HỌC VIỆN CÔNG NGHỆ BƯU CHÍNH VIỄN THÔNG

NGUYỄN THỊ PHƯƠNG

PHÂN CỤM MỜ SỬ DỤNG LÝ THUYẾT ĐẠI SỐ GIA TỬ

Chuyên ngành: KHOA HỌC MÁY TÍNH

Mã số: 60.48.01

TÓM TẮT LUẬN VĂN THẠC SỸ KĨ THUẬT

HÀ NỘI – NĂM 2012

Luận văn được hoàn thành tại:

HỌC VIỆN CÔNG NGHỆ BƯU CHÍNH VIỄN THÔNG

Người hướng dẫn khoa học: TS. NGUYỄN MẠNH HÙNG

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

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

Công nghệ Logic mờ được giáo sư Lotfi Zadeh công bố lần đầu tiên tại Mỹ vào năm 1965. Sự bùng nổ của thời đại thông tin như hiện nay, lượng thông tin được tạo ra hàng ngày là rất lớn. Nhu cầu cần thiết đến các quá trình tự động tìm kiếm thông tin hữu ích, các quan hệ phát hiện các tri thức. Để làm được điều đó các nhà nghiên cứu đã đề xuất và nghiên cứu lĩnh vực này như phân lớp và nhận dạng mẫu, hồi quy và dự báo, phân cụm… dựa trên tập mờ.

Lý thuyết tập mờ được coi là nền tảng của lập luận xấp xỉ, nhưng lý thuyết tập mờ vẫn chưa mô phỏng đầy đủ, hoàn chỉnh cấu trúc ngôn ngữ mà con người vẫn sử dụng. Vì thế năm 1990 N.C.Ho & W.Wechler đã khởi xướng phương pháp tiếp cận đại số dựa trên miền giá trị của biến ngôn ngữ.

Với ý nghĩa như vậy mục tiêu của luận văn đặt ra cụ thể như sau:

- Trình bày về tập mờ, logic mờ
- Trình bày thuật toán FCM
- Trình bày về Đại số gia tử
- Ứng dụng đại số gia tử
- Giải thuật di truyền để tối ưu bộ số gia tử

Về bố cục luận văn được chia làm 4 chương:

Chương 1: Trình bày các vấn đề về logic mờ và bài toán phân cụm. Trong đó sẽ đi tìm hiểu giải thuật Fuzzy C-Means, so sánh với K-Means để thấy được ưu/nhược điểm của thuật toán.

Chương 2: Trong chương này sẽ trình bày về đại số gia tử, tìm hiểu cấu trúc, định lý, tính mờ của một ngôn ngữ. Sử dụng đại số gia tử sẽ sửa đổi khoảng cách từ mẫu tới tâm cụm, đo độ mờ của giá trị ngôn ngữ

Chương 3: Là chương phân tích thiết kế và cài đặt thử nghiệm. Bộ hoa Iris là tập dữ liệu đầu vào, qua chương trình sẽ đánh giá tính hiệu năng của thuật toán, thấy được tỉ lệ nhận dạng đúng khi phân loại bộ hoa Iris.

Chương 4: Đánh giá kết quả và cài đặt tối ưu. Để có được tỉ lệ nhận dạng cao, sử dụng giải thuật di truyền để tối ưu bộ số gia tử.

II. NỘI DUNG

Chương 1: LOGIC MỜ VÀ BÀI TOÁN PHÂN CỤM

Thực tế cho thấy khái niệm mờ luôn luôn luôn tồn tại, ứng dụng trong các bài toán và ngay cả trong cách thức suy luận của con người. Bằng các phương pháp tiếp cận khác nhau các nhà nghiên cứu đã đưa ra kết quả về lý thuyết cũng như ứng dụng trong các bài toán điều khiển mờ, hệ hỗ trợ quyết định… Vậy để làm được những điều đó luận văn sẽ đi trình bày những ngữ nghĩa của thông tin mờ, tìm cách biểu diễn chúng bằng khái niệm toán học là tập mờ và xét bài toán phân cụm.

