

**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 $\epsilon$ FCM( $\epsilon$ - Insensitive Fuzzy C-means) .....	46
3.2.2.1.	Hàm mục tiêu .....	46
3.2.2.2.	Thuật toán $\epsilon$ 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 <i>(Mạng Noron đa khớp phục vụ cho phân cụm mờ)</i>
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

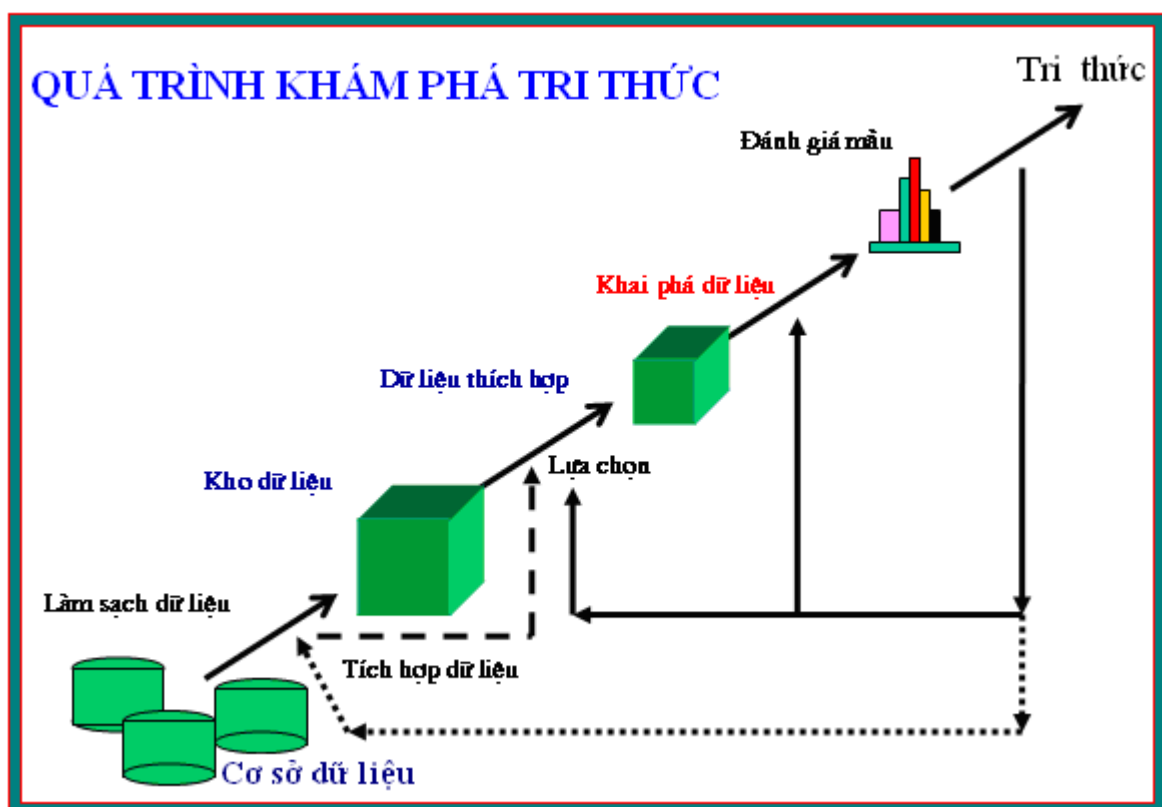
Nếu cho rằng, điện tử và truyền thông chính là bản chất của khoa học điện tử, thì dữ liệu, thông tin, và tri thức hiện đang là tiêu điểm của một lĩnh vực mới để nghiên cứu và ứng dụng, đó là khám phá tri thức và khai phá dữ liệu.

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 Intergation):** 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.

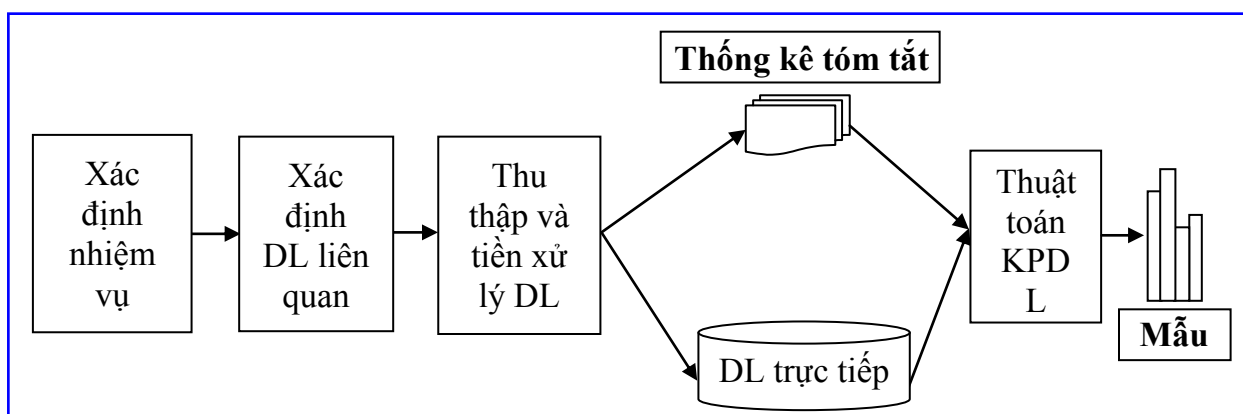
### 1.3. Quá trình khai phá dữ liệu

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

- *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 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 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 (*Visualization*)