BỘ GIÁO DỤC VÀ ĐÀO TẠO

ĐẠI HỌC ĐÀ NẴNG

THÁI THỊ BÍCH THỦY

ỨNG DỤNG MẠNG NƠRON TRUYỀN THẲNG PHÂN TÍCH NHẬT KÝ MOODLE DỰ BÁO KẾT QUẢ HỌC TẬP TRỰC TUYẾN

Chuyên ngành: Khoa học Máy tính

Mã số: 60.48.01

TÓM TẮT LUẬN VĂN THẠC SĨ KỸ THUẬT

Đà Nẵng - Năm 2011

Công trình được hoàn thành tại

ĐẠI HỌC ĐÀ NẴNG

Người hướng dẫn khoa học: PGS. TS. Lê Văn Sơn

Phản biện 1: PGS.TS. Trần Quốc Chiến

Phản biện 2: TS. Nguyễn Mậu Hân

Luận văn được bảo vệ trước hội đồng chấm Luận văn tốt nghiệp Thạc sĩ Kỹ thuật họp tại Đại học Đà Nẵng vào ngày 11 tháng 9 năm 2011

Có thể tìm hiểu luận văn tại:

- Trung tâm Thông tin - Học liệu, Đại học Đà Nẵng
- Trung tâm Học liệu, Đại học Đà Nẵng

MỞ ĐẦU

1. Lý do chọn đề tài

Sự bùng nổ và phát triển của Công nghệ thông tin đã mang lại nhiều hiệu quả đối với khoa học cũng như các hoạt động thực tế, trong đó khai phá dữ liệu là một lĩnh vực đem đến hiệu quả thiết thực cho con người. Khai phá dữ liệu đã giúp người sử dụng thu được những tri thức hữu ích từ những cơ sở dữ liệu hoặc các kho dữ liệu khổng lồ khác nhau. Để khai thác có hiệu quả những kho dữ liệu khổng lồ này, đã có nhiều công cụ được xây dựng để thỏa mãn nhu cầu khai thác dữ liệu mức cao, chẳng hạn công cụ khai thác dữ liệu Oracle Discoverer của hãng Oracle, hay là việc xây dựng các hệ chuyên gia, các hệ thống dựa trên một cơ sở tri thức của các chuyên gia để có thể dự báo được khuynh hướng phát triển của dữ liệu, thực hiện các phân tích trên các dữ liệu của tổ chức. Mặc dù các công cụ, các hệ thống đó hoàn toàn có thể thực hiện được phần lớn các công việc nêu trên, chúng vẫn yêu cầu một độ chính xác, đầy đủ nhất định về mặt dữ liệu.

Hiện nay, xu hướng học trực tuyến đang phát triển rất mạnh mẽ ở trên thế giới. Tại Việt Nam, e-learning trong giáo dục cũng đã được ứng dụng rộng rãi trong những năm gần đây và có nhiều sản phẩm có sẵn phục vụ cho mục đích này. Với những ưu thế của mình, hệ thống mã nguồn mở Moodle hiện nay vẫn được sử dụng rộng rãi nhất tại Việt Nam. Tuy nhiên đi kèm với mô hình đào tạo này là vấn

đề quản lý và sử dụng nguồn tài nguyên của hệ thống sao cho hiệu quả nhất.

Hệ thống Moodle có sẵn nhiều công cụ đánh giá và theo dõi quá trình học của học viên, tuy nhiên các công cụ này phần lớn mang tính chất thống kê là chính. Vậy tại sao không xây dựng một công cụ phân tích tập hợp các hành vi của học viên trên hệ thống e-learning nhằm đánh giá sự tiến bộ của họ? Công cụ này sẽ sử dụng nguồn dữ liệu giám sát hành vi từ hệ thống e-learning (các tập tin nhật ký) làm dữ liệu đầu vào kết hợp với các giải thuật tiên tiến của trí tuệ nhân tạo để dự báo dữ liệu trong tương lai. Cụ thể hơn, công cụ này sẽ giúp giảng viên dự báo kết quả học tập cuối khóa của học viên, từ đó sẽ có những điều chỉnh kịp thời đối với các học viên có khả năng không đạt kết quả tốt theo dự báo.

Luận văn này được thực hiện với mục đích tìm hiểu một số khía cạnh về mạng Nơron truyền thẳng nhiều lớp, thuật toán lan truyền ngược và ứng dụng chúng trong giải quyết bài toán dự báo kết quả học tập trực tuyến qua các dữ liệu thống kê thu thập được từ tập tin nhật ký Moodle.

2. Mục tiêu và nhiệm vụ

Mục tiêu của đề tài là xây dựng một công cụ sử dụng giải thuật khai phá dữ liệu dự báo kết quả học tập của học viên tham gia các khóa học trực tuyến. Nguồn dữ liệu dự báo được trích xuất từ tập tin nhật ký của hệ thống CMS dựa trên nền tảng mã nguồn mở Moodle.

Nhiệm vụ 1 – Nghiên cứu cơ bản

Nhiệm vụ 2 ─ Nghiên cứu ứng dụng

3. Đối tượng và phạm vi nghiên cứu

Đề tài hướng đến đối tượng nghiên cứu chủ yếu là các giải thuật khai phá dữ liệu nhằm áp dụng cho việc khám phá tri thức giáo dục.

Do còn hạn chế về thời gian, nguồn kinh phí và những hạn chế chủ quan của tác giả luận văn nên đề tài chỉ **tập trung nghiên cứu việc áp dụng mạng Nơron truyền thẳng nhiều lớp sử dụng thuật toán lan truyền ngược** cho quá trình khai phá dữ liệu giáo dục từ hệ thống CMS.

