ĐẠI HỌC THÁI NGUYÊN
TRƯỜNG ĐẠI HỌC CNTT & TRUYỀN THÔNG

---------- ☙ ❡ ❧ ----------

Vũ Thị Hiền

PHÂN LỚP CÁC MẪU VỚI ỨNG DỤNG CỦA MẠNG NƠRON NHÂN TẠO

LUẬN VĂN THẠC SĨ KHOA HỌC MÁY TÍNH

Thái Nguyên - 2012

MỞ ĐẦU

Phân lớp các mẫu là một vấn đề thú vị và bổ ích. Đây là vấn đề rất hay gặp trong cuộc sống.

Các nhà băng cần phân lớp khách hàng theo các mức rủi do để trong từng hoàn cảnh cụ thể, với những khách hàng cụ thể nhân viên nhà băng quyết định có cho vay hay không hoặc cho vay nhiều hay cho vay ít.

Các nhà quản lý cần xếp cán bộ, nhân viên dưới quyền vào các lớp để qui hoạch xây dựng đội ngũ. Mỗi cán bộ, nhân viên với những giá trị của những thông số khác nhau cần phải được đưa vào những lớp khác nhau.

Các nhà giáo dục cần phân lớp các em học sinh theo những tiêu chí khác nhau. Khi biết đối tượng dạy học của mình thuộc lớp nào ta sẽ có những phương pháp giáo dục thích hợp.

Các trường chuyên nghiệp cần phân lớp các học sinh theo các tham số khác nhau tương ứng với những mức học bổng khác nhau.

Với những kho dữ liệu khổng lồ, phân lớp là thao tác giúp ta khai phá dữ liệu, tìm kiếm tri thức được nhanh chóng và hiệu quả hơn.

Các đối tượng cần phân lớp thường được biểu diễn bởi một vectơ, trong đó mỗi thuộc tính có thể có những thứ nguyên khác nhau vì thế việc phân lớp rất khó khăn. Ví dụ cần phân lớp cán bộ theo các tiêu chí tài và đức. Thật khó đánh giá xem ai ở lớp trên, ai ở lớp dưới khi giá trị trung bình của hai tiêu chí này của họ là xấp xỉ như nhau. Nếu việc phân lớp không chính xác, cũng giống như đánh giá không chính xác tất yếu sẽ dẫn đến những hậu quả tai hại.

Đã có nhiều người quan tâm đến vấn đề phân lớp. Nhưng các phương pháp đã có thường chịu ảnh hưởng nhiều của cảm tính, hoặc chịu ảnh hưởng nhiều của yếu tố tâm lý, của chủ thể phân lớp. Ngay cả với những phương pháp toán học, do ranh giới phân lớp nhiều trường hợp không phải là tuyến tính nên bài toán phân lớp thường có độ phức tạp tính toán lớn và độ chính xác không cao.

Mạng nơron nhân tạo là mô hình tính toán mô phỏng hoạt động của não người. Do có tính mềm dẻo, linh hoạt và khả năng dung thứ lỗi, mạng nơron có thể xấp xỉ mọi hàm với độ chính xác cao nên việc phân lớp bằng mạng nơron sẽ đưa đến kết quả không những khách quan mà còn đảm bảo kết quả tốt.

Về mặt lý thuyết, phân lớp nhờ mạng nơron nhân tạo đã được nghiên cứu và khẳng định là một khả năng tiềm tàng của mô hình tính toán này. Những khảo nghiệm sâu về các giải thuật với mạng phân lớp và nghiên cứu thử nghiệm mô hình này còn chưa nhiều. Vì thế, trong khuôn khổ của một luận văn thạc sĩ tôi chọn đề tài: "**Phân lớp các mẫu với ứng dụng của mạng nơron nhân tạo**" nhằm tìm hiểu thêm về mạng nơron nhân tạo với việc phân lớp các con số. Từ đó rút ra những kết luận cần thiết cho việc xây dựng các ứng dụng cụ thể về sau.

Em xin cảm ơn sự giúp đỡ nhiệt tình của các thầy, đặc biệt là TS. Nguyễn Tân Ân.

Luận văn không thể tránh khỏi sai xót, em mong nhận được sự đóng góp ý kiến của các thầy và các bạn để luận văn được hoàn thiện hơn.

