pháp mới và công cụ mới. Trước nhu cầu thực tế đó, luận án nghiên cứu đề xuất một framework dựa trên các kỹ thuật khám phá tri thức để làm cơ sở xây dựng một số phương pháp nhận diện và phân tích rủi ro nằm tiềm ẩn trong các cơ sở dữ liệu của doanh nghiệp. Các phương pháp nhận diện và phân tích rủi ro này được xây dựng dựa trên việc cải tiến và áp dụng các kỹ thuật khám phá tri thức trong ngành khoa học máy tính.

1.2 Mục tiêu nghiên cứu

Với mục tiêu chung là phát triển và ứng dụng các kỹ thuật khám phá tri thức vào bài toán quản lý các đối tượng rủi ro nằm tiềm ẩn trong dữ liệu. Các mục tiêu nghiên cứu cụ thể của luận án đó là:

- 1) Đề xuất một framework dựa trên các kỹ thuật khám phá tri thức để làm cơ sở định hướng chung cho việc xây dựng các phương pháp nhận diện và phân tích rủi ro tiềm ẩn trong dữ liệu.
- 2) Xây dựng một số phương pháp nhận diện và phân tích các đối tượng rủi ro nằm tiềm ẩn trong dữ liệu bằng các kỹ thuật khám phá tri thức dựa trên framework đã đề xuất.
- 3) Phát triển một số thuật toán nhằm hỗ trợ cho việc xây dựng các phương pháp nhận diện và phân tích các đối tượng rủi ro nằm tiềm ẩn trong dữ liệu.

Việc hiện thực các mục tiêu trên được trình bày trong các chương sau.

1.3 Đóng góp của luận án

Theo kết quả nghiên cứu, các đóng góp chính của luận án được tóm tắt như sau:

- 1) Đã đề xuất một framework gồm hai giai đoạn dựa trên các kỹ thuật khám phá tri thức để làm cơ sở định hướng chung cho việc xây dựng các phương pháp nhận diện và phân tích rủi ro tiềm ẩn trong dữ liệu. Kết quả của đóng góp này được công bố trong [CT2].
- 2) Dựa trên framework đề xuất, xây dựng phương pháp nhận diện và phân tích rủi ro rõ bằng cách phát triển và áp dụng các kỹ thuật khám phá tri thức. Phương pháp này đáp ứng một phần cho mục tiêu 2) của luận án và được công bố trong [CT2], [CT4].
- 3) Xây dựng phương pháp nhận diện và phân tích rủi ro mờ bằng cách tích hợp logic mờ và các kỹ thuật khám phá tri thức đặc biệt là gom cụm mờ. Phương pháp này đáp ứng phần còn lại cho mục tiêu 2) của luận án và được công bố trong [CT3], [CT5].
- 4) Phát triển thuật toán gom cụm mờ cải tiến FCM-E bằng cách xây dựng chỉ số xác định tự động số cụm thích hợp và tích hợp vào thuật toán FCM và công việc này đáp ứng một phần cho mục tiêu 3) của luận án và được công bố trong [CT1], [CT6].
- 5) Phát triển thuật toán gom cụm mờ cải tiến FCM-R bằng cách xây dựng một độ đo xếp hạng mức độ rủi ro của các cụm thu được sau quá trình gom cụm và tích hợp vào thuật toán FCM-E. Công việc này đáp ứng phần còn lại cho mục tiêu 3) của luận án và được công bố trong [CT2], [CT4]. Thuật toán FCM-R là nền tảng quan trọng trong việc xây dựng các phương pháp nhận diện và phân tích rủi ro.

1.4 Bố cục của luận án

Luận án được tổ chức thành 6 chương như sau:

Chương 1 giới thiệu tổng quan về luận án bao gồm động cơ nghiên cứu, mục tiêu nghiên cứu, các đóng góp của luận án, và bố cục của luận án.

Chương 2 giới thiệu về rủi ro, quản lý rủi ro trong doanh nghiệp, bài toán quản lý rủi ro bằng các kỹ thuật của ngành khoa học máy tính.

Chương 3 đề xuất framework và xây dựng một số phương pháp nhận diện và phân tích rủi ro.

Chương 4 trình bày việc phát triển các thuật toán cải tiến FCM-E và FCM-R làm cơ sở cho việc xây dựng các phương pháp nhận diện và phân tích rủi ro.

Chương 5 tiến hành thực nghiệm hai phương pháp nhận diện và phân tích rủi ro rõ và mờ.

Chương 6 là các kết luận và định hướng phát triển của luận án.

CHƯƠNG 2 RỦI RO VÀ BÀI TOÁN QUẢN LÝ RỦI RO

2.1 Quản lý rủi ro trong doanh nghiệp

2.1.1. *Định nghĩa rủi ro*

Một định nghĩa rủi ro khá phổ biến đã được ISO (International Organization for Standardization) đưa ra như sau: "Rủi ro là sự ảnh hưởng của sự không chắc chắn trên các mục tiêu". Định nghĩa này được chọn làm cơ sở cho việc nghiên cứu.

2.1.2. *Phân loại rủi ro trong tổ chức kinh tế*

Có bốn loại rủi ro mà một doanh nghiệp phải đối mặt, đó là: Tai họa, Tài chính, Hoạt động và Chiến lược.

Luận án chọn loại rủi ro tài chính để nghiên cứu và sự lựa chọn này được trình bày trong mục 2.3.1 của chương này.

2.1.3. Định nghĩa quản lý rủi ro và bộ tiêu chuẩn quản lý rủi ro

ISO đưa ra định nghĩa: "quản lý rủi ro là việc phối hợp các hoạt động để chỉ đạo và giám sát một tổ chức đối với rủi ro". Cùng với đó tổ chức này cũng đưa ra bộ tiêu chuẩn quản lý rủi ro ISO 31000. Bộ tiêu chuẩn này cũng được chọn làm định hướng cho việc nghiên cứu.

2.1.4. Các giai đoạn của quá trình quản lý rủi ro

Theo bộ tiêu chuẩn ISO 31000, các giai đoạn của quá trình quản lý rủi ro bao gồm:

1. Thiết lập bối cảnh
2. Xác định rủi ro
3. Phân tích rủi ro
4. Đánh giá rủi ro
5. Xử lý rủi ro

2.1.5. Các kỹ thuật nhận diện rủi ro và phân tích rủi ro trong doanh nghiệp

Trong kinh tế, các kỹ thuật nhận diện rủi ro, phân tích và đánh giá rủi ro thường được sử dụng là: Bảng câu hỏi và kiểm mục, Hội thảo và thảo luận góp ý, Thanh tra và kiểm toán, Các sơ đồ và phân tích sự phụ thuộc, HAZOP, FMEA, SWOT, PESTLE, ...

Có thể nhận xét rằng các kỹ thuật này không dựa nhiều vào dữ liệu lịch sử hoạt động có sẵn. Điểm yếu không dựa nhiều vào dữ liệu lịch sử hoạt động có sẵn của các kỹ thuật này là điểm mạnh của các kỹ thuật khám phá tri thức trong ngành khoa học máy tính.

2.2 Bài toán quản lý rủi ro bằng các kỹ thuật của ngành khoa học máy tính

Các kỹ thuật của ngành khoa học máy tính, nhất là các kỹ thuật khám phá tri thức từ dữ liệu đã được áp dụng phổ biến trong nhiều lĩnh vực của cuộc sống và gần đây cũng xâm nhập vào bài toán quản lý rủi ro.