4. Giả thiết nghiên cứu

Mạng Nơron truyền thẳng sử dụng thuật toán lan truyền ngược có khả năng sử dụng như là một mô hình dự báo nhằm đánh giá khả năng hoàn thành khóa học của học viên hay không? Thông qua các nghiên cứu và thực nghiệm xây dựng ứng dụng, đề tài nhằm kiểm định tính hợp lý của giả thiết trên.

5. Phương pháp nghiên cứu

- Phương pháp nghiên cứu tài liệu
- Phương pháp thực nghiệm tự nhiên
- Phương pháp quan sát

6. Ý nghĩa khoa học và thực tiễn của đề tài

Về mặt ý nghĩa khoa học, đề tài đạt được các kết quả như sau:

▪ Đã hệ thống hóa các nội dung cơ bản khi giải quyết bài toán dự báo sử dụng mạng nơron nói chung và mạng truyền thẳng lan truyền ngược nói riêng.

▪ Đã đề xuất và hiện thực phương pháp tìm kiếm các tham số quan trọng của mạng nơron truyền thẳng lan truyền ngược từ bài toán thực tiễn tại đơn vị công tác.

▪ Đã đề xuất quy trình tổng quát giải quyết bài toán dự báo kết quả tương lai từ dữ liệu quá khứ sử dụng thuật toán lan truyền ngược.Quy trình được thực nghiệm thông qua việc giải quyết bài toán cụ thể: dự báo kết quả học tập của học viên trực tuyến thông qua dữ liệu thu thập được từ tập tin nhật ký Moodle.

Về giá trị thực tiễn, sau khi hoàn tất, sản phẩm của đề tài là khả năng dự báo kết quả học tập, qua đó góp phần hỗ trợ giảng viên trong công tác dự báo, đánh giá học viên.

7. Bố cục của luận văn

Luận văn gồm ba chương:

Chương 1 - TỔNG QUAN VỀ MẠNG NƠRON VÀ VẤN ĐỀ DỰ BÁO SỬ DỤNG MẠNG NƠRON

Chương 2 - MẠNG NƠRON TRUYỀN THẲNG LAN TRUYỀN NGƯỢC VÀ ỨNG DỤNG TRONG DỰ BÁO DỮ LIỆU

Chương 3 - XÂY DỰNG GIẢI PHÁP KỸ THUẬT CHO PHÉP DỰ BÁO KẾT QUẢ HỌC TẬP TRỰC TUYẾN

CHƯƠNG 1 - TỔNG QUAN VỀ MẠNG NƠRON VÀ VẤN ĐỀ DỰ BÁO SỬ DỤNG MẠNG NƠRON

Khoa học trí tuệ nhân tạo có thể được chia làm ba nhánh chính: Hệ chuyên gia, Logic mờ và Mạng Nơron. Trong đó, hệ chuyên gia là công cụ thích hợp để xử lý tín hiệu dưới dạng phi số; Logic mờ là công cụ mạnh để xử lý dữ liệu dưới dạng khái quát, mô tả không rõ ràng; còn mạng Nơron được sử dụng trong công tác xử lý số liệu dưới dạng số (các bài toán phân loại, nhận dạng,..). Mạng Nơron nhân tạo là một hệ thống xử lý thông tin được xây dựng trên cơ sở tổng quát hóa mô hình toán học của Nơron sinh học và phỏng theo cơ chế làm việc của bộ não con người.

1.1 Tổng quan về mạng Nơron

1.1.1. Lịch sử phát triển

1.1.2. Mô hình mạng Nơron

1.1.3. Các luật học

Luật học là một trong các yếu tố quan trọng tạo nên một mạng Nơron nhân tạo. Có hai vấn đề cần học đối với mỗi mạng Nơron nhân tạo, đó là học tham số và học cấu trúc. Học tham số là việc thay đổi trọng số của các liên kết giữa các Nơron trong một mạng; còn học cấu trúc là việc điều chỉnh cấu trúc của mạng bao gồm thay đổi số lớp Nơron, số Nơron của mỗi lớp và cách liên kết giữa chúng. Hai vấn đề này có thể được thực hiện đồng thời hoặc tách biệt.

1.1.4. Hình trạng mạng

Các mạng về tổng thể được chia thành hai loại dựa trên cách thức liên kết các đơn vị.

1.1.4.1. Mạng truyền thẳng

Dòng dữ liệu giữa đơn vị đầu vào và đầu ra chỉ truyền thẳng theo một hướng. Việc xử lý dữ liệu có thể mở rộng ra thành nhiều lớp, nhưng không có các liên kết phản hồi. Điều đó có nghĩa là không tồn tại các liên kết mở rộng từ các đơn vị đầu ra tới các đơn vị đầu vào trong cùng một lớp hay các lớp trước đó.

1.1.4.2. Mạng quay lui (mạng hồi quy)

1.1.5. Ứng dụng của mạng Nơron

Trong quá trình phát triển, mạng Nơron đã được ứng dụng thành công trong rất nhiều lĩnh vực như hàng không vũ trụ, điều khiển tự động, ngân hàng, trong quốc phòng, trong y học,…

1.2 Ứng dụng mạng Nơron trong dự báo dữ liệu

1.2.1 Khái quát về lĩnh vực dự báo

1.2.1.1 Khái niệm dự báo

Dự báo là một khoa học và nghệ thuật tiên đoán những sự việc sẽ xảy ra trong tương lai trên cơ sở phân tích khoa học về các dữ liệu đã thu thập được. Khi tiến hành dự báo cần căn cứ vào việc thu thập, xử lý số liệu trong quá khứ và hiện tại để xác định xu hướng vận động của các hiện tượng trong tương lai nhờ vào một số mô hình toán học (định lượng).

