Kỷ yếu Hội nghị Quốc gia lần thứ VIII về Nghiên cứu cơ bản và ứng dụng Công nghệ thông tin (FAIR); Hà Nội, ngày 9-10/7/2015
DOI: 10.15625/vap.2015.000182

MỘT SỐ VẤN ĐỀ VỀ DỰ BÁO DỮ LIỆU CHUỖI THỜI GIAN

Trần Đức Minh (*), Trần Huy Dương (*), Vũ Đức Thi (**)
(*) Phòng Công nghệ phần mềm trong quản lý, Viện CNTT, Viện Hàn lâm Khoa học và Công nghệ Việt Nam
(**) Viện CNTT, Đại học Quốc gia Hà Nội

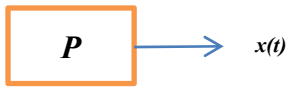
TÓM TẮT *- Dự báo dữ liệu chuỗi thời gian (time series prediction) là một bài toán khá phức tạp, bao gồm nhiều kỹ thuật áp dụng trong thực tế. Trong bài báo này chúng tôi phân tích các cách tiếp cận lựa chọn mô hình và quy trình áp dụng dự báo dữ liệu chuỗi thời gian tập trung vào ứng dụng mạng nơron trong việc dự báo dữ liệu dạng này.*

I. GIỚI THIỆU

Dữ liệu chuỗi thời gian (time series) được hiểu là một dãy các vector (hoặc số thực) phụ thuộc vào thời gian:

$\{x(t_0), x(t_1),\ldots, x(t_{i-1}), x(t_i), x(t_{i+1}), \ldots\}$

Trong đó, việc phân tích dữ liệu chuỗi thời gian trong bài báo này là việc tìm ra một hộp đen *P*, có khả năng tạo ra các giá trị *x(t)* dựa trên các dữ liệu đã thu thập trước đó [2].



Trong thực tế, có thể thấy có nhiều ví dụ về dữ liệu chuỗi thời gian như: dữ liệu sử dụng điện của một thành phố, quốc gia; số lượng trẻ em mới sinh trong khoảng thời gian; dữ liệu sử dụng băng thông của nhà cung cấp dịch vụ internet,… Về cơ bản có thể chia dữ liệu chuỗi thời gian thành hai dạng: rời rạc hoặc liên tục.

Các dữ liệu rời rạc, chỉ các chuỗi dữ liệu có thời gian thu thập dữ liệu không liền mạch, chẳng hạn như dữ liệu đóng cửa sàn giao dịch chứng khoán. Các dữ liệu liện tục được thu thập theo khoảng thời gian liên tục, bằng nhau, chẳng hạn dữ liệu sử dụng băng thông của nhà cung cấp dịch vụ internet.

Trong trường hợp dữ liệu liên tục, *t* là thời gian thực và *x(t)* là các dữ liệu liên tục, để lựa chọn chuỗi *x(t)*, ta phải lấy dữ liệu tại các điểm rời rạc. Nếu lấy mẫu đồng bộ (uniform), giả sử thời gian lấy mẫu là Δt thì chuỗi thời gian được biểu diễn như sau:

$\{x[t]\} = \{x(0), x(\Delta t) , x(2\Delta t), x(3\Delta t),\ldots\}$

Để đảm bảo *x(t)* có thể nhận được từ *x[t]*, Δt cần được lựa chọn tuân theo *Nyquist sampling theorem* [11].

Bên cạnh đó, dữ liệu chuỗi thời gian cũng có thể phân loại theo dạng đơn điệu (deterministic) hoặc không đơn điệu (stochastic) hay tuyến tính hoặc phi tuyến tính,…

Mạng nơron được coi như là bộ xấp xỉ đa năng, có khả năng giải quyết các bài toán dự báo trong thực tế [1]. Đặc điểm của mạng nơron cho phép hoạt động trên các dữ liệu phi tuyến tính, không cần hiểu biết trước về các mỗi quan hệ của dữ liệu đầu vào.

Trong bài báo này chúng tôi phân tích các cách lựa chọn mô hình cũng như phương pháp dự báo, tập trung vào sử dụng mạng nơron giải quyết bài toán dự báo chuỗi thời gian. Chúng tôi cũng phần tích làm rõ các khía cạnh thực tế khi áp dụng phương pháp này.

II. BÀI TOÁN DỰ BÁO CHUỖI THỜI GIAN SỬ DỤNG MẠNG NƠ RON

Giả sử ta có chuỗi thời gian $\{x[t]), x[t-1],\ldots\}$ tính đến thời điểm t , nhiệm vụ của chúng ta là dự báo giá trị của x tại một thời điểm trong tương lai.

$$x_{db}[t+s] = \boldsymbol{f}(x[t], x[t-1],\cdots)$$

s: khoảng dự đoán (horizon of prediction)

trong trường hợp $s = 1$, nghĩa là ta chỉ dự báo 01 giá trị tại tương lai, khi đó, bài toán rơi vào trường hợp tìm ra một hàm xấp xỉ (function approximation) biểu diễn chuỗi thời gian, nói cách khác là dự đoán giá trị tương lai từ các giá trị đã thu thập trước đó trong chuỗi thời gian.

Để giải quyết bài toán dự báo chuỗi thời gian nói chung và sử dụng mạng nơron nói riêng, cần thực hiện các bước tổng quát sau:

- ⇨ chọn mô hình tổng quát
- ⇨ với mỗi $x[t_i]$ trong quá khứ, huấn luyện mô hình với đầu vào là các giá trị trước đó và đầu ra mong muốn, là chính t_i.
- ⇨ sau khi huấn luyện mô hình, chạy mô hình với chuỗi $\{x[t], x[t-1],\cdots\}$ để thu được giá trị dự đoán $x_{db}[t+s]$.

III. MỘT SỐ MÔ HÌNH ỨNG DỤNG

Trong thời điểm ban đầu, việc giải bài toán dự báo chuỗi thời gian, dự báo được thực hiện bằng phương pháp làm trơn và ngoại suy chuỗi dữ liệu thời gian thông qua việc làm khớp toàn cục (*global fit*) trên miền thời gian. Sau này, phương pháp nói trên được thay thế bởi sự xuất hiện các mô hình chuỗi thời gian tuyến tính (linear) với các đặc điểm tích cực: dễ hiểu để phân tích dữ liệu và rất dễ để thực hiện. Điểm chưa tốt là chúng làm việc không hiệu quả với các chuỗi thời gian phi tuyến (non-linear) [2]. Do vậy, các mô hình phi tuyến dần được nghiên cứu và áp dụng đối với các chuỗi thời gian phi tuyến tính, với mức độ phức tạp cao.

3.1. Mô hình tuyến tính

Đối với các hệ thống tuyến tính (Linear systems), thuộc phạm vi nghiên cứu của lĩnh vực xử lý tín hiệu số (Digital Signal Processing - DSP). DSP quan tâm đến các thao tác tuyến tính, chuyển dịch trạng thái trên dòng dữ liệu. Các thao tác này được thực hiện bởi các bộ lọc. Việc phân tích, thiết kế các bộ lọc một cách hiệu quả là cốt lõi của lĩnh vực này.

Các mô hình tuyến tính biểu diễn chuỗi thời gian như một tổ hợp tuyến tính của các biến thời gian trễ và có thể có hoặc không có việc kết hợp thêm một đại lượng khác là tổ hợp tuyến tính của các số hạng của quá trình nhiễu trắng (*white noise)*. Các mô hình tuyến tính tiêu biểu bao gồm: AR (auto regressive – tự hồi quy), MA (moving average – trung bình trượt) và ARMA (autoregressive-moving average – Tự hồi quy và trung bình trượt).

a. Mô hình tự hồi quy (AR)

Trong mô hình tự hồi quy, chuỗi thời gian $\{X_t\}$ được mô tả bởi phương trình sau:

$$X_t = \phi_1 X_{t-1} + \phi_2 X_{t-2} + \cdots + \phi_p X_{t-p} + \varepsilon_t$$

Trong đó:

$\phi_{i:1\to p}$ là các tham số của mô hình

ε_t: nhiễu trắng (white noise)

Phương trình này được gọi là phương trình biểu diễn của mô hình tự hồi quy bậc p (AR($\boldsymbol{p}$)).

b. Mô hình trung bình di động (MA)

Chuỗi thời gian $\{X_t\}$ được gọi là quá trình trung bình di động bậc q (MA($\boldsymbol{q}$)) nếu như mỗi quan sát X_t của quá trình MA(q) được viết dưới dạng như sau:

$$X_t = \varepsilon_t + \theta_1 \varepsilon_{t-1} + \theta_2 \varepsilon_{t-2} + \cdots + \theta_q \varepsilon_{t-q}$$

Với $\{\varepsilon_t\}$ là một quá trình nhiễu trắng (white noise) với trung bình bằng 0, $\theta_{i:1\to q}$ là các tham số của mô hình.

Phương trình trên cho thấy mô hình MA hoạt động mà không cần thông tin phản hồi. Có nhiều chuỗi thời gian được làm khớp dựa hoàn toàn trên các thông tin phản hồi, điều này được thực hiện thông qua mô hình tự hồi quy AR.

c. Mô hình tự hồi quy và trung bình trượt (ARMA)

Các chuỗi thời gian đôi khi không thể mô hình hóa được bằng MA hay AR do chúng có đặc tính của cả hai quá trình này. Khi đó, để biểu diễn, người ta sử dụng mô hình ARMA, là pha trộn của cả hai mô hình MA và AR.

Khi đó, quá trình ARMA*(p,q)* được mô tả như sau:

$$X_t = \phi_1 X_{t-1} + \cdots + \phi_p X_{t-p} + \varepsilon_t + \theta_1 \varepsilon_{t-1} + \cdots + \theta_q \varepsilon_{t-q}$$

Lúc này, việc dự báo có thể thực hiện được nhờ xác định *p và q*. Việc xác định này được thực hiện bởi người thực hiện dự báo thông qua kinh nghiệm. Trong đó, *p* được xác định dựa trên việc vẽ các hàm tự tương quan một phần (partial autocorrelation functions), đồng thời *q* được xác định thông qua các hàm tự tương quan (autocorrelation functions). Điều quan trọng là các mô hình này có thể giải thích được kết quả dự báo thông qua các công cụ trình diễn trên máy tính.

3.2. Mô hình phi tuyến tính

Để mô tả các quá trình phi tuyến tính, các mô hình này giả thiết dữ liệu chuỗi thời gian là phi tuyến tính. Điều này phù hợp với thực tế rằng các chuỗi thời gian không thể biết trước chúng có đặc tính là tuyến tính hay phi tuyến tính. Tuy nhiên, đặc điểm của mô hình này là sử dụng rất nhiều tham số xây dựng mô hình và do đó, rất khó giải thích quá trình xác định các tham số của mô hình. Vì đặc tính này, các mô hình phi tuyến tính được coi như quá trình hộp đen.

Dưới đây trình bày một số mô hình tiêu biểu sử dụng để dự báo dữ liệu chuỗi thời gian, theo [2].

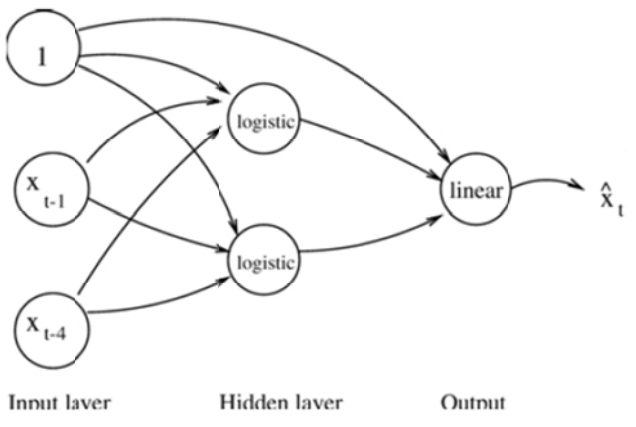
a. Mô hình Markov ẩn (Hidden Markov Model)

Mô hình Markov ẩn (HMM) cũng được sử dụng để dự báo dữ liệu chuỗi thời gian [5]. Tuy vậy, mô hình này không thích hợp để giải quyết các vấn đề liên quan đến dữ liệu liên tục. Do vậy, các mô hình HMM đã được hiệu chỉnh để sử dụng trong giải quyết bài toán dự báo chuỗi thời gian. Theo đó, mô hình toán học của nó trở nên quá phức tạp để áp dụng thuật toán forward-backward xác định các tham số, độ phức tạp của giải thuật này là O(N2), nên rất khó mở rộng cho các tập dữ liệu kích thước lớn.

Cũng có vài phương pháp khác không thông dụng để dự báo phi tuyến. Một trong số đó được gọi phương pháp Analogues [6]. Cách tiếp cận này khá đơn giản và chỉ có vài tham số tự do nhưng chỉ áp dụng cho các chu kỳ thời gian ngắn.

b. Mạng nơron nhân tạo

Việc sử dụng mạng nơron nhân tạo để dự báo chuỗi thời gian đã được nghiên cứu nhiều, do đặc điểm rất phù hợp với các dữ liệu phi tuyến tính. Có nhiều vấn đề trong việc xây dựng mạng nơron nhân tạo áp dụng trong dự báo dữ liệu như được nêu ở [1][7][8][9]. Trong phạm vi bài báo này, chúng tôi mô tả cách xây dựng mô hình sử dụng mạng nơron để thực hiện dự báo chuỗi thời gian.



Theo đó, các quan sát *x[t-s]* được sử dụng làm đầu vào để dự báo giá trị $x_{db}[t]$. Người ta sẽ xây dựng tập dữ liệu huấn luyện mạng bằng phương pháp như sau:

- ⇨ Chuẩn hóa dữ liệu.
- ⇨ Xác định khoảng dự báo (horizon of prediction) *s*.
- ⇨ Chia tập dữ liệu ban đầu thành các tập: huấn luyện (training) (> 50% số mẫu), kiểm tra (test) (10% -> 30% số mẫu) và tập kiểm định (validation).
- ⇨ Xây dựng tập dữ liệu với mẫu đầu tiên có đầu ra là *x[s]*, các đầu vào là các *x[s-1], x[s-2],..., x[1]*.
- ⇨ Xây dựng mô hình mạng nơ ron áp dụng cho dự báo. Việc xác định cấu trúc tối ưu cần quá trình thử-sai.
- ⇨ Huấn luyện mạng với các thông số khởi tạo trên các tập dữ liệu training, xác định lỗi với tập dữ liệu test để xác định khả năng tổng quát hóa.
- ⇨ Sau khi huấn luyện, thực hiện kiểm định độ chính xác của mô hình với tập validation.

Một kiến trúc khác của ANN cho dự báo chuỗi thời gian gọi là mạng nơron thời gian trễ [3] [4], trong đó độ trễ thời gian được gắn vào cấu trúc mạng. Phân loại về các kiến trúc mạng nơron cho xử lý chuỗi thời gian có thể xem ở [11]. Các phương pháp này đều gặp phải các vấn đề của một mạng nơron: thời gian huấn luyện lâu, số lượng tham số nhiều. Thực tế, trong trường hợp giải thuật của Wan [12], có 1105 tham số để khớp vào 1000 điểm dữ liệu. Nghĩa là rủi ro về quá khớp (overfitting) trong quá trình học của mạng là rất lớn.

