HỌC VIỆN CÔNG NGHỆ BƯU CHÍNH VIỄN THÔNG

Vũ Hải Thuyết

NGHIÊN CỨU MỘT SỐ GIẢI THUẬT PHÂN CỤM TRONG KHAI PHÁ DỮ LIỆU

Chuyên ngành: Truyền dữ liệu và mạng máy tính
Mã số: 60.48.15

TÓM TẮT LUẬN VĂN THẠC SĨ

HÀ NỘI - 2012

Luận văn được hoàn thành tại:

HỌC VIỆN CÔNG NGHỆ BƯU CHÍNH VIỄN THÔNG

Người hướng dẫn khoa học: **PGS.TS Đoàn Văn Ban**

Phản biện 1: ……..…

Phản biện 2: ……..……

Luận văn sẽ được bảo vệ trước Hội đồng chấm luận văn thạc sĩ tại Học viện Công nghệ Bưu chính Viễn thông

Vào lúc: ..... giờ ..... ngày ..... tháng ..... năm .....

Có thể tìm hiểu luận văn tại:

- Thư viện của Học viện Công nghệ Bưu chính Viễn thông

I. MỞ ĐẦU

- **Lý do chọn đề tài**

Nhu cầu về tìm kiếm và xử lý thông tin, cùng với yêu cầu về khả năng kịp thời khai thác chúng để mang lại những năng suất và chất lượng cho công tác quản lý, hoạt động kinh doanh,… đã trở nên cấp thiết.

Với những lý do như vậy, các phương pháp quản trị và khai thác cơ sở dữ liệu truyền thống ngày càng không đáp ứng được thực tế đã làm phát triển một khuynh hướng kỹ thuật mới đó là Kỹ thuật phát hiện tri thức và khai phá dữ liệu (KDD – Knowledge Discovery and Data Mining).

Kỹ thuật phát hiện tri thức và khai phá dữ liệu đã và đang được nghiên cứu, ứng dụng trong nhiều lĩnh vực khác nhau. Bước quan trọng nhất của quá trình này là Khai phá dữ liệu (Data Mining), giúp người sử dụng thu được những tri thức hữu ích từ những cơ sở dữ liệu hoặc các nguồn dữ liệu khổng lồ khác. Rất nhiều doanh nghiệp và tổ chức trên thế giới đã ứng dụng kĩ thuật khai phá dữ liệu vào hoạt động sản xuất, kinh doanh và đã thu được những lợi ích to lớn. Nhưng để làm được điều đó, sự phát triển của các mô hình toán học và các giải thuật hiệu quả

là chìa khoá quan trọng. Do đó, tôi đã chọn đề tài ***"Nghiên cứu một số giải thuật phân cụm trong khai phá dữ liệu***".

- **Mục đích đề tài**
  - Nghiên cứu các phương pháp khai phá dữ liệu.
  - Nghiên cứu các kỹ thuật phân cụm dữ liệu và khả năng ứng dụng trong khai phá dữ liệu và phát triển tri thức.
- **Phương pháp nghiên cứu**

Nghiên cứu các tài liệu về khai phá dữ liệu, kỹ thuật phân cụm của các tác giả trong và ngoài nước, các bài báo, thông tin trên mạng.

- **Đối tượng và phạm vi nghiên cứu**

Tập trung nghiên cứu các thuật toán phân cụm dữ liệu.

- **Cấu trúc luận văn**

Ngoài các phần mở đầu, mục lục, danh mục hình vẽ, danh mục từ viết tắt, kết luận, tài liệu tham khảo, luận văn được chia làm 3 phần như sau:

Chương 1: Khai phá dữ liệu và phát hiện trí thức. Trình bày về khai phá dữ liệu, các khái niệm cơ bản, các kỹ thuật khai phá dữ liệu và ứng dụng khai phá dữ liệu.

Chương 2: Chương này trình bày một số phương pháp phân cụm dữ liệu phổ biến như phân cụm phân cấp, phân cụm dựa trên lưới, phân cụm dựa vào cụm trung tâm và phương pháp tiếp cận mới trong PCDL là phân cụm mờ.

Chương 3: Đánh giá và thử nghiệm. Phần này trình bày một số kết quả đã đạt được khi tiến hành áp dụng các giải thuật khai phá dữ liệu để khai thác thông tin dữ liệu mẫu.