1.2.1.2 Đặc điểm của dự báo

Không có cách nào để xác định tương lai là gì một cách chắc chắn, đó là tính không chính xác của dự báo.

Luôn có điểm mù trong các dự báo, không thể dự báo một cách chính xác hoàn toàn điều gì sẽ xảy ra trong tương tương lai.

1.2.1.3 Các phương pháp dự báo

1.2.2 Sử dụng mạng Nơron như công cụ dự báo

1.2.2.1 Lĩnh vực áp dụng

a) Bài toán phân lớp: loại bài toán này đòi hỏi giải quyết vấn đề phân loại các đối tượng quan sát được thành các nhóm dựa trên những đặc điểm của các nhóm đối tượng đó. Đây là dạng bài toán cơ sở của rất nhiều bài toán trong thực tế: nhận dạng chữ viết, tiếng nói, phân loại gen, phân loại chất lượng sản phẩm,…

b) Bài toán dự báo: mạng Nơron nhân tạo đã được ứng dụng thành công trong việc xây dựng các mô hình dự báo sử dụng tập dữ liệu trong quá khứ để dự báo số liệu trong tương lai. Đây là nhóm bài toán khó và rất quan trọng trong nhiều ngành khoa học.

c) Bài toán điều khiển và tối ưu hóa: nhờ khả năng học và xấp xỉ hàm mà mạng Nơron nhân tạo đã được sử dụng trong nhiều hệ thống điều khiển tự động cũng như góp phần giải quyết những bài toán tối ưu trong thực tế.

1.2.2.2 Ứng dụng trong giáo dục

Riêng trong lĩnh vực giáo dục, các ứng dụng của mạng Nơron nói riêng và khai phá dữ liệu nói chung đã và đang được áp dụng rộng rãi. Tuy nhiên, ở Việt Nam, việc ứng dụng trí tuệ nhân tạo trong các hệ thống quản lý học tập và công tác giảng dạy chưa được quan tâm nghiên cứu và áp dụng nhiều trong thực tế.

CHƯƠNG 2 - MẠNG NƠRON TRUYỀN THẲNG LAN TRUYỀN NGƯỢC VÀ ỨNG DỤNG TRONG DỰ BÁO DỮ LIỆU

2.1 Mạng Nơron truyền thẳng lan truyền ngược

2.1.1 Khái niệm

Một mạng Nơron lan truyền ngược điển hình có một lớp vào, một lớp ra và ít nhất một lớp ẩn. Trong một ứng dụng mạng lan truyền ngược, có hai quá trình tính toán phân biệt nhau, đó là quá trình lan truyền thẳng và quá trình lan truyền ngược.

Trong quá trình lan truyền thẳng, tất cả các trọng số không thay đổi, các tín hiệu hàm được tính toán từ trái qua phải từ Nơron này qua Nơron kia.

Trong quá trình lan truyền ngược, tín hiệu lỗi xuất phát từ lớp xuất lan truyền ngược về phía trái. Trong khi lan truyền các trọng số được cập nhật theo chiều hướng làm giá trị đầu ra xích gần giá trị mong muốn hơn.

2.1.2 Hướng tiếp cận của mạng Nơron lan truyền ngược

Mạng Nơron lan truyền ngược chỉ đạt kết quả tốt trong các trường hợp nhất định:

▪ Một số lượng lớn dữ liệu đầu vào/ra là có sẵn, nhưng ta không chắc chắn chúng có liên quan đến đầu ra như thế nào.

▪ Dễ dàng để tạo ra một số ví dụ về các hành vi đúng.

▪ Các giải pháp cho vấn đề này có thể thay đổi theo thời gian, trong phạm vi của các tham số các đầu vào, đầu ra đã cho.

▪ Kết quả có thể là "mờ", hay ở dạng phi số.

Sau đây là một số kinh nghiệm khi nào không nên sử dụng mạng Nơron lan truyền ngược:

▪ Với vấn đề cần giải quyết mà có thể vẽ một biểu đồ hoặc công thức mô tả chính xác vấn đề, hãy sử dụng lập trình truyền thống.

▪ Nếu có thể sử dụng phần cứng hoặc phần mềm để giải quyết những dự định làm với mạng Nơron lan truyền ngược thì không nên dùng mạng Nơron.

▪ Nếu mong muốn các chức năng "tiến hóa" theo hướng không được xác định trước, hãy cân nhắc sử dụng một thuật toán di truyền.

▪ Có thể dễ dàng để tạo ra một số lượng đáng kể các đầu vào/đầu ra minh họa cho các hành vi mong muốn hay không? Nếu không thực hiện được điều này ta sẽ không thể huấn luyện mạng Nơron để thực hiện bất cứ điều gì.

▪ Các giá trị đầu ra yêu cầu phải là các con số chính xác? Mạng Nơron không tốt trong việc đưa ra câu trả lời là các con số chính xác.

2.2 Thuật toán lan truyền ngược

2.2.1 Giới thiệu thuật toán

Nguyên tắc huấn luyện mạng Nơron đa lớp sử dụng thuật toán lan truyền ngược gồm hai giai đoạn chính: lan truyền thẳng (tính toán đầu ra của các Nơron) và lan truyền ngược qua mạng.