Chương 1 *Bài toán phân lớp*

1.1 Khái niệm phân lớp

1.1.1 Khái niệm phân lớp

Phân lớp dữ liệu là kĩ thuật dựa trên tập huấn luyện và những giá trị hay nhãn của lớp trong một thuộc tính phân lớp và sử dụng nó trong việc phân lớp dữ liệu mới. Phân lớp cũng là tiên đoán loại lớp của nhãn.

1.1.2 Bài toán phân lớp

- **Mục đích:** để dự đoán những nhãn phân lớp cho các bộ dữ liệu/mẫu mới
- **Đầu vào:** một tập các mẫu dữ liệu huấn luyện, với một nhãn phân lớp cho mỗi mẫu dữ liệu
- **Đầu ra:** mô hình (bộ phân lớp) dựa trên tập huấn luyện và những nhãn phân lớp

1.2 Một số ứng dụng phân lớp tiêu biểu

- Tín dụng
- Tiếp thị
- Chẩn đoán y khoa
- Phân tích hiệu quả điều trị

1.3 Quy trình phân lớp

Bao gồm 2 bước: Xây dựng mô hình và sử dụng mô hình.

- Bước 1. Xây dựng mô hình: là mô tả một tập những lớp được định nghĩa trước. Trong đó, mỗi bộ hoặc mẫu được gán thuộc về một lớp được định nghĩa trước như là được xác định bởi thuộc tính nhãn lớp, tập hợp của những bộ được sử dụng trong việc sử dụng mô hình được gọi là tập huấn luyện. Mô hình được biểu diễn là những luật phân lớp, cây quyết định và những công thức toán học.

- Bước 2. Sử dụng mô hình: Việc sử dụng mô hình phục vụ cho mục đích phân lớp dữ liệu trong tương lai hoặc phân lớp cho những đối tượng chưa biết đến. Trước khi sử dụng mô hình người ta thường phải đánh giá tính chính xác của mô hình, trong đó nhãn được biết của mẫu kiểm tra được so sánh với kết quả

phân lớp của mô hình, độ chính xác là phần trăm của tập hợp mẫu kiểm tra mà phân loại đúng bởi mô hình, tập kiểm tra là độc lập với tập huấn luyện.

1.4 Các kỹ thuật phân lớp

1.4.1 Phân lớp bằng phương pháp qui nạp cây quyết định

1.4.1.1 Khái niệm cây quyết định

Cây quyết định là một flow-chart giống cấu trúc cây, nút bên trong biểu thị một kiểm tra trên một thuộc tính, nhánh biểu diễn đầu ra của kiểm tra, nút lá biểu diễn nhãn lớp hoặc sự phân bố của lớp.

Việc tạo cây quyết định bao gồm 2 giai đoạn: Tạo cây và tỉa cây.

Để tạo cây ở thời điểm bắt đầu tất cả những ví dụ huấn luyện ở gốc sau đó phân chia ví dụ huấn luyện theo cách đệ qui dựa trên thuộc tính được chọn.

Việc tỉa cây là xác định và xóa những nhánh mà có phần tử hỗn loạn hoặc những phần tử nằm ngoài (những phần tử không thể phân vào một lớp nào đó).

Việc sử dụng cây quyết định như sau: Kiểm tra những giá trị thuộc tính của mẫu đối với cây quyết định.

1.4.1.2 Thuật toán qui nạp cây quyết định

Giải thuật cơ bản (giải thuật tham lam) được chia thành các bước như sau:

1. Cây được xây dựng đệ qui từ trên xuống dưới (top-down) và theo cách thức chia để trị (divide-conquer).
2. Ở thời điểm bắt đầu, tất cả những ví dụ huấn luyện ở gốc.
3. Thuộc tính được phân loại (nếu là giá trị liên tục chúng được rời rạc hóa)
4. Những ví dụ huấn luyện được phân chia đệ qui dựa trên thuộc tính mà nó chọn lựa.
5. Kiểm tra những thuộc tính được chọn dựa trên nền tảng của heristic hoặc của một định lượng thống kê.

