

ĐẠI HỌC THÁI NGUYÊN
KHOA CÔNG NGHỆ THÔNG TIN

AN HỒNG SƠN

**NGHIÊN CỨU MỘT SỐ PHƯƠNG PHÁP
PHÂN CỤM MỜ VÀ ỨNG DỤNG**

CHUYÊN NGÀNH: KHOA HỌC MÁY TÍNH
MÃ SỐ: 60 48 01

LUẬN VĂN THẠC SĨ KHOA HỌC

HƯỚNG DẪN KHOA HỌC: PGS.TS NGÔ QUỐC TẠO

THÁI NGUYÊN - 2008

MỤC LỤC

DANH MỤC CÁC TỪ VIẾT TẮT	4
DANH MỤC CÁC HÌNH MINH HOẠ	5
Chương 1 - TỔNG QUAN VỀ KHÁM PHÁ TRI THỨC VÀ KPDL	6
1.1. Giới thiệu chung về khám phá tri thức và khai phá dữ liệu	6
1.2. Quá trình khám phá tri thức	7
1.3. Quá trình khai phá dữ liệu	8
1.4. Các phương pháp khai phá dữ liệu	9
1.5. Các lĩnh vực ứng dụng thực tiễn của KPDL	10
1.6. Các hướng tiếp cận cơ bản và kỹ thuật áp dụng trong KPDL	11
1.7. Các thách thức - khó khăn trong KPTT và KPDL.....	12
1.8. Kết luận	12
Chương 2 - PHÂN CỤM DỮ LIỆU VÀ CÁC THUẬT TOÁN TRONG PCDL .	13
2.1. Khái niệm và mục tiêu của phân cụm dữ liệu	13
2.2. Các ứng dụng của phân cụm dữ liệu	15
2.3. Các yêu cầu của phân cụm	16
2.4. Những kỹ thuật tiếp cận trong phân cụm dữ liệu	18
2.4.1. Phương pháp phân cụm phân hoạch	19
2.4.2. Phương pháp phân cụm phân cấp	19
2.4.3. Phương pháp phân cụm dựa trên mật độ	20
2.4.4. Phương pháp phân cụm dựa trên lưới	21
2.4.5. Phương pháp phân cụm dựa trên mô hình	22
2.4.6. Phương pháp phân cụm có dữ liệu ràng buộc	22
2.5. Một số thuật toán cơ bản trong phân cụm dữ liệu	24
2.5.1. Các thuật toán phân cụm phân hoạch	24
2.5.2. Các thuật toán phân cụm phân cấp	26
2.5.3. Các thuật toán phân cụm dựa trên mật độ	29
2.5.4. Các thuật toán phân cụm dựa trên lưới	32

2.5.5.	Các thuật toán phân cụm dựa trên mô hình	35
2.5.6.	Các thuật toán phân cụm có dữ liệu ràng buộc	36
Chương 3 -	KỸ THUẬT PHÂN CỤM DỮ LIỆU MỜ	37
3.1.	Tổng quan về phân cụm mờ	37
3.2.	Các thuật toán trong phân cụm mờ	38
3.2.1.	Thuật toán FCM(Fuzzy C-means)	39
3.2.1.1.	Hàm mục tiêu	39
3.2.1.2.	Thuật toán FCM	42
3.2.2.	Thuật toán ϵ FCM(ϵ - Insensitive Fuzzy C-means)	46
3.2.2.1.	Hàm mục tiêu	46
3.2.2.2.	Thuật toán ϵ FCM	48
3.2.3.	Thuật toán FCM Cải tiến	49
3.2.3.1.	Thuật toán 1: Thuật toán lựa chọn các điểm dữ liệu làm ứng viên cho việc chọn các trung tâm của các cụm	49
3.2.3.2.	Thuật toán 2: Thuật toán lược bớt các ứng viên	51
3.2.3.3.	Thuật toán 3: Thuật toán chọn các ứng viên làm cực tiểu hàm mục tiêu	51
3.2.3.4.	Thuật toán 4: Gán các trung tâm có liên kết “gần gũi” vào một cụm	52
3.2.3.5.	Tổng kết thuật toán FCM-Cải tiến	56
Chương 4 -	MÔ HÌNH MẠNG NƠON ĐA KHỚP DÙNG CHO PCM	58
4.1.	Tổng quan về mạng Nơon	58
4.2.	Cấu trúc mạng Nơon	61
4.2.1.	Hàm kích hoạt	61
4.2.2.	Liên kết mạng	61
4.2.3.	Bài toán huấn luyện mạng	61
4.3.	Mạng HOPFIELD	62
4.3.1.	Huấn luyện mạng	62
4.3.2.	Sử dụng mạng	63