Tóm tắt thuật toán lan truyền ngược:

- Khởi tạo trọng số (thường là khởi tạo ngẫu nhiên)
- Đối với mỗi mẫu dữ liệu *e* trong tập huấn luyện
  - *Lan truyền thẳng*: tính O = giá trị đầu ra của mạng;
  - Với T = giá trị đầu ra mong muốn của e, tính toán lỗi tại đơn vị đầu ra (T - O)
  - *Lan truyền ngược*:
    - tính giá trị delta_wi cho tất cả các trọng số từ lớp ẩn đến lớp ra;
    - tính giá trị delta_wi cho tất cả các trọng số từ lớp vào đến lớp ẩn;
  - Cập nhật trọng số của mạng.
- Kết thúc thuật toán.

2.2.2 Một số yếu tố ảnh hưởng đến quá trình học

- *Khởi tạo các trọng số*
- *Hằng số học η*

2.3 Phát biểu bài toán dự báo kết quả học tập trực tuyến

Học trực tuyến e-Learning đáp ứng được những tiêu chí giáo dục mới: học mọi nơi, học mọi lúc, học theo sở thích, và học suốt đời. E-Learning tồn tại song song và bổ sung cho cách học tập truyền thống. Nhìn chung, hệ thống E-Learning bao gồm:

- Hệ thống quản lý học tập (LMS) giúp xây dựng các lớp học trực tuyến hiệu quả.

▪ Hệ thống quản lí nội dung học tập (LCMS) cho phép tạo và quản lý nội dung học tập.

▪ Công cụ làm bài giảng một cách sinh động, dễ dùng và đầy đủ multimedia.

Điều quan trọng hơn là E-Learning đã được thế giới chuẩn hoá nên các bài giảng có thể trao đổi với nhau trên toàn thế giới cũng như giữa các trường học ở Việt Nam.

2.3.1 Khái quát hệ thống quản lý học tập sử dụng Moodle

Moodle là một hệ thống quản lý học tập mã nguồn mở. Moodle là một thành phần quan trọng của hệ thống E-learning, hỗ trợ học tập trực tuyến.

- Moodle nổi bật là thiết kế hướng tới giáo dục.
- Moodle phù hợp với nhiều cấp học và hình thức đào tạo.
- Moodle rất đáng tin cậy, có trên 10 000 site trên thế giới (thống kê tại Moodle.org) đã dùng Moodle tại 138 quốc gia và đã được dịch ra trên 70 ngôn ngữ khác nhau.

2.3.2 Phát biểu bài toán

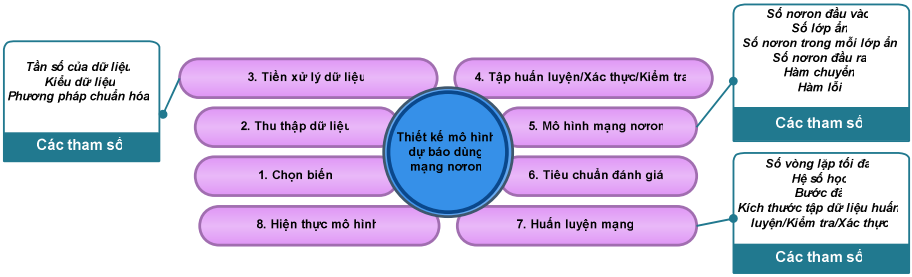
Luận văn này tập trung tìm hiểu hướng tiếp cận sử dụng mạng Noron truyền thẳng lan truyền ngược để phát triển và thử nghiệm với dữ liệu thu thập là các tri thức của sinh viên khi tham gia học môn Tin tại trường Cao đẳng Kỹ thuật Y tế II trong năm 2010-2011 từ tập tin nhật ký của hệ thống Moodle. Các tri thức này sẽ được kết hợp với kết quả đánh giá các bài thi tại lớp (theo phương thức học truyền thống) nhằm xây dựng mô hình có khả năng dự báo khả năng hoàn tất khóa học của sinh viên.

CHƯƠNG 3 - XÂY DỰNG GIẢI PHÁP KỸ THUẬT ĐỂ DỰ BÁO KẾT QUẢ HỌC TẬP TRỰC TUYẾN

Để đơn giản và tránh hiểu nhầm, thuật ngữ "mạng Nơron" được dùng trong chương 3 này được hiểu là mạng Nơron truyền thẳng nhiều lớp lan truyền ngược.

3.1 Phân tích bài toán

Theo Kaastra and Boyd (1996), các bước chính cần thực hiện khi thiết kế mô hình mạng Nơron sử dụng cho bài toán dự báo nói chung, bao gồm tám bước như Hình 3.1.



Hình 3.1 Các bước thiết kế mô hình mạng Nơron dự báo dữ liệu

Trong quá trình thực hiện, không nhất thiết phải thực hiện theo đúng thứ tự các bước trên mà có thể quay về các bước trước đó, đặc biệt là bước huấn luyện và lựa chọn các biến.

Các vấn đề chủ yếu cần giải quyết khi xây dựng mạng Nơron truyền thẳng lan truyền ngược dự báo kết quả học tập là:

- *Tiền xử lý dữ liệu*
  - Xác định tần số của dữ liệu: hàng ngày, hàng tuần,…
  - Kiểu của dữ liệu

- Phương thức chuẩn hóa dữ liệu: công thức Max/Min hay độ lệch trung bình,…

▪ *Cấu trúc mạng*

- Số đầu vào
- Số lớp ẩn và số Nơron trong mỗi lớp ẩn
- Số Nơron đầu ra
- Hàm chuyển
- Hàm lỗi

▪ *Huấn luyện mạng*

- Hệ số học
- Bước đà
- Số chu kỳ huấn luyện tối đa
- Khởi tạo trọng số
- Kích thước tập huấn luyện/kiểm tra/xác thực

Việc sử dụng mạng Nơron khám phá tri thức trong tập tin nhật ký Moodle hướng đến việc giải quyết các câu hỏi như:

▪ Có thể sử dụng mạng Nơron như một mô hình dự báo nhằm phát hiện các học sinh tham gia học trực tuyến cần phải được bổ sung kiến thức khi kết thúc khóa học hay không?