IV. ĐẶC ĐIỂM ỨNG DỤNG

Các nghiên cứu về dự báo dữ liệu chuỗi thời gian sử dụng mạng nơron cho thấy khi áp dụng có một số điểm đặc trưng:

- ⇨ Quá trình dự báo dữ liệu là một quá trình hộp đen.
- ⇨ Số lượng tham số của mô hình, các trọng số của các nơron, là rất lớn phụ thuộc vào đặc trưng của bài toán thực tế. Do vậy khó giải thích quá trình dẫn đến kết quả.
- ⇨ Thích hợp với nhiều dạng chuỗi thời gian do coi tất cả dữ liệu thuộc dạng phi tuyến tính. Đặc biệt đối với các tập dữ liệu lớn, phức tạp.

- ⇨ Khi lựa chọn các thông số cho mạng nơron, cần quá trình thử-sai khi thực hiện các chu kỳ huấn luyện – kiểm tra và kiểm định kết quả.
- ⇨ Đôi khi kết quả dự báo trên các tập dữ liệu chuỗi thời gian tuyến tính cho kết quả không tốt bằng các phương pháp tuyến tính.

V. KẾT LUẬN

Dự báo dữ liệu chuỗi thời gian là một bài toán gặp rất nhiều trong thực tế. Làm chủ các kỹ thuật phân tích và giải quyết các bài toán dự báo chuỗi thời gian sử dụng mạng nơron là một phương pháp tốt dựa trên thực tế rằng các dạng dữ liệu chuỗi thời gian thường khó có thể nhận biết chúng có các đặc điểm quá trình là tuyến tính hay phi tuyến tính, đặc biệt đối với các dữ liệu lớn, phức tạp.

Quy trình áp dụng nêu trong bài báo chỉ mang tính tổng quát, nêu lên các bước cần thiết khi áp dụng mạng nơron trong dự báo dữ liệu chuỗi thời gian. Trong nghiên cứu sắp tới, chúng tôi sẽ xây dựng phần mềm ứng dụng các kỹ thuật nêu trong bài và thực hiện đánh giá các kết quả nhận được khi áp dụng mạng nơron trên một số tập dữ liệu chuỗi thời gian.

VI. REFERENCES

[1] Lê Hải Khôi & Trần Đức Minh, Về một phương pháp dự báo dữ liệu sử dụng mạng nơron. (Tạp chí Tin học và Điều khiển học 20 (2004), N2).

[2] G.E.P.Box, G.M.Jenkins and G.C.Reinsel. Time Series Analysis: Forecasting and Control, San Francisco: Holden-Day, 1994.

[3] K. Lang and G. Hilton. A time-delay neural network architecture for speech recognition. Technical Report CMU-CS-88-152, Carnegie Mellon University, Pittsburgh, PA, 1988.

[4] A.Waibel. Modular construction of time-delay neural networks for speech recognition. Neur. Comp., 1(1):39-46, 1989.

[5] A.M.Fraser and A.Dimitriadis. Forecasting Probability Densities by Using Hidden Markov Models with Mixed States. 1993.

[6] E.J.Kostelich and D.P.Lathrop. Time Series Prediction by Using the Method of Analogues. 1993.

[7] Kaastra, I., Boyd, M. - Designing a neural network for forecasting financial and economic time series - *Neurocomputing* **10** (1996), pp 215-236.

[8] Morioka Y., Sakurai K., Yokoyama A. Sekine Y., Next day peak load forecasting using a Multilayer neural network with an additional learning, *IEEE,* 0-7803-1217-1/93, 1993.

[9] Takashi O., Next day's peak load forecasting using an artificial neural network, *IEEE 0-7803-1217-1/93*, pp 284-289, 1993.

[10] Wikipedia, Nyquist–Shannon sampling theorem, https://en.wikipedia.org/wiki/Nyquist%E2%80%93Shannon_sampling_theorem.

[11] M.C. Mozer. Neural Network Architectures for Temporal Sequence Processing, pages 243-264. Addison Wesley, 1993.

[12] E.A.Wan. Time Series Prediction by Using a Connectionist Network with Internal Delay Line, pages 195-217. Addison Wesley, 1993.