4.4.	Mạng Noron đa khớp dùng cho phân cụm	63
4.4.1.	Xây dựng lớp mạng Layer1 cho tối ưu các trung tâm cụm	65
4.4.2.	Xây dựng lớp mạng Layer2 cho tối ưu các độ thuộc	68
4.5.	Sự hội tụ của FBACN	72
4.5.1.	Chứng minh sự hội tụ của FBACN	72
4.5.2.	Sự hội tụ FBACN liên tục của Layer1	74
4.6.	Giải thuật của FBACN và FBACN với việc học	75
Chương 5 - CÀI ĐẶT THỬ NGHIỆM VÀ ỨNG DỤNG		79
5.1.	Cài đặt thử nghiệm thuật toán FCM	79
5.2.	Ứng dụng thuật toán FCM-Cải tiến vào nhận dạng ảnh	82
KẾT LUẬN		86
TÀI LIỆU THAM KHẢO		87

DANH MỤC CÁC TỪ VIẾT TẮT

CNTT	Công nghệ thông tin
CSDL	Cơ sở dữ liệu
CEF	Computational Energy Function
DL	Dữ liệu
FBACN	Fuzzy Bi-directional Associative Clustering Network (Mạng Noron đa khớp phục vụ cho phân cụm mờ)
FCM	Fuzzy C-Means
HMT	Hàm mục tiêu
KPDL	Khai phá dữ liệu
KPTT	Khám phá tri thức
LKM	Liên kết mạng
MH	Mô hình
NDA	Nhận dạng ảnh
NN	Neural Network
PCM	Phân cụm mờ
PCDL	Phân cụm dữ liệu
TLTK	Tài liệu tham khảo
TT	Thuật toán
XLA	Xử lý ảnh

DANH MỤC CÁC HÌNH MINH HOẠ

Hình 1.1	Quá trình Khám phá tri thức	7
Hình 1.2	Quá trình Khai phá dữ liệu	9
Hình 2.1	Mô tả tập dữ liệu vay nợ được phân thành 3 cụm	14
Hình 2.2	Các chiến lược phân cụm phân cấp	20
Hình 2.3	Cấu trúc phân cấp	21
Hình 2.4	Các cách mà các cụm có thể đưa ra	23
Hình 2.5	Các thiết lập để xác định ranh giới các cụm ban đầu	24
Hình 2.6	Tính toán trọng tâm của các cụm mới	25
Hình 2.7	Khái quát thuật toán CURE	27
Hình 2.8	Các cụm dữ liệu được khám phá bởi CURE	27
Hình 2.9	Hình dạng các cụm được khám phá bởi TT DBSCAN	30
Hình 3.1	Mô phỏng về tập dữ liệu đơn chiều	44
Hình 3.2	Hàm thuộc với trọng tâm của cụm A trong k-means	44
Hình 3.3	Hàm thuộc với trọng tâm của cụm A trong FCM	45
Hình 3.4	Các cụm khám phá được bởi thuật toán FCM	46
Hình 4.1	Mô hình mạng Noron	60
Hình 4.2	Mô hình học có giám sát	62
Hình 4.3	Mô hình FBACN	64
Hình 4.4	Mô hình Lớp Layer1 của FBACN	65
Hình 4.5	Mô hình Lớp Layer2 của FBACN	69
Hình 5.1	Giao diện của thuật toán FCM khi khởi động	80
Hình 5.2	Giao diện của thuật toán FCM khi làm việc	81
Hình 5.3	Giao diện của chương trình khi khởi động	83
Hình 5.4	Giao diện của chương trình khi chọn ảnh để phân cụm	84
Hình 5.5	Giao diện của chương trình khi thực hiện phân cụm	85

CHƯƠNG 1

TỔNG QUAN VỀ KHÁM PHÁ TRI THỨC VÀ KHAI PHÁ DỮ LIỆU

1.1.	Giới thiệu chung về khám phá tri thức và khai phá dữ liệu	6
1.2.	Quá trình khám phá tri thức	7
1.3.	Quá trình khai phá dữ liệu	8
1.4.	Các phương pháp khai phá dữ liệu	9
1.5.	Các lĩnh vực ứng dụng thực tiễn của KPDL	10
1.6.	Các hướng tiếp cận cơ bản và kỹ thuật áp dụng trong KPDL	11
1.7.	Các thách thức - khó khăn trong KPTT và KPDL	12
1.8.	Kết luận	12