CHƯƠNG I. KHAI PHÁ DỮ LIỆU VÀ PHÁT HIỆN TRI THỨC

1.1 Giới thiệu chung

Từ vài thập niên trở lại đây, với những tác động mạnh mẽ của các tiến bộ trong công nghệ phần cứng và truyền thông, các hệ thống dữ liệu phục vụ cho các lĩnh vực kinh tế xã hội phát triển bùng nổ, lượng dữ liệu được tạo ra ngày càng lớn.

Sự bùng nổ này đã dẫn tới một yêu cầu cấp thiết là cần có những kỹ thuật và công cụ mới để tự động chuyển đổi lượng dữ liệu khổng lồ kia thành các tri thức có ích, phục vụ cho việc ra quyết định.

1.2 Phát hiện tri thức và khai phá dữ liệu là gì?

Khai phá dữ liệu là một tập hợp các kỹ thuật được sử dụng để tự động khai thác và tìm ra các mối quan hệ lẫn nhau của dữ liệu trong một tập hợp dữ liệu khổng lồ và phức tạp, đồng thời cũng tìm ra các mẫu tiềm ẩn trong tập dữ liệu đó.

1.3 Các bước của quá trình khai phá dữ liệu

Quá trình khai phá dữ liệu gồm 6 bước:

1. Gom cụm dữ liệu.
2. Trích lọc dữ liệu.
3. Làm sạch, tiền xử lý và chuẩn bị trước dữ liệu.
4. Chuyển đổi dữ liệu.
5. Khai phá dữ liệu.
6. Đánh giá các luật và biểu diễn tri thức.

1.4 Các kỹ thuật áp dụng trong khai phá dữ liệu

Thường được chia thành 2 nhóm chính sau:

1.4.1 Kỹ thuật khai phá dữ liệu mô tả

Có nhiệm vụ mô tả về các tính chất hoặc các đặc tính chung của dữ liệu trong CSDL hiện có.

1.4.2 Kỹ thuật khai phá dữ liệu dự đoán

Kỹ thuật khai phá dữ liệu dự đoán: có nhiệm vụ đưa ra các dự đoán dựa vào suy các suy diễn trên dữ liệu hiện thời.

1.5 Ứng dụng của khai phá dữ liệu

1.5.1 Ứng dụng của khai phá dữ liệu

- Học máy: khai phá dữ liệu có thể sử dụng với các CSDL chứa nhiều nhiễu, dữ liệu không đầy đủ hoặc biến đổi liên tục.

- Phương pháp hệ chuyên gia: Phương pháp này khác với khai phá dữ liệu ở chỗ các ví dụ của chuyên gia thường ở mức chất lượng cao hơn rất nhiều so với các dữ liệu trong CSDL và chúng thường chỉ bao quát được các trường hợp quan trọng.

- Phương pháp thống kê: Khai phá dữ liệu tự động hóa quá trình thống kê một cách hiệu quả, vì vậy làm nhẹ bớt công việc của người dùng cuối, tạo ra một công cụ dễ sử dụng hơn.

1.5.2 Những thách thức trong khai phá dữ liệu

- Các cơ sở dữ liệu lớn hơn rất nhiều.
- Số chiều cao.
- Thay đổi dữ liệu (dữ liệu luôn động).
- Dữ liệu thiếu và bị nhiễu.
- Mối quan hệ phức tạp giữa các trường dữ liệu.
- Tính dễ hiểu của các mẫu.

- Người dùng tương tác và tri thức có sẵn.
- Tích hợp với các hệ thống khác.

CHƯƠNG II. PHÂN CỤM DỮ LIỆU TRONG KHAI PHÁ DỮ LIỆU

2.1 Phân cụm dữ liệu

Phân cụm dữ liệu là xử lý một tập các đối tượng vào trong các lớp các đối tượng giống nhau được gọi là phân cụm. Một cụm là một tập hợp các đối tượng dữ liệu giống nhau trong phạm vi cùng một cụm và không giống nhau với các đối tượng trong các cụm khác. Số các cụm dữ liệu được phân ở đây có thể được xác định trước theo kinh nghiệm hoặc có thể được tự động xác định của phương pháp phân cụm.

2.2 Các kiểu dữ liệu và độ đo tương tự trong phép phân cụm

2.2.1 Phân loại dữ liệu dựa trên kích thước miền

- Thuộc tính liên tục.
- Thuộc tính rời rạc.