▪ Kết quả bài thi khóa học của sinh viên như thế nào?...

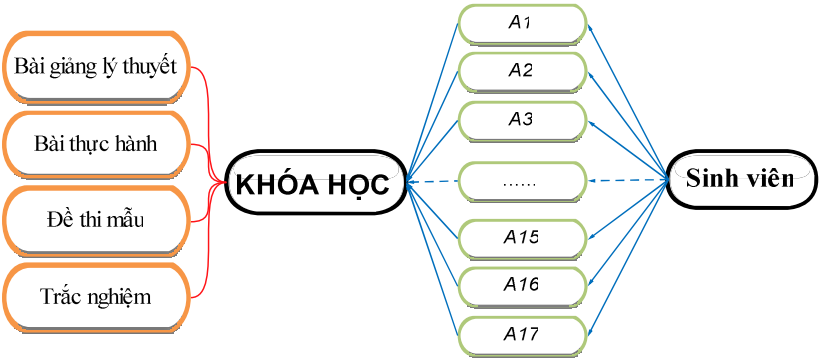
3.2 Xây dựng giải pháp kỹ thuật dự báo kết quả học tập trực tuyến

Luận văn này sử dụng hướng tiếp cận từ Kaastra và Boyd (1996) nhưng có một số thay đổi được thực hiện cho phù hợp với khuôn khổ của bài toán cần giải quyết. Đó là bài toán "**Ứng dụng mạng Nơron truyền thẳng lan truyền ngược phân tích tập tin nhật ký Moodle dự báo kết quả học tập trực tuyến**".

Bài toán dự báo kết quả học tập trực tuyến hình thành từ bài báo "Dự đoán kết quả thi sinh viên tại trường đại học mở Hellenic – Hy Lạp" của hai tác giả Sotiris B. Kotsiantis và Panayiotis E. Pintelas. Tuy có điểm chung là dự báo kết quả học tập của sinh viên nhưng hướng tiếp cận lại hoàn toàn khác nhau.

3.2.1 Bước 1 – Lựa chọn biến đầu vào

Mục đích của luận văn là dự báo kết quả của sinh viên từ các dữ liệu truy cập của họ vào hệ thống tài nguyên, vì vậy các tác động của sinh viên tới hệ thống tài nguyên sau sẽ được giữ lại, bao gồm:



Hình 3.2 Tác động của sinh viên đến khóa học

- "*Bài giảng lý thuyết*": nguồn tài nguyên chính chứa các bài giảng cần thiết cho khóa học
- "*Bài thực hành*": các bài tập phụ trợ, bổ sung kiến thức cho phần bài giảng lý thuyết
- "*Đề thi mẫu*": tập hợp các đề thi mẫu của khóa học đã được thực hiện trước đây. Mục đích cho sinh viên làm quen cấu trúc bài thi
- "*Các câu hỏi trắc nghiệm*": dùng cho mục đích củng cố kiến thức nhận được từ khóa học.

Bảng 3.1 Các biến chính phục vụ dự báo

Mã số	**Mô tả**
A1	Họ tên (tên đăng nhập hoặc tên đầy đủ)
A2	Số lần đăng ký tham gia khóa học (thi lần 1/lần 2);
A3	Tổng thời gian truy cập trong suốt khóa học, từ 9/2010 đến 12/2010
A4	Tổng thời gian truy cập với mục đích chỉ xem tài nguyên
A5	Tỷ lệ A4 / A3
A6	Số lần truy cập tài nguyên "Lý thuyết"
A7	Số lần truy cập tài nguyên "Đề thi mẫu"
A8	Số lần truy cập tài nguyên "Bài thực hành"
A9	Số lượng câu trắc nghiệm đã thực hiện
A10	Tổng thời gian đã thực hiện thi trắc nghiệm
A11	Số lượng câu trắc nghiệm đã trả lời đúng
A12	Số lượng câu trắc nghiệm đã trả lời sai
A13	Số lần gửi bài viết lên diễn đàn
A14	Số lần đọc bài viết trên diễn đàn
A15	Các ngày trong tuần
A16	Ngày cuối tuần
A17	Thời gian đăng nhập

Với các biến đầu vào và đầu ra như đã trình bày trong Bảng 3.1, dữ liệu chính sử dụng là tập tin nhật ký của 100 học sinh thuộc hai lớp Cao đẳng Điều dưỡng tại trường Cao đẳng Kỹ thuật Y tế II, môn Tin học, trong thời gian bốn tháng cuối năm 2010. Tập tin nhật ký được lấy từ hệ thống Moodle chạy trên mạng LAN của trường. Hiện tại hệ thống tài nguyên sử dụng Moodle của trường chỉ phục vụ cho mạng nội bộ nên học sinh có thể tham khảo các khóa học trực tuyến vào bất kỳ thời gian nào từ 7g30 – 11g30, từ 13g – 17g và từ 17g30 – 21g (dành cho các lớp ban đêm tại trung tâm) của các ngày trong tuần, trừ ngày lễ và chủ nhật.

