

ĐẠI HỌC THÁI NGUYÊN
TRƯỜNG ĐH CNTT VÀ TRUYỀN THÔNG

HÀ THỊ THƯ

MẠNG KOHONEN-SOM VÀ ỨNG DỤNG
PHÂN CỤM ĐIỂM HỌC SINH THPT

CHUYÊN NGÀNH: KHOA HỌC MÁY TÍNH

MÃ SỐ: 60 48 01

LUẬN VĂN THẠC SĨ KHOA HỌC

HƯỚNG DẪN KHOA HỌC: TS NGUYỄN VĂN TẢO

THÁI NGUYÊN - 2012

ĐẠI HỌC THÁI NGUYÊN
TRƯỜNG ĐH CNTT VÀ TRUYỀN THÔNG

HÀ THỊ THƯ

**MẠNG KOHONEN-SOM VÀ ỨNG DỤNG
PHÂN CỤM ĐIỂM HỌC SINH THPT**

CHUYÊN NGÀNH: KHOA HỌC MÁY TÍNH

MÃ SỐ: 60 48 01

LUẬN VĂN THẠC SĨ KHOA HỌC

HƯỚNG DẪN KHOA HỌC

TS NGUYỄN VĂN TẢO

THÁI NGUYÊN - 2012

LỜI CAM ĐOAN

Tôi xin cam đoan luận văn là công trình nghiên cứu của cá nhân tôi, không sao chép của ai. Nội dung lý thuyết trong luận văn có sự tham khảo và sử dụng của một số tài liệu, thông tin được đăng tải trên các tác phẩm, tạp chí và các trang web theo danh mục tài liệu của luận văn. Các số liệu, chương trình phần mềm và những kết quả trong luận văn là trung thực và chưa được công bố trong bất kỳ một công trình nào khác.

Thái Nguyên, ngày 15 tháng 9 năm 2012

Học viên thực hiện

HÀ THỊ THU'

MỤC LỤC

CHƯƠNG 1: TỔNG QUAN VỀ MẠNG KOHONEN-SOM

| | |
|---|-----------|
| 1.1. Sơ lược về mạng neural | 2 |
| 1.1.1. Lịch sử phát triển | 2 |
| 1.1.2. Ứng dụng..... | 3 |
| 1.1.3. Căn nguyên sinh học | 4 |
| 1.2. Tổng quan về mạng Kohonen-SOM | 5 |
| 1.2.1. Vecto Quantization – VQ..... | 6 |
| 1.2.2. Learning Vector Quantization – LVQ | 6 |
| 1.2.3. Bản đồ tự tổ chức – SOM..... | 7 |
| 1.3. Cấu trúc mạng neural Kohonen | 7 |
| 1.3.1. Mạng neural Kohonen..... | 7 |
| 1.3.2. Cấu trúc của mạng neural Kohonen | 8 |
| 1.4. Thực thi mạng neural Kohonen | 12 |
| 1.5. Kết luận | 13 |
| 2.1. Phân cụm dữ liệu: | 14 |
| 2.1.1. Khái niệm: | 14 |
| 2.1.2. Các bước cơ bản trong phân cụm: | 15 |
| 2.2. Những kỹ thuật tiếp cận trong phân cụm dữ liệu | 15 |
| 2.2.1. Phương pháp phân cụm phân hoạch | 15 |
| 2.2.2. Phương pháp phân cụm phân cấp | 16 |
| 2.2.3. Phương pháp phân cụm dựa trên mật độ..... | 16 |
| 2.2.4. Phương pháp phân cụm dựa trên lưới | 17 |
| 2.2.5. Phương pháp phân cụm dựa trên mô hình..... | 18 |
| 2.2.6. Phương pháp phân cụm có dữ liệu ràng buộc | 18 |
| 2.3. Một số thuật toán cơ bản trong phân cụm dữ liệu | 19 |
| 2.3.1. Thuật toán phân cụm phân hoạch | 19 |
| 2.3.2. Thuật toán phân cụm phân cấp | 21 |
| 2.3.3. Thuật toán phân cụm dựa trên mật độ..... | 23 |

| | |
|--|-----------|
| 2.3.4. Thuật toán phân cụm dựa trên lưới | 24 |
| 2.3.5. Các thuật toán phân cụm dựa trên mô hình..... | 25 |
| 2.4. Dùng mạng neural trong phân cụm: | 26 |
| 2.5. SOM –Bài toán phân cụm: | 28 |
| 2.5.1. Thuật toán SOM:..... | 28 |
| 2.5.2. Sử dụng SOM trong khai phá dữ liệu..... | 31 |
| 2.5.3. SOM –Thách thức phân nhóm | 36 |
| 2.5.4. SOM –Thách thức tốc độ giải thuật | 37 |
| 2.5.5. SOM –Vấn đề số lượng nhóm..... | 38 |
| 2.6. SOM –Giải quyết những vấn đề tồn tại khi phân cụm..... | 38 |
| 2.6.1 Giải pháp phân cụm tự nhiên trong quá trình học..... | 38 |
| 2.6.2 Mạng thu gọn và tăng tốc giải thuật..... | 40 |
| 2.6.3 Giới hạn điều chỉnh số lượng nhóm trên mạng SOM | 42 |
| 2.7. Đánh giá kết quả phân cụm..... | 43 |
| 2.8. Kết luận : | 46 |
| 3.1. Mô tả bài toán..... | 48 |
| 3.1.1. Dữ liệu vào:..... | 48 |
| 3.1.2. Dữ liệu ra: | 48 |
| 3.1.3.Quá trình cài đặt:..... | 48 |
| 3.1.4. Mục đích- Yêu cầu..... | 48 |
| 3.2. Phân tích thiết kế hệ thống cho ứng dụng..... | 49 |
| 3.2.1. Xác định các tác nhân và các Use case | 49 |
| 3.2.2. Biểu đồ Use Case | 50 |
| 3.3. Chương trình ứng dụng..... | 54 |
| 3.3.1. Giao diện tổng quan cho ứng dụng | 54 |
| 3.3.2 Một số tính năng cho ứng dụng..... | 54 |
| 3.3.2 Hướng dẫn sử dụng, chạy thử nghiệm | 59 |
| 3.4. Kết luận..... | 62 |

DANH MỤC VIẾT TẮT

| | |
|------|------------------|
| CSDL | Cơ sở dữ liệu |
| KPDL | Khai phá dữ liệu |
| FCM | Fuzzy C-Means |
| PCDL | Phân cụm dữ liệu |

DANH MỤC CÁC HÌNH

| | | |
|-----------|--|----|
| Hình 1.1 | Cấu tạo tế bào neural..... | 4 |
| Hình 1.2 | Giáo sư Kohonen và mạng neural..... | 6 |
| Hình 1.3 | Cấu trúc mạng Kohonen..... | 8 |
| Hình 1.4 | Những bức tường trong Kohonen..... | 12 |
| Hình 2.1 | Các chiến lược phân cụm phân phân cấp..... | 16 |
| Hình 2.2 | Cấu trúc phân cấp..... | 17 |
| Hình 2.3 | Các cách mà các cụm có thể đưa ra..... | 18 |
| Hình 2.4 | Các thiết lập để xác định ranh giới các cụm ban đầu..... | 20 |
| Hình 2.5 | Tính toán trọng tâm của các cụm mới..... | 20 |
| Hình 2.6 | Khái quát thuật toán CURE..... | 22 |
| Hình 2.7 | Các cụm dữ liệu được khám phá bởi CURE..... | 22 |
| Hình 2.8 | Hình dạng các cụm được khám phá bởi thuật toán DBSCAN..... | 23 |
| Hình 2.9 | Đơn vị xử lý ganh đua SOM..... | 28 |
| Hình 2.10 | Không gian ban đầu và SOM..... | 29 |
| Hình 2.11 | BMU..... | 31 |
| Hình 2.12 | Vecto chiến thắng liên tục đối với SOM | 35 |
| Hình 2.13 | Định nghĩa một U-Matrix..... | 36 |
| Hình 2.14 | U-Matrix của SOM..... | 36 |
| Hình 2.15 | Quan hệ giữa 2 cụm..... | 44 |
| Hình 3.1 | Mô hình Use Case tổng thể của bài toán..... | 47 |
| Hình 3.2 | Biểu đồ trình tự chọn CSDL..... | 47 |
| Hình 3.3 | Biểu đồ trình tự tạo và huấn luyện mạng neural..... | 48 |
| Hình 3.4 | Biểu đồ trình tự tạo biểu đồ theo nhóm..... | 48 |
| Hình 3.5 | Biểu đồ trình tự phân cụm học sinh theo nhóm..... | 49 |
| Hình 3.6 | Biểu đồ trình tự cấu hình mạng neural..... | 49 |

| | | |
|-----------|--|----|
| Hình 3.7 | Biểu đồ trình tự chọn loại biểu đồ..... | 50 |
| Hình 3.8 | Biểu đồ trình tự chọn và hiển thị biểu đồ..... | 50 |
| Hình 3.9 | Giao diện chương trình..... | 51 |
| Hình 3.10 | Tab hỗ trợ nhập và chọn dữ liệu..... | 52 |
| Hình 3.11 | Một số biểu đồ trợ giúp dạng 3D..... | 53 |
| Hình 3.12 | Một số biểu đồ trợ giúp dạng 2D..... | 54 |
| Hình 3.13 | Nhóm và cấu hình nhóm..... | 55 |
| Hình 3.14 | Một phần dữ liệu và khả năng phân nhóm..... | 56 |
| Hình 3.15 | Các phần của ứng dụng..... | 57 |