1.1. Giới thiệu chung về khám phá tri thức và khai phá dữ liệu

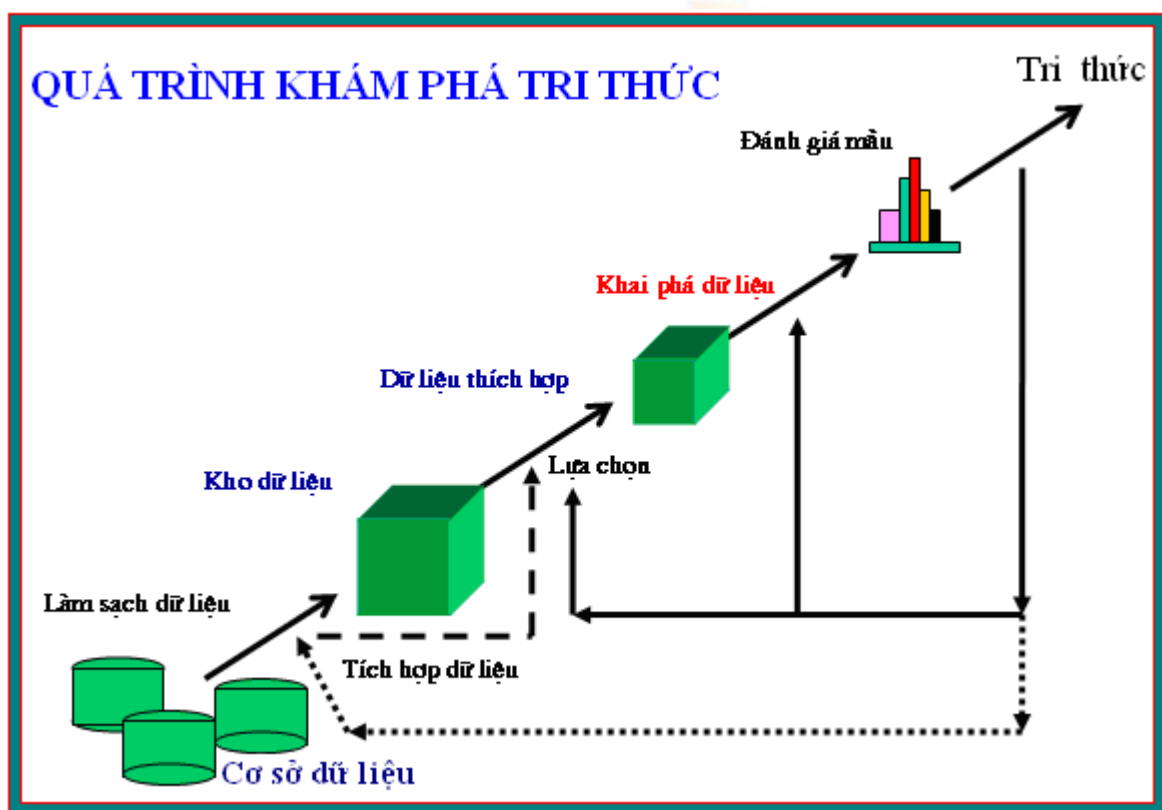
Nếu cho rằng, điện tử và truyền thông chính là bản chất của khoa học điện tử, thì dữ liệu, thông tin, và tri thức hiện đang là tiêu điểm của một lĩnh vực mới để nghiên cứu và ứng dụng, đó là khám phá tri thức và khai phá dữ liệu.

Thông thường, chúng ta coi *dữ liệu* như là một chuỗi các bits, hoặc các số và các ký hiệu hay là các “đối tượng” với một ý nghĩa nào đó khi được gửi cho một chương trình dưới một dạng nhất định. Các bits thường được sử dụng để đo *thông tin*, và xem nó như là dữ liệu đã được loại bỏ phần tử thừa, lặp lại, và rút gọn tới mức tối thiểu để đặc trưng một cách cơ bản cho dữ liệu. *Tri thức* được xem như là các thông tin tích hợp, bao gồm các sự kiện và mối quan hệ giữa chúng, đã được nhận thức, khám phá, hoặc nghiên cứu. Nói cách khác, tri thức có thể được coi là dữ liệu ở mức độ cao của sự trừu tượng và tổng quát.

Khám phá tri thức hay phát hiện tri thức trong CSDL là một quy trình nhận biết các mẫu hoặc các mô hình trong dữ liệu với các tính năng: Phân tích, tổng hợp, hợp thức, khả ích và có thể hiểu được.

Khai phá dữ liệu là một bước trong quá trình khám phá tri thức, gồm các thuật toán khai thác dữ liệu chuyên dùng dưới một số qui định về hiệu quả tính toán chấp nhận được để tìm ra các mẫu hoặc các mô hình trong dữ liệu. Nói cách khác, mục tiêu của Khai phá dữ liệu là tìm kiếm các mẫu hoặc mô hình tồn tại trong CSDL nhưng ẩn trong khối lượng lớn dữ liệu.

1.2. Quá trình khám phá tri thức



Hình 1.1: Quá trình KPTT

Bao gồm các bước sau:

Làm sạch dữ liệu (Data Cleaning): Loại bỏ dữ liệu nhiễu và dữ liệu không nhất quán.

Tích hợp dữ liệu (Data Intergration): Dữ liệu của nhiều nguồn có thể được tổ hợp lại.

Lựa chọn dữ liệu (Data Selection): Lựa chọn những dữ liệu phù hợp với nhiệm vụ phân tích trích rút từ cơ sở dữ liệu.

