NGHIÊN CỨU MÔ HÌNH MẠNG NƠRON KOHONEN VÀ ỨNG DỤNG TRONG BÀI TOÁN PHÂN CỤM DỮ LIỆU

TỔNG QUAN

Phân cụm dữ liệu (PCDL) là quá trình nhóm một tập các đối tượng tương tự nhau trong tập dữ liệu vào các cụm sao cho các đối tượng thuộc cùng một cụm là tương đồng còn các đối tượng thuộc các cụm khác nhau sẽ không tương đồng. Phân cụm dữ liệu là một ví dụ của phương pháp học không có thầy. Không giống như phân lớp dữ liệu, phân cụm dữ liệu không đòi hỏi phải định nghĩa trước các mẫu dữ liệu huấn luyện. Vì thế, có thể coi phân cụm dữ liệu là một cách học bằng quan sát, trong khi phân lớp dữ liệu là học bằng ví dụ…

Hiện nay, các phương pháp phân cụm trên đã và đang được phát triển [6] và áp dụng nhiều trong các lĩnh vực khác nhau và đã có một số nhánh nghiên cứu được phát triển trên cơ sở của các phương pháp đó như: Phân cụm thống kê, phân cụm khái niệm, phân cụm mờ.

Phân cụm sử dụng mạng Kohonen SOM (Self-Organizing Maps): Loại phân cụm này dựa trên khái niệm của các mạng nơron. Mạng SOM có tầng nơron vào và các tầng nơron ra. Mỗi nơron của tầng vào tương ứng với mỗi thuộc tính của bản ghi, mỗi một nơron vào kết nối với tất cả các nơron của tầng ra. Mỗi liên kết được gắn liền với một trọng số nhằm xác định vị trí của nơron ra tương ứng.

Trong số này, SOM là một giải thuật được phát triển bởi Kohonen, nó có thể được áp dụng cho nhiều lớp bài toán khác nhau. Giaì thuật SOM ban đâù được phat triển cho mục đích phân loại tiếng noi, tuy nhiên SOM coǹ cơ thể ap dụng được trong nhiêù lĩ nh vực khac như điêù khiển tự động (Control Engineering) [14], nhận dạng tiếng noi (Kohonen, 1989, robotics (Ritter et al., 1989), maý ao (Oja, 1992), tôi ưu tổ hợp (Fort, 1988), phân lớp (Kohonen, 1984), hoá - sinh trăc học (Biomedical Sciences and Chemistry), phân tích taì chính (Financial Analysis) [16][20] va` xử lý ngôn ngữ tự nhiên (Natural Language Processing)

Tại Việt Nam, học phần mạng noron cũng đã được đưa vào chương trình đào tạo hệ đại học và sau đại học. Hiện nay cũng đã có rất nhiều bài báo, công trình nghiên cứu về mạng noron nhân tạo, trong số đó có một vài bài có đề cập đến loại mạng SOM do Kohonen đề xuất và các ứng dụng trong thực tế của loại mạng này.

Bài báo của tác giả Đỗ Phúc [18] đề cập đến vấn đề gom cụm đồ thị và ứng dụng vào việc rút trích nội dung chính của khối thông điệp trên diễn đàn thảo luận, tác giả đã biểu diễn các thông điệp bằng đồ thị, và chọn giải pháp gom cụm đồ thị bằng mạng Kohonen. Ngoài ra, mạng Kohonen cũng có thể được áp dụng để giải quyết các bài toán về dự báo [19] như: dự báo thông tin thị trường, thời thiết, phụ tải điện,…

MỤC TIÊU

- Đưa ra mô hình mạng nơ ron Kohonen cải tiến (về thuật học hoặc cấu trúc) đảm bảo hiệu quả và tốc độ hội tụ của mạng, ứng dụng trong bài toán phân cụm dữ liệu tự động.

NỘI DUNG

- Một số về vấn đề và thuật toán cơ bản ứng dụng cho phân cụm dữ liệu hiện nay

Một số thuật toán phân cụm rõ

Một số thuật toán phân cụm mờ

Đánh giá ưu nhược điểm của các thuật toán

- Vấn đề phân cụm dữ liệu sử dụng mạng nơron Kohonen

Mô hình mạng nơron Kohonen
Đề xuất một số giải pháp cải tiến mạng kohonen về thuật học hoặc cấu trúc
Đánh giá hiệu quả của giải pháp đề xuất so với mô hình cũ.

- Xây dựng chương trình phần mềm thử nghiệm

Chạy thử trên tập dữ liệu thực tế (dự kiến sử dụng tập dữ liệu điểm của sinh viên Khoa CNTT, hoặc phân vùng ảnh,...)
Phân tích kết quả, đánh giá

PHƯƠNG PHÁP NGHIÊN CỨU

- Nghiên cứu lý thuyết và xây dựng mô hình ứng dụng cho bài toán thực tế
- Thu thập số liệu thực tế để thử nghiệm trên mô hình
- Xây dựng chương trình thử nghiệm

HIỆU QUẢ KTXH

- Bổ sung nguồn tài liệu nghiên cứu về lĩnh vực khai phá dữ liệu và tính toán thông minh cho chương trình đào tạo Công nghệ thông tin tại Khoa CNTT
- Có thể sử dụng phần mềm đã cài đặt để thực hiện việc phân cụm trên nhiều các dữ liệu của nhiều bài toán thực tế, đặc biệt đối với những bài toán cần xem xét trên các tập dữ liệu lớn. Qua đó giúp người dùng thu nhận được các đặc trưng trên tập dữ liệu cần phân tích để từ đó có những biện pháp giải quyết phù hợp mà ít tốn kém về thời gian, kinh phí.

Ví dụ có thể phân cụm dựa trên tập dữ liệu điểm của sinh viên khoa CNTT. Điều này có ý nghĩa quan trọng trong công tác quản lý, giúp đơn vị đào tạo có thể thấy được toàn cảnh về các môn học trong chương trình, từ đó có sự đánh giá đúng đắn về tính chất môn học và khả năng nhận thức của người học, đưa ra biện pháp tác động phù hợp để cải thiện chất lượng dạy và học. Ngoài ra, việc phân cụm cũng có thể giúp thấy được khả năng học tập của mỗi sinh viên với từng môn học, từ đó có biện pháp áp dụng phương pháp dạy học cá biệt hóa thích hợp (phương pháp dạy học đặc trưng của học chế tín chỉ), dựa trên kết quả phân cụm này cũng có thể giúp người học có có định hướng phù hợp cho quá trình học tập tiếp theo.

Như vậy, nội dung nghiên cứu của đề tài là giải quyết bước đầu trong vấn đề về khai phá dữ liệu, có vài trò cũng như đóng góp rất quan trọng trong việc nâng cao hiệu quả cho các bước phân tích dữ liệu phía sau.

ĐƠN VỊ SỬ DỤNG