2.2.2 Phân loại dữ liệu dựa trên hệ đo

- Thuộc tính định danh, thuộc tính thứ tự, thuộc tính khoảng, thuộc tính tỷ lệ,...

2.3 Các yêu cầu đối với kỹ thuật phân cụm

1 Khả năng mở rộng.

2 Thích nghi với các kiểu dữ liệu khác nhau.

3 Khám phá ra các cụm với hình thù bất kỳ.

4 Tối thiểu lượng tri thức cần cho xác định tham số vào.

5 Ít nhạy cảm với thứ tự của dữ liệu vào.

6 Thích nghi với dữ liệu nhiễu cao.

7 Ít nhạy cảm với tham số đầu vào.

8 Thích nghi với dữ liệu đa chiều.

9 Dễ hiểu, dễ cài đặt và khả thi.

2.4 Một số phương pháp phân cụm chính trong khai phá dữ liệu

2.4.1 Phương pháp phân cụm dữ liệu dựa trên phân cụm phân cấp

Phương pháp phân cụm phân cấp làm việc bằng cách nhóm các đối tượng vào trong một cây các cụm.

2.4.1.1 Phân cụm phân cấp tích đống và phân ly

Phân cụm phân cấp tích đống: bắt đầu bằng cách đặt mỗi đối tượng vào trong cụm của bản thân nó, sau đó kết nhập các cụm nguyên tử này vào trong các cụm ngày

càng lớn hơn cho tới khi tất cả các đối tượng nằm trong một cụm đơn hay cho tới khi thỏa mãn điều kiện dừng cho trước.

Phân cụm phân cấp phân ly: phương pháp này ngược lại bằng cách bắt đầu với tất cả các đối tượng trong một cụm, chia nhỏ nó vào trong các phần ngày càng nhỏ hơn cho tới khi mỗi đối tượng hình thành nên một cụm hay thỏa mãn một điều kiện dừng cho trước.

2.4.1.2 Thuật toán BIRCH (Balanced Iterative Reducing and Clustering using Hierarchies): phân hoạch các đối tượng dùng cấu trúc cây theo độ co giãn của phân giải.

Thuật toán BIRCH có 2 pha:

Pha 1: quét tất cả các đối tượng trong cở sở dữ liệu để xây dựng một cây CF (Clustering Feature) bộ nhớ trong ban đầu.

Cây CF đặc trưng bởi hai tham số:

1. Hệ số phân nhánh *B* (Braching Factor – B): Nhằm xác định tối đa các nút con của một nút lá trong cây.

2. Ngưỡng T (Threshold - T): khoảng cách tối đa giữa bất kỳ một cặp đối tượng trong nút lá của cây, khoảng cách này được gọi là đường kính của các cụm con được lưu trong nút lá.

Các đối tượng lần lượt được chèn vào nút lá gần nhất của cây CF. Sau khi chèn xong tất cả các nút trong cây CF được cập nhật thông tin. Nếu đường kính cụm con sau khi chèn lớn hơn ngưỡng T thì nút lá được tách. Quá trình lặp lại cho đến khi tất cả các đối tượng trong cây chỉ được đọc một lần.

Pha 2: Lựa chọn một thuật toán phân cụm để phân cụm các nút lá của cây CF.

2.4.1.3 Thuật toán CURE: phân cụm sử dụng đại diện

CURE (Clustering Using REpresentatives) cung cấp một giải thuật phân cụm theo vị trí giữa dựa trên trọng tâm và tất cả các cực điểm. Thay vì sử dụng một trọng tâm đơn đại diện một cụm, CURE ấn định một số lượng các điểm đại diện được lựa chọn để miêu tả một cụm. Các điểm đại diện này được sinh ra bằng cách trước tiên lựa chọn các điểm rải rác đều trong cụm, sau đó co chúng lại

về phía tâm cụm bởi một phân số (hệ số co). Các cụm với các cặp các điểm đại diện gần nhất sẽ được kết nhập tại mỗi bước của giải thuật.

CURE đưa ra các cụm chất lượng cao với sự hiện hữu của các outlier, các hình dạng phức tạp của các cụm với các kích thước khác nhau. Nó có khả năng mở rộng tốt cho các cơ sở dữ liệu lớn mà không cần hy sinh chất lượng phân cụm.