Các dữ liệu lịch sử được chọn lọc theo Bảng 3.1 và được xử lý theo các nguyên tắc sau:

1) Họ và tên/mã số sinh viên/tên đăng nhập (biến A1): giá trị biến này lấy theo số thứ tự của sinh viên khi được đăng ký tham gia hệ thống. Đây là dữ liệu dạng số nguyên và có thể lấy chính giá trị thực của nó

2) Số lần đăng ký tham gia khóa học: mỗi học sinh được thi hai lần cho mỗi môn học. Đây là dữ liệu dạng số nguyên chỉ có hai giá trị 1 hoặc 2 nên có thể biểu diễn bằng chính nó.

3) Các biến tính theo thời gian (tổng thời gian truy cập trong suốt khóa học, tổng thời gian truy cập với mục đích chỉ xem tài nguyên, tổng thời gian đã thực hiện thi trắc nghiệm): được biểu diễn bằng chính nó và đơn vị tính theo phút.

4) Số lần truy cập tài nguyên “Lý thuyết”/“Đề thi mẫu”/“Bài thực hành”, số lượng câu trắc nghiệm đã thực hiện, số câu trắc

nghiệm đã trả lời đúng/sai, số lần gửi/đọc bài viết trên diễn đàn: biểu diễn bằng giá trị thực của chính nó.

5) Ngày trong tuần: thể hiện bằng các số từ 0 – 6 tương ứng các ngày từ Chủ nhật, thứ hai,…đến thứ bảy.

6) Ngày cuối tuần: các ngày từ thứ hai đến thứ sáu thể hiện bởi giá trị 0 và thứ bảy được biểu diễn bằng 1.

7) Thời gian đăng nhập: thể hiện 24 tiếng trong ngày: 0, 1, 2,…,23.

Rõ ràng các hiệu ứng trong 1), 2), 3) và 4) là các biến có thứ tự. Giá trị thực của chúng có thể đưa vào mạng như chúng vốn có. Các hiệu ứng còn lại là các biến phân loại.

Luận văn sử dụng phương pháp chọn số đầu vào theo phương pháp one-perfect-one-unit. Mặc dù phương thức này có khả năng tạo ra một trật tự nhân tạo trên các giá trị nhưng ngược lại, số lượng đầu vào cũng sẽ giảm đi và mô hình thực hiện cũng đơn giản hơn.

3.2.2 Bước 2 – Thu thập dữ liệu

3.2.3 Bước 3 – Tiền xử lý dữ liệu

▪ ***Chuẩn hóa dữ liệu***

Do tính chất hỗn loạn của dữ liệu, các giá trị của chúng có thể sai lệch rất nhiều trong khoảng thời gian rất ngắn. Điều này có thể gây ra khó khăn rất lớn để các mạng Nơron thực hiện công việc của mình. Hơn nữa, hàm kích hoạt sử dụng bởi mạng Nơron là bị chặn, do đó sẽ gây ra tình trạng không thống nhất trong cả hai giai đoạn huấn luyện và dự báo. Để tránh gặp những khó khăn tiềm ẩn như

vậy, dữ liệu thường được thu nhỏ trong khoảng giữa 0 và 1 hoặc -1 và 1, vì như vậy sẽ phù hợp với các hàm chuyển được sử dụng.

3.2.4 Bước 4 – Phân hoạch dữ liệu

Phân hoạch là quá trình chia dữ liệu thành các tập *huấn luyện*, *tối ưu* và *tập thử nghiệm*. Theo định nghĩa, tập tối ưu được sử dụng để xác định kiến trúc của mạng; các tập huấn luyện dùng để cập nhật các trọng số của mạng; các tập thử nghiệm được dùng để kiểm tra hiệu năng của mạng sau khi luyện.

Vì tập dữ liệu tối ưu là tùy ý, bên cạnh đó, các tham số của mạng Nơron sẽ được xác định thông qua thực nghiệm nên luận văn không sử dụng tập tối ưu khi phân hoạch dữ liệu. Như vậy tập dữ liệu đầu vào sẽ được chia thành hai tập dữ liệu huấn luyện và thử nghiệm theo tỉ lệ mặc định 80% và 20%.

3.2.5 Bước 5 – Xây dựng mô hình mạng Nơron

3.2.5.1 Số lượng lớp ẩn

Về mặt lý thuyết, một mạng Nơron với chỉ một lớp ẩn cùng với số Nơron ẩn hợp lý là có đủ khả năng xấp xỉ bất kỳ một hàm liên tục nào. Trong thực tế, mạng Nơron có từ một và đôi khi có hai lớp ẩn được sử dụng rộng rãi và đạt được hiệu quả tốt.

3.2.5.2 Số Nơron trong lớp ẩn

Cho đến hiện nay, vẫn không có công thức chung nhất cho việc xác định số Nơron trong mỗi lớp ẩn. Hầu hết các nhà nghiên cứu sử dụng kinh nghiệm kết hợp với phương pháp thử-sai để tìm ra kết quả khả dĩ nhất.

3.2.5.3 Hàm chuyển (hàm kích hoạt)

Trong luận văn sử dụng hàm chuyển là hàm Bipolar Sigmoid (tansig). Hàm Bipolar Sigmoid có các thuộc tính tương tự hàm Sigmoid. Nó làm việc tốt với các ứng dụng có yêu cầu đầu ra trong khoảng [-1, 1].

3.2.6 Bước 6 – Tiêu chuẩn đánh giá

Luận văn sử dụng giá trị tổng bình phương lỗi học (SSE) để đánh giá các tham số của mạng.

