NGHIÊN CỨU ỨNG DỤNG MÔ HÌNH KHAI PHÁ DỮ LIỆU ĐỂ PHÂN TÍCH DỮ LIỆU THỜI GIAN THỰC

Vũ Khương Duy

Ngành: Công Nghệ Thông Tin

ĐH Công Nghệ- ĐHQG Hà Nội

Khóa luận chương trình cử nhân Công Nghệ Thông Tin

05/2012

Lời cảm ơn

Lời đầu tiên, tôi xin gửi lời cảm ơn và lòng biết ơn sâu sắc nhất tới **PGS.TS Nguyễn Hà Nam**, **ThS. Nguyễn Thu Trang** đã tận tình hướng dẫn tôi trong suốt quá trình thực hiện khoá luận tốt nghiệp.

Tôi chân thành cảm ơn các thầy, cô đã tạo cho tôi những điều kiện thuận lợi để tôi học tập và nghiên cứu tại trường Đại học Công Nghệ.

Cuối cùng, tôi muốn được gửi lời cảm ơn vô hạn tới gia đình và bạn bè, những người thân yêu luôn bên cạnh và động viên tôi trong suốt quá trình thực hiện khóa luận tốt nghiệp.

Tôi xin chân thành cảm ơn !

NGHIÊN CỨU ỨNG DỤNG MÔ HÌNH KHAI PHÁ DỮ LIỆU ĐỂ PHÂN TÍCH DỮ LIỆU THỜI GIAN THỰC

Vũ Khương Duy

Khóa QH-2008-I/CQ, ngành Công nghệ Thông tin

Tóm tắt Khóa luận tốt nghiệp:

Put your abstract or summary here, if your university requires it.

TIME SERIES ANALYSIS USING DATA MINING TECHNIQUES

VU KHUONG DUY

QH-2008-I/CQ, Information Technology

Abstract

Put your abstract or summary here, if your university requires it.

Lời cam đoan

Tôi xin cam đoan kết quả đạt được trong luận văn là sản phẩm của riêng cá nhân tôi, không sao chép lại của người khác. Trong toàn bộ nội dung luận văn, những điều đã được trình bày hoặc là của riêng cá nhân tôi, hoặc là được tổng hợp từ nhiều nguồn tài liệu. Tất cả các nguồn tài liệu tham khảo được dùng đều có xuất xứ rõ ràng, được trích dẫn hợp pháp.

Tôi xin chịu hoàn toàn trách nhiệm và chịu mọi hình thức kỉ luật theo quy định cho lời cam đoan của mình.

Hà Nội, ngày 18 tháng 03 năm 2012

Sinh Viên

Mục lục

Danh sách hình vẽ

DANH SÁCH HÌNH VẼ

Danh sách bảng

DANH SÁCH BẢNG

1

Tính chất của chuỗi thời gian

1.1 Chuỗi thời gian

Trong quá trình theo dõi sự thay đổi của các quá trình, ta đòi hỏi có một số lượng lớn các quan sát cho các đại lượng thích hợp để nghiên cứu các mối quan hệ giữa các đại lượng này. Các quan sát này có thể được tiến hành đều đặn qua các thời kỳ chẳng hạn theo từng ngày, từng tuần, từng tháng, từng quý hoặc từng năm. Dãy các quan sát này gọi là chuỗi thời gian. Như vậy chuỗi thời gian là tập hợp các quan sát được ghi nhận tại thời điểm t với $t \in T$. Chuỗi thời gian được gọi là rời rạc nếu T là tập các điểm rời rạc (Thí dụ các quan sát được thực hiện cách nhau một khoảng thời gian đều đặn như doanh thu cước phí điện thoại hàng tháng của một trạm bưu điện từ tháng 1 năm 1990 đến tháng 12 năm 2000). Ngược lại nếu T là một khoảng thì chuỗi thời gian là liên tục. Biểu đồ ghi nhịp tim của một bệnh nhân trong 3 giờ là một ví dụ minh họa cho chuỗi thời gian liên tục với T = [0,3].

Hầu hết các thủ tục thống kê đều dùng những số liệu xuất phát từ một loạt các quan sát độc lập tập hợp thành một mẫu quan sát và ký hiệu là $X = \{x_1, x_2, ..., x_n\}$(ứng với n