2.4.1.4 Thuật toán ROCK

Nó đo độ tương đồng của 2 cụm bằng cách so sánh toàn bộ liên kết nối của 2 cụm dựa trên mô hình liên kết nối tĩnh được chỉ định bởi người dùng, tại đó liên kết nối của hai cụm C_1 và C_2 được định nghĩa bởi số lượng các liên kết chéo giữa hai cụm và liên kết link(p_i,p_j) là số lượng các láng giềng chung giữa hai điểm p_i và p_j .

ROCK trước tiên xây dựng đồ thị thưa từ một ma trận tương đồng dữ liệu cho trước, sử dụng một ngưỡng tương đồng và khái niệm các láng giềng chia sẻ và sau đó biểu diễn một giải thuật phân cụm phân cấp trên đồ thị thưa.

2.4.1.5 Thuật toán CHAMELEON

CHAMELEON miêu tả các đối tượng dựa trên tiếp cận đồ thị được dùng phổ biến: k-láng giềng gần nhất. CHAMELEON trước tiên sử dụng một giải thuật phân chia đồ thị để phân cụm các mục dữ liệu vào trong một số lượng lớn các cụm con tương đối nhỏ. Sau đó dùng giải thuật phân cụm phân cấp tập hợp để tìm ra các cụm xác thực bằng cách lặp lại việc kết hợp các cụm này với nhau. Để xác định các cặp cụm con giống nhau nhất, cần đánh giá cả liên kết nối cũng như độ chặt của các cụm, đặc biệt là các đặc tính nội tại của bản thân các cụm. Do vậy nó không tuỳ thuộc vào một mô hình tĩnh được cung cấp bởi người dùng và có thể tự động thích ứng với các đặc tính nội tại của các cụm đang được kết nhập.

CHAMELEON chỉ rõ sự tương đồng giữa mỗi cặp các cụm C_i, C_j theo liên kết nối tương ứng $RI(C_i , C_j)$ giữa hai cụm C_i và C_j được định nghĩa như liên kết nối tuyệt đối giữa C_i và C_j. CHAMELEON có nhiều khả năng khám phá ra các cụm có hình dạng tùy ý với chất lượng cao hơn CURE.

CHAMELEON sử dụng thuật toán phân cụm phân cấp để tìm các cụm xác thực bằng cách lặp lại nhiều lần kết hợp hoặc hòa nhập các cụm con.

2.4.2 Phương pháp phân cụm dữ liệu dựa vào dữ liệu mờ

2.4.2.1 Thuật toán FCM (Fuzzy C-means)

Kỹ thuật này phân hoạch một tập n vectơ đối tượng dữ liệu $X = \{x_1,x_2,\ldots,x_n\} \subset R^s$ thành c các nhóm mờ dựa trên tính toán tối thiểu hóa hàm mục tiêu để đo chất lượng của phân hoạch và tìm trung tâm cụm trong mỗi nhóm, sao cho chi phí hàm độ đo độ phi tương tự là nhỏ nhất.

Tuy nhiên, thuật toán này vẫn mang những nhược điểm của thuật toán K-means.

2.4.2.2 Thuật toán εFCM (ε- Insensitive Fuzzy C-means)

Thuật toán εFCM là một mở rộng của thuật toán FCM nhằm khắc phục các nhược điểm của thuật toán FCM.

2.4.3 Phương pháp phân cụm dữ liệu dựa trên lưới.

2.4.3.1 Thuật toán STING

STING là kỹ thuật phân cụm đa phân giải dựa trên lưới, trong đó vùng không gian dữ liệu được phân rã thành số hữu hạn các cells chữ nhật. Điều này có ý nghĩa là các cells lưới được hình thành từ các cells lưới con để thực hiện phân cụm. Có nhiều mức của các cells chữ nhật tương ứng với các mức khác nhau của phân giải trong cấu trúc lưới, các cells này hình thành cấu trúc phân cấp: mỗi cells ở mức cao được phân hoạch thành các số các cells nhỏ ở mức thấp hơn tiếp theo trong cấu trúc phân cấp. Các điểm dữ liệu được nạp từ CSDL, giá trị của các tham số thống kê cho các thuộc tính của đối tượng dữ liệu trong mỗi ô lưới được tính toán từ dữ liệu và lưu trữ thông qua các tham số thống kê ở các cell mức thấp hơn. Các giá trị của các tham số thống kê gồm: số trung bình – mean, số tối đa – max, số tối thiểu – min, số đếm –count , độ lệch chuẩn – s,…

Các đối tượng dữ liệu lần lượt được chèn vào lưới và các tham số thống kê ở trên được tính trực tiếp thông

qua các đối tượng dữ liệu này. Các truy vấn không gian được thực hiện bằng cách xét các cells thích hợp tại mỗi mức phân cấp. STING có khả năng mở rộng cao, nhưng do sử dụng phương pháp đa phân giải nên nó phụ thuộc chặt chẽ vào trọng tâm của mức thấp nhất.