1.1. Logic mờ

1.1.1. Lý thuyết tập mờ

Lý thuyết tập mờ lần đầu tiên được Lotfi.A.Zadeh, một giáo sư thuộc trường Đại học Caliornia, Berkley giới thiệu trong một công trình nghiên cứu vào năm 1965. Lý thuyết tập mờ bao gồm logic mờ, số học mờ, quy hoạch toán học mờ, hình học tôpô mờ, lý thuyết đồ thị mờ, và phân tích dữ liệu mờ, mặc dù thuật ngữ logic mờ thường được dùng chung cho tất cả.

Không giống như tập rõ mà ta đã biết trước đây, mỗi phần tử luôn xác định hoặc thuộc hoặc không thuộc nó, thì với tập mờ chỉ có thể xác định một phần tử liệu thuộc vào nó là nhiều hay ít, tức mỗi một đối tượng chỉ là phần tử của tập mờ với một khả năng nhất định mà thôi.

Trọng tâm của lý thuyết tập mờ là việc đề xuất khái niệm tập mờ (fuzzy sets). Về mặt toán học, một tập mờ A là một hàm số (gọi là hàm thuộc (membership function)) xác định trên khoảng giá trị số mà đối số x có thể chấp nhận (gọi là tập vũ trụ (universe of discourse)) X, cho bởi:

$$\mu_A(x) : X \rightarrow [0.0;1.0]$$

Trong đó, A là nhãn mờ của biến x, thường mang một ý nghĩa ngôn ngữ nào đó, mô tả định tính thuộc tính của đối tượng, chẳng hạn như cao, thấp, nóng, lạnh, sáng, tối …

Một khái niệm cơ bản khác được đưa ra - biến ngôn ngữ (linguistic variables). Biến ngôn ngữ là biến nhận các giá trị ngôn ngữ (linguistic terms) chẳng hạn như

"già ", " trẻ " và "trung niên ", trong đó, mỗi giá trị ngôn ngữ thực chất là một tập mờ xác định bởi một hàm thuộc và khoảng giá trị số tương ứng, chẳng hạn giá trị ngôn ngữ "trung niên" là một tập mờ có hàm thuộc dạng hình tam giác cân xác định trong khoảng độ tuổi [25 , 55]. Logic mờ cho phép các tập này có thể xếp phủ lên nhau (chẳng hạn, một người ở tuổi 50 có thể trực thuộc cả tập mờ " trung niên ” lẫn tập mờ " già ", với mức độ trực thuộc với mỗi tập là khác nhau).

1.1.2. Logic mờ

Trong logic rõ thì mệnh đề là một câu phát biểu đúng, sai. Trong logic mờ thì mỗi mệnh đề mờ là một câu phát biểu không nhất thiết là đúng hoặc sai. Mệnh đề mờ được gán cho một giá trị trong khoảng từ 0 đến 1 để chỉ mức độ đúng (độ thuộc) của nó.

Các phép toán mệnh đề trong logic mờ được định nghĩa nhưsau:

- Phép phủ định : v(Pphủ định) = 1 - v(P).
- Phép tuyển :v(P1∨P2) = max(v(P1), v(P2)).
- Phép hội :v(P1∧P2) = min(v(P1), v(P2))
- Phép kéo theo: v(P→Q) = v(Pphủ định∨Q) = max(v(Pphủ định), v(Q))

Xét cho cùng, tập mờ là một công cụ toán học cho phép chuyển đổi từ giá trị định lượng sang giá trị định tính

Như vậy có thể nói, sự ra đời của lý thuyết tập mờ đã mở ra một nhánh quan trọng trong việc biểu diễn tri thức và ý nghĩ của con người. Đây chính là công cụ toán học và logic để tiến hành xây dựng ứng dụng phân cụm mờ sẽ được cụ thể hóa trong các chương tiếp theo.

1.2. Bài toán phân cụm mờ

Bài toán phân cụm mờ được ứng dụng rất nhiều như trong việc nhận dạng mẫu (vân tay, ảnh), xử lí ảnh, y học (phân loại bệnh lí, triệu chứng)…

Tuy nhiên với giải thuật thứ 2, tức là sử dụng logic mờ để phân cụm dữ liệu mềm dẻo hơn rất nhiều (so với giải thuật K-means). Nó cho phép một đối tượng có thể thuộc vào một hay nhiều phân vùng khác nhau được biểu diễn thông qua khái niệm hàm thuộc hay mức độ thuộc.

1.2.1. Phân cụm rõ

Phương pháp đơn giản và dễ hiểu này vẫn được dùng khá phổ biến trong nhiều ứng dụng. Với giải thuật này, việc phân cụm sẽ được thực hiện qua 2 bước:

- Tính toán tâm cụm
- Sắp xếp lại các đối tượng sao cho gần với tâm vùng nhất.

1.2.2. Phân cụm mờ

Tập các đối tượng sẽ được phân vùng

X={x1,…,xN} ; (k=1,2,…,N)

Việc đánh giá quan hệ không đồng dạng trong 1 không gian cho trước thường sử dụng nhiều đến khái niệm metric, metric giữa 2 đối tượng x,y là m(x,y) cần thỏa mãn:

Khái niệm gần gũi chúng ta nhất là khoảng cách Euclid:

$$D2(x,y) = \sqrt{\sum_{j=1}^{p}(x^j - y^j)^2} = \|x - y\|_2$$

Với những ứng dụng xây dựng trong không gian Euclid, hàm quan hệ đánh giá mức độ không đồng dạng D(X,Y) chúng ta dùng (được mô tả dưới đây) được xác định bằng bình phương khoảng cách Euclid :

$$D(x,y) = d^2_2(x,y) = \|x - y\|_2^2 = \sum_{j=1}^{p}(x^j - y^j)^2$$

Tiến hành phân chia X={x1,…, xN} vào c phân vùng Gi (i=1,2,…,c). Trong mỗi vùng, giá trị tâm vùng là xác định

Thuật toán có thể được mô tả như sau :

- Bước 1: tạo ngẫu nhiên c phân vùng với c tâm vùng Vi tương ứng
- Bước 2: sắp xếp các đối tượng sao cho gần tâm vùng nhất, điều này có nghĩa là:

$$x_k \in G_i \qquad D(x_k, v_i) = \min_{1\le j\le c} D(x_k, v_j)$$

- Bước 3: Tính toán lại tâm vùng:

$$v_i = \frac{1}{G_i} \sum_{x_k \in G_i} x_k$$

- Bước 4: Dừng nếu vùng hội tụ, quay lại bước 2 trong trường hợp khác

Như vậy với việc đưa vào G,V và hàm mục tiêu J, ta có thể mô tả lại việc xác định tâm vùng và gom cụm như sau:

- Bước 2 : Tối thiểu hàm J với G trong khi V được cố định
- Bước 3 : Tối thiểu J với V trong khi G được cố định

Bằng việc xây dựng ma trận U (NxC)

$$U = (U_{ki})$$

$$U_{ki} = \begin{matrix} 1 & x_k \in G_i \\ 0 & (x_k \notin G_i) \end{matrix},$$

Trong đó N là số đối tượng, C là số phân vùng, chúng ta viết lại hàm mục tiêu J như sau:

$$J_0(U,V) = \sum_{i=1}^{c} \sum_{k=1}^{N} u_{ki} D(x_k, v_i)$$

Nhược điểm lớn nhất của Fuzzy C- Means là việc xử lí gặp khó khăn khi tập dữ liệu lớn, tập dữ liệu nhiều chiều, nhạy cảm đối với nhiễu và phần tử ngoại lai trong dữ liệu, tức là các trung tâm cụm có thể sẽ nằm xa so với trung tâm thực của cụm. Để giải quyết vấn đề này, đã có nhiều phương pháp được đề xuất như phân cụm dựa trên xác suất (Keller, 1993), phân cụm nhiễu mờ (Dave, 1991), thuật toán Є – Intensitive Fuzzy C- Means và FCM cải tiến.

1.3. Kết luận chương 1

Như vậy qua chương 1 luận văn đã trình bày cơ sở lý thuyết về logic mờ cũng như khái niệm ban đầu về giải thuật phân cụm.Trong chương tiếp theo luận văn sẽ đề cập tới lý thuyết đại số gia tử và áp dụng lý thuyết này vào bài toán phân cụm dữ liệu.