

**ĐẠI HỌC THÁI NGUYÊN
TRƯỜNG ĐẠI HỌC CÔNG NGHỆ THÔNG TIN VÀ TRUYỀN THÔNG**

NGUYỄN ĐÔNG HUY

**MỘT SỐ KỸ THUẬT PHÂN CỤM DỮ LIỆU VÀ ỨNG DỤNG
PHÂN LOẠI KHÁCH HÀNG SỬ DỤNG DỊCH VỤ VIỄN THÔNG**

LUẬN VĂN THẠC SĨ CÔNG NGHỆ THÔNG TIN

Thái Nguyên - 2014

LỜI CẢM ƠN

Trong quá trình làm luận văn, bản thân em đã nhận được nhiều sự giúp đỡ chỉ bảo tận tình của các thầy cô giáo, sự giúp đỡ, tạo điều kiện của gia đình, bạn bè để hoàn thành khóa luận đúng tiến độ.

Em xin trân trọng cảm ơn thầy giáo TS. Nguyễn Huy Đức đã trực tiếp hướng dẫn nhiệt tình, chỉ bảo cặn kẽ trong quá trình làm luận văn.

Em cũng xin gửi lời cảm ơn chân thành tới Ban lãnh đạo nhà trường, các cán bộ giảng viên của trường Đại học Công nghệ Thông tin và Truyền thông – Đại học Thái Nguyên đã tạo điều kiện thuận lợi để em hoàn thành tốt khóa luận.

Học viên

Nguyễn Đông Huy

LỜI CAM ĐOAN

Em xin cam đoan những kiến thức trình bày trong luận văn này là do em tìm hiểu, nghiên cứu và trình bày lại theo cách hiểu của em. Trong quá trình làm luận văn em có tham khảo các tài liệu liên quan và đã ghi rõ nguồn tài liệu tham khảo đó. Phần lớn những kiến thức do em trình bày trong luận văn này chưa được trình bày hoàn chỉnh trong bất cứ tài liệu nào.

Thái Nguyên, ngày 10 tháng 4 năm 2014

Học viên

Nguyễn Đông Huy

MỤC LỤC

LỜI CẢM ƠN	2
LỜI CAM ĐOAN.....	3
MỤC LỤC.....	4
DANH SÁCH HÌNH VẼ	6
DANH SÁCH BẢNG BIỂU	8
DANH MỤC CÁC TỪ VIẾT TẮT.....	9
LỜI MỞ ĐẦU	10
CHƯƠNG 1: TỔNG QUAN VỀ KHAI PHÁ DỮ LIỆU	10
1.1. Khai phá dữ liệu và phát hiện tri thức.....	11
1.1.1. Khai phá dữ liệu.....	11
1.1.2. Quá trình khám phá tri thức.....	12
1.1.3. Khai phá dữ liệu và các lĩnh vực liên quan.....	13
1.1.4. Các kỹ thuật áp dụng trong khai phá dữ liệu.....	13
1.1.5. Ứng dụng khai phá dữ liệu.....	15
1.2. Kỹ thuật phân cụm trong khai phá dữ liệu.....	16
1.2.1. Tổng quan về kỹ thuật phân cụm.....	16
1.2.2. Ứng dụng của phân cụm dữ liệu.....	18
1.2.3. Các yêu cầu kỹ thuật đối với phân cụm dữ liệu.....	19
1.3. Tổng kết chương 1	20
CHƯƠNG 2: MỘT SỐ KỸ THUẬT PHÂN CỤM DỮ LIỆU	21
2.1. Phân cụm phân hoạch.....	21
2.1.1 Thuật toán k-means.....	22
2.1.2 Thuật toán PAM.....	24
2.1.3 Thuật toán CLARA.....	28
2.1.4 Thuật toán CLARANS.....	29
2.2. Phân cụm phân cấp.....	31
2.2.1. Thuật toán BIRCH.....	32
2.2.2. Thuật toán CURE.....	35
2.3. Phân cụm dựa trên mật độ.....	37