2.4.3.2 Thuật toán CLIQUE

CLIQUE phân chia không gian dữ liệu m chiều thành các unit hình chữ nhật không chồng lên nhau, nhận biết các unit dày đặc, và tìm ra các cụm trong toàn bộ các không gian con của không gian dữ liệu gốc, sử dụng phương pháp phát sinh candidate (ứng cử) giống với giải thuật Apriori cho khai phá các luật kết hợp.

CLIQUE thực hiện phân cụm đa chiều theo hai bước:

1. CLIQUE nhận biết các cụm bằng cách xác định các unit dày đặc trong toàn bộ các không gian con của các interest và sau đó xác định các unit dày đặc có kết nối trong toàn bộ các không gian con của các interest. Một heuristic quan trọng mà CLIQUE thông qua đó là nguyên lý Apriori trong phân cụm số chiều cao.

2. CLIQUE sinh ra mô tả tối thiểu cho các cụm như sau: Trước tiên nó xác định các miền tối đa phủ một cụm các unit dày đặc có kết nối cho mỗi cụm và sau đó xác định phủ tối thiểu cho mỗi cụm. CLIQUE tự động tìm các không gian con số chiều cao nhất để các cụm mật độ cao tồn tại trong các không gian con này.

2.4.3.3 Thuật toán WaveCluster

WaveCluster là một tiếp cận phân cụm đa phân giải, trước tiên tóm tắt dữ liệu bằng cách lợi dụng cấu trúc lưới đa phân giải trên không gian dữ liệu, sau đó biến đổi không gian đặc trưng gốc bằng phép biến đối WaveCluster và tìm các miền đông đúc trong không gian đã biến đổi.

Trong tiếp cận này, mỗi ô lưới tóm tắt thông tin của một nhóm các điểm, thông tin tóm tắt này vừa đủ để đưa vào trong bộ nhớ chính cho phép biến đổi wavelet đa phân giải và phép phân tích cụm sau đó.

Phép biến đổi WaveCluster là một kỹ thuật xử lý tín hiệu, nó phân tích một tín hiệu vào trong các dải tần số con. Mô hình WaveCluster cũng làm việc trên các tín hiệu n chiều bằng cách áp dụng phép biến đổi 1 chiều n lần.

Trong phép biến đổi WaveCluster, dữ liệu không gian được chuyển đổi vào trong miền tần số. Kết hợp với một hàm nòng cốt thích hợp cho kết quả trong một không gian biến đổi, tại đó các cụm tự nhiên trong dữ liệu trở nên dễ phân biệt hơn. Các cụm sau đó có thể được nhận biết bằng cách tìm ra các miền đông đúc trong vùng biến đổi.

2.4.4 Phương pháp phân cụm dựa vào cụm trung tâm (K-means, K-medoids)

2.4.4.1 Phương pháp K-means

Thuật toán phân cụm K-mean do Macqueen đề xuất trong lĩnh vực thống kê năm 1967, mục đích của thuật toán là sinh ra k cụm dữ liệu $\{C_1, C_2, \ldots, C_k\}$ từ một tập dữ liệu ban đầu gồm n đối tượng trong không gian d chiều $X_i = (x_{i_1}, x_{i_2}, \ldots, x_{i_d})$ i= $(\overline{1,n})$, sao cho hàm tiêu chuẩn $E = \sum_{i=1}^{i} \sum_{x \in C_i} D^2(x - m_i)$ đạt giá trị cực tiểu. Trong đó, m_i là trọng tâm của cụm C_i. D là khoảng cách giữa hai đối tượng.

Ưu điểm của thuật toán k-means: *Đây là một phương pháp đơn giản, hiệu quả, tự tổ chức, được sử dụng trong tiến trình khởi tạo trong nhiều thuật toán khác,*

hiệu xuất tương đối, thường kết thúc ở tối ưu cục bộ, có thể tìm được tối ưu toàn cục.