2.2.1. Phân loại rủi ro trong nghiên cứu quản lý rủi ro của ngành khoa học máy tính

Tsumoto và Hong² nghiên cứu rủi ro trong bối cảnh của ngành khoa học máy tính, phân chia rủi ro theo bốn loại sau: rủi ro không mong muốn, rủi ro lợi ích, rủi ro tiềm ẩn, rủi ro toán học. Trong đó rủi ro tiềm ẩn là để chỉ các đối tượng rủi ro có thể ẩn chứa trong các cơ sở dữ liệu hay nói vắn tắt là rủi ro tiềm ẩn trong dữ liệu. Một loại rủi ro theo cách phân loại này có thể sẽ chứa nhiều loại rủi ro theo cách phân loại trong doanh nghiệp. Ví dụ rủi ro tài chính, rủi ro hoạt động, ... đều có thể ẩn chứa trong cơ sở dữ liệu doanh nghiệp nên chúng thuộc loại rủi ro tiềm ẩn theo cách phân loại của Tsumoto và Hong. Do đó loại rủi ro chọn nghiên cứu là giao của hai cách phân loại trong kinh tế và trong khoa học máy tính. Chẳng hạn với việc chọn rủi ro Tài chính theo phân loại trong kinh tế và rủi ro tiềm ẩn theo phân loại trong khoa học máy tính thì rủi ro cần nghiên cứu là Rủi ro tài chính tiềm ẩn trong dữ liệu.

2.2.2. Tình hình nghiên cứu quản lý rủi ro bằng các kỹ thuật khám phá tri thức của ngành khoa học máy tính

Luận án đã tiến hành khảo sát các công trình ứng dụng kỹ thuật khám phá tri thức vào quá trình quản lý rủi ro. Các nhóm tác giả thường sử dụng các kỹ thuật như gom cụm, phân lớp, logic mờ trong việc nhận diện hoặc phân tích các loại rủi ro trong đó có rủi ro tài

² S. Tsumoto and T.-P. Hong, "Special issue on data mining for decision making and risk management," *J Intell Inf Syst*, vol. 36, p. 249–251, 2011.

chính của khách hàng tiềm ẩn trong cơ sở dữ liệu của các doanh nghiệp hoặc tổ chức tài chính.

2.3 Một số lựa chọn làm cơ sở nghiên cứu của luận án

2.3.1. Chọn loại rủi ro phù hợp để tiến hành nghiên cứu

Luận án chọn loại rủi ro tài chính tiềm ẩn trong dữ liệu để nghiên cứu. Rủi ro tài chính là rủi ro liên quan đến khả năng thanh toán công nợ của khách hàng, các khoản vay, tín dụng, ... và dữ liệu liên quan về chúng thường được các doanh nghiệp ghi nhận bằng các giao dịch trong các cơ sở dữ liệu của các phần mềm quản lý.

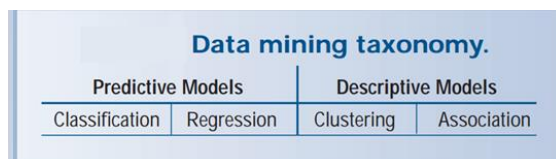
2.3.2. Chọn giai đoạn phù hợp của quá trình quản lý rủi ro để tiến hành nghiên cứu

Hai giai đoạn được chọn để nghiên cứu là Nhận diện rủi ro và Phân tích rủi ro. Giai đoạn Nhận diện rủi ro là quá trình tìm kiếm, đoán nhận ra rủi ro và giai đoạn Phân tích rủi ro là quá trình thấu hiểu bản chất của rủi ro từ đó xác định mức độ của rủi ro. Tìm kiếm thông tin, tri thức trong đó có rủi ro tiềm ẩn từ dữ liệu và phân tích, tính toán mức độ của chúng là thế mạnh của các kỹ thuật khám phá tri thức nên có thể nói hai giai đoạn này là phù hợp để tiến hành nghiên cứu.

2.3.3. Chọn kỹ thuật của ngành khoa học máy tính phù hợp cho bài toán quản lý rủi ro

Khai thác dữ liệu được phân ra hai mô hình³: dự báo và mô tả như trong Hình 2.1, trong đó mô hình dự báo dùng để nhận diện và mô hình mô tả dùng để phân tích hiểu dữ liệu.

³ R. Gerritsen, "Assessing Loan Risks: A Data Mining Case Study," *IT Professional*, vol. 1, no. 6, pp. 16-21, 1999.



Hình 2.1 Hai mô hình chính của khai thác dữ liệu

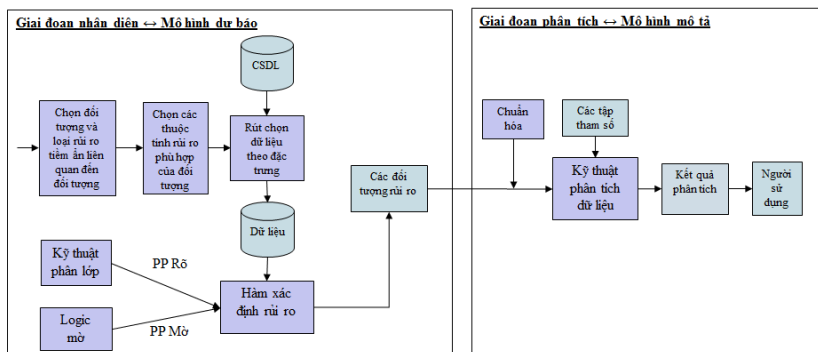
Do đó lựa chọn kỹ thuật của mô hình dự báo như phân lớp để sử dụng cho giai đoạn nhận diện và kỹ thuật của mô hình mô tả như gom cụm để sử dụng cho giai đoạn phân tích hiểu dữ liệu là phù hợp.

CHƯƠNG 3 ĐỀ XUẤT MỘT FRAMEWORK VÀ XÂY DỰNG MỘT SỐ PHƯƠNG PHÁP NHẬN DIỆN VÀ PHÂN TÍCH RỦI RO

3.1 Đề xuất framework dựa trên các kỹ thuật khám phá tri thức để nhận diện và phân tích rủi ro tiềm ẩn trong dữ liệu

Qua quá trình nghiên cứu chức năng của hai mô hình chính của kỹ thuật khai thác dữ liệu là mô hình dự báo và mô hình mô tả, cùng với kinh nghiệm thực tế phát triển một số phương pháp nhận diện và phân tích rủi ro tiềm ẩn trong dữ liệu bằng kỹ thuật khám phá tri thức, luận án đề xuất một framework dựa trên các kỹ thuật khám phá tri thức như trong Hình 3.1.

Mục tiêu của việc đề xuất framework này là nhằm giới thiệu một hướng tiếp cận chung cho việc xây dựng các phương pháp nhận diện và phân tích rủi ro tiềm ẩn trong dữ liệu bằng các kỹ thuật khám phá tri thức của ngành khoa học máy tính. Framework này đã được công bố trong công trình [CT2].



Hình 3.1 Framework nhận diện và phân tích rủi ro tiềm ẩn trong dữ liệu

3.2 Xây dựng phương pháp nhận diện và phân tích rủi ro rõ

Dựa trên định hướng của framework, mục này đề xuất việc xây dựng một phương pháp gồm hai giai đoạn để nhận diện và phân tích rủi ro rõ.

a) Giai đoạn 1

- 1) Chọn đối tượng và loại rủi ro tiềm ẩn liên quan đến đối tượng nghiên cứu.
- 2) Chọn các thuộc tính rủi ro phù hợp của đối tượng
- 3) Tập hợp dữ liệu và tiền xử lý dữ liệu
- 4) Xây dựng hàm xác định rủi ro

Áp dụng một trong số các kỹ thuật phân lớp trên các tập dữ liệu đã chuẩn bị ở bước trên để xây dựng một hàm xác định rủi ro rõ. Sử dụng hàm xác định rủi ro rõ này để xác định rủi ro của các đối tượng bằng cách phân lớp các đối tượng mới vào hai nhóm: rủi ro và không rủi ro.

- 5) Chọn các đối tượng rủi ro làm dữ liệu đầu vào cho giai đoạn tiếp theo của phương pháp.

b) Giai đoạn 2

Trong giai đoạn 2 sẽ tiến hành việc phân tích các đối tượng rủi ro

nhận được từ giai đoạn 1 như sau:

- 1) Chuẩn hóa dữ liệu
- 2) Chọn thuộc tính xây dựng độ đo mức độ rủi ro cho các cụm.
- 3) Phân tích rủi ro bằng cách áp dụng thuật toán cải tiến FCM-R trên tập dữ liệu đã chuẩn hóa.

