TRƯỜNG ĐẠI HỌC CÔNG NGHIỆP HÀ NỘI

Giáo trình

# KHAI PHÁ DỮ LIỆU

NHÀ XUẤT BẢN THỐNG KÊ

TRƯỜNG ĐẠI HỌC CÔNG NGHIỆP HÀ NỘI
Chủ biên: Trần Hùng Cường
Trần Thanh Hùng

Giáo trình

# KHAI PHÁ DỮ LIỆU



NHÀ XUẤT BẢN THỐNG KÊ - 2017

# LỜI NÓI ĐẦU

*Sự phát triển của công nghệ thông tin và việc ứng dụng công nghệ này trong hầu hết các lĩnh vực của đời sống, kinh tế, xã hội trong nhiều năm qua cũng đồng nghĩa với lượng dữ liệu được thu thập và lưu trữ ngày càng nhiều đã dẫn đến sự bùng nổ dữ liệu. Mặt khác trong môi trường cạnh tranh, người ta ngày càng cần thông tin có độ tin cậy cao để trợ giúp việc ra quyết định. Từ dữ liệu lớn thu thập được, ta cần phải tìm ra tri thức và trả lời cho những câu hỏi mang tính khái quát, định hướng,... Với những lý do như vậy, các phương pháp quản trị và khai thác cơ sở dữ liệu truyền thống không còn đáp ứng đầy đủ những yêu cầu và những thách thức mới. Do vậy một khuynh hướng mới được ra đời đó là kỹ thuật phát hiện tri thức trong cơ sở dữ liệu.*

*Trong tiến trình phát hiện tri thức từ cơ sở dữ liệu, người ta đặc biệt quan tâm đến pha khai phá dữ liệu. Khai phá dữ liệu chính là sử dụng những kỹ thuật, những phương pháp để đưa ra những thông tin có cấu trúc, những tri thức tiềm ẩn trong lượng dữ liệu lớn. Kỹ thuật phát hiện tri thức được thực hiện qua nhiều giai đoạn và sử dụng nhiều phương pháp như: phân lớp, gom cụm, phân tích sự tương tự, tổng hợp, phát hiện luật kết hợp, mẫu tuần tự, ...*

*Khai phá dữ liệu là một giai đoạn trong tiến trình phát hiện tri thức từ cơ sở dữ liệu bao gồm các bước, các thuật giải nhằm tìm ra các mẫu, các mô hình và tri thức còn tiềm ẩn trong dữ liệu. Các tri thức này phục vụ cho nhiệm vụ mô tả, dự báo và hỗ trợ ra quyết định. Với yêu cầu như vậy, giáo trình* ***"Khai phá dữ liệu"*** *phần nào giúp độc giả bước đầu làm quen và tiến tới triển khai các ứng dụng của khai phá dữ liệu trong thực tế.*

*Giáo trình gồm 5 chương:*

*- Chương 1: Cung cấp cho bạn đọc tổng quan về khai phá dữ liệu, các kỹ thuật cơ bản trong khai phá dữ liệu cũng như ứng dụng của kỹ thuật này trong thực tế.*

*- Chương 2: Giới thiệu kỹ thuật khai phá dữ liệu đầu tiên đó là luật kết hợp, kỹ thuật này cho phép khám phá mối quan hệ giữa các tập mục trong các cơ sở dữ liệu giao dịch.*

*- Chương 3: Đề cập đến kỹ thuật phân lớp dựa trên cây quyết định, xác suất và mạng nơron nhân tạo. Kỹ thuật này cho phép gán nhãn cho các đối tượng chưa được xếp lớp.*

*- Chương 4: Trình bày kỹ thuật gom cụm. Kỹ thuật gom cụm nhằm nhóm các đối tượng có mức độ tương tự nhau dựa trên một tiêu chuẩn nào đó thành một nhóm. Một số phương pháp gom dựa trên phân hoạch và dựa trên phân cấp với các thuật toán tiêu biểu: thuật toán K - means, thuật toán gộp.*