***Nhược điểm của thuật toán này**: Số cụm k phải được xác định trước,chỉ áp dụng được khi xác định được trị trung bình, không thể xử lý nhiễu và outliers, không thích hợp nhằm khám phá các dạng không lồi hay các cụm có kích thước khác nhau, đây là thuật toán độc lập tuyến tính.*

2.4.4.2 Phương pháp K-medoids

PAM (Partition Arount medoids) – phân chia xung quanh các medoid

PAM sử dụng các đối tượng medoid (k-medoids lấy một đối tượng đại diện trong cụm gọi là medoid, nó là điểm đại diện được định vị trung tâm nhất trong cụm) để biểu diễn cho các cụm dữ liệu, một đối tượng medoid là đối tượng đặt tại vị trí trung tâm nhất bên trong của mỗi cụm.

Để xác định các medoid, PAM bắt đầu bằng cách lựa chọn k đối tượng medoid bất kỳ. Sau mỗi bước thực hiện, PAM cố gắng hoán chuyển giữa đối tượng medoid O_m và một đối tượng O_p không phải là medoid, quá trình

này kết thúc khi chất lượng phân cụm không thay đổi. Chất lượng phân cụm được đánh giá thông qua hàm tiêu chuẩn, chất lượng phân cụm tốt nhất khi hàm tiêu chuẩn đạt giá trị tối thiểu. Khi có sự hiện diện của nhiễu và các outlier, phương pháp k-medoids mạnh hơn k-means. Tuy nhiên, xử lý của nó có chi phí tốn kém hơn phương pháp k-means và nó cũng cần người dùng chỉ ra k - số cụm.

2.5 Kết luận

Chương này trình bày một số phương pháp phân cụm dữ liệu phổ biến như phân cụm phân cấp, phân cụm dựa trên lưới, phân cụm dựa vào cụm trung tâm và phương pháp tiếp cận mới trong PCDL là phân cụm mờ.

Phương pháp phân cụm dữ liệu dựa vào cụm trung tâm dựa trên ý tưởng ban đầu tạo ra k cụm, sau đó lặp lại nhiều lần để phân bố lại các đối tượng dữ liệu giữa các cụm nhằm cải thiện chất lượng phân cụm. Một số thuật toán điển hình như K-means, PAM,...

Phương pháp phân cụm phân cấp dựa trên ý tưởng cây phân cấp để phân cụm dữ liệu. Có hai cách tiếp cận đó là phân cụm dưới lên (Bottom up) và phân cụm trên xuống

(Top down). Một số thuật toán điển hình như BIRCH, CURE,..

Phương pháp phân cụm dựa trên lưới, ý tưởng của nó là đầu tiên lượng hoá không gian đối tượng vào một số hữu hạn các ô theo một cấu trúc dưới dạng lưới, sau đó thực hiện phân cụm dựa trên cấu trúc lưới đó. Một số thuật toán tiêu biểu của phương pháp này là STING, CLIQUE,...

Một cách tiếp cận khác trong PCDL đó là hướng tiếp cận mờ, trong phương pháp phân cụm mờ phải kể đến các thuật toán như FCM, εFCM,...

CHƯƠNG III. ĐÁNH GIÁ VÀ THỬ NGHIỆM

3.1 Chuẩn bị dữ liệu

Dữ liệu đưa vào chương trình là các tệp văn bản và được chia thành hai loại:

- Tệp định dạng dữ liệu (*.name): Định nghĩa tên các lớp, tên các thuộc tính, các giá trị của từng thuộc tính, kiểu thuộc tính.
- Tệp mẫu dữ liệu (*.data): Gồm các mẫu dữ liệu chứa đầy đủ thông tin giá trị của các thuộc tính và giá trị lớp.

3.1.1 Tệp định dạng dữ liệu

- Dòng 1: Liệt kê các giá trị lớp. Các giá trị này cách nhau bởi dấu “,” và kết thúc bằng dấu chấm “.”.

- Từ dòng 2:

+ Mỗi mẫu một dòng

+ Bắt đầu bằng tên một thuộc tính, dấu “:”, sau đó là các giá trị rời rạc của thuộc tính (nếu thuộc tính là xác thực hay nhị phân) hoặc kiểu thuộc tính (nếu thuộc tính có kiểu liên tục)

- Tất cả chú thích được đặt sau dấu “|”.