3.2.7 Bước 7 – Huấn luyện mạng

3.2.7.1 Số chu kỳ huấn luyện

Việc xác định điểm dừng của quá trình huấn luyện trong nghiên cứu này tuân theo quan điểm thử nghiệm hàng loạt các phân đoạn huấn luyện - đào tạo khác nhau, kết hợp với hai quá trình tìm kiếm nhị phân và tuyến tính nhằm tiết kiệm thời gian xác định số chu kỳ huấn luyện tối ưu.

3.2.7.2 Hệ số học và bước đà

Việc lựa chọn được giá trị tối ưu của hệ số học hay bước đà là khó khăn . Theo Rumelhart và cộng sự, với hệ số học giữa 0,05 và 0,5 cho kết quả tốt trong nhiều trường hợp thực tế trong khi bước đà thường tiến gần đến 1, ví dụ 0.9. Luận văn lựa chọn bước đà và hệ số học theo phương thức thử-và-sai qua đồ thị kết quả dự báo.

3.2.8 Bước 8 – Hiện thực mô hình

Chương trình được xây dựng dựa trên hệ điều hành Microsoft Windows XP, Windows 7, công cụ phát triển Microsoft Visual C#

2008, Excel 2007, bộ công cụ nguồn mở Aforge.NET 2.1.5 và Edraw Max phiên bản 5.2, dùng thử 30 ngày.

Bảng 3.2 Tổng kết các bước xây dựng mô hình dự báo trên thực tế

Mã số	Mô tả
Bước 1	Chuẩn bị dữ liệu huấn luyện/kiểm tra
Bước 2	Xây dựng mạng truyền thẳng lan truyền ngược gồm: - Ba lớp: lớp vào 14 nút, lớp ra 1 nút và số nút ẩn khởi đầu là 4 (theo Masters) - Hệ số học và bước đà tương ứng là 0,05 và 0 (theo Rumelhart) - Sigmoid's Alpha = 1 (theo khuyến cáo của Aforge.NET)
Bước 3	***Tìm số chu kỳ huấn luyện (epochs)*** - Huấn luyện mạng với dữ liệu thu nhận ở *bước 1*, số chu kỳ huấn luyện bắt đầu từ 1000, tăng cấp số nhân đến tối đa 128 000 chu kỳ - Chọn mạng có tổng bình phương lỗi học nhỏ nhất
Bước 4	Lưu thông tin mạng Nơron, số chu kỳ *epochs*.
Bước 5	***Xác định số nút ẩn*** - Sử dụng số chu kỳ *epochs* đã tìm được ở *bước 4* - Khởi tạo mạng với số nút ẩn bắt đầu từ hai - Huấn luyện mạng và lưu thông tin lỗi
Bước 6	Thực hiện lại *bước 5* cho đến khi số nút ẩn đạt 32 nút
Bước 7	Chọn dải mạng ứng với hai giá trị lỗi nhỏ nhất (trong tất cả năm mạng với số nút ẩn bắt dầu từ 2 đến 32), giả sử có số nút ẩn lần lượt là $n_{h1} = x$ và $n_{h2} = y$
Bước 8	Thực hiện lại *bước 5* với các điều kiện sau: - Số nút ẩn khởi đầu bằng n_{h1}, và tăng tuyến tính số lượng nút ẩn đến khi bằng n_{h2}. - Các tham số khác của mạng không đổi
Bước 9	Chọn ra mạng có giá trị lỗi nhỏ nhất trong nhóm mạng chọn ở *bước 8* Đây chính là mạng có số nút ẩn $\boldsymbol{n_h}$ sẽ sử dụng cho mô hình dự báo

Bước 10	***Tìm bộ số (hệ số học, bước đà) bằng thực nghiệm*** Cấu hình mạng gồm 14 nút vào, 1 nút ra, n_h nút ẩn, số chu kỳ ***epochs***, hệ số khác không đổi: - Hệ số học $LR \in \{0,05; 0,5\}$ (theo Rumelhart) - Bước đà $momentum \in \{0; 0,9\}$ Lặp lại bước huấn luyện để tìm mạng có bình phương lỗi bé nhất
Bước 11	Lưu thông tin mạng cuối cùng. Thực hiện huấn luyện với tập dữ liệu ở *bước 1* và lưu mạng phục vụ giai đoạn kiểm tra
Bước 12	Sử dụng mạng đã lưu ở *bước 11*, tiến hành dự báo với mẫu dữ liệu đầu vào x_t và dữ liệu dự báo y_t lấy từ tập dữ liệu kiểm tra
Bước 13	So sánh, đối chiếu độ tin cậy dự báo của mạng và tiến hành điều chỉnh các tham số nếu cần thiết. Huấn luyện mạng lần nữa với tập dữ liệu ở bước 1 cộng thêm mẫu dữ liệu (x_t, y_t) ở *bước 12*
Bước 14	Tiến hành dự báo mẫu dữ liệu (x_{t+1}, y_{t+1}) dùng mạng đã được huấn luyện
Bước 15	Thực hiện lại *bước 12* và *14* đến khi hết tập dữ liệu kiểm tra

3.3 Kết quả thực hiện

3.3.1 Biến đầu vào

3.3.2 Số chu kỳ huấn luyện

3.3.3 Xác định số nút ẩn

3.3.4 Lựa chọn hệ số học và bước đà

3.3.5 Kết quả huấn luyện và dự báo dữ liệu

3.4 Một số nhận xét

KẾT LUẬN

Mạng Nơron truyền thẳng nhiều lớp đã và đang được sử dụng trong rất nhiều bài toán dự báo trong các lĩnh vực khác nhau như dự báo lượng sử dụng điện, nước, thị trường chứng khoán, lưu lượng giao thông,.. Tuy vậy, không tồn tại một mô hình chung thích hợp cho tất cả bài toán dự báo trong thực tế. Đối với mỗi một bài toán, cần thực hiện phân tích cặn kẽ, cụ thể các dữ liệu trong phạm vi đang xét và sử dụng các tri thức thu thập được nhằm hướng đến xây dựng một mô hình thích hợp. Các phân tích và các tri thức thu thập được luôn có ích trong việc lựa chọn các đầu vào, mã hóa các đầu vào này hoặc quyết định cấu trúc của mạng, điều này đặc biệt hữu ích khi mà dữ liệu trong lĩnh vực đó chỉ có giới hạn.