MỞ ĐẦU

Sự phát triển mạnh mẽ của Công nghệ nói chung và Công nghệ thông tin nói riêng đã tạo nên nhiều hệ thống thông tin phục vụ việc tự động hoá mọi hoạt động kinh doanh cũng như quản lý trong xã hội. Nhiều hệ quản trị cơ sở dữ liệu mạnh với các công cụ phong phú và thuận tiện đã giúp con người khai thác có hiệu quả các nguồn tài nguyên dữ liệu lớn. Trong đó, khai phá dữ liệu (Data Mining) là quá trình chính trong phát hiện tri thức. Sử dụng các kỹ thuật và các khái niệm của các lĩnh vực đã được nghiên cứu từ trước như học máy, nhận dạng, thống kê, hồi quy, xếp loại, phân nhóm, đồ thị, mạng nơron, mạng Bayes,... được sử dụng để khai phá dữ liệu nhằm phát hiện ra các mẫu mới, tương quan mới, các xu hướng có ý nghĩa.

Luận văn với đề tài **“Mạng Kohonen-SOM và ứng dụng phân cụm điểm học sinh THPT”** khảo sát lĩnh vực KPDL dùng mạng nơron. Luận văn tập trung vào phương pháp học mạng nơron không có giám sát, dùng thuật toán SOM để giải quyết bài toán phân cụm theo mô hình mạng nơron.

Luận văn được thực hiện dưới sự hướng dẫn khoa học của TS. Nguyễn Văn Tảo. Tôi xin chân thành cảm ơn sâu sắc tới Thầy đã chỉ dẫn tận tình giúp tôi có thể hoàn thành bản luận văn này. Tôi cũng vô cùng cảm ơn sự giúp đỡ và động viên khích lệ của người thân trong gia đình tôi, bạn bè và các đồng nghiệp trong trường THPT Ngô Sĩ Liên trong suốt quá trình thực hiện luận văn.

Bắc Giang, ngày 15 tháng 9 năm 2012

Hà Thị Thu

CHƯƠNG 1

TỔNG QUAN VỀ MẠNG KOHONEN-SOM

Chương này đề cập các vấn đề sau:

- 1.1. Sơ lược về mạng neural*
 - 1.2. Tổng quan về mạng KOHONEN-SOM*
 - 1.3. Cấu trúc của mạng neural KOHONEN*
 - 1.4. Thực thi của mạng neural KOHONEN*
 - 1.5. Kết luận*
-

1.1. Sơ lược về mạng neural

1.1.1. Lịch sử phát triển

Mạng neural nhân tạo được xây dựng từ những năm 1940 nhằm mô phỏng một số chức năng của bộ não người. Dựa trên quan điểm cho rằng bộ não người là bộ điều khiển. Mạng neural nhân tạo được thiết kế tương tự như neural sinh học sẽ có khả năng giải quyết hàng loạt các bài toán như tính toán tối ưu, điều khiển, công nghệ robot...

Dưới đây là các mốc đáng chú ý trong lịch sử phát triển của mạng neural:

- *Giai đoạn 1:* Có thể tính từ nghiên cứu của William (1890) về tâm lý học với sự liên kết các neural thần kinh. Năm 1940 Mc Culloch và Pitts đã cho biết neural có thể mô hình hoá như thiết bị ngưỡng (Giới hạn) để thực hiện các phép tính logic và mô hình mạng neural của Mc Culloch – Pitts cùng với giải thuật huấn luyện mạng của Hebb ra đời năm 1943.

- *Giai đoạn 2:* vào khoảng gần những năm 1960, một số mô hình neural hoàn thiện hơn đã được đưa ra như: Mô hình Perceptron của Rosenblatt (1958), Adalile của Widrow (1962). Trong đó mô hình Perceptron rất được quan tâm vì nguyên lý đơn giản, nhưng nó cũng có hạn chế vì như Marvin Minsky và Seymour papert của MIT (Massachurehs Insritute of Technology) đã chứng minh nó không dùng được cho các hàm logic phức (1969). Còn Adaline là mô hình tuyến tính, tự