Chuyển đổi dữ liệu (Data Transformation): Dữ liệu được chuyển đổi hay được hợp nhất về dạng thích hợp cho việc khai phá.

Khai phá dữ liệu (Data Mining): Đây là một tiến trình cốt yếu trong đó các phương pháp thông minh được áp dụng nhằm trích rút ra mẫu dữ liệu.

Đánh giá mẫu (Pattern Evaluation): Dựa trên một độ đo nào đó xác định lợi ích thực sự, độ quan trọng của các mẫu biểu diễn tri thức.

Biểu diễn tri thức (Knowledge Presentation): Ở giai đoạn này các kỹ thuật biểu diễn và hiển thị được sử dụng để đưa tri thức lấy ra cho người dùng.

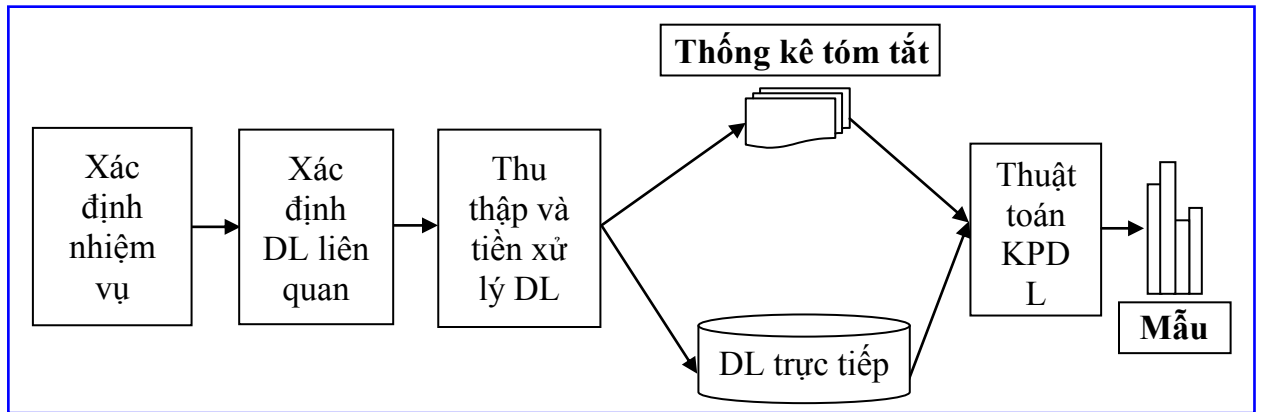
1.3. Quá trình khai phá dữ liệu

KPDL là một giai đoạn quan trọng trong quá trình KPTT. Về bản chất, nó là giai đoạn duy nhất tìm ra được thông tin mới, thông tin tiềm ẩn có trong CSDL chủ yếu phục vụ cho mô tả và dự đoán.

Mô tả dữ liệu là tổng kết hoặc diễn tả những đặc điểm chung của những thuộc tính dữ liệu trong kho dữ liệu mà con người có thể hiểu được.

Dự đoán là dựa trên những dữ liệu hiện thời để dự đoán những quy luật được phát hiện từ các mối liên hệ giữa các thuộc tính của dữ liệu trên cơ sở đó chiết xuất ra các mẫu, dự đoán được những giá trị chưa biết hoặc những giá trị tương lai của các biến quan tâm.

Quá trình KPDL bao gồm các bước chính được thể hiện như Hình 1.2 sau:



Hình 1.2: Quá trình KPD L

- *Xác định nhiệm vụ:* Xác định chính xác các vấn đề cần giải quyết.
- *Xác định các dữ liệu liên quan:* Dùng để xây dựng giải pháp.
- *Thu thập và tiền xử lý dữ liệu:* Thu thập các dữ liệu liên quan và tiền xử lý chúng sao cho thuật toán KPD L có thể hiểu được. Đây là một quá trình rất khó khăn, có thể gặp phải rất nhiều các vướng mắc như: dữ liệu phải được sao ra nhiều bản (nếu được chiết xuất vào các tệp), quản lý tập các dữ liệu, phải lặp đi lặp lại nhiều lần toàn bộ quá trình (nếu mô hình dữ liệu thay đổi), v.v..
- *Thuật toán khai phá dữ liệu:* Lựa chọn thuật toán KPD L và thực hiện việc PKDL để tìm được các mẫu có ý nghĩa, các mẫu này được biểu diễn dưới dạng luật kết hợp, cây quyết định... tương ứng với ý nghĩa của nó.

1.4. Các phương pháp khai phá dữ liệu

Với hai mục đích khai phá dữ liệu là Mô tả và Dự đoán, người ta thường sử dụng các phương pháp sau cho khai phá dữ liệu:

- Luật kết hợp (*association rules*)
- Phân lớp (*Classification*)
- Hồi qui (*Regression*)
- Trực quan hóa (*Visualiztion*)

- Phân cụm (*Clustering*)
- Tổng hợp (*Summarization*)
- Mô hình ràng buộc (*Dependency modeling*)
- Biểu diễn mô hình (*Model Evaluation*)
- Phân tích sự phát triển và độ lệch (*Evolution and deviation analyst*)
- Phương pháp tìm kiếm (*Search Method*)

Có nhiều phương pháp khai phá dữ liệu được nghiên cứu ở trên, trong đó có ba phương pháp được các nhà nghiên cứu sử dụng nhiều nhất đó là: Luật kết hợp, Phân lớp dữ liệu và Phân cụm dữ liệu.

1.5. Các lĩnh vực ứng dụng thực tiễn của KPDL

KPDL là một lĩnh vực mới phát triển nhưng thu hút được khá nhiều nhà nghiên cứu nhờ vào những ứng dụng thực tiễn của nó. Sau đây là một số lĩnh vực ứng dụng thực tế điển hình của KPDL:

- Phân tích dữ liệu và hỗ trợ ra quyết định
- Phân lớp văn bản, tóm tắt văn bản, phân lớp các trang Web và phân cụm ảnh màu
- Chuẩn đoán triệu chứng, phương pháp trong điều trị y học
- Tìm kiếm, đối sánh các hệ Gene và thông tin di truyền trong sinh học
- Phân tích tình hình tài chính, thị trường, dự báo giá cổ phiếu trong tài chính, thị trường và chứng khoán
- Phân tích dữ liệu marketing, khách hàng.
- Điều khiển và lập lịch trình
- Bảo hiểm
- Giáo dục.....

1.6. Các hướng tiếp cận cơ bản và kỹ thuật áp dụng trong KPDL.

Vấn đề khai phá dữ liệu có thể được phân chia theo lớp các hướng tiếp cận chính sau:

- **Phân lớp và dự đoán (classification & prediction):** Là quá trình xếp một đối tượng vào một trong những lớp đã biết trước (ví dụ: phân lớp các bệnh nhân theo dữ liệu hồ sơ bệnh án, phân lớp vùng địa lý theo dữ liệu thời tiết...). Đối với hướng tiếp cận này thường sử dụng một số kỹ thuật của học máy như cây quyết định (decision tree), mạng nơron nhân tạo (neural network),... Hay lớp bài toán này còn được gọi là học có giám sát - Học có thầy (supervised learning).

- **Phân cụm (clustering/segmentation):** Sắp xếp các đối tượng theo từng cụm dữ liệu tự nhiên, tức là số lượng và tên cụm chưa được biết trước. Các đối tượng được gom cụm sao cho mức độ tương tự giữa các đối tượng trong cùng một cụm là lớn nhất và mức độ tương tự giữa các đối tượng nằm trong các cụm khác nhau là nhỏ nhất. Lớp bài toán này còn được gọi là học không giám sát - Học không thầy (unsupervised learning).

- **Luật kết hợp (association rules):** Là dạng luật biểu diễn tri thức ở dạng khá đơn giản (Ví dụ: 80% sinh viên đăng ký học CSDL thì có tới 60% trong số họ đăng ký học Phân tích thiết kế hệ thống thông tin). Hướng tiếp cận này được ứng dụng nhiều trong lĩnh vực kinh doanh, y học, tin sinh học, giáo dục, viễn thông, tài chính và thị trường chứng khoán,...

- **Phân tích chuỗi theo thời gian (sequential/temporal patterns):** Cũng tương tự như khai phá dữ liệu bằng luật kết hợp nhưng có thêm tính thứ tự và tính thời gian. Một luật mô tả mẫu tuần tự có dạng tiêu biểu $X \rightarrow Y$, phản ánh sự xuất hiện của biến cố X sẽ dẫn đến việc xuất hiện biến cố Y. Hướng tiếp cận này được ứng dụng nhiều trong lĩnh vực tài chính và thị trường chứng khoán bởi chúng có tính dự báo cao.

- **Mô tả khái niệm (concept description & summarization):** Lớp bài toán này thiên về mô tả, tổng hợp và tóm tắt khái niệm (Ví dụ: tóm tắt văn bản).

1.7. Các thách thức - khó khăn trong KPTT và KPDL

KPTT và KPDL liên quan đến nhiều ngành, nhiều lĩnh vực trong thực tế, vì vậy các thách thức và khó khăn ngày càng nhiều, càng lớn hơn. Sau đây là một số các thách thức và khó khăn cần được quan tâm:

- + Các cơ sở dữ liệu lớn, các tập dữ liệu cần xử lý có kích thước cực lớn, Trong thực tế, kích thước của các tập dữ liệu thường ở mức tera-byte (*hàng ngàn giga-byte*).
- + Mức độ nhiễu cao hoặc dữ liệu bị thiếu
- + Số chiều lớn
- + Thay đổi dữ liệu và tri thức có thể làm cho các mẫu đã phát hiện không còn phù hợp
- + Quan hệ giữa các trường phức tạp

1.8. Kết luận

KPDL là lĩnh vực đã và đang trở thành một trong những hướng nghiên cứu thu hút được sự quan tâm của nhiều chuyên gia về CNTT trên thế giới. Trong những năm gần đây, rất nhiều các phương pháp và thuật toán mới liên tục được công bố. Điều này chứng tỏ những ưu thế, lợi ích và khả năng ứng dụng thực tế to lớn của KPDL. Chương này đã trình bày một số kiến thức tổng quan về KPTT, những khái niệm và kiến thức cơ bản nhất về KPDL.

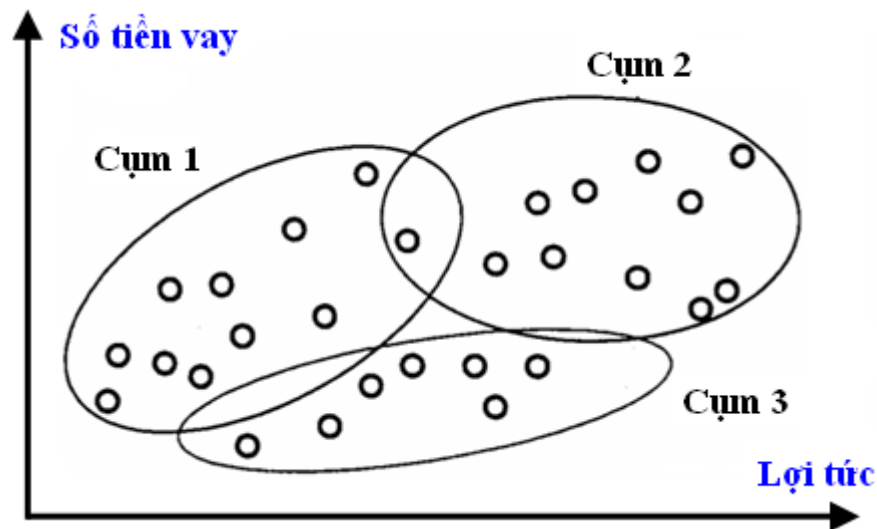
CHƯƠNG 2

PHÂN CỤM DỮ LIỆU VÀ CÁC THUẬT TOÁN TRONG PHÂN CỤM DỮ LIỆU

2.1.	Khái niệm và mục tiêu của phân cụm dữ liệu	13
2.2.	Các ứng dụng của phân cụm dữ liệu	15
2.3.	Các yêu cầu của phân cụm	16
2.4.	Những kỹ thuật tiếp cận trong phân cụm dữ liệu	18
2.4.1.	Phương pháp phân cụm phân hoạch	19
2.4.2.	Phương pháp phân cụm phân cấp	19
2.4.3.	Phương pháp phân cụm dựa trên mật độ	20
2.4.4.	Phương pháp phân cụm dựa trên lưới	21
2.4.5.	Phương pháp phân cụm dựa trên mô hình	22
2.4.6.	Phương pháp phân cụm có dữ liệu ràng buộc	22
2.5.	Một số thuật toán cơ bản trong phân cụm dữ liệu	24
2.5.1.	Các thuật toán phân cụm phân hoạch	24
2.5.2.	Các thuật toán phân cụm phân cấp	26
2.5.3.	Các thuật toán phân cụm dựa trên mật độ	29
2.5.4.	Các thuật toán phân cụm dựa trên lưới	32
2.5.5.	Các thuật toán phân cụm dựa trên mô hình	35
2.5.6.	Các thuật toán phân cụm có dữ liệu ràng buộc	36

2.1. Khái niệm và mục tiêu của phân cụm dữ liệu

Phân cụm dữ liệu là quá trình nhóm một tập các đối tượng tương tự nhau trong tập dữ liệu vào các cụm sao cho các đối tượng thuộc cùng một cụm là tương đồng còn các đối tượng thuộc các cụm khác nhau sẽ không tương đồng. Phân cụm dữ liệu là một ví dụ của phương pháp học không có thầy. Không giống như phân lớp dữ liệu, phân cụm dữ liệu không đòi hỏi phải định nghĩa trước các mẫu dữ liệu huấn luyện. Vì thế, có thể coi phân cụm dữ liệu là một cách học bằng quan sát, trong khi phân lớp dữ liệu là học bằng ví dụ... Ngoài ra phân cụm dữ liệu còn có thể được sử dụng như một bước tiền xử lý cho các thuật toán khai phá dữ liệu khác như là phân loại và mô tả đặc điểm, có tác dụng trong việc phát hiện ra các cụm.



Hình 2.1: Mô tả tập dữ liệu vay nợ được phân thành 3 cụm.

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.... Với tư cách là một chức năng khai phá dữ liệu, phân tích 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.