2.3.1. Thuật toán DBSCAN.....	38
2.3.2. Thuật toán OPTICS.....	42
2.3.3. Thuật toán DENCLUE.....	43
2.4. Phân cụm trên lưới	44
2.4.1. Thuật toán STING.....	45
2.4.2. Thuật toán CLIQUE.....	46
2.5. Phân cụm dữ liệu dựa trên mô hình	47
2.5.1 Thuật toán EM	48
2.5.2 Thuật toán COBWEB	49
2.6. Phân cụm dữ liệu mờ	49
2.7. Tổng kết chương 2	50
CHƯƠNG 3: ỨNG DỤNG PHÂN CỤM DỮ LIỆU ĐỂ PHÂN LOẠI KHÁCH HÀNG SỬ DỤNG DỊCH VỤ VIỄN THÔNG	52
3.1 Đặt vấn đề bài toán.....	52
3.2 Cài đặt Cơ sở dữ liệu.....	52
3.3 Cài đặt thuật toán	56
3.4 Đánh giá kết quả phân cụm bằng thuật toán PAM	60
3.5 Kết luận chương 3	61
KẾT LUẬN	62
TÀI LIỆU THAM KHẢO	63
PHỤ LỤC	65

DANH SÁCH HÌNH VẼ

Hình 1.1. Quá trình khám phá tri thức.....	11
Hình 1.2. Các lĩnh vực liên quan đến khám phá tri thức trong CSDL.....	13
Hình 1.3. Trực quan hóa kết quả KPD L trong Oracle.....	15
Hình 1.4. Mô phỏng sự PCDL.....	16
Hình 2.1. Thuật toán k-means.....	22
Hình 2.2. Hình dạng cụm dữ liệu được khám phá bởi k-means.....	23
Hình 2.3. Trường hợp $C_{jmp} = d(O_j, O_m, 2) - d(O_j, O_m)$ không âm.....	25
Hình 2.4. Trường hợp $C_{jmp} = (O_j, O_p) - d(O_j, O_m)$ có thể âm hoặc dương.....	26
Hình 2.5. Trường hợp C_{jmp} bằng không.....	26
Hình 2.6. Trường hợp $C_{jmp} = (O_j, O_p) - d(O_j, O_m, 2)$ luôn âm.....	27
Hình 2.7. Thuật toán PAM.....	27
Hình 2.8. Thuật toán CLARA.....	28
Hình 2.9. Thuật toán CLARANS.....	31
Hình 2.10. Các chiến lược phân cụm phân cấp.....	32
Hình 2.11. Cây CF được sử dụng bởi thuật toán BIRCH.....	34
Hình 2.12. Thuật toán BIRCH.....	35
Hình 2.13. Ví dụ về kết quả phân cụm bằng thuật toán BIRCH.....	35
Hình 2.14. Các cụm dữ liệu được khám phá bởi CURE.....	37
Hình 2.15. Thuật toán CURE.....	37
Hình 2.16. Một số hình dạng khám phá bởi phân cụm dựa trên mật độ.....	38
Hình 2.17. Lân cận của P với ngưỡng Eps.....	39
Hình 2.18. Mật độ-đến được trực tiếp.....	40
Hình 2.19. Mật độ đến được.....	40
Hình 2.20. Mật độ liên thông.....	41
Hình 2.21. Cụm và nhiễu.....	41

Hình 2.22. Thuật toán DBSCAN.....	42
Hình 2.23. Thứ tự phân cụm các đối tượng theo OPTICS.....	43
Hình 2.24. DENCLUE với hàm phân phối Gaussian.....	45
Hình 2.25. Mô hình cấu trúc dữ liệu lưới.....	46
Hình 2.26. Thuật toán CLIQUE	48
Hình 2.27. Quá trình nhận dạng các cô của CLIQUE	48

DANH SÁCH BẢNG BIỂU

Hình 3.1. Các trường khai báo dữ liệu.....	54
Hình 3.2.Dữ liệu khách hàng.....	55
Hình 3.3.Dữ liệu khách hàng trong SQL Server.....	56
Hình 3.4.Giao diện chính của chương trình nhập dữ liệu.....	57
Hình 3.5.Giao diện chọn các tham số cho thuật toán.....	58
Hình 3.6.Giao diện phân cụm theo thời lượng cuộc gọi.....	58
Hình 3.7.Danh sách các khách hàng thuộc cụm 1 theo thời lượng cuộc gọi.....	59
Hình 3.8.Danh sách các khách hàng thuộc cụm 2 theo thời lượng cuộc gọi.....	59
Hình 3.9.Danh sách các khách hàng thuộc cụm 3 theo thời lượng cuộc gọi.....	59
Hình 3.10.Giao diện phân cụm theo tiền dịch vụ.....	60
Hình 3.11.Danh sách các khách hàng thuộc cụm 1 theo tiền dịch vụ.....	60
Hình 3.12.Danh sách các khách hàng thuộc cụm 2 theo tiền dịch vụ.....	61
Hình 3.13.Danh sách các khách hàng thuộc cụm 3 theo tiền dịch vụ.....	61

DANH MỤC CÁC TỪ VIẾT TẮT

Stt	Viết tắt	Cụm từ tiếng Anh	Cụm từ tiếng Việt
1	CNTT	Information Technology	Công nghệ thông tin
2	CSDL	Database	Cơ sở dữ liệu
3	KDD	Knowledge Discovery in Database	Khám phá tri thức trong cơ sở dữ liệu
4	KPDL	Data Mining	Khai phá dữ liệu
5	KPVB	Text Mining	Khai phá văn bản
6	PCDL	Data Clustering	Phân cụm dữ liệu

LỜI MỞ ĐẦU

Trong những năm gần đây cùng với phát triển nhanh chóng của khoa học kỹ thuật là sự bùng nổ về tri thức. Khodữ liệu, nguồn tri thức của nhân loại cũng trở nên đồ sộ, vô tận là mchovấn đề khai thác các nguồn tri thức đó ngày càng trở nên nóng bỏng và đặt ra thách thức lớn chọn công nghệ thông tin thế giới.

Đối với một doanh nghiệp thông tin đi động việc phát triển thuê bao mới để kiếm tìm lợi nhuận vào thời điểm hiện tại đã không còn đem lại hiệu quả. Thay vào đó là một phương án kinh doanh tiến đến phát triển chất lượng dịch vụ và cung cấp thêm nhiều dịch vụ giá trị gia tăng. Tuy nhiên các dịch vụ truyềnthôngnhu tho ại, nhấntin vẫn có thể đem lại nguồn lợi nhuận cao hơn nếu kích thích được nhu cầu sử dụng của khách hàng. Để thực hiện được điều đó, các doanh nghiệp phải không ngừng giữ vững được khách hàng hiện có mà còn phải đưa ra được các chiến lược phát triển kinh doanh dài hạn, phân loại được các nhóm khách hàng đang sử dụng để từ đó có chính sách phân khúc thị trường hợp lý. Vì vậy, em dựa vào thực trạng như trên và kết hợp với kỹ thuật phân cụm trong khai phá dữ liệu để thực hiện đề tài: *“Một số kỹ thuật phân cụm dữ liệu và ứng dụng phân loại khách hàng sử dụng dịch vụ Viễn thông”*

Bố cục luận văn gồm 3 chương:

Chương 1: Trình bày một cách tổng quan các kiến thức cơ bản về khai phá dữ liệu và phát hiện tri thức, các kỹ thuật phân cụm trong khai phá dữ liệu.

Chương 2: Giới thiệu một số dữ liệu phân cụm phổ biến thường được sử dụng trong khai phá dữ liệu và phát hiện tri thức.

Chương 3: Sử dụng kỹ thuật phân cụm để ứng dụng vào phân loại khách hàng sử dụng dịch vụ viễn thông. Trong chương này cũng trình bày chương trình mô phỏng áp dụng kỹ thuật phân cụm để phân loại sử dụng dịch vụ Viễn thông.

Phần kết luận của luận văn tổng kết lại những vấn đề đã nghiên cứu, đánh giá kết quả nghiên cứu, hướng phát triển của đề tài.

CHƯƠNG 1: TỔNG QUAN VỀ KHAI PHÁ DỮ LIỆU