3.1.2 *Tệp dữ liệu mẫu*

Mỗi mẫu một dòng. Các giá trị thuộc tính của mẫu ghi trước, cuối cùng là giá trị lớp. Mỗi một giá trị này cách nhau bởi dấu ",".

3.1.3 Nguồn dữ liệu

Dữ liệu mẫu được lấy từ địa chỉ website:

ftp://ftp.ics.uci.edu/pub/machine-learning-databases

3.2 Kết quả thực nghiệm phân cụm dữ liệu bằng giải thuật K-means, K-medoids và đánh giá

3.2.1 Các bước tiến hành thực nghiệm

- Phân cụm dữ liệu bằng giải thuật Kmeans và Kmedoids

- Gắn nhãn cho các cụm, đánh giá, so sánh hiệu quả gắn nhãn giữa hai giải thuật trên cho các bộ số liệu UCI (chỉ dùng các dữ liệu có thuộc tính liên tục).

3.2.2 Kết quả thực nghiệm

3.3 Kết luận

Sau khi tiến hành thực nghiệm trên một số bộ dữ liệu của UCI ta nhận thấy kết quả phân loại các dữ liệu có thuộc tính liên tục của Kmeans tốt hơn của K-medoids. Với dữ liệu có thuộc tính hỗn hợp, K-means không xử lý

được. K-medoids với phương pháp tính độ tương đồng giữa hai mẫu do Ducker (1965) đề xuất, Kaufman và Rousseeuw cải tiến (1990) đã xử lý được dữ liệu này với độ chính xác trên trung bình với độ phức tạp tính toán là $O(k(n-k)^2)$.

Đối với các giá trị n và k lớn, độ phức tạp tính toán sẽ cao. Vậy nên cải tiến độ chính xác và tốc độ tính toán là hướng phát triển sau này.

KẾT LUẬN

Luận văn tập trung nghiên cứu lý thuyết và áp dụng một số kỹ thuật khai phá dữ liệu trên bộ dữ liệu của UCI. Đây là bước khởi đầu trong quá trình tìm hiểu những vấn đề cần quan tâm khi giải quyết các bài toán khai phá dữ liệu trong thực tế.

Những kết quả mà luận văn đã đạt được

- Về lý thuyết: luận văn tập trung tìm hiểu kỹ thuật phân cụm truyền thống và phương pháp cải tiến chúng. Ngoài ra còn tìm hiểu thêm các ứng dụng vào các lĩnh vực khoa học thực tế.
- Về thực tiễn: luận văn cài đặt hai thuật toán K-means, K-medoid và so sánh đánh giá chúng.

Qua quá trình nghiên cứu lý thuyết và thực nghiệm có thể đưa ra một số kết luận sau:

- Mỗi một giải thuật phân cụm áp dụng cho một số mục tiêu và kiểu dữ liệu nhất định.
- Mỗi một giải thuật có độ chính xác riêng và khả năng thực hiện trên từng kích thước dữ liệu là khác nhau. Điều này còn tùy thuộc vào cách tổ chức dữ liệu ở bộ nhớ chính, bộ nhớ ngoài,… của các giải thuật.

- Khai phá dữ liệu sẽ hiệu quả hơn khi bước tiền xử lý, lựa chọn thuộc tính, mô hình được giải quyết tốt.

Với những gì mà luận văn đã thực hiện, hướng phát triển sau này của luận văn:

- Độ chính xác, kết quả phụ thuộc nhiều yếu tố như chất lượng dữ liệu, thuật toán cài đặt, phương pháp tính độ tương đồng của các đối tượng dữ liệu. Ngoài ra, các giá trị khuyết hay các thuộc tính dư thừa cũng phần nào làm ảnh hưởng đến chúng. Vì vậy, hướng phát triển sau này là xử lý các giá trị khuyết, phát hiện và loại bỏ các thuộc tính dư thừa, cái tiến phương pháp tính toán độ tương đồng,… nhằm nâng cao chất lượng và kết quả phân cụm
- Tiến hành cài đặt và tiếp tục nghiên cứu nhiều kỹ thuật khai phá dữ liệu hơn nữa, đặc biệt là triển khai giải quyết các bài toán cụ thể trong thực tế.