*- Chương 5: Giới thiệu một số phần mềm sử dụng kỹ thuật khai phá dữ liệu cho các bài toán trong thực tế như Weka và Microsoft SQL Server. Bên cạnh đó, chương này cũng đề cập đến việc sử dụng ngôn ngữ C# để cài đặt các thuật toán khai phá dữ liệu.*

*Tập thể tác giả bày tỏ lòng biết ơn đến Khoa Công nghệ Thông tin - Trường Đại học Công nghiệp Hà Nội và gửi lời cảm ơn đến TS. Nguyễn Mạnh Cường cùng các đồng nghiệp đã động viên và có những đóng góp quý báu giúp chúng tôi hoàn thành giáo trình này. Chúng tôi mong muốn nhận được những ý kiến đóng góp của bạn đọc gần xa để cuốn sách ngày càng hoàn thiện. Mọi ý kiến đóng góp xin gửi về Khoa Công nghệ Thông tin - Trường Đại học Công nghiệp Hà Nội.*

**TẬP THỂ TÁC GIẢ**

# MỤC LỤC


Trang

LỜI NÓI ĐẦU 3

**Chương 1. TỔNG QUAN VỀ KHAI PHÁ DỮ LIỆU** 7

1.1. Khai phá dữ liệu và quá trình phát hiện tri thức 7

1.2. Quá trình khám phá tri thức từ cơ sở dữ liệu 8

1.3. Khái quát các kỹ thuật khai phá dữ liệu 10

1.4. Ứng dụng của khai phá dữ liệu 12

1.5. Những thách thức trong khai phá dữ liệu 13

**Chương 2. KHAI PHÁ LUẬT KẾT HỢP** 18

2.1. Các khái niệm cơ bản 18

2.2. Bài toán khai phá luật kết hợp 20

2.3. Một số tính chất 21

2.4. Thuật toán Apriori 22

2.5. Thuật toán FP - growth 30

2.5.1. Giới thiệu 30

2.5.2. Thuật toán FP - growth 30

2.6. Tìm luật kết hợp 39

**Chương 3. KỸ THUẬT PHÂN LỚP** 43

3.1. Giới thiệu 43

3.2. Phân lớp dựa trên cây quyết định 45

3.2.1. Cây quyết định 45

3.2.2. Độ lợi thông tin 46

3.2.3. Thuật toán ID3 48

3.2.4. Sinh luật 53

3.2.5. Thuật toán C4.5 53

3.3. Phân lớp dựa trên xác suất 58
3.3.1. Một số khái niệm 58
3.3.2. Định lý Bayes 60
3.3.3. Phân lớp Naïve Bayes 61
3.4. Phân lớp dựa trên mạng nơron 65
3.4.1. Mạng nơron nhân tạo 65
3.4.2. Kiến trúc mạng Perceptron 68
3.4.3. Thuật giải huấn luyện mạng Perceptron 69

**Chương 4. KỸ THUẬT GOM CỤM** 79
4.1. Giới thiệu 79
4.2. Phân loại các kỹ thuật gom cụm 80
4.3. Một số khoảng cách 81
4.4. Gom cụm dữ liệu bằng phân hoạch 84
4.4.1. Giới thiệu 84
4.4.2. Thuật toán K - means 84
4.5. Gom cụm dữ liệu bằng phân cấp 89
4.5.1. Giới thiệu 89
4.5.2. Một số khoảng cách giữa các cụm 89
4.5.3. Thuật toán gộp 90

**Chương 5. MỘT SỐ PHẦN MỀM KHAI PHÁ DỮ LIỆU** 100
5.1. Khai phá dữ liệu với Weka 100
5.1.1. Giới thiệu Weka 100
5.1.2. Dữ liệu trong Weka 101
5.1.3. Môi trường chính trong Explorer 102
5.1.4. Các kỹ thuật học máy chính trong Weka 103
5.1.5. Khai phá luật kết hợp trong Weka 104
5.1.6. Phân lớp dữ liệu trong Weka 110
5.2. Khai phá dữ liệu với Microsoft SQL Server 113
5.2.1. Giới thiệu 113
5.2.2. Gom cụm trong SQL Server 114
5.3. Khai phá dữ liệu với ngôn ngữ C# 131
PHỤ LỤC 136
TÀI LIỆU THAM KHẢO 145