

386096

ĐẠI HỌC QUỐC GIA THÀNH PHỐ HỒ CHÍ MINH

TRƯỜNG ĐẠI HỌC CÔNG NGHỆ THÔNG TIN

ĐOOR



PGS.TS Đỗ Phúc

GIÁO TRÌNH

KHAI PHÁ DỮ LIỆU

(Data Mining)

VV17.8285

THƯ VIỆN QUỐC GIA  
VIỆT NAM

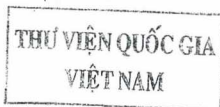
NHÀ XUẤT BẢN ĐẠI HỌC QUỐC GIA  
THÀNH PHỐ HỒ CHÍ MINH - 2016

ĐẠI HỌC QUỐC GIA THÀNH PHỐ HỒ CHÍ MINH  
TRƯỜNG ĐẠI HỌC CÔNG NGHỆ THÔNG TIN  
2002

PGS.TS Đỗ Phúc

GIÁO TRÌNH  
KHAI PHÁ DỮ LIỆU  
(Data Mining)

VV17.8285



NHÀ XUẤT BẢN ĐẠI HỌC QUỐC GIA  
THÀNH PHỐ HỒ CHÍ MINH - 2016

**GIÁO TRÌNH  
KHAI PHÁ DỮ LIỆU**

PGS. TS ĐỖ PHÚC

Bản tiếng Việt ©, TRƯỜNG ĐẠI HỌC CÔNG NGHỆ THÔNG TIN (ĐHQG-HCM),  
NXB ĐHQG-HCM và TÁC GIẢ.

Bản quyền tác phẩm đã được bảo hộ bởi Luật Xuất bản và Luật Sở hữu trí tuệ Việt Nam.  
Nghiêm cấm mọi hình thức xuất bản, sao chụp, phát tán nội dung khi chưa có sự đồng ý  
của tác giả và Nhà xuất bản.

ĐỂ CÓ SÁCH HAY, CẢN CHUNG TAY BẢO VỆ TÁC QUYỀN!

## LỜI NÓI ĐẦU

Khai phá dữ liệu (data mining) nhằm tìm kiếm và nhận dạng các mẫu hữu ích tiềm ẩn trong khối dữ liệu lớn. Hiện nay, công nghệ thông tin (CNTT) được áp dụng rộng rãi trong mọi lĩnh vực của đời sống xã hội, do vậy nhiều cơ sở dữ liệu (CSDL) khổng lồ đã được tạo lập như: các CSDL trong siêu thị, các CSDL lưu trữ các thông tin trên Internet, các CSDL về trình tự sinh học, các CSDL lưu trữ các tín hiệu thu được từ các sensor,... Nhu cầu phát triển các phương pháp cho phép tìm kiếm, nhận dạng các mẫu tiềm ẩn trong dữ liệu là cấp thiết và có nhiều ứng dụng. Ví dụ tìm các quy luật về hành vi mua hàng của khách hàng trong siêu thị, tìm các chuỗi bất thường trong dữ liệu chuỗi thời gian, tìm các quy luật phân lớp khách hàng trong kinh doanh, phân loại xếp hạng doanh nghiệp theo các đặc điểm của doanh nghiệp,... Đây là các bài toán có rất nhiều ứng dụng trong thực tế.

*Giáo trình Khai phá dữ liệu* tập trung trình bày các nội dung cơ bản liên quan đến khai phá dữ liệu như: tiền xử lý dữ liệu, bài toán tìm tập phổ biến và luật kết hợp, bài toán luật dãy, bài toán tìm luật phân lớp, lý thuyết tập thô và ứng dụng, bài toán phân tích gom cụm, xử lý văn bản. Bên cạnh đó, chúng tôi còn trình bày giải pháp khai phá dữ liệu của SQL Server và phần mềm khai phá dữ liệu Weka nhằm mục đích cụ thể hóa các thực hành khai phá dữ liệu. Sau mỗi chương của giáo trình, chúng tôi chuẩn bị phần bài tập để sinh viên có thể thực hành khai phá dữ liệu.

Mặc dù đã cố gắng trong quá trình biên soạn, nhưng chắc chắn giáo trình còn có những thiếu sót, chúng tôi rất mong nhận được góp ý của người đọc để sách ngày càng hoàn thiện hơn.

Trân trọng

PGS.TS ĐỖ Phúc



## MỤC LỤC

LỜI NÓI ĐẦU .....	iii
<b>CHƯƠNG 1: TỔNG QUAN VỀ KHAI PHÁ DỮ LIỆU .....</b>	<b>1</b>
1.1. MỞ ĐẦU .....	1
1.1.1. Khai phá dữ liệu .....	1
1.1.2. Lịch sử phát triển KPD L .....	3
1.1.3. Tại sao dùng khai phá dữ liệu .....	4
1.2. CÁC CÔNG ĐOẠN KHÁM PHÁ TRI THỨC TỪ CSDL .....	5
1.2.1. Giai đoạn chọn lọc dữ liệu.....	6
1.2.2. Giai đoạn làm sạch dữ liệu .....	8
1.2.3. Giai đoạn mã hóa dữ liệu.....	11
1.2.4. Khai phá dữ liệu .....	12
1.2.5. Trình diễn dữ liệu.....	12
1.3. ỨNG DỤNG CỦA KPD L .....	12
1.4. KHÁI QUÁT CÁC KỸ THUẬT KPD L .....	13
1.5. NHỮNG THÁCH THỨC TRONG KPD L .....	15
1.6. BÀI TẬP .....	15
<b>CHƯƠNG 2: TẬP PHỔ BIẾN VÀ LUẬT KẾT HỢP .....</b>	<b>17</b>
2.1. MỞ ĐẦU .....	17
2.2. BÀI TOÁN TÌM TẬP PHỔ BIẾN .....	17
2.2.1. Các khái niệm cơ bản.....	17
2.2.2. Phát biểu bài toán và độ phức tạp .....	21

2.3. PHÁT TRIỂN GIẢI THUẬT KHÔNG TĂNG CƯỜNG ĐỂ TÌM TẬP PHỔ BIẾN .....	22
2.3.1. Các khái niệm cơ bản.....	22
2.3.2. Phát triển giải thuật không tăng cường để tìm tập phổ biến.....	24
2.4. TÌM TẬP PHỔ BIẾN TỐI ĐẠI .....	29
2.4.1. Tìm tập phổ biến tối đại bằng đồ thị liên kết các tập phổ biến.....	29
2.4.2. Quan hệ giữa bao đóng và tập phổ biến tối đại.....	31
2.4.3. Dùng bao đóng để giảm số lượng ứng viên cần tính độ phổ biến.....	33
2.5. PHÁT TRIỂN GIẢI THUẬT TĂNG CƯỜNG ĐỂ TÌM TẬP PHỔ BIẾN .....	38
2.5.1. Các khái niệm cơ bản.....	38
2.5.2. Dùng giải thuật tạo dần khái niệm để tìm tập phổ biến.....	41
2.5.3. Duyệt dần khái niệm tìm tập phổ biến bị đóng.....	47
2.5.4. Tìm các tập không phổ biến từ dần khái niệm.....	47
2.6. PHÁT TRIỂN GIẢI THUẬT TÌM LUẬT KẾT HỢP.....	47
2.6.1. Các khái niệm cơ bản.....	47
2.6.2. Phát biểu bài toán tìm luật kết hợp.....	48
2.6.3. Phát triển giải thuật tìm luật kết hợp.....	48
2.7. BÀI TẬP .....	49
<b>CHƯƠNG 3: DÃY PHỔ BIẾN.....</b>	<b>52</b>
3.1. MỞ ĐẦU .....	52

3.2. DÂY PHỔ BIẾN TRONG MỘT CHUỖI .....	52
3.2.1. Giới thiệu .....	52
3.2.2. Các khái niệm cơ bản.....	53
3.2.3. Dây phổ biến trong một chuỗi.....	53
3.2.4. Các loại episode .....	53
3.2.5. Tiếp cận WINEPI.....	54
3.2.6. Tần suất.....	55
3.2.7. Luật Episode .....	55
3.2.8. Giải thuật WINEPI.....	56
3.3. DÂY PHỔ BIẾN TRONG NHIỀU CHUỖI .....	59
3.3.1. Giới thiệu .....	59
3.3.2. Bài toán tìm dây phổ biến trong nhiều chuỗi .....	59
3.3.3. Giải thuật AprioriAll.....	60
3.4. BÀI TẬP .....	64
<b>CHƯƠNG 4: PHÂN LỚP DỮ LIỆU .....</b>	<b>66</b>
4.1. MỞ ĐẦU .....	66
4.2. PHÂN LỚP QUY NẠP TRÊN CÂY QUYẾT ĐỊNH.....	67
4.2.1. Rút gọn cây quyết định và tập luật suy dẫn.....	74
4.2.2. Phân lớp với chỉ số Gini .....	74
4.3. PHƯƠNG PHÁP PHÂN LỚP BAYES .....	75
4.3.1. Sự phân hoạch và công thức Bayes .....	75
4.3.2. Bộ phân lớp Naive Bayes .....	76
4.4. PHÂN LỚP BẢNG MẠNG NƠON LAN TRUYỀN NGƯỢC...78	



4.5. CÁC PHƯƠNG PHÁP PHÂN LỚP KHÁC .....	80
4.5.1. Phân lớp dựa trên luật kết hợp.....	80
4.5.2. Giải thuật di truyền.....	80
4.5.3. Tiếp cận lý thuyết tập thô .....	80
4.6. BÀI TẬP .....	80
<b>CHƯƠNG 5: LÝ THUYẾT TẬP THÔ.....</b>	<b>83</b>
5.1. MỞ ĐẦU .....	83
5.2. CÁC HỆ THỐNG TIN .....	83
5.2.1. Hệ thống tin.....	83
5.2.2. Hệ quyết định (decision system).....	84
5.3. QUAN HỆ BẤT KHẢ PHÂN BIỆT .....	85
5.4. XẤP XỈ TẬP HỢP .....	86
5.5. RÚT GỌN .....	90
5.5.1. Định nghĩa rút gọn .....	92
5.5.2. Ma trận phân biệt .....	92
5.5.3. Hàm phân biệt .....	92
5.6. PHỤ THUỘC THUỘC TÍNH .....	97
5.7. BÀI TẬP .....	98
<b>CHƯƠNG 6: GOM CỤM DỮ LIỆU.....</b>	<b>100</b>
6.1. MỞ ĐẦU .....	100
6.2. ĐỘ ĐO KHOẢNG CÁCH.....	102
6.2.1. Biến trị khoảng.....	103
6.2.2. Biến nhị phân đối xứng.....	104

6.2.3. Biến nhị phân bất đối xứng .....	105
6.2.4. Biến định danh (nominal variable) .....	106
6.2.5. Biến thứ tự .....	107
6.2.6. Biến tỷ lệ theo khoảng .....	108
6.2.7. Biến có kiểu hỗn hợp .....	108
<b>6.3. CÁC PHƯƠNG PHÁP GOM CỤM .....</b>	<b>109</b>
6.3.1. Các phương pháp phân hoạch .....	109
6.3.2. Các phương pháp phân cấp .....	110
6.3.3. Các phương pháp dựa trên mật độ .....	112
6.3.4. Các phương pháp dựa trên mô hình .....	115
6.3.5. Các phương pháp dựa trên lưới .....	115
<b>6.4. GIẢI THUẬT GOM CỤM K-MEANS .....</b>	<b>115</b>
6.4.1. Giới thiệu .....	115
6.4.2. Giải thuật k-means .....	116
6.4.3. Ưu điểm và nhược điểm của giải thuật .....	119
6.4.4. Các biến thể và cải tiến của k-means .....	120
<b>6.5. BÀI TẬP .....</b>	<b>127</b>
<b>CHƯƠNG 7: KHAI PHÁ VĂN BẢN .....</b>	<b>129</b>
<b>7.1. MỞ ĐẦU .....</b>	<b>129</b>
7.1.1. Khai thác văn bản và khai thác dữ liệu .....	129
7.1.2. Ý nghĩa của khai phá văn bản .....	129
<b>7.2. KIẾN TRÚC CỦA KHAI PHÁ VĂN BẢN .....</b>	<b>130</b>
7.2.1. Lựa chọn tài nguyên .....	131