ĐẠI HỌC THÁI NGUYÊN
TRƯỜNG ĐẠI HỌC CNTT&TT

Điêu Thiện Chiến

MỘT SỐ KỸ THUẬT MÔ HÌNH HÓA VÀ ÁP DỤNG CHO BÀI TOÁN DỰ BÁO KẾT QUẢ TUYỂN SINH ĐẠI HỌC

LUẬN VĂN THẠC SĨ: KHOA HỌC MÁY TÍNH

Người hướng dẫn: GS. TS Nguyễn Thanh Thủy

Thái Nguyên, tháng 01 năm 2014

MỤC LỤC

DANH MỤC CÁC KÝ HIỆU VÀ CHỮ VIẾT TẮT

THPT: Trung học phổ thông

ĐH: Đại học

TK: Thế kỷ

TSĐH: tuyển sinh Đại học

School_Cla: trường/lớp

Time_X: thời gian tự học

Time_Y: thời gian học thêm

Pressure: áp lực

Point_School: tổng điểm 3 môn thi Đại học ở bậc THPT

Mark: tổng điểm thi Đại học

Group: nhóm

MLP: mạng nơron truyền thẳng nhiều lớp

DANH MỤC CÁC BẢNG

DANH MỤC CÁC HÌNH

MỞ ĐẦU

Bài toán dự báo tuyển sinh Đại học ngày càng được các trường Cao đẳng và Đại học quan tâm trong bối cảnh, nhiều trường được thành lập và gặp khó khăn về tuyển sinh đầu vào. Học sinh sau khi tốt nghiệp THPT lựa chọn ngành dự thi còn chưa phù hợp với năng lực của mình cũng như nhu cầu việc làm của xã hội. Sinh viên ra trường chưa có việc làm đúng ngành nghề của mình còn nhiều.

Một số kỹ thuật mô hình hóa dữ liệu được áp dụng nhằm dự báo kết quả tuyển sinh Đại học sẽ giúp cho học sinh THPT biết được năng lực và khả năng thi tuyển đầu vào của các trường Đại học, Cao đẳng.

Mạng nơron truyền thẳng và thuật toán lan truyền ngược được áp dụng để dự báo kết quả tuyển sinh Đại học. Số liệu thu thập từ các sinh viên trường Đại học Hùng Vương Phú Thọ. Bài toán dự báo dựa vào các yếu tố có ảnh hưởng đến kết quả tuyển sinh Đại học của thí sinh.

NỘI DUNG

Chương 1: Các phương pháp mô hình hóa dữ liệu.

1.1 Phương pháp trực quan.[1]

1.1.1 Quan sát các hoạt động không theo chủ quan

Kỹ thuật khai phá dữ liệu trực quan cung cấp cho người khai phá khả năng đầy đủ để quan sát các hoạt động mà không theo định kiến cá nhân nào cả. Điều đó có nghĩa là ta không cần phải biết là cần phải tìm kiếm cái gì trong thời gian sáp tới. Hơn thế, bạn có thể bắt dữ liệu chỉ ra cho bạn thấy cái gì là quan trọng.

1.1.2 Trực quan và đòi hỏi của nhận thức

Có thể sự mở rộng lớn nhất trong việc sử dụng trực quan trong các phương pháp khai phá dữ liệu là phương pháp trực quan cốt để làm nổi bật khả năng nhận thức, kinh nghiệm của con người có thể làm tốt và một số công việc khác lại làm rất tốt. Việc lựa chọn phương pháp nghiên cứu thường phải có sự cân nhắc về kiểu xử lý thông tin mà người đó đòi hỏi trong suất quá trình nghiên cứu.

1.1.3 Vẽ sơ đồ dữ liệu trên lược đồ trực quan

Khi đưa dữ liệu vào trong một môi trường trực quan, bạn phải quyết định làm sao để trình bày dữ liệu theo một kiểu cách có ý nghĩa. Hoạt động này tập trung vào sử dụng những thuộc tính của các phần tử dữ liệu đã được định nghĩa trong mô hình để xác định làm sao thông tin sẽ được nhìn thấy và cảm nhận. bạn có thể chọn những giải thuật xác định vị trí như gộp nhóm, phân cụm, …

1.2 Phương pháp truyền thống

1.2.1 Phương pháp thống kê

Trong phương pháp này, ta sử dụng những thông tin được thống kê để suy luận và miêu tả xa hơn trong phân tích dữ liệu.

Trong hệ thống hỗ trợ quyết định thì việc dùng phương pháp thống kê là rất phổ biến.

1.2.2 Phương pháp cây quyết định và luật

Cây quyết định là công cụ phân tích để khám phá ra các luật và mối quan hệ bằng phương pháp phân tích thống kê phân chi thành các phần nhỏ các thông tin chứa trong tập dữ liệu.

Cây quyết định là một mô tả tri thức dạng đơn giản nhằm phân các đối tượng dữ liệu thành một số lớp nhất định. Các nút của cây được gán nhãn là tên các thuộc tính, các cạnh được gán các giá trị cụ thể các thuộc tính, các lá miêu tả các lớp khác nhau. Các đối tượng được phân lớp theo các đường đi trên cây, các cạnh tương ứng với giá trị các thuộc tính của đối tượng tới lá.

1.2.3 Các luật kết hợp

Những luật kết hợp được dẫn xuất ra từ sự phân tích các thông tin trùng hợp. Phương pháp luật kết hợp này cho phép khám phá những tương quan, hoặc những biến cố trong giao dịch là các sự kiện.

Các luật kết hợp là một dạng biểu diễn tri thức, hay chính xác hơn là dạng mẫu của hình thành tri thức. Phương pháp này nhằm phát hiện ra các luật kết hợp giữa các hình thành phân dữ liệu trong cơ sở dữ liệu. Mẫu đầu ra của giải thuật khai phá dữ liệu là các tập luật kết hợp tìm được.

1.2.4 Mạng Nơron

Mạng nơron là một hệ thống bao gồm rất nhiều phần tử xử lý đơn giản cùng hoạt động song song. Tính năng hoạt động của hệ thống này phụ thuộc vào cấu trúc của hệ thống, vào cường độ liên kết giữa các phần tử trong hệ thống và dựa vào quá trình xử lý bên trong các phần tử đó. Hệ thống này có thể học từ các dữ liệu có khả năng tổng quát hóa các dữ liệu đó.

1.2.5 Giải thuật di truyền

Giải thuật di truyền được phát triển mô phỏng lại hệ thống tiến hóa trong tự nhiên, chính xác hơn đó là giải thuật chỉ ra tập các cá thể được hình

thành, được ước lượng và biến đổi như thế nào. Giải thuật cũng mô phỏng lại yếu tố gen trong nhiễm sắc thể sinh học trên máy tính để có thể giải quyết nhiều bài toán thực tế khác nhau.

Giải thuật di truyền dựa trên ba cơ chế cơ bản: Chọn lọc, tương giao chéo và đột biến.**[1]**

1.3 Phương pháp khác

1.3.1 Phân nhóm và phân đoạn

Phương pháp phân nhóm và phân đoạn là những kỹ thuật phân chia dữ liệu sao cho mỗi phần hoặc một nhóm giống nhau theo một tiêu chuẩn nào đó.

1.3.2 Phương pháp suy diễn và quy nạp

Một cơ sở dữ liệu là một kho thông tin những các thông tin quan trọng hơn cũng có thể được suy diễn từ kho thông tin đó. Có hai kỹ thuật chính để thực hiện việc này là suy diễn và quy nạp.

Phương pháp suy diễn: Nhằm rút ra những thông tin là kết quả logic của các thông tin trong cơ sở dữ liệu, dựa trên các quan hệ trong dữ liệu.

Phương pháp quy nạp: Nhằm suy ra các thông tin được sinh ra từ cơ sở dữ liệu.

1.3.3 Các phương pháp dựa trên mẫu

Sử dụng các mẫu miêu tả ừ cơ sơ dữ liệu để tạo nên mọt mô hình dự đoán các mẫu mới dằng cách rút ra các thuộc tính tương tự như các mẫu đã biết trong mo hình. Ở đây, nhiệm vụ chính là phải xác định được độ đo giống nhau giữa các mẫu, sau đó mới rạo ra mẫu dự đoán.**[1]**

Chương 2: Mạng Nơron truyền thẳng và thuật toán lan truyền ngược

2.1 Tổng quan về mạng Nơron

2.1.1 Lịch sử phát triển

Khái niệm mạng nơ-ron được đề xuất nhằm mô tả hoạt động của nơron trong bộ não con người. Ý tưởng này bắt đầu được nêu ra trong mô hình tính toán mạng Perceptron. (Perceptron là một bộ phân loại tuyến tính dành cho