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 lý 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... Một trong các kỹ thuật xử lý 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 phần tử 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 CSDL, 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.


Mục tiêu của phân cụm là xác định được bản chất nhóm trong tập DL chưa có nhãn. Nhưng để có thể quyết định được cái gì tạo thành một cụm tốt. Nó có thể được chỉ ra rằng không có tiêu chuẩn tuyệt đối “tốt” mà có thể không phụ thuộc vào kq phân cụm. Vì vậy, nó đòi hỏi người sử dụng phải cung cấp tiêu chuẩn này, theo cách mà kết quả phân cụm sẽ đáp ứng yêu cầu.

Theo các nghiên cứu cho thấy thì hiện nay chưa có một phương pháp phân cụm tổng quát nào có thể giải quyết trọn vẹn cho tất cả các dạng cấu trúc CDL. Hơn nữa, các phương pháp phân cụm cần có cách thức biểu diễn cấu trúc của các CDL, với mỗi cách thức biểu diễn khác nhau sẽ có tương ứng một thuật toán phân cụm phù hợp. 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 KPDL.

2.2. Các ứng dụng của phân cụm dữ liệu

Phân cụm dữ liệu có thể được ứng dụng trong nhiều lĩnh vực như:

 **Thương mại:** Tìm kiếm nhóm các khách hàng quan trọng có đặc trưng tương đồng và những đặc tả họ từ các bản ghi mua bán trong CSDL

 **Sinh học:** Phân loại các gen với các chức năng tương đồng và thu được các cấu trúc trong mẫu

✚ **Thư viện:** Phân loại các cụm sách có nội dung và ý nghĩa tương đồng nhau để cung cấp cho độc giả

✚ **Bảo hiểm:** Nhận dạng nhóm tham gia bảo hiểm có chi phí bồi thường cao, nhận dạng gian lận thương mại

✚ **Quy hoạch đô thị:** Nhận dạng các nhóm nhà theo kiểu và vị trí địa lí,... nhằm cung cấp thông tin cho quy hoạch đô thị

✚ **Nghiên cứu trái đất:** Phân cụm để theo dõi các tâm động đất nhằm cung cấp thông tin cho nhận dạng các vùng nguy hiểm

✚ **WWW:** Có thể khám phá các nhóm tài liệu quan trọng, có nhiều ý nghĩa trong môi trường Web. Các lớp tài liệu này trợ giúp cho việc KPTT từ dữ liệu.

2.3. Các yêu cầu của phân cụm

Phân cụm là một thách thức trong lĩnh vực nghiên cứu ở chỗ những ứng dụng tiềm năng của chúng được đưa ra ngay chính trong những yêu cầu đặc biệt của chúng. Sau đây là những yêu cầu cơ bản của phân cụm trong KPDL:

✚ **Có khả năng mở rộng:** Nhiều thuật toán phân cụm làm việc tốt với những tập dữ liệu nhỏ chứa ít hơn 200 đối tượng, tuy nhiên, một CSDL lớn có thể chứa tới hàng triệu đối tượng. Việc phân cụm với một tập dữ liệu lớn có thể làm ảnh hưởng tới kết quả. Vậy làm cách nào để chúng ta có thể phát triển các thuật toán phân cụm có khả năng mở rộng cao đối với các CSDL lớn ?

✚ **Khả năng thích nghi với các kiểu thuộc tính khác nhau:** Nhiều thuật toán được thiết kế cho việc phân cụm dữ liệu có kiểu khoảng (kiểu số). Tuy nhiên, nhiều ứng dụng có thể đòi hỏi việc phân cụm với nhiều kiểu dữ liệu khác nhau, như kiểu nhị phân, kiểu tường minh (định danh -

không thứ tự), và dữ liệu có thứ tự hay dạng hỗn hợp của những kiểu dữ liệu này.

- ✚ **Khám phá các cụm với hình dạng bất kỳ:** Nhiều thuật toán phân cụm xác định các cụm dựa trên các phép đo khoảng cách Euclidean và khoảng cách Manhattan. Các thuật toán dựa trên các phép đo như vậy hướng tới việc tìm kiếm các cụm hình cầu với mật độ và kích cỡ tương tự nhau. Tuy nhiên, một cụm có thể có bất cứ một hình dạng nào. Do đó, việc phát triển các thuật toán có thể khám phá ra các cụm có hình dạng bất kỳ là một việc làm quan trọng.
- ✚ **Tối thiểu lượng tri thức cần cho xác định các tham số đầu vào:** Nhiều thuật toán phân cụm yêu cầu người dùng đưa vào những tham số nhất định trong phân tích phân cụm (như số lượng các cụm mong muốn). Kết quả của phân cụm thường khá nhạy cảm với các tham số đầu vào. Nhiều tham số rất khó để xác định, nhất là với các tập dữ liệu có lượng các đối tượng lớn. Điều này không những gây trở ngại cho người dùng mà còn làm cho khó có thể điều chỉnh được chất lượng của phân cụm.
- ✚ **Khả năng thích nghi với dữ liệu nhiễu:** Hầu hết những CSDL thực đều chứa đựng dữ liệu ngoại lai, dữ liệu lỗi, dữ liệu chưa biết hoặc dữ liệu sai. Một số thuật toán phân cụm nhạy cảm với dữ liệu như vậy và có thể dẫn đến chất lượng phân cụm thấp.
- ✚ **Ít nhạy cảm với thứ tự của các dữ liệu vào:** Một số thuật toán phân cụm nhạy cảm với thứ tự của dữ liệu vào, ví dụ như với cùng một tập dữ liệu, khi được đưa ra với các thứ tự khác nhau thì với cùng một thuật toán có thể sinh ra các cụm rất khác nhau. Do đó, việc quan trọng là phát triển các thuật toán mà ít nhạy cảm với thứ tự vào của dữ liệu.
- ✚ **Số chiều lớn:** Một CSDL hoặc một kho dữ liệu có thể chứa một số chiều hoặc một số các thuộc tính. Nhiều thuật toán phân cụm áp dụng

tốt cho dữ liệu với số chiều thấp, bao gồm chỉ từ hai đến 3 chiều. Người ta đánh giá việc phân cụm là có chất lượng tốt nếu nó áp dụng được cho dữ liệu có từ 3 chiều trở lên. Nó là sự thách thức với các đối tượng dữ liệu cụm trong không gian với số chiều lớn, đặc biệt vì khi xét những không gian với số chiều lớn có thể rất thưa và có độ nghiêng lớn.

✚ **Phân cụm ràng buộc:** Nhiều ứng dụng thực tế có thể cần thực hiện phân cụm dưới các loại ràng buộc khác nhau. Một nhiệm vụ đặt ra là đi tìm những nhóm dữ liệu có trạng thái phân cụm tốt và thỏa mãn các ràng buộc.

✚ **Dễ hiểu và dễ sử dụng:** Người sử dụng có thể chờ đợi những kết quả phân cụm dễ hiểu, dễ lý giải và dễ sử dụng. Nghĩa là, sự phân cụm có thể cần được giải thích ý nghĩa và ứng dụng rõ ràng.

Với những yêu cầu đáng lưu ý này, nghiên cứu của ta về phân tích phân cụm diễn ra như sau: Đầu tiên, ta nghiên cứu các kiểu dữ liệu khác và cách chúng có thể gây ảnh hưởng tới các phương pháp phân cụm. Thứ hai, ta đưa ra một cách phân loại chung trong các phương pháp phân cụm. Sau đó, ta nghiên cứu chi tiết mỗi phương pháp phân cụm, bao gồm các phương pháp phân hoạch, phân cấp, dựa trên mật độ,... Ta cũng khảo sát sự phân cụm trong không gian đa chiều và các biến thể của các phương pháp khác.

2.4. Những kỹ thuật tiếp cận trong phân cụm dữ liệu

Các kỹ thuật phân cụm có rất nhiều cách tiếp cận và các ứng dụng trong thực tế, nó đều hướng tới hai mục tiêu chung đó là chất lượng của các cụm khám phá được và tốc độ thực hiện của thuật toán. Hiện nay, các kỹ thuật phân cụm có thể phân loại theo các cách tiếp cận chính sau :

2.4.1. Phương pháp phân cụm phân hoạch

Kỹ thuật này phân hoạch một tập hợp dữ liệu có n phần tử thành k nhóm cho đến khi xác định số các cụm được thiết lập. Số các cụm được thiết lập là các đặc trưng được lựa chọn trước. Phương pháp này là tốt cho việc tìm các cụm hình cầu trong không gian Euclidean. Ngoài ra, phương pháp này cũng phụ thuộc vào khoảng cách cơ bản giữa các điểm để lựa chọn các điểm dữ liệu nào có quan hệ là gần nhau với mỗi điểm khác và các điểm dữ liệu nào không có quan hệ hoặc có quan hệ là xa nhau so với mỗi điểm khác. Tuy nhiên, phương pháp này không thể xử lý các cụm có hình dạng kỳ quặc hoặc các cụm có mật độ các điểm dày 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 PCDL, 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 (Greedy) để tìm kiếm nghiệm.

2.4.2. Phương pháp phân cụm phân cấp

Phương pháp này xây dựng một phân cấp trên cơ sở các đối tượng dữ liệu đang xem xét. Nghĩa là sắp xếp một tập dữ liệu đã cho thành một cấu trúc có dạng hình cây, cây phân cấp này được xây dựng theo kỹ thuật đệ quy. Có hai cách tiếp cận phổ biến của kỹ thuật này đó là:

- * Hòa nhập nhóm, thường được gọi là tiếp cận Bottom-Up
- * Phân chia nhóm, thường được gọi là tiếp cận Top-Down