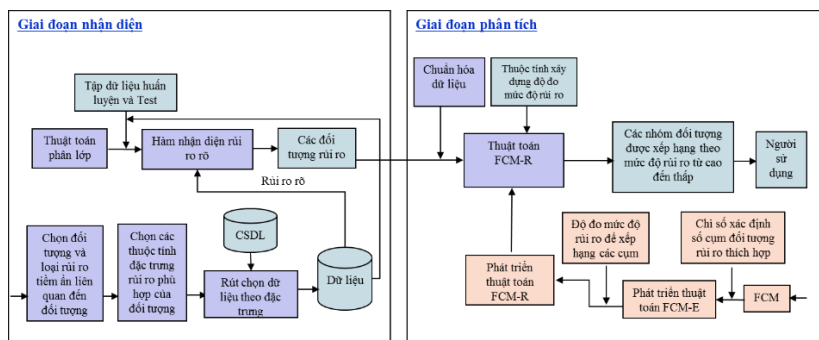
Thuật toán FCM-R sẽ thực hiện các công việc sau:

- Xác định số cụm thích hợp đối với tập dữ liệu.
- Gom các đối tượng vào số cụm thích hợp đã xác định.
- Tính toán độ đo mức độ rủi ro của các cụm nhận được.
- Sắp xếp các cụm theo độ đo mức độ rủi ro của các cụm từ cao đến thấp.

4) Người sử dụng xử lý kết quả nhận được

Dựa trên các kết quả thu được, người sử dụng sẽ đưa ra các chính sách xử lý, điều tiết thích hợp đối với các đối tượng rủi ro tùy theo mức độ rủi ro trong mỗi cụm nhằm quản lý tốt nhất rủi ro.

Các bước xây dựng phương pháp nhận diện và phân tích rủi ro rõ được mô tả trong Hình 3.2, và được công bố trong công trình [CT2], [CT4].



Hình 3.2 Lưu đồ mô tả phương pháp nhận diện và phân tích rủi ro rõ

3.3 Xây dựng phương pháp nhận diện và phân tích rủi ro mờ

Mục này sẽ đề xuất việc xây dựng phương pháp nhận diện và phân tích rủi ro mờ bằng cách tích hợp logic mờ và các kỹ thuật khám phá tri thức.

a) Giai đoạn 1

- 1) Chọn đối tượng và loại rủi ro tiềm ẩn liên quan đến đối tượng nghiên cứu.
- 2) Chọn các thuộc tính rủi ro phù hợp của đối tượng
- 3) Tập hợp dữ liệu và tiền xử lý dữ liệu
- 4) Xây dựng hàm xác định rủi ro mờ

Dựa trên logic mờ để xây dựng hàm xác định rủi ro mờ nên ở bước này cần phải tiến hành các công việc sau:

- Thiết lập các biến ngôn ngữ
- Xác định hàm thành viên cho các tập mờ
- Thiết lập các tập luật

Tiếp theo sử dụng các bước hoạt động của logic mờ: Mờ hóa, Lập luận mờ, Giải mờ để xây dựng hàm xác định rủi ro mờ. Theo lý thuyết logic mờ, hàm xác định rủi ro mờ này có dạng như sau:

$$\mu_R(z) = \sum_{i=1}^n W_i \mu_{Ri}(z) \quad (3.1)$$

Ở đây n là số luật, W_i là trọng số của luật thứ i , $\mu_{Ri}(z)$ là hàm kết luận rủi ro của luật thứ i .

- 5) Chọn các đối tượng rủi ro mờ

Sử dụng hàm xác định rủi ro mờ để xác định mức độ rủi ro của các đối tượng. Các đối tượng có mức độ rủi ro lớn hơn một ngưỡng xác định cho trước nào đó sẽ được chọn làm dữ liệu đầu vào cho giai đoạn tiếp theo.

b) Giai đoạn 2

Trong giai đoạn này sẽ tiến hành việc phân tích các đối tượng khách

hàng có mức độ rủi ro lớn hơn một ngưỡng xác định cho trước, nhận được từ giai đoạn 1 như sau:

- 1) Chuẩn hóa dữ liệu
- 2) Chọn thuộc tính xây dựng độ đo mức độ rủi ro cho các cụm.
- 3) Phân tích rủi ro bằng cách áp dụng thuật toán cải tiến FCM-R trên tập dữ liệu đã chuẩn hóa.

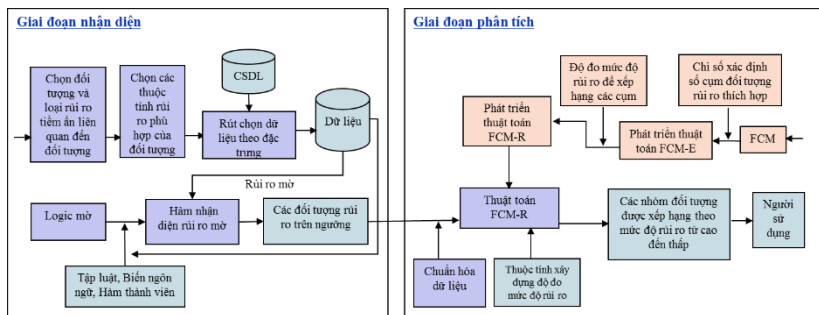
Thuật toán FCM-R sẽ thực hiện các công việc sau:

- Xác định số cụm thích hợp đối với tập dữ liệu.
- Gom các đối tượng vào số cụm thích hợp đã xác định.
- Tính toán độ đo mức độ rủi ro của các cụm nhận được.
- Sắp xếp các cụm theo độ đo mức độ rủi ro của các cụm từ cao đến thấp.

- 4) Người sử dụng xử lý kết quả nhận được

Dựa trên các kết quả thu được, người sử dụng sẽ đưa ra chính sách xử lý các đối tượng rủi ro tùy theo mức độ rủi ro.

Các bước xây dựng phương pháp nhận diện và phân tích rủi ro mờ được mô tả trong Hình 3.3, và được công bố trong công trình [CT3], [CT5].



Hình 3.3 Lưu đồ mô tả phương pháp nhận diện và phân tích rủi ro mờ

CHƯƠNG 4 PHÁT TRIỂN MỘT SỐ THUẬT TOÁN GOM CỤM HỖ TRỢ CHO VIỆC PHÂN TÍCH RỦI RO

4.1 Phát triển thuật toán gom cụm mờ cải tiến FCM-E

4.1.1. Một số khái niệm

Mục này đưa ra một số khái niệm liên quan đến các cụm tạo ra trong thuật toán gom cụm.

- Phần tử biên: một cụm chứa các đối tượng và mỗi đối tượng có các thuộc tính của nó, tùy theo số thuộc tính mà cụm có hình dáng hình học trong không gian đó. Phần tử biên là đối tượng nằm trên biên của cụm theo hình dáng hình học của cụm đó.
- Phần tử cực biên: là đối tượng nằm trên biên của cụm và có ít nhất một thuộc tính có giá trị lớn nhất hoặc nhỏ nhất.
- Khoảng cách cực đại giữa tất cả các phần tử cực biên trong cụm là khoảng cách giữa hai phần tử cực biên có khoảng cách lớn nhất được ký hiệu $\bar{d}_{max}$.
- Khoảng cách trung bình giữa tất cả các phần tử cực biên trong một cụm là tỷ số của tổng tất cả khoảng cách giữa các phần tử cực biên và số lượng khoảng cách, được ký hiệu $\bar{d}_{avr}$.
- Khoảng cách cực tiểu từ trung tâm của cụm này đến tất cả các phần tử cực biên của cụm khác là khoảng cách đến phần tử cực biên gần nhất, ký hiệu $\bar{\phi}_{min}$.
- Khoảng cách trung bình từ trung tâm của cụm này đến tất cả các phần tử cực biên của cụm khác là tỷ số của tổng tất cả các khoảng cách và số lượng khoảng cách, ký hiệu $\bar{\phi}_{avr}$.

4.1.2. Xây dựng các hệ số và phân tích

Thiết lập hệ số trong cụm $\bar{\alpha}$ phản ánh sự biến dạng bên trong cụm như sau:

$$\bar{\alpha} = \frac{\bar{d}_{\max}}{\bar{d}_{\text{avr}}} = \frac{\max_q \left(\sqrt{\sum_{l=1}^k w_l (x_{il} - x_{jl})^2} \right)}{\sum_q \left(\sqrt{\sum_{l=1}^k w_l (x_{il} - x_{jl})^2} \right)} \quad (4.1)$$

q

Ở đây $j = 1, \dots, p$; $i = 1, \dots, p$ với $i \neq j$ và p là số phần tử cực biên trong mỗi cụm. W_1 là véc tơ trọng số, $l = 1, \dots, k$, với k là số chiều (số thuộc tính), q là số khoảng cách giữa các phần tử cực biên trong mỗi cụm. x_{il} là thuộc tính thứ l của phần tử cực biên x_i và x_{jl} là thuộc tính thứ l của phần tử cực biên x_j .

Khi một cụm có ít nhất hai phần tử, nó có $\bar{\alpha} \geq 1$, còn khi cụm đó chỉ có một phần tử, chấp nhận qui ước $\bar{\alpha} = 0$.

Gọi $\bar{\alpha}_{\max}$ là hệ số lớn nhất của các hệ số $\bar{\alpha}$ của tất cả các cụm sau một quá trình gom cụm:

$$\bar{\alpha}_{\max} = \max_c (\bar{\alpha}) \quad (4.2)$$

Ở đây c là số cụm.

Thiết lập hệ số ngoài cụm $\bar{\beta}$ phản ánh tỷ số khoảng cách giữa hai cụm, từ trung tâm cụm j đến các phần tử cực biên trong cụm x như sau:

$$\bar{\beta} = \frac{\bar{\phi}_{\min}}{\bar{\phi}_{\text{avr}}} = \frac{\min_p \left(\sqrt{\sum_{l=1}^k w_l (x_{il} - c_{jl})^2} \right)}{\sum_p \left(\sqrt{\sum_{l=1}^k w_l (x_{il} - c_{jl})^2} \right)} \quad (4.3)$$

p

Ở đây C_j là trung tâm của cụm j , $j = 1, \dots, c$; $i = 1, \dots, p$ với c là số lượng cụm, p là số phần tử cực biên trong cụm x . W_1 là véc tơ trọng số, $l = 1, \dots, k$, với k là số chiều (số thuộc tính). x_{il} là thuộc tính thứ l của phần tử cực biên x_i trong cụm x và c_{jl} là thuộc tính thứ l của trung tâm cụm j .

Rõ ràng từ biểu thức trên ta thấy $\bar{\beta} \leq 1$. Khi chỉ có một cụm thì chấp nhận qui ước $\bar{\beta} = 0$. Gọi $\bar{\beta}_{max}$ là hệ số lớn nhất của các hệ số $\bar{\beta}$ của tất cả các cụm sau một quá trình gom cụm:

$$\bar{\beta}_{max} = \max_{c(c-1)} (\bar{\beta}) \quad (4.4)$$

Ở đây c là số cụm.

Thông qua thực nghiệm, nhận thấy rằng $\bar{\alpha}_{max}$ chỉ định số lượng cụm tốt nhất tại lân cận mà nó đạt một giá trị cực tiểu địa phương sau khi nó đã giảm nhanh. Ngoài ra $\bar{\beta}_{max}$ chỉ định số cụm tốt nhất tại lân cận mà nó tiến gần đến 1 sau khi đã tăng nhanh.

$$\text{Đặt:} \quad \bar{\gamma} = \bar{\alpha}_{max} + (1 - \bar{\beta}_{max}) \quad (4.5)$$

Chỉ số $\bar{\gamma}$ là hàm đồng biến với $\bar{\alpha}_{max}$ và nghịch biến với $\bar{\beta}_{max}$. Do đó có thể dự báo $\bar{\gamma}$ đạt cực tiểu khi $\bar{\alpha}_{max}$ đạt giá trị ổn định, ít biến động và có giá trị càng nhỏ, đồng thời $\bar{\beta}_{max}$ ($0 \leq \bar{\beta}_{max} \leq 1$) nhận một giá trị lân cận nào đó của 1 khi đang trên đường tiến về giá trị 1 tức là $\bar{\gamma}$ ở trạng thái mà ở đó các cụm đạt mức tối ưu, không quá lớn cần phải phân tách và không quá nhỏ để phải ghép lại. Và ở trạng thái đó có thể dự báo $\bar{\gamma}$ sẽ chỉ định số lượng cụm tốt nhất.

Việc tiến hành thực nghiệm để chứng minh cho nhận xét này và đánh giá hiệu quả của nó đã được thực hiện với một số tập dữ liệu.

4.1.3. Phát triển thuật toán cải tiến FCM-E

Tích hợp việc tính toán các hệ số $\bar{\alpha}$, $\bar{\beta}$, $\bar{\alpha}_{max}$, $\bar{\beta}_{max}$ ở trên và chỉ số $\bar{\gamma}$ vào thuật toán FCM và sử dụng chỉ số $\bar{\gamma}$ làm chỉ số xác định tự động số lượng cụm thích hợp trong quá trình gom cụm. Thuật toán FCM mở rộng được gọi là FCM-E được trình bày như sau:

Thuật toán FCM-E

Đầu vào: Tập dữ liệu n đối tượng x_i .

Đầu ra: c_{best} cụm dữ liệu với c_{best} là số cụm thích hợp nhất

- 1) Đưa vào n đối tượng x_i , tham số mờ $m > 1$, số epsilon đủ nhỏ.
- 2) Đưa vào véc tơ trọng số w .
- 3) Đưa vào $c_{\min}$ và $c_{\max}$ ($c_{\min} \geq 1, c_{\min} < c_{\max} < n$).
- 4) Cho c chạy từ $c_{\min}$ đến $c_{\max}$
 - a) Khởi tạo ma trận thành viên $U_{c \times n}$
 - b) Tính trung tâm C_j của cụm j ($j = 1, \dots, c$)
 - c) Cập nhật ma trận khoảng cách D ($c \times n$)
 - d) Cập nhật ma trận thành viên U
 - e) Nếu sự thay đổi của ma trận U là đủ nhỏ ($< \epsilon$) so với bước trước đó thì nhảy đến bước f), ngược lại nhảy đến bước b)
 - f) Dựa trên ma trận U , x_i được sắp xếp vào các cụm theo qui tắc: x_i sẽ thuộc vào cụm nào mà nó có độ thuộc lớn nhất
 - g) Ghi c cụm dữ liệu tại bước c vào đĩa
 - h) Lấy tất cả các phần tử cực biên của tất cả các cụm
 - i) Tính chỉ số $\bar{\gamma}$ để chỉ định số cụm thích hợp dựa trên (4.1), (4.2), (4.3), (4.4), (4.5).
- 5) Chọn c_{best} tại lân cận khi $\bar{\gamma}$ đạt giá trị cực tiểu địa phương
- 6) Lấy c_{best} cụm dữ liệu tại bước c_{best} từ đĩa

4.1.4. Đánh giá độ phức tạp tính toán của thuật toán FCM-E

Độ phức tạp tính toán của thuật toán FCM là $O(tcn)$. Ở đây n là số đối tượng trong tập dữ liệu, c là số lượng cụm, t là số lần lặp. Độ phức tạp tính toán của thuật toán cải tiến FCM-E được tính như sau⁴:

Đặt q ($q=c_{\max}$) là số lần lặp để lựa chọn số lượng cụm thích hợp.

$$O(\text{FCM-E}) = [O(\text{FCM}) + O(\text{Step h}) + O(\bar{\gamma})] * q$$

$$O(\text{FCM-E}) = [O(\text{FCM}) + O(\text{Step h}) + O(\bar{\alpha}) + O(\bar{\beta})] * q$$

⁴ Ở đây dùng ký hiệu * thay cho phép nhân

Ở đây: $O(\text{FCM})$ là độ phức tạp tính toán của thuật toán FCM.