Điều kiện để dừng việc phân chia:

1. Tất cả những mẫu huấn luyện đối với một nút cho trước thuộc về cùng một lớp.

2. Không còn thuộc tính còn lại nào để phân chia tiếp.

3. Không còn lại mẫu nào.

1.4.1.3 Nội dung giải thuật học cây quyết định cơ bản ID3

ID3 là một giải thuật học cây quyết định được phát triển bởi Ross Quinlan (1983). Ý tưởng cơ bản của giải thuật ID3 là để xây dựng cây quyết định bằng việc sử dụng một cách tìm kiếm từ trên xuống trên những tập hợp cho trước để kiểm tra mỗi thuộc tính tại mỗi nút của cây. Để chọn ra thuộc tính hữu ích nhất cho sự phân loại trên những tập hợp cho trước, chúng ta sẽ đưa ra một hệ đo độ lợi thông tin.

Để tìm ra một cách tối ưu để phân loại một tập hợp thông tin, vấn đề đặt ra là chúng ta cần phải làm tối thiểu hóa (chẳng hạn, tối thiểu chiều cao của cây). Như vậy chúng ta cần một số chức năng có thể đánh giá trường hợp nào cho ra một sự phân chia cân bằng nhất. Hệ đo độ lợi thông tin sẽ là hàm như vậy.

1.4.1.4 Những thiếu sót của giải thuật ID3

Trường hợp thiếu sót thứ nhất:

Một thiếu sót quan trọng của ID3 là không gian phân chia hợp lệ tại một nút là cạn kiệt. Một sự phân chia là sự phân hoạch của mỗi trường hợp của không gian mà kết quả đạt được từ việc thử nghiệm tại một nút quyết định ID3 và con cháu của nó cho phép sự kiểm tra tại một thuộc tính đơn và nhánh trong kết quả cho ra từ sự kiểm tra này.

Trường hợp thiếu sót thứ hai:

ID3 dựa rất nhiều vào số lượng của những tập hợp dữ liệu đưa vào. Quản lý sự tạp nhiễu của tập dữ liệu vào vô cùng quan trọng khi chúng ta ứng dụng giải thuật học cây quyết định vào thế giới thực. Ví dụ, khi có sự lẫn tạp trong tập dữ liệu đưa vào hoặc khi số lượng ví dụ đưa vào quá nhỏ để tạo ra một ví dụ điển hình của hàm mục tiêu đúng. ID3 có thể dẫn đến việc tạo quyết định sai.

Có rất nhiều những mở rộng từ giải thuật ID3 cơ bản đã phát triển để áp dụng những luật học cây quyết định vào thế giới thực, như những post-pruning

tree, quản lý những thuộc tính giá trị thực, liên quan đến việc thiếu những thuộc tính, sử dụng những tiêu chuẩn chọn lựa thuộc tính khác hơn thu thập thông tin.

1.4.1.5 Sự phân lớp cây quyết định trong cơ sở dữ liệu lớn

Sự phân lớp là một vấn đề cổ điển được nguyên cứu một cách mở rộng bởi những nhà thống kê và những nhà nguyên cứu máy học. Hướng phát triển hiện nay của việc phân lớp là phân lớp những tập dữ liệu với hàng tỉ những mẫu thử và hàng trăm thuộc tính với tốc độ vừa phải.

Qui nạp cây quyết định được đánh giá cao trong khai phá dữ liệu lớn vì:

- Tốc độ học tương đối nhanh hơn so với những phương pháp phân loại khác.
- Có thể hoán chuyển được thành những luật phân lớp đơn giản và dễ hiểu.
- Có thể sử dụng truy vấn SQL để truy xuất cơ sở dữ liệu.
- Sự chính xác phân lớp có thể so sánh được với những phương pháp khác.

1.4.2 Phương pháp phân lớp Bayesian (Bayesian classifier)

1.4.2.1 Đặc điểm

Lý thuyết Bayesian cung cấp một tiếp cận theo xác xuất để suy diễn. Nó dựa trên giả thuyết rằng số lượng của khuynh hướng bị chi phối bởi phân bố xác suất và quyết định tối ưu có thể được tạo bởi sự suy luận về những xác suất đi liền với dữ liệu được quan sát. Đây là vấn đề quan trọng của máy học bởi vì nó cung cấp một tiếp cận định lượng cho việc xem xét cẩn thận bằng chứng hỗ trợ những giả thuyết thay đổi.

Lý thuyết Bayesian cung cấp giải thuật học cơ bản mà vận dụng những xác suất như là một khung làm việc cho sự phân tích sự hoạt động của những giải thuật mà không thể vận dụng rõ ràng.

Học theo xác suất: Tính xác suất xuất hiện cho giả thuyết, trong số những tiếp cận thực dụng nhất cho các kiểu chắc chắn của những vấn đề học.

Tính tăng dần: mỗi ví dụ huấn luyện có thể gia tăng việc tăng hoặc giảm mà không gian giả thuyết đúng. Kiến thức trước có thể kết hợp với dữ liệu được quan sát.

Tiên đoán xác suất: Tiên đoán nhiều không gian giả thuyết, được đo bởi xác suất của nó.

Tiêu chuẩn: Thậm chí khi phương thức Bayesian khó tính toán, chúng cũng cung cấp một tiêu chuẩn tốt nhất cho việc tạo quyết định.

1.4.2.2 Khó khăn của phương pháp phân lớp Bayesian

Khó khăn thực tế của phương pháp phân lớp Bayesian ở chỗ:

- Đòi hỏi kiến thức khởi tạo của nhiều khả năng có thể xảy ra, và
- Chi phí tính toán đáng kể.

1.4.2.3 Sự độc lập của giả thuyết:

Những giả thuyết độc lập nhau sẽ giúp cho việc tính toán trở nên dễ dàng. Độ lợi phân lớp tốt nhất đạt được rất ít trong thực tế vì những thuộc tính (biến) thường liên quan với nhau.

Để vượt qua những giới hạn này người ta giải quyết bằng 2 cách:

- Dùng mạng Bayesian, đây chính là sự kết hợp của lý luận và quan hệ nhân quả giữa những thuộc tính.
- Cây quyết định mà suy luận trên một thuộc tính ở thời điểm xem xét những thuộc tính quan trọng đầu tiên .

1.4.2.4 Mạng Bayesian Tin cậy (Bayesian belief network) :

Bayesian belief network cho phép một tập con của những biến độc lập theo điều kiện.

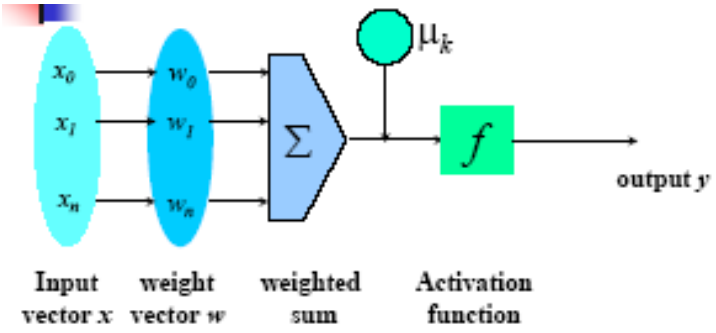
Trong Bayesian belief người ta sử dụng mô hình đồ thị của quan hệ nhân quả. Có nhiều cách học của Bayesian belief networks như sau:

- Cho trước cả cấu trúc mạng và những biến: đây là cách dễ dàng.
- Cho trước cấu trúc mạng nhưng chỉ có một vài biến chứ không phải là tất cả.

- Cấu trúc mạng hoàn toàn không được biết.

1.4.3 Phương pháp phân lớp bằng mạng lan truyền ngược (mạng Nơron)

1.4.3.1 Cấu trúc của một neural như sau:



Hình 1 - *Cấu trúc của một Nơron*

Vector x n chiều được ánh xạ vào biến y dựa trên tích vô hướng và một hàm ánh xạ phi tuyến .

1.4.3.2 Mạng huấn luyện:

a) Mục tiêu cơ bản của việc huấn luyện

Đạt được một tập hợp của những trọng số mà có thể làm cho hầu hết tất cả những bộ trong tập huấn luyện được phân lớp đúng.

b) Những bước của quá trình huấn luyện

- Khởi tạo trọng số với những giá trị ngẫu nhiên.
- Lần lượt đưa mỗi bộ vào trong mạng.
- Đối với mỗi đơn vị:

- Tính toán mạng input cho mỗi đơn vị như một sự kết hợp tuyến tính của tất cả những input đối với đơn vị.

- Tính toán giá trị output sử dụng hàm kích hoạt.

- Tính toán lỗi.

- Cập nhật trọng số và khuynh hướng.

1.4.3.3 Mạng thu giảm và rút trích luật

Mạng thu giảm: