ỨNG DỤNG CÁC THUẬT TOÁN MÁY HỌC HIỆN ĐẠI TRONG PHÂN TÍCH DỰ BÁO

Vũ Thanh Nguyên[1], Lư Nhật Vinh[1], Phạm Đại Xuân[1], Lê Phấn Ninh[2]
[1]Sở Khoa học và Công nghệ Thành phố Hồ Chí Minh
[2] Trường Đại học Khoa học Xã hội và Nhân văn – ĐHQG-HCM
(Bài nhận ngày 24 tháng 3 năm 2004)

TÓM TẮT: *Bài báo này đề cập đến một số phương pháp máy học hiện đại, như: phương pháp ứng dụng mạng nơron; phương pháp ứng dụng lý thuyết tập mờ; phương pháp ứng dụng các thuật toán di truyền; .... Việc áp dụng các phương pháp này nhằm đưa ra những dự báo trên cơ sở phân tích các tham số liên quan đến đối tượng nghiên cứu.*

I. Tổng quan về máy học:

Hiện nay trên thế giới, máy học đã được ứng dụng rộng rãi trong nhiều lĩnh vực khác nhau như dự đoán kết quả bầu cử, phân tích dự báo đầu tư, lên kế hoạch phát triển kinh tế, đánh giá kết quả phát triển trong các lãnh vực nông nghiệp, công nghiệp, định hướng phát triển thị trường, xây dựng các mô hình kinh doanh, v.v.. Ở nước ta, lĩnh vực này đang từng bước hình thành và phát triển.

Một số lợi ích của máy học:

- Máy học có thể giúp xử lý và dự báo các thông tin bằng cách tạo ra các luật sản xuất từ dữ liệu thu thập.
- Ở những nơi không có chuyên gia, máy học có thể giúp tạo ra được các quyết định từ các dữ liệu có được.
- Các thuật toán máy học có thể giúp xử lý khi dữ liệu không đầy đủ, không chính xác.
- Máy học giúp thiết kế hệ thống huấn luyện tự động (mạng nơrôn nhân tạo) và giải mã mối liên hệ giữa các tri thức được lưu trữ trong mạng từ dữ liệu.

Máy học là một cụm từ dùng để chỉ khả năng một chương trình máy tính làm tăng tính thực thi dựa trên những kinh nghiệm đã trải qua, hoặc để chỉ khả năng một chương trình có thể phát sinh ra một cấu trúc dữ liệu mới khác với các cấu trúc dữ liệu cũ.

Lợi điểm của các phương pháp máy học là nó phát sinh ra các luật tường minh, có thể được sửa đổi, hoặc được huấn luyện trong một giới hạn nhất định. Các phương pháp máy học hoạt động trên các dữ liệu có đặc tả thông tin. Các thông tin được trình bày theo một cấu trúc gồm 4 mức được gọi là tri thức kim tự tháp (pyramid knowledge). Mức thấp nhất là một tập dữ liệu lớn dùng để lưu thông tin, thường nó rất khó hiểu với con người. Do đó, cần phải trích các thông tin hữu ích từ kho dữ liệu bằng cách sử dụng các phương pháp tiền xử lý. Kết quả của việc tiền xử lý sẽ tạo ra thông tin ở mức thứ 2. Các tri thức ở mức thứ 3 thường được tạo ra do các luật sản xuất. Các siêu tri thức được gọi là tri thức của tri thức.

Máy học là sự tự động của quy trình học và việc học thì tương đương với việc xây dựng những luật dựa trên việc quan sát trạng thái trên cơ sở dữ liệu và những sự chuyển hóa của chúng. Đây là lĩnh vực rộng lớn không chỉ bao gồm việc học từ mẫu, mà còn học tăng cường, học với "thầy",... Các thuật toán học lấy bộ dữ liệu và những thông tin quen thuộc của nó khi nhập và trả về một kết quả, một khái niệm để diễn tả những kết quả học. Máy học kiểm tra những ví dụ trước đó và kiểm tra luôn cả những kết quả của chúng khi xuất và học làm cách nào để tái tạo lại những kết quả này và tạo nên những sự tổng quát hóa cho những trường hợp mới.

Nói chung, máy học sử dụng một tập hữu hạn dữ liệu được gọi là tập huấn luyện. Tập này chứa những mẫu dữ liệu mà nó được viết bằng mã theo một cách nào đó để máy có thể đọc và hiểu được. Tuy nhiên, tập huấn luyện bao giờ cũng hữu hạn do đó không phải toàn bộ dữ liệu sẽ được học một cách chính xác.

Một tiến trình máy học gồm 2 giai đoạn:

- *Giai đoạn học (studying):* hệ thống phân tích dữ liệu và nhận ra sự mối quan hệ (có thể là phi tuyến hoặc tuyến tính) giữa các đối tượng dữ liệu. Kết quả của việc học có thể là: nhóm các đối tượng vào trong các lớp, tạo ra các luật, tiên đoán lớp cho các đối tượng mới.
- *Giai đoạn thử nghiệm (testing):* mối quan hệ (các luật, lớp...) được tạo ra phải được kiểm nghiệm lại bằng một số hàm tính toán thực thi trên một phần của tập dữ liệu huấn luyện hoặc trên một tập dữ liệu lớn.

Các thuật toán máy học được chia làm 3 loại: học giám sát, học không giám sát và học nửa giám sát.

Học có giám sát:

Đây là cách học từ những mẫu dữ liệu mà ở đó các kỹ thuật máy học giúp hệ thống xây dựng cách xác định những lớp dữ liệu. Hệ thống phải tìm một sự mô tả cho từng lớp (đặc tính của mẫu dữ liệu). Nhờ vậy người ta có thể sử dụng các luật phân loại hình thành trong quá trình học và phân lớp để có thể sử dụng dự báo các lớp dữ liệu sau này.

Thuật toán học có giám sát gồm tập dữ liệu huấn luyện M cặp:

$S = \{(x_i, c_j) | i=1,\ldots,M; j=1,\ldots,C\}$

Các cặp huấn luyện này được gọi là mẫu, với

xi là vector n-chiều còn gọi là vector đặc trưng,

cj là lớp thứ j đã biết trước.

Thuật toán máy học giám sát tìm kiếm không gian của những giả thuyết có thể, gọi là H. Đối với một hay nhiều giả thuyết, mà ước lượng tốt nhất hàm không được biết chính xác $f : x \rightarrow c$. Đối với công việc phân lớp có thể xem giả thuyết như một tiêu chí phân lớp. Thuật toán máy học tìm ra những giả thuyết bằng cách khám phá ra những đặc trưng chung của những ví dụ mẫu thể hiện cho mỗi lớp. Kết quả nhận được thường ở dạng luật (Nếu ... thì).

Tùy thuộc vào mức độ của thuật toán học giám sát, người ta có những mô hình học giám sát như sau:

- *Học vẹt: hệ thống luôn luôn được "dạy" những luật đúng.*
- *Học bằng phép loại suy:* hệ thống được dạy phản hồi đúng cho một công việc tương tự, nhưng không xác định. Vì thế hệ thống phải hiệu chỉnh phản hồi trước đó bằng cách tạo ra một luật mới có thể áp dụng cho trường hợp mới.
- *Học dựa trên trường hợp:* trong trường hợp này hệ thống học lưu trữ tất cả các trường hợp, cùng với kết quả đầu ra của chúng. Khi bắt gặp một trường hợp mới, nó sẽ cố gắng hiệu chỉnh đến trường hợp mới này cách xử lý trước đó của nó đã được lưu trữ.
- *Học dựa* trên *sự giải thích:* hệ thống sẽ phân tích tập hợp những giải pháp nhằm chỉ ra tại sao mỗi phương pháp là thành công hay không thành công. Sau khi những giải thích này được tạo ra, chúng sẽ được dùng để giải quyết những vấn đề mới.

Học Không giám sát:

Đây là việc học từ quan sát và khám phá. Hệ thống khai thác dữ liệu được ứng dụng với những đối tượng nhưng không có lớp được định nghĩa trước, mà để nó phải tự hệ thống quan sát những mẫu và nhận ra mẫu. Hệ thống này dẫn đến một tập lớp, mỗi lớp có một tập mẫu được khám phá trong tập dữ liệu.

Học không giám sát còn gọi là học từ quan sát và khám phá. Trong trường hợp chỉ có ít, hay gần như không có tri thức về dữ liệu đầu vào, khi đó một hệ thống học không giám sát sẽ khám phá ra những phân lớp của dữ liệu, bằng cách tìm ra những thuộc tính, đặc trưng chung của những mẫu hình thành nên tập dữ liệu.

Một thuật toán máy học giám sát luôn có thể biến đổi thành một thuật toán máy học không giám sát. Đối với một bài toán mà những mẫu dữ liệu được mô tả bởi n đặc trưng, người ta có thể chạy thuật toán học giám sát *n* -lần, mỗi lần với một đặc trưng khác nhau đóng vai trò thuộc tính lớp, mà chúng ta đang tiên đoán. Kết quả sẽ là n tiêu chí phân lớp (n bộ phân lớp), với hy vọng là ít nhất một trong n bộ phân lớp đó là đúng.

Học nửa giám sát:

Học nửa giám sát là các thuật toán học tích hợp từ học giám sát và học không giám sát. Việc học nửa giám sát tận dụng những ưu điểm của việc học giám sát và học không giám sát và loại bỏ những khuyết điểm thường gặp trên hai kiểu học này.

II. Các thuật toán máy học hiện đại áp dụng trong công tác dự báo:

Nhóm nghiên cứu đã tiến hành nghiên cứu các thuật toán hiện đại của máy học bao gồm: mô hình hệ thống mạng nơrôn mờ hồi quy, mô hình hệ thống nơrôn dạng hàm radial, mô hình hệ luật mờ và sự kết hợp giữa các mô hình mạng nơrôn với thuật toán di truyền cụ thể như sau:

1. Mô hình mạng nơrôn mờ hồi quy:

a. Giới thiệu:

Mô hình mạng nơrôn này kết hợp từ lý thuyết tập mờ và mô hình mạng nơrôn tận dụng những ưu điểm như có khả năng xấp xỉ một hàm liên tục với độ chính xác cho trước (mạng nơrôn) và khai thác khả năng xử lý những tri thức như con người (lý thuyết tập mờ). Mô hình này rất thích hợp do yêu cầu đặt ra có những đầu vào và đầu ra phụ thuộc theo thời gian. Mạng nơ ron mờ hồi quy tỏ ra đạt hiệu quả cao cho những ứng dụng như: dự báo chuỗi thời gian, nhận dạng và điều khiển những hệ phi tuyến.

b. Cấu trúc mạng:

Một mạng gồm 4 lớp như sau

Lớp 1: Là lớp nhập gồm N dữ liệu nhập (input).

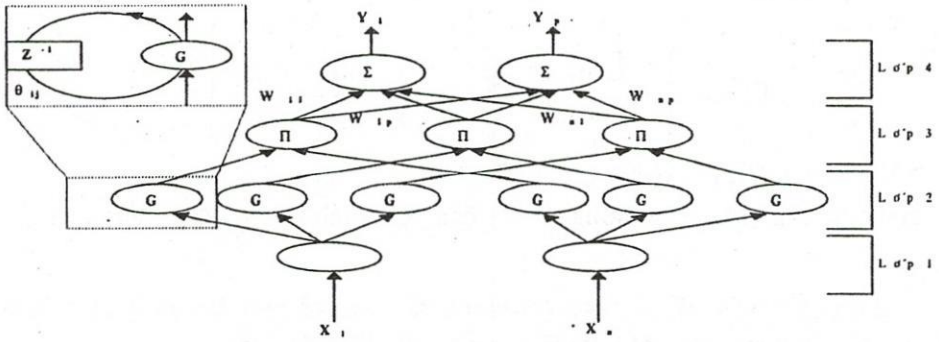
Lớp 2: Gọi là lớp các hàm thành viên. Các nút trong lớp này thực hiện việc mờ hoá. Lớp này dùng để tính giá trị hàm thành viên theo hàm phân phối Gauss. Số nút trong lớp 2 là N x M, trong đó M là số luật mờ (số nút của lớp 3)

Lớp 3: Lớp các luật mờ. Các nút trong lớp này tạo thành cơ sở luật mờ gồm M nút. Liên kết giữa lớp 2 và lớp 3 biểu diễn giả thiết của luật mờ. Liên kết giữa lớp 3 và lớp 4 biểu diễn kết luận của luật mờ.

Lớp 4: Lớp xuất gồm P nút.

Liên kết giữa lớp 3 và lớp 4 được gán trọng số w_{jk}.

Như vậy số nút của mô hình là: N + (N x M) + M + P



Cấu trúc mạng

c. Hoạt động của mô hình.

Ký hiệu $u_i^{(k)}$ và $O_i^{(k)}$ tương ứng là đầu vào và đầu ra của nút thứ i trong lớp k.

Lớp 1

$$O_i^{(1)} = u_i^{(1)} = x_i(t) \qquad i = 1 \div N$$

Lớp 2

$$O_{ij}^{(2)} = \exp\left[-\frac{\left(u_{ij}^{(2)} - m_{ij}\right)^2}{\left(\sigma_{ij}\right)}\right] \qquad i = 1 \div N,\ j = 1 \div M$$

Trong đó m_{ij} và σ_{ij} tương ứng là trọng tâm và độ rộng của hàm thành viên theo phân bố Gauss.

Hơn nữa, đầu và) của các nút này là

$$u_{ij}^{(2)}(t) = O_i^{(1)} + \theta_{ij} O_{ij}^{(2)}(t-1) \qquad i = 1 \div N, j = 1 \div M$$

Trong đó biểu diễn trọng số cho các nút hồi tiếp.

Dễ dàng thấy rằng đầu vào của các nút trong lớp này có chứa toán hạng $O_{ij}^{(2)}(t-1)$ lưu thông tin trước đó của mô hình. Và đây chính là sự khác biệt giữa mạng nơron mờ và mạng nơrôn hồi quy mờ.

Như vậy

$$O_{ij}^{(2)} = exp\left[-\frac{\left[O_i^{(1)} + \theta_{ij} O_{ij}^{(2)}(t-1) - m_{ij}\right]^2}{(\sigma_{ij})^2}\right] \qquad i = 1 \div N, j = 1 \div M$$

$$= exp\left[-\frac{\left[x_i(t) + \theta_{ij} O_{ij}^{(2)}(t-1) - m_{ij}\right]^2}{(\sigma_{ij})^2}\right]$$

Mỗi nút trong lớp này có 3 thông số là m_{ij}, σ_{ij} và θ_{ij}.

Lớp 3: Các nút trong lớp này thực hiện phép toán AND

$$O_j^{(3)} = \prod_{i=1}^{N} O_{ij}^{(2)} \qquad i = 1 \div N, j = 1 \div M$$

$$= \prod_{i=1}^{N} exp\left[-\frac{\left[x_i(t) + \theta_{ij} O_{ij}^{(2)}(t-1) - m_{ij}\right]^2}{(\sigma_{ij})^2}\right]$$

Lớp 4: Các nút trong lớp này thực hiện việc giải mờ.

$$y_k = O_k^{(4)}$$

$$= \sum_{j=1}^{M} u_{jk}^{(4)} w_{jk}$$

$$= \sum_{j=1}^{M} O_j^{(3)} w_{jk}$$

$$= \sum_{j=1}^{M} w_{jk} \prod_{i=1}^{N} exp\left[-\frac{\left[x_i(t) + \theta_{ij} O_{ij}^{(2)}(t-1) - m_{ij}\right]^2}{(\sigma_{ij})^2}\right]$$

Với i = 1÷N, j = 1÷M, k = 1÷P

Như vậy, trong mô hình này, các thông số cần phải xác định là θ_{ij}, σ_{ij}, θ_{ij} và w_{jk}.

d. Lập luận mờ.

Giả sử cho hệ mạng nơrôn hồi quy mờ với nhiều đầu vào và một đầu ra. Gọi xi là biến ngôn ngữ thứ i và α_j là giá trị kích hoạt của luật j, wj là trọng số của kết nối thứ j.

Một luật suy diễn mờ được biểu diễn như sau

Rj : Nếu u1j là A1j, u2j là A2j, .. , unj là Anj thì y=wj

Trong đó

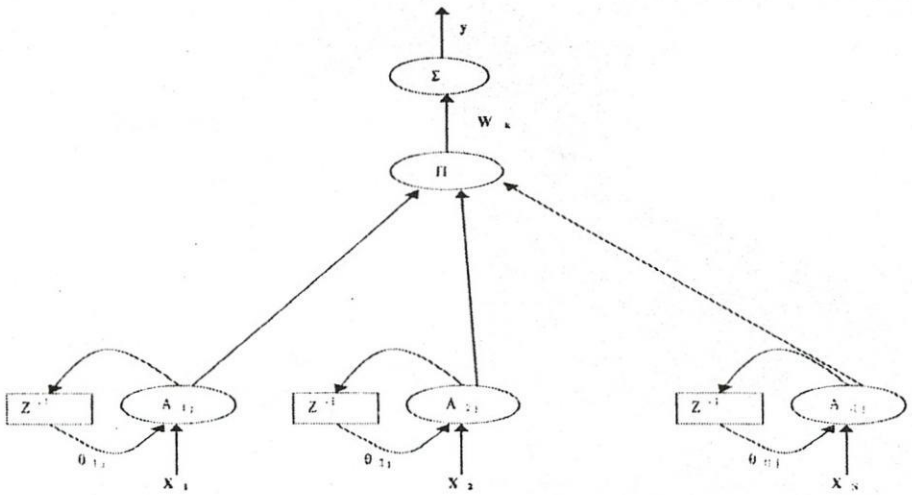
i = 1, 2, .. , n. n là số đầu vào

uij = xi + θ_{ij} * oij$^{(2)}$(t-1)

Aij là các tập mờ

Wj là trọng số kết nối

Đầu vào của mỗi hàm thành viên là đầu vào xi của mạng cộng với số hạng oij(2)θ_{ij}. Sơ đồ kết nối dựa trên luật suy diễn mờ như hình sau



Một cấu trúc kết nối trên cơ sở luật mờ thứ j

Hệ thống mờ với những thành phần nhớ (Feedback unit) có thể được xem như là một hệ suy luận mờ động và giá trị suy luận được tính bởi:

Với M là số luật

$$y^* = \sum_{j=1}^{M} w_j \alpha_j \quad \text{và} \quad \alpha_j = \prod_{i=1}^{N} \mu A_{ij}(u_{ij})$$

Từ mô tả trên, mạng nơrôn hồi quy mờ là một hệ suy luận mờ có các phần tử nhớ. Sau khi huấn luyện xong, các thông số trong mạng cùng với các phần tử nhớ đã xác định tri thức.

e. Giải thuật học cho mô hình.

RFNNs sử dụng giải thuật học lan truyền ngược (Back Propagation) để xác định các thông số trong mô hình.

Mục đích của việc học là cực tiểu hóa hàm lỗi

$$E = \frac{1}{2}\sum_t \left(y^{(d)}(t) - y\ (t)\right)^2 = \frac{1}{2}\sum_t \left(y^{(d)}(t) - O^{(4)}(t)\right)^2$$

Trong đó $y^{(d)}(t)$ là đầu ra mong muốn và $y\ (t) = O^{(4)}(t)$ là đầu ra hiện tại của mô hình tại mẫu dữ liệu thứ t.

Như đã biết, trong giải thuật lan truyền ngược, các thông số sẽ được cập nhật như sau:

$$W\ (t+1) = W\ (t) + \Delta W\ (t) = W\ (t) + \eta\left(-\frac{\partial E\ (t)}{\partial W}\right)$$

Trong đó W là vector thông số của mô hình và η là tốc độ học.

Đặt $e(t) = y^{(d)}(t) - y\ (t)$, ta có

$$\frac{\partial E\ (t)}{\partial W} = -e(t)\frac{\partial y(t)}{\partial W} = -e(t)\frac{\partial O^{(4)}(t)}{\partial W}$$

Do đó các thông số m_{ij}, σ_{ij}, θ_{ij} và w_{jk} sẽ được cập nhật tương ứng như sau

$$w_{jk}(t+1) = w_{jk}(t) - \eta^w \frac{\partial E}{\partial w_{jk}}$$

$$m_{ij}(t+1) = m_{ij}(t) - \eta^m \frac{\partial E}{\partial m_{ij}}$$

$$\sigma_{ij}(t+1) = \sigma_{ij}(t) - \eta^\sigma \frac{\partial E}{\partial \sigma_{ij}}$$

$$\theta_{ij}(t+1) = \theta_{ij}(t) - \eta^{\theta} \frac{\partial E}{\partial \theta_{ij}}$$

Trong đó

$$\frac{\partial E}{\partial w_{jk}} = -e(t) O_j^{(3)}$$

$$\frac{\partial E}{\partial m_{ij}} = -e(t) \sum_{j=1}^{M} w_{jk} \frac{\partial O_j^{(3)}}{\partial m_{ij}}$$

$$= -e(t) \sum_{j=1}^{M} w_{jk} O_j^{(3)} \frac{2\left[x_i(t) + O_{ij}^{(2)}(t-1)\theta_{ij} - m_{ij}\right]}{(\sigma_{ij})^2}$$

$$\frac{\partial E}{\partial \sigma_{ij}} = -e(t) \sum_{j=1}^{M} w_{jk} \frac{\partial O_j^{(3)}}{\partial \sigma_{ij}}$$

$$= -e(t) \sum_{j=1}^{M} w_{jk} O_j^{(3)} \frac{2\left[x_i(t) + O_{ij}^{(2)}(t-1)\theta_{ij} - m_{ij}\right]^2}{(\sigma_{ij})^3}$$

$$\frac{\partial E}{\partial \theta_{ij}} = -e(t) \sum_{j=1}^{M} w_{jk} \frac{\partial O_j^{(3)}}{\partial \theta_{ij}}$$

$$-e(t) \sum_{j=1}^{M} w_{jk} \frac{-2\left[x_i(t) + O_{ij}^{(2)}(t-1)\theta_{ij} - m_{ij}\right] O_{ij}^{(2)}(t-1)}{(\sigma_{ij})^2}$$

Giải thuật học cho mạng như sau:

Học có giám sát với tập mẫu {(Xs, Ys)}

Mỗi khi đưa một mẫu Xs = (x1, x2, .., xn) vào mạng, các công việc lần lược được thực hiện như sau:

- Lan truyền mẫu Xs qua mạng để tính giá trị đầu ra (Output) theo công thức:

$$y_k = \sum_{j=1}^{M} w_{jk} \prod_{i=1}^{N} \exp\left[- \frac{\left[x_i(t) + \theta_{ij} O_{ij}^{(2)}(t-1) - m_{ij}\right]^2}{(\sigma_{ij})^2} \right]$$

- Tính lỗi tại mẫu học dựa trên sai lệch

$$e(t) = y^{(d)}(t) - y(t)$$

- Cập nhật các thông số mij, σ_{ij}, θ_{ij} và wij theo công thức

$$w_{jk}(t+1) = w_{jk}(t) - \eta^{w} \frac{\partial E}{\partial w_{jk}}$$

$$m_{ij}(t+1) = m_{ij}(t) - \eta^{m} \frac{\partial E}{\partial m_{ij}}$$

$$\sigma_{ij}(t+1) = \sigma_{ij}(t) - \eta^{\sigma} \frac{\partial E}{\partial \sigma_{ij}}$$

$$\theta_{ij}(t+1) = \theta_{ij}(t) - \eta^{\theta} \frac{\partial E}{\partial \theta_{ij}}$$

2. Mô hình hệ thống mạng nơrôn dạng hàm radial.

a. Giới thiệu.

Hệ thống Mạng Nơrôn dạng hàm radial sẽ giải quyết vấn đề xấp xỉ một hàm liên tục n biến trên một miền compact. Mô hình này tiến hành lấy đặc trưng cục bộ của hàm, và như vậy sẽ dễ dàng khởi tạo và huấn luyện dữ liệu khi học.

b. Cấu trúc mạng.

Một mạng RBFNNs gồm có 3 lớp

- Lớp đầu vào
- Lớp các hàm Gauss (số nút là do người sử dụng quy định)
- Lớp đầu ra
- Các liên kết từ tầng lớp đầu vào đến tầng các hàm gauss không có trọng số
- Các liên kết ở tầng các hàm Gauss đến tầng lớp đầu ra có trọng số.
- Mỗi node ở tầng các hàm gauss có các thông số cần xác định là: trọng tâm (xác định trọng tâm hàm gauss) , thông số sigma (xác định độ lệch chuẩn của hàm gauss). Xác định các thông số ở tầng này dùng để phân lớp.
- Các trọng số trên đường liên kết từ tầng các hàm Gauss đến tầng lớp đầu ra được xác định thông qua cách học bình thường: phương pháp học lan truyền ngược, phương pháp học tuyến tính, phương pháp học theo vết cũ.

c. Hoạt động của mô hình.

Mỗi mẫu dữ liệu nhập sẽ qua k hàm gauss (giả sử ở tầng này có k nút hàm gauss), hay có thể hiểu là có k lớp, xem mẫu thuộc vào lớp nào qua tính xác xuất phân bố chuẩn (là hàm gauss của các lớp). Các giá trị tính được này được tổ hợp tuyến tính (tính trung bình có trọng số).

d. Giải thuật học cho mạng.

Giải thuật 1 (ước lượng mạng)

Bước 1: đặt y=0

Bước 2: Thực hiện bước lặp khi i từ1 tới n

{

a. ui = int[pi(xi)]
b. nếu ui < 1, đặt ui=1
c. nếu ui > di – 1, đặt ui = di - 1

}

Bước 3: Thực hiện bước lặp khi j từ 0 tới 2n – 1

{

a. Chuyển hoá j từ hệ mười sang hệ nhị phân n-bit dạng c = cn-1cn-2..c0
b. For i=1 to n tính qi=ui + c_{i-1}
c. Chuyển hoá q to k
d. Tính y = y + (wk + vkx)$\phi_k(x)$

}

Giải thuật 2 (huấn luyện mạng)

Bước 1: Đặt Emax >=0 và pmax >=1. Khởi tạo V và w. Đặt p = 0.

Bước 2: Thực hiện vòng lặp như sau

{

a. Set E = 0 and p = p +1
b. Thực hiện bước lặp khi h từ 1 tới ns

{

i. Tính E = f(sh) – f_1(sh)
ii. Tính ui = int[pi(sih)], 1<=i<=n
iii. Thực hiện bước lặp khi j từ 0 tới 2^n-1

{

a) Chuyển hoá số j từ dạng hệ mười sang hệ hai n-bit, biểu diễn dưới dạng c = $cn_{-1}cn_{-2}..c_0$

b) Tính qi = ui + c_{i-1} khi 1<=i<=n

c) Chuyển số q tới k

d) Tính

{

$wk = wk + \mu E\phi_k(sh)$

$vk = vk + \mu E\, \phi_k(sh)(sh - xk)^T$

}

}

c. Thực hiện bước lặp khi k từ 1 tới nt

Tính

$E = E = \max\{E, |f(tk)-f_1(tk)|\}$

}

}

Bước 3: Chuyển sang bước 2 khi (E>Emax) và (p<pmax)

III. Các chương trình dự báo:

Nhóm nghiên cứu tiến hành cài đặt các chương trình máy tính dựa trên các thuật toán máy học hiện đại đã được giới thiệu ở trên và những dữ liệu thử nghiệm cho các chương trình này.Các chương trình có các chức năng như huấn luyện dữ liệu, thử nghiệm dữ liệu, dự báo dữ liệu trong khoảng thời gian thực. Các chương trình cho phép biểu diễn lỗi, dữ liệu dưới dạng đồ thị hoặc dạng số khi huấn luyện, thử nghiệm hoặc dự báo. Các chương trình dự báo mức độ tăng trưởng cao nhất, thấp nhất và tốc độ tăng trưởng trung bình của dữ liệu cần dự báo.

Ngoài ra để kiểm tra và so sánh khả năng dự báo của các phương pháp máy học so với các phương pháp đang được ứng dụng rộng rãi trong kinh tế, các chương trình được cài đặt thêm phương pháp dự báo chuỗi ARIMA, là mô hình phân tích dự báo kinh tế cổ điển đang được sử dụng rộng rãi trong các ngành dự báo của kinh tế ở Việt Nam và trên thế giới, nhằm đối chiếu và so sánh các phương pháp dự báo của máy học với phương pháp dự báo chuỗi ARIMA.

Dữ liệu thực nghiệm.

Chúng tôi lựa chọn phương án thử nghiệm phương pháp dự báo này bằng cách sử dụng dữ liệu của giá cả một số mặt hàng quan trọng, cụ thể như: ***vàng, đô la, gạo, cà phê, xi măng.*** Đây là các mặt hàng quan trọng trong nền kinh tế VN, sự biến động giá của chúng có tác động rất lớn đến các hoạt động kinh tế khác (nguồn cung cấp từ Viện Kinh Tế Thành Phố Hồ Chí Minh).

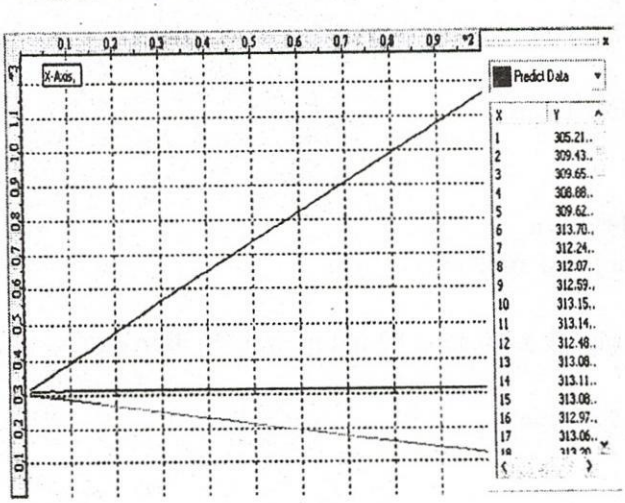
Dữ liệu về giá cả các mặt hàng này được cập nhật hàng ngày, và có thể sử dụng dữ liệu từ nhiều năm trước, từ năm 1994 đến 2001, tức bao gồm khoảng hơn 2000 số liệu cho mỗi bộ dữ liệu của mỗi loại mặt hàng thử nghiệm.

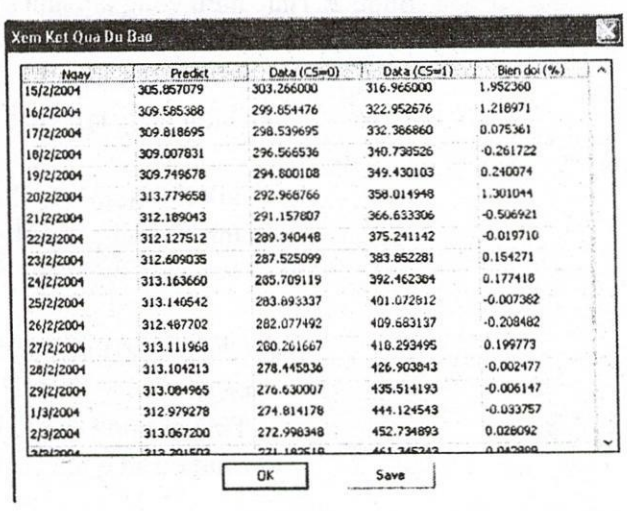
1. Kết quả thử nghiệm.

a. Ứng dụng mô hình mạng nơrôn hồi quy mờ.

- Kết quả thử nghiệm (có đối chiếu với mô hình ARIMA)

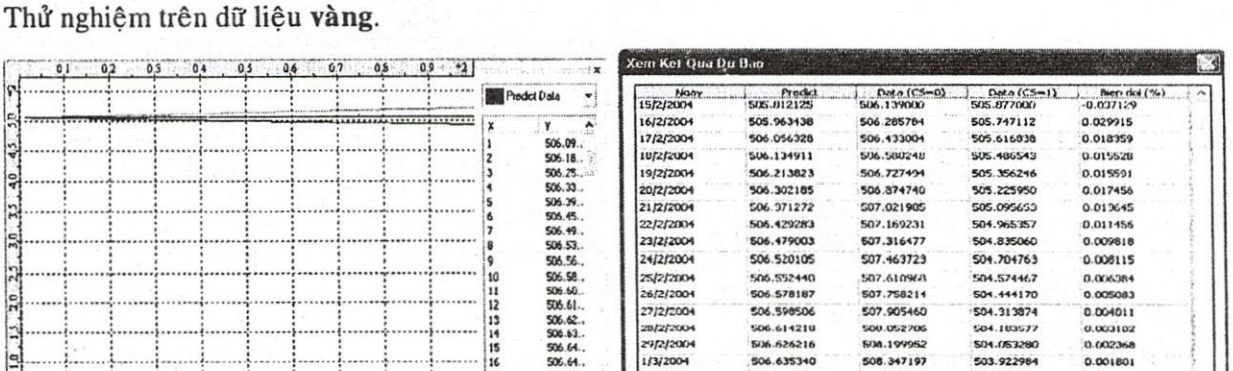
Thử nghiệm trên dữ liệu **café**.





(*Lưu ý:* các đường màu xanh biểu diễn mô hình dữ liệu dự báo theo phương pháp ARIMA, đường màu đỏ theo phương pháp dự báo của máy học).

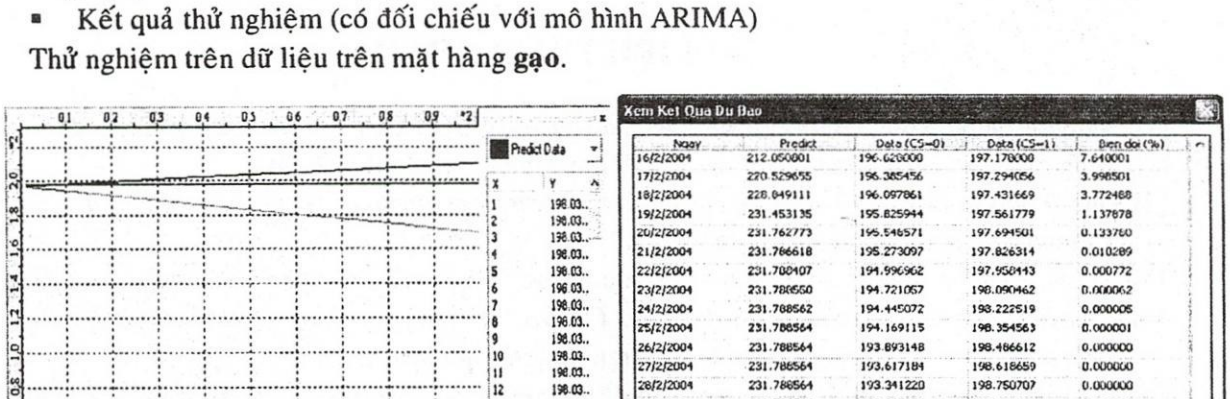
Thử nghiệm trên dữ liệu **vàng**.



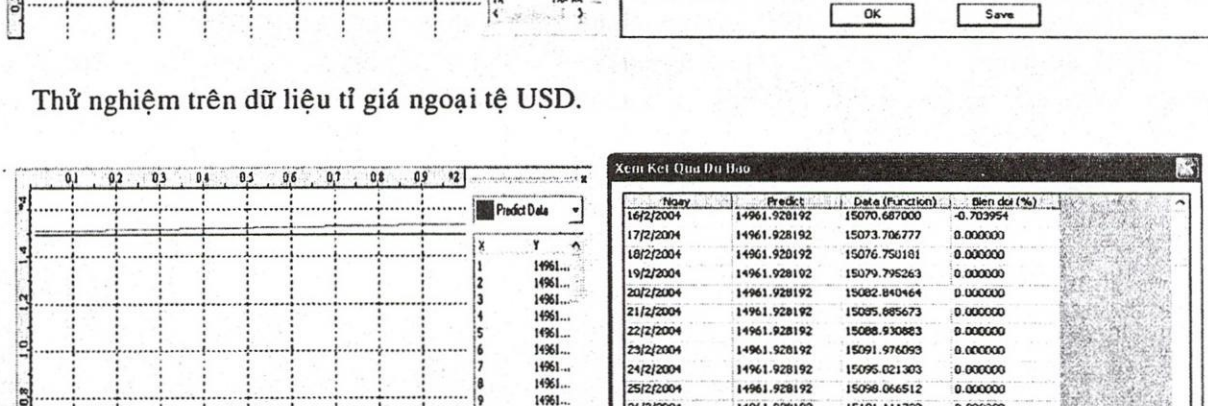
b. Ứng dụng mô hình nơrôn dạng hàm radial.

- Kết quả thử nghiệm (có đối chiếu với mô hình ARIMA)

Thử nghiệm trên dữ liệu trên mặt hàng **gạo**.



Thử nghiệm trên dữ liệu tỉ giá ngoại tệ USD.



IV. Kết luận:

Kết quả thử nghiệm các loại dữ liệu trên so sánh với các phương pháp dự báo đã được sử dụng trước đây (như phương pháp ARIMA, các phương pháp định tính và định lượng khác) có thể chấp nhận được. Tuy nhiên để có thể kiểm tra thêm các thuật toán máy học hiện đại ứng dụng cho công tác

phân tích dự báo đối với trung hạn và dài hạn cần phải có thêm nhiều dữ liệu hơn nữa mới có thể đánh giá hết được tính ổn định và độ chính xác của các thuật toán máy học hiện đại đang thử nghiệm.

APPLICATION OF MODERN LEARNING MACHINE ALGORITHMS IN ANALYZING FORECASTS

Vu Thanh Nguyen[(1)], Lu Nhat Vinh[(1)], Pham Dai Xuan[(1)], Le Phan Ninh[(2)]
[(1)]Department of Sciences and Technology – HoChiMinh City
[(2)]University of Social Sciences and Humanities – VNU-HCM

ABSTRACT: *This paper mentions a number of modern learning machine methods such as neuron – network method; fuzzy – set method; genetic algorithm methods; etc .... These methods are applied to present forecasts based on analyzing parameters which relate to studying objects.*

TÀI LIỆU THAM KHẢO

[1] Nguyễn Quốc Tòng. Báo cáo về các phương pháp định tính và định lượng được ứng dụng trong các công tác phân tích dự báo của Viện Kinh tế thành phố. 2000.

[2] Duc Truong Pham, Liu Xing. *Neural Networks for Identication, Prediction and Control.* Springer – Verlag London Limited 1998.

[3] Nguyễn Thống. Phân tích dữ liệu và áp dụng vào dự báo. Nhà xuất bản Thanh niên.1999.

[4] C-H. Lee, C.-C Teng. *Identìication and Control of Dynamic Systems Using Recurent Fuzzy Neural Networks.* p.349 – p.366. IEEE Transactions on Fuzzy System 08/2000.

[5] F-J.Lin, R.-J. Wai. *Hybird Control Using Recurrent Fuzzy Neural Network for Linear Induction Motor Servo Drive.* p.102 – p.115. IEEE Transactions on Fuzzy System 02/2001.

[6] P.Tino, C.Schittenkopf, G.Dorffner. *Financial Volatility Trading Using Recurrent Neural Network.* p. 865 – p.874. IEEE Transactions on Neural Networks 07/2001.

[7] S. Seshagiri, H. K. Khalil. *Output Feedback Control of Nonlinear Systems Using RBF Neural Networks.* p. 69 – p.79. IEEE Transactions on Neural Networks 01/2000.