Thuật toán lan truyền ngược chuẩn được sử dụng trong việc huấn luyện mạng Nơron truyền thẳng nhiều lớp đã chứng tỏ khả năng rất tốt, thậm chí đối với cả các bài toán hết sức phức tạp. Dù vậy, để đạt kết quả mong muốn, cần mất rất nhiều thời gian để huấn luyện, điều chỉnh các tham số của mạng (thậm chí cả đối với các bài toán có cấu trúc hết sức đơn giản). Điều này luôn là trở ngại đối với các bài toán trong thực tế. Do đó, các thuật toán cải tiến cần được áp dụng để tăng khả năng hội tụ của mạng khi huấn luyện. Khả năng hội tụ của mạng phụ thuộc vào các tham số khởi đầu, còn khả năng tổng quát hóa thì lại phụ thuộc rất nhiều vào dữ liệu đầu vào. Nếu dữ liệu đầu vào quá nhiều thì có thể dẫn tới tình trạng luyện mạng mất rất nhiều thời gian và khả năng tổng quát hóa kém, ngược lại, nếu quá ít dữ liệu thì sai số sẽ tăng lên.

❖ ***Những đóng góp khoa học của luận văn***

▪ Nghiên cứu và hệ thống hóa các nội dung cơ bản khi giải quyết bài toán dự báo sử dụng mạng nơron nói chung và mạng truyền thẳng lan truyền ngược nói riêng.

▪ Đề xuất và hiện thực phương pháp tìm kiếm các tham số quan trọng của mạng nơron truyền thẳng lan truyền ngược từ bài toán thực tiễn tại đơn vị công tác.

▪ Đề xuất quy trình tổng quát giải quyết bài toán dự báo kết quả tương lai từ dữ liệu quá khứ sử dụng thuật toán lan truyền ngược. Quy trình được thực nghiệm thông qua việc giải quyết bài toán cụ thể, bài toán dự báo kết quả học tập của học viên trực tuyến thông qua dữ liệu thu thập được từ tập tin nhật ký Moodle. Kết quả thực nghiệm dự báo kết quả học tập của sinh viên có sự sai lệch không quá lớn (dưới 10%) giữa giá trị thực tế và dự báo trên tập dữ liệu thử nghiệm.

❖ ***Hướng phát triển***

Những kết quả nghiên cứu về ứng dụng mạng Nơron nhân tạo với thuật toán học lan truyền ngược trong bài toán dự báo kết quả học tập trực tuyến đã chứng tỏ rằng đây là một mô hình có thể ứng dụng hiệu quả đối với bài toán này. Ngoài thuật toán lan truyền ngược, các phương pháp học máy khác như Support Vector Machine, Decision Tree...cũng có thể được sử dụng để giải quyết bài toán trên. Vì vậy, hướng phát triển tiếp theo của đề tài là nghiên cứu, cải tiến và thử nghiệm các phương pháp học máy tiên tiến khác để có thể nâng cao độ tin cậy dự báo và thời gian dự báo.

Việc nghiên cứu kết hợp thuật toán di truyền và lan truyền ngược trong quá trình khởi tạo các trọng số của mạng cũng là một hướng phát triển của luận văn. Trong sự kết hợp giữa hai thuật toán này, thuật toán di truyền được sử dụng như một bộ khởi tạo cho lan truyền ngược. Tập các trọng số được mã hóa thành các nhiễm sắc thể và được tiến hóa nhờ di truyền. Kết thúc quá trình tiến hóa, bộ trọng số tốt nhất tương ứng với cá thể ưu việt nhất trong quần thể được lựa chọn làm những trọng số khởi tạo cho thuật toán lan truyền ngược. Tất nhiên phải có một số thay đổi trong thuật toán lan truyền ngược cho phù hợp với sự kết hợp này.

Về mặt ứng dụng, do hạn chế lớn nhất của ứng dụng là quá trình tập hợp dữ liệu từ hệ thống Moodle vẫn đang tiến hành theo phương pháp thủ công cần nghiên cứu xây dựng bộ công cụ add-in cho Moodle nhằm tự động hóa quá trình thu thập và tiền xử lý dữ liệu. Bên cạnh đó cần cải tiến hình thức trình bày kết quả dự báo cho phù hợp và trực quan hơn.

Một hướng phát triển khác nữa của đề tài là xây dựng thành một công cụ cho phép học viên khi tham gia khóa học bất kỳ có thể tự dự báo kết quả học tập của mình, qua đó định hướng phương pháp học tập cụ thể nhằm hy vọng đạt kết quả tốt khi kết thúc khóa học.

Trong thực tế, sự đầy đủ hay không của tập số liệu đầu vào có ảnh hưởng rất lớn đến kết quả dự báo. Vì thế, cùng với hướng phát triển trên, tôi sẽ tiến hành thu thập số liệu tập tin nhật ký Moodle ở một số website học tập trên mạng để thử nghiệm và đánh giá kết quả dự báo nhằm cải thiện độ tin cậy dự báo của mô hình.