$O(\text{Step } h)$ là độ phức tạp tính toán của bước h .

$O(\bar{\alpha})$ là độ phức tạp tính toán của $\bar{\alpha}$.

$O(\bar{\beta})$ là độ phức tạp tính toán của $\bar{\beta}$.

Về giá trị trung bình, số đối tượng trong mỗi cụm là $\frac{n}{c}$.

Đặt p là số phần tử cực biên trong mỗi cụm. Về trung bình có thể cho rằng $p \approx k$, với k là số chiều. Khi đó:

$$O(\text{Step } h) = c * (n/c) * k = nk = O(nk)$$

$$O(\bar{\alpha}) = c * p * (p-1) = c * k * (k-1) = ck^2 - ck \leq ck^2. \text{ Do đó}$$

$$O(\bar{\alpha}) \approx O(ck^2)$$

$$O(\bar{\beta}) = c * c * p = c * c * k = c^2k = O(kc^2)$$

Trong thực tế số cụm c thường nhỏ hơn n , số lần lặp t, q cũng thường nhỏ hơn n và số chiều k cũng thường nhỏ hơn n . Có thể đặt $t = \max(t, q, k, c)$.

$$\text{Khi đó: } O(\text{FCM-E}) = [O(tcn) + O(nk) + O(ck^2) + O(kc^2)] * q$$

$$= [O(tcnq) + O(nkq) + O(cqk^2) + O(kqc^2)]$$

$$\leq [O(nt^3) + O(nt^2) + O(t^4) + O(t^4)]$$

$$\approx [O(nt^3) + O(nt^2)] \approx O(nt^3)$$

Độ phức tạp tính toán của thuật toán cải tiến FCM-E tương đương $O(nt^3)$ là lớn hơn độ phức tạp tính toán của thuật toán FCM nhưng không nhiều.

4.1.5. Thực nghiệm thuật toán FCM-E

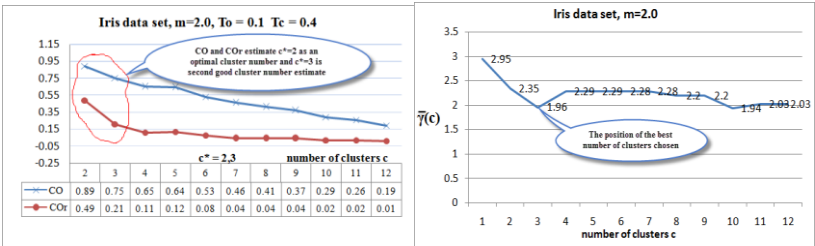
Tiến hành cài đặt, thực nghiệm và đánh giá thuật toán FCM-E trên một số tập dữ liệu: Iris, Wine, Breast Cancer Wisconsin, and Glass (được lấy từ UCI Machine Learning Repository). Chạy thuật toán FCM-E với các tham số $m = 2$, $\text{epsilon} = 0.0001$ trên các tập dữ liệu này và so sánh việc xác định số cụm với một số tác giả khác cho thấy kết quả của phương pháp đề xuất là tốt hơn như sau:

Bảng 4.1 So sánh kết quả của phương pháp đề xuất và H. Capitaine và C. Frélicot

Tập dữ liệu	SOI T (Phương pháp của H. Capitaine và C. Frélicot)				$\bar{\gamma}$ (Phương pháp đề xuất)		Số cụm trong thực tế
	C_{max}	S	A	Ho	C_{max}	Số lượng cụm	
Iris	10	2	2	2	12	3	3
Wine	10	3	3	3	10	3	3

Bảng 4.2 So sánh kết quả của phương pháp đề xuất và Yan M. and Ye K.

Method	Iris (G = 3)	Breast Cancer (G = 2)
Gap/uni	6/8	9
Gap/pc	4	9
Gap/uni	6	2
Gap/pc	4	2
DDGap/uni	2	2
DDGap/pc	2	2
Multilayer/pc	3	3
$\bar{\gamma}$ (Phương pháp đề xuất)	3	3



Chỉ số hợp lệ cụm CO and COr của Zalik

Chỉ số $\bar{\gamma}$ của phương pháp đề xuất

Hình 4.1 So sánh kết quả chỉ số $\bar{\gamma}$ của phương pháp đề xuất với chỉ số hợp lệ cụm của Zalik trên Iris

4.2 Xây dựng độ đo mức độ rủi ro để xếp hạng các cụm và tích hợp vào thuật toán gom cụm mờ cải tiến FCM-R

4.2.1. Xây dựng định nghĩa độ đo xếp hạng cụm

Cho một tập hợp O gồm n đối tượng $o_1, o_2, \dots, o_n$. Khi áp dụng thuật toán gom cụm trên tập hợp O ta thu được k cụm $c_1, c_2, \dots, c_k$.

Định nghĩa 1: Độ đo sức mạnh để xếp hạng cụm hay vắn tắt là độ đo xếp hạng cụm là một hàm μ ánh xạ k cụm thu được sau khi gom cụm vào tập số thực R tức là $\mu(c_i) \rightarrow R$ với $i=1, 2, \dots, k$.

Mục tiếp theo sẽ giới thiệu việc xây dựng một độ đo xếp hạng cụm về mức độ rủi ro.

4.2.2. Xây dựng một độ đo mức độ rủi ro của các cụm và tích hợp vào thuật toán gom cụm

Giả sử kết quả của thuật toán gom cụm là c cụm và G là tập các thuộc tính của đối tượng đã được chuẩn hóa mà tương quan dương với tiêu chuẩn mức độ rủi ro. Đề xuất việc xây dựng độ đo mức độ rủi ro của các cụm như sau:

$$R_i = \frac{\sum_{l=1}^{n_i} \sum_{j \in G} A_j}{n_i}, \quad i = 1..c \quad (4.6)$$

Ở đây R_i là giá trị của độ đo xếp hạng mức độ rủi ro của cụm i , A_j là các thuộc tính thuộc về nhóm thuộc tính G , n_i là số phần tử của cụm i và c là số cụm.

Kết quả của việc xây dựng độ đo mức độ rủi ro để xếp hạng các cụm và tích hợp vào thuật toán gom cụm được công bố trong công trình [CT4].

4.2.3. Tích hợp độ đo mức độ rủi ro để xây dựng thuật toán gom cụm mờ cải tiến FCM-R hỗ trợ phân tích rủi ro

Thuật toán FCM-R

Đầu vào: Tập dữ liệu n đối tượng x_i .

Đầu ra: c_{best} cụm dữ liệu được xếp hạng với c_{best} là số cụm thích hợp nhất

- 1) Đưa vào n đối tượng x_i , tham số mờ $m > 1$, số epsilon đủ nhỏ.
- 2) Đưa vào véc tơ trọng số w .
- 3) Đưa vào c_{min} và c_{max} ($c_{min} \geq 1, c_{min} < c_{max} < n$).
- 4) Đưa vào nhóm thuộc tính rủi ro G
- 5) Cho c chạy từ c_{min} đến c_{max}
 - a) Khởi tạo ma trận thành viên U_{cxn}
 - b) Tính trung tâm C_j của cụm j ($j = 1, \dots, c$)
 - c) Cập nhật ma trận khoảng cách D ($c \times n$)
 - d) Cập nhật ma trận thành viên U
 - e) Nếu sự thay đổi của ma trận U là đủ nhỏ ($< \epsilon$) so với bước trước đó thì nhảy đến bước f), ngược lại nhảy đến bước b)
 - f) Dựa trên ma trận U , x_i được sắp xếp vào các cụm theo qui tắc: x_i sẽ thuộc vào cụm nào mà nó có độ thuộc lớn nhất
 - g) Ghi c cụm dữ liệu tại bước c vào đĩa
 - h) Lấy tất cả các phần tử cực biên của tất cả các cụm
 - i) Tính chỉ số $\bar{\gamma}$ để chỉ định số cụm thích hợp dựa trên (4.1), (4.2), (4.3), (4.4), (4.5) như trong mục 4.1.2.
- 6) Chọn c_{best} tại lân cận khi $\bar{\gamma}$ đạt giá trị cực tiểu địa phương.
- 7) Lấy c_{best} cụm dữ liệu tại bước c_{best} từ đĩa
- 8) Cho i chạy từ 1 đến c_{best}

$$\text{Tính toán } R_i = \frac{\sum_{j=1}^{n_i} \sum_{j \in G} A_j}{n_i}, \quad i = 1..c \quad (4.6)$$

9) Sắp xếp các cụm dựa trên R_i , ($i=1..c_{best}$)

Kết quả của quá trình xây dựng thuật toán cải tiến FCM-R và thực nghiệm nó được công bố trong công trình [CT2], [CT4].

4.2.4. Đánh giá độ phức tạp tính toán của thuật toán FCM-R

Độ phức tạp tính toán của thuật toán FCM là $O(tcn)$. Ở đây n là số đối tượng trong tập dữ liệu, c là số lượng cụm, t là số lần lặp. Độ phức tạp tính toán của thuật toán cải tiến FCM-R được tính như sau:

$$O(\text{FCM-R}) = [O(\text{FCM}) + O(\text{Step h}) + O(\bar{\gamma})] * c_{\max} + O(R_i) * c_{\text{best}}$$

Do $c_{\text{best}} \leq c_{\max}$:

$$O(\text{FCM-R}) = [O(\text{FCM}) + O(\text{Step h}) + O(\bar{\gamma})] * c_{\max} + O(R_i) * c_{\max}$$

$$O(\text{FCM-R}) = [O(\text{FCM}) + O(\text{Step h}) + O(\bar{\gamma}) + O(R_i)] * c_{\max}$$

$$O(\text{FCM-R}) = [O(\text{FCM}) + O(\text{Step h}) + O(\bar{\alpha}) + O(\bar{\beta}) + O(R_i)] * c_{\max}$$

Ở đây: $O(\text{FCM})$ là độ phức tạp tính toán của thuật toán FCM.

$O(\text{Step h})$ là độ phức tạp tính toán của bước h.

$O(\bar{\alpha})$ là độ phức tạp tính toán của $\bar{\alpha}$.

$O(\bar{\beta})$ là độ phức tạp tính toán của $\bar{\beta}$.

$O(R_i)$ là độ phức tạp tính toán của R_i trong (4.6).

Về giá trị trung bình, số đối tượng trong mỗi cụm là $\frac{n}{c}$.

Đặt p là số phần tử cực biên trong mỗi cụm. Về trung bình có thể cho rằng $p \approx k$, với k là số chiều. Khi đó:

$$O(\text{Step h}) = c * (n/c) * k = n * k = O(nk)$$

$$O(\bar{\alpha}) = c * p * (p-1) = c * k * (k-1) = ck^2 - ck \leq ck^2.$$

Như vậy $O(\bar{\alpha}) \approx O(ck^2)$

$$O(\bar{\beta}) = c * c * p = c * c * k = c^2 k = O(c^2 k)$$

$O(R_i) = n_i * |G|$ với n_i là số đối tượng trong cụm i .

Do G là tập các thuộc tính của đối tượng mà tương quan dương với tiêu chuẩn mức độ rủi ro, cho nên $|G| \leq k$. Từ đó: $O(R_i) = \frac{n}{c} * k$.

Trong thực tế số cụm c thường nhỏ hơn n , số lần lặp t , q cũng thường nhỏ hơn n và số chiều k cũng thường nhỏ hơn n . Có thể đặt $t = \max(t, q, k, c)$. Khi đó thay t vào ta có:

$$\begin{aligned}
 O(\text{FCM-R}) &= [O(tcn) + O(nk) + O(ck^2) + O(c^2k) + O(\frac{n}{c} * k)] * c_{\max} \\
 &= [O(tcn) + O(nk) + O(ck^2) + O(c^2k) + O(\frac{n}{c} * k)] * \sqrt{n} \\
 &\approx [O(nt^2) + O(nt) + O(t^3) + O(t^3) + O(\frac{n}{t} * t)] * \sqrt{n} \\
 &\approx [O(nt^2) + O(nt) + O(t^3) + O(t^3) + O(n)] * \sqrt{n} \\
 &\approx O(nt^2) * \sqrt{n} \approx O(nt^2\sqrt{n}).
 \end{aligned}$$

Ở đây, có thể nhận thấy rằng độ phức tạp tính toán của FCM-R là cao hơn độ phức tạp tính toán của FCM nhưng không nhiều.

4.2.5. Thực nghiệm thuật toán cải tiến FCM-R

Thuật toán FCM-R được xây dựng từ FCM-E bằng cách tích hợp thêm độ đo mức độ rủi ro. Việc thực nghiệm thuật toán FCM-E đã tiến hành khá kỹ ở mục 4.1.5 nên việc thực nghiệm FCM-R thực chất chỉ cần thực nghiệm thêm độ đo mức độ rủi ro. Do các phương pháp nhận diện và phân tích rủi ro được xây dựng trong luận án đều sử dụng FCM-R ở bước phân tích rủi ro nên thực nghiệm các phương pháp này cũng đồng thời thực nghiệm FCM-R. Việc tiến hành thực nghiệm các phương pháp nhận diện và phân tích rủi ro và FCM-R được trình bày trong chương 5.

CHƯƠNG 5 THỰC NGHIỆM CÁC PHƯƠNG PHÁP NHẬN DIỆN VÀ PHÂN TÍCH RỦI RO

5.1 Thực nghiệm trên tập dữ liệu thực của doanh nghiệp

5.1.1. *Thực nghiệm phương pháp nhận diện và phân tích rủi ro rõ*

a) *Thực nghiệm*

Thực nghiệm phương pháp nhận diện và phân tích rủi ro rõ trên tập dữ liệu thực trích xuất từ CSDL của nhóm Công ty Hồng Ký tại Thành phố Hồ Chí Minh. Kết quả thực nghiệm thu được bảng sau:

Bảng 5.1 Các cụm khách hàng được xếp hạng theo độ đo mức độ rủi ro về
thanh toán công nợ từ cao đến thấp.

Độ đo mức độ rủi ro	Cụm	Mã khách hàng (O ₁₀)	Tổng nợ (O ₁₁)	Trung bình số ngày nợ quá hạn (O ₁₂)	Tổng nợ quá hạn (O ₁₃)	Tổng nợ xấu (O ₁₄)
42.98 (Rủi ro rất cao)	4	C00375	871,842,518	53	749,311,518	605,225,418
16.23 (Rủi ro cao)	3	C00402	100630198	757	100630198	63505000
		C00428	248053729	638	248053729	248053729
		C00429	291433992	638	291433992	291433992
		C00430	975000	1583	975000	975000
		C00433	92229600	496	92229600	92229600
		C00444	2500000	638	2500000	2500000
9.03 (Rủi ro trung bình)	2	C00440	454610465	7	149540506	0
		C00448	413814591	23	242944391	52635091
		C00443	384591982	17	131148182	49229182
		C00454	1067174224	14	527971524	0
		C00431	395218413	28	207261353	98518953
		C00395	227663183	39	111872183	111872183
		C00420	128112416	72	128112416	106017416

Độ đo mức độ rủi ro	Cụm	Mã khách hàng (O ₁₀)	Tổng nợ (O ₁₁)	Trung bình số ngày nợ quá hạn (O ₁₂)	Tổng nợ quá hạn (O ₁₃)	Tổng nợ xấu (O ₁₄)
						
-0.97 (Rủi ro thấp)	1	C00358	34846000	76	34846000	34630000
		C00361	195812060	11	46373020	0
		C00363	52316500	13	11300500	0
		C00364	33171800	9	20346800	0
		C00366	56389935	1	13009935	0
						

b) Phân tích kết quả thực nghiệm

Trong Bảng 5.1, nhận thấy rằng cụm 4 (độ đo xếp hạng mức độ rủi ro = 42.98) có rủi ro rất cao. Khách hàng trong cụm này có giá trị các thuộc tính (Tổng nợ, Tổng nợ quá hạn, Tổng nợ xấu) cao hơn hẳn các khách hàng trong các cụm khác. Đối với các khách hàng trong cụm 4, doanh nghiệp cần phải có những chính sách xử lý mạnh mẽ tránh những thất thoát có thể xảy ra. Cụm 3 (độ đo xếp hạng mức độ rủi ro = 16.23) cũng là cụm có các khách hàng rủi ro cao, mặc dù mức độ rủi ro thấp hơn cụm 4, nhưng các khách hàng trong cụm này có thuộc tính *Trung bình số ngày nợ quá hạn* cao. Đối với các khách hàng trong cụm này người quản lý cũng cần chú ý đến đặc điểm kéo dài thời gian trả nợ này để đưa ra những quyết định thích hợp như chỉ bán hàng bằng tiền mặt hoặc hợp đồng cung cấp hàng với điều kiện thời hạn trả nợ ngắn. Tóm lại dựa trên các kết quả thu được, người sử dụng sẽ có chính sách xử lý và điều tiết phù hợp đối với các khách hàng trong mỗi cụm nhằm quản lý tốt nhất rủi ro.

5.1.2. Thực nghiệm phương pháp nhận diện và phân tích rủi ro mờ

a) Thực nghiệm

Thực nghiệm phương pháp nhận diện và phân tích rủi ro mờ trên cùng tập dữ liệu cho kết quả:

Bảng 5.2 Các cụm được xếp hạng theo độ đo mức độ rủi ro từ cao đến thấp của phương pháp mờ

Độ đo mức độ rủi ro	Cụm	Mã khách hàng (O_{i0})	Tổng nợ (O_{i1})	Trung bình số ngày nợ quá hạn (O_{i2})	Tổng nợ quá hạn (O_{i3})	Tổng nợ xấu (O_{i4})
92.63 (Rủi ro rất cao)	5	C00221	2682995725	28	1862595125	799114485
42.57 (Rủi ro cao)	4	C00375	871842518	53	749311518	605225418
		C00082	688600005	36	688600005	688600005
20.26 (Rủi ro vừa)	3	C00228	58080177	796	58080177	58080177
		C00198	60200000	1583	60200000	60200000
		C00402	100630198	757	100630198	63505000
		C00138	175164000	1583	175164000	175164000
16.11 (Rủi ro thấp)	2	C00091	442697944	25	373096944	142171144
		C00250	344272500	65	230158500	149935500
		C00356	367735748	45	270853748	190078748
		C00235	303825573	36	296400573	235875573
		C00151	209267850	87	209267850	209267850
		...				
4.63 (Rủi ro rất thấp)	1	C00440	454610465	7	149540506	0
		C00167	90978440	85	84318440	56418440
		C00246	110505675	51	110505675	85760075
		C00283	255893629	27	108098629	78207829
		C00443	384591982	17	131148182	49229182
		...				

b) Phân tích kết quả thực nghiệm

Trong Bảng 5.2 nhận thấy cụm 5 (độ đo xếp hạng mức độ rủi ro =92.63) có rủi ro rất cao. Đối tượng khách hàng trong cụm này có giá trị các thuộc tính *Tổng nợ*, *Tổng nợ quá hạn*, *Tổng nợ xấu* lớn hơn hẳn so với các đối tượng trong các cụm khác. Đối với khách hàng trong cụm 5, doanh nghiệp cần phải có những chính sách xử lý mạnh mẽ tránh mất mát có thể xảy ra. Cụm 4 (độ đo xếp hạng mức độ rủi ro= 42.57) cũng là cụm có các đối tượng khách hàng rủi ro cao mặc dù độ đo mức độ rủi ro chỉ gần bằng nửa của cụm 5. Trong cụm này giá trị các thuộc tính *Tổng nợ*, *Tổng nợ quá hạn* cũng chưa bằng một phần ba của của cụm 5 nhưng giá trị thuộc tính *Tổng nợ xấu* lại gần bằng Cụm 5. Do đó đối với các khách hàng của Cụm 4 này, cần phải quan tâm đến việc xử lý vấn đề nợ xấu.

5.2 Thực nghiệm trên tập dữ liệu UCI Machine Learning Repository

a) Thực nghiệm

Thực nghiệm phương pháp nhận diện và phân tích rủi ro rõ trên tập dữ liệu đã được công bố được lấy từ UCI Machine Learning Repository⁵. Tập dữ liệu được chọn thực nghiệm để dự đoán việc thanh toán thẻ tín dụng không đúng hạn của khách hàng phát hành năm 2016 nằm trong lĩnh vực kinh doanh. Kết quả thu được như sau:

⁵ <https://archive.ics.uci.edu/ml/datasets/default+of+credit+card+clients>

Bảng 5.3 Các cụm khách hàng từ tập dữ liệu trong UCI được xếp hạng theo độ đo mức độ rủi ro từ cao đến thấp

Độ đo mức độ rủi ro	C ụm	ID	X1	X12	X13	X14	X15	X16	X17
38.53 (Rủi ro rất rất cao)	8	C20993	460000	282944	288729	295114	299739	304319	299400
		C21050	490000	427020	370428	337432	286247	256618	188154
		C22626	380000	315912	322199	331518	337703	343612	351089
		C22657	350000	309250	314607	322032	327193	333278	340791
		...							
22.96 (Rủi ro rất cao)	7	C23240	330000	238795	243621	259517	253863	257535	262753
		C23603	300000	279065	273067	278910	283920	253113	246560
		C23733	280000	284128	284204	284041	283396	213294	213384
		C24449	320000	283743	289508	302561	283356	275361	268871
		...							
13.22 (Rủi ro cao)	6	C20012	170000	170348	166348	162319	165592	167482	171106
		C20425	200000	177998	174350	158632	154522	135375	142526
		C20465	150000	164576	160500	170405	176048	161874	144124
		C20687	200000	208012	199367	202668	111486	118951	89077
		...							
6.56 (Rủi ro trung bình cao)	5	C22537	130000	105500	107761	109965	110797	107529	114104
		C22691	140000	119715	122283	118913	120435	139795	142950
		C22331	150000	107603	109835	110255	110247	110955	105145
		C22417	130000	108879	111115	107185	108569	112541	114885
		...							
1.18 (Rủi ro trung bình)	4	C20336	90000	86228	85437	90635	91745	87466	85662
		C20131	80000	68612	69971	71846	72372	73475	74842
		C20246	80000	79822	77747	77733	78017	29599	26927
		C20256	80000	76414	78009	78616	49662	48867	48278
		...							
-2.82 (Rủi ro)	3	C22169	50000	45004	48119	42958	37812	15617	9625
		C22112	60000	58501	54932	85672	41483	19394	18893
		C22125	80000	45088	46040	46998	47929	46860	51931

Độ đo mức độ rủi ro	Cụm	ID	X1	X12	X13	X14	X15	X16	X17
trung bình thấp)		C21801	50000	40959	43870	42891	43582	44561	45592
		...							
-4.18 (Rủi ro thấp)	2	C20255	350000	2400	2400	2400	2400	2400	2400
		C20578	310000	400	400	400	400	20334	19050
		C20811	240000	2500	0	0	0	0	0
		C20919	200000	1810	3388	-12	1081	1884	2089
		...							
-6.35 (Rủi ro rất thấp)	1	C23920	20000	15662	16004	17046	16676	16704	16984
		C23935	20000	18292	17710	17137	19830	19240	18419
		C23937	50000	31708	32439	28302	13881	14376	15261
		C24004	20000	21673	21051	20440	19709	20113	19840
		...							

b) Phân tích kết quả thực nghiệm

Trong Bảng 5.3 nhận thấy rằng cụm 8 (độ đo xếp hạng mức độ rủi ro = 38.53) có rủi ro cao nhất. Khách hàng trong cụm này hầu hết có hạn mức tín dụng trong khoảng 400,000 đến 500,000 và tổng số tiền trên hóa đơn của các tháng hầu hết nằm trong khoảng 300,000 đến 400,000. Khách hàng trong cụm 8 vừa có hạn mức tín dụng thuộc loại lớn nhất vừa có giá trị tiền trên các hóa đơn mua hàng cũng thuộc loại lớn nhất nên nếu thanh toán tiền trễ hạn sẽ gây ra rủi ro lớn nhất. Đối với nhóm này cần có chính sách đối phó sớm như ngăn chặn việc ra tiếp hóa đơn dù cho hạn mức tín dụng vẫn còn và đôn đốc khách thanh toán đúng hạn nhằm hạn chế mất mát có thể xảy ra. Phân tích tương tự đối với các cụm còn lại để đưa các quyết định phù hợp.

CHƯƠNG 6 KẾT LUẬN

6.1 Kết luận

Với mục tiêu chung là phát triển và ứng dụng các kỹ thuật khám phá tri thức để xây dựng các phương pháp nhận diện và phân tích rủi ro tiềm ẩn trong dữ liệu. Luận án đã hoàn thành các mục tiêu đặt ra ban đầu và các đóng góp chính luận án đã đạt được bao gồm:

- 1) Đề xuất một framework để làm cơ sở định hướng chung cho việc xây dựng các phương pháp nhận diện và phân tích rủi ro. Công việc này được công bố trong [CT2].
- 2) Xây dựng hai phương pháp nhận diện và phân tích rủi ro tiềm ẩn trong dữ liệu dựa trên framework đề xuất bằng cách tích hợp một số kỹ thuật khám phá tri thức của ngành khoa học máy tính:
 - Phương pháp thứ nhất là phương pháp nhận diện và phân tích rủi ro rõ bằng cách tích hợp một thuật toán phân lớp và thuật toán gom cụm mờ cải tiến FCM-R. Việc xây dựng phương pháp rõ được công bố trong [CT2], [CT4].
 - Phương pháp thứ hai là phương pháp nhận diện và phân tích rủi ro mờ bằng cách tích hợp logic mờ và thuật toán gom cụm mờ cải tiến FCM-R.. Việc xây dựng phương pháp mờ được công bố trong [CT3], [CT5].
- 3) Ngoài ra, luận án cũng đã phát triển một số thuật toán nhằm hỗ trợ cho việc xây dựng một số phương pháp nhận diện và phân tích rủi ro tiềm ẩn trong dữ liệu ở trên:
 - Phát triển thuật toán gom cụm mờ FCM-E với mục tiêu cải tiến là xây dựng thuật toán mới có thể xác định tự động được số cụm thích hợp của tập dữ liệu đầu vào trong quá trình gom cụm. Việc phát triển thuật toán này được công bố trong [CT1], [CT6].

- Phát triển thuật toán gom cụm mờ FCM-R bằng cách xây dựng một độ đo mức độ rủi ro nhằm mục đích xếp hạng các cụm thu được sau quá trình gom cụm và tích hợp vào FCM-E. Sử dụng FCM-R làm công cụ cho giai đoạn phân tích rủi ro của một số phương pháp nhận diện và phân tích rủi ro ở chương 3. Việc xây dựng độ đo mức độ rủi ro để xếp hạng cụm và thuật toán FCM-R được công bố trong [CT2], [CT4].

Thực nghiệm các phương pháp nhận diện và phân tích rủi ro đề xuất trên hai tập dữ liệu: một tập dữ liệu thực của doanh nghiệp và một tập dữ liệu đã được công bố trên UCI Machine Learning Repository. Kết quả thực nghiệm thu được cho thấy sự hiệu quả cũng như có ý nghĩa về mặt ứng dụng trong thực tế của các phương pháp đề xuất.

6.2 Hướng phát triển

Bên cạnh những kết quả luận án đã đạt được vẫn còn những vấn đề cần được nghiên cứu mở rộng như sau:

- (1) Nghiên cứu triển khai việc ứng dụng các phương pháp nhận diện và phân tích rủi ro ở trên cho các đối tượng khác và ở các lĩnh vực khác.
- (2) Phát triển thêm các phương pháp nhận diện và phân tích rủi ro dựa trên việc khai thác các nguồn dữ liệu phi cấu trúc như văn bản, âm thanh, hình ảnh...
- (3) Cần nghiên cứu thêm việc xây dựng chỉ số xác định số cụm và độ đo xếp hạng cụm cho các đối tượng có các thuộc tính khác giá trị số.
- (4) Nghiên cứu cải tiến các thuật toán gom cụm khác thuật toán FCM để bổ sung chỉ số xác định số lượng cụm thích hợp và độ đo xếp hạng cụm.

DANH MỤC CÁC CÔNG TRÌNH ĐÃ CÔNG BỐ

Tạp chí chuyên ngành quốc tế

- [CT1] **H. Doan** and T. D. Nguyen, "A Method for Finding the Appropriate Number of Clusters," The International Arab Journal of Information Technology (IAJIT), Vol. 15, No. 4, (2018) pp. 675-682 (**ISI**, IF: 0.724, SCImago: **Q2**).
- [CT2] **H. Doan**, T. D. Nguyen and Q. B. Ho, "A Framework for Identifying and Analyzing the Latent Risk Based on the Knowledge Discovery Techniques", International Journal of Applied Engineering Research (IJAER), Volume 13, Number 2, (2018) pp. 1336-1343. (**Scopus**: 2010-2017).
- [CT3] **H. Doan** and T. D. Nguyen, " A Method for Identifying and Analyzing the Risk based on the Fuzzy Logic and the Algorithm FCM-R,“, Advances in Information Sciences and Service Sciences (**AISS**): An International Journal of Research and Innovation, Korea. Vol 9, No 2, pp. 19-28. 2017.

Hội nghị chuyên ngành quốc tế

- [CT4] **H. Doan**, T. D. Nguyen and Q. B. Ho, "Building a Measure to Integrate Into a Hybrid Data Mining Method to Analyze the Risk of Customer," in Advanced Computer and Communication Engineering Technology, vol. 362 of the series Lecture Notes in Electrical Engineering, **Springer** International Publishing Switzerland, 2015, pp. 843-851.

- [CT5] **H. Doan** and T. D. Nguyen, "Building a Hybrid Method for Analyzing the Risk by Integrating the Fuzzy Logic and the Improved Fuzzy Clustering Algorithm FCM-R," in *Information Science and Applications (ICISA) 2016*, vol. 376 of the series *Lecture Notes in Electrical Engineering*, **Springer** Science+Business Media Singapore, 2016, pp. 923-932.
- [CT6] **H. Doan** and T. D. Nguyen, "An Adaptive Method to Determine the Number of Clusters in Clustering Process," in *The International Conference on Computer & Information Sciences, ICCOINS 2014*, Kuala Lumpur, **IEEE**, ISBN 978-1-4799-4391-3., 2014. pp. 142-147.
- [CT7] T. Nguyen and **H. Doan**, "An Approach to determine the Number of Clusters for Clustering Algorithms," in *Computational Collective Intelligence. Technologies and Applications-ICCCI 2012*, Vols. LNAI 7653 of the series *Lecture Notes in Computer Science, Part I*, **Springer-Verlag** Berlin Heidelberg, 2012, p. 485–494.

Hội thảo chuyên ngành trong nước

- [CT8] **Đoàn Huân** và Nguyễn Đình Thuận, "Một hướng tiếp cận của bài toán phân tích rủi ro," trong *Hội thảo Quốc gia lần thứ XVIII "Một số vấn đề chọn lọc của Công nghệ thông tin và Truyền thông*, HCM, 2015. Nhà xuất bản Khoa học và Kỹ thuật, 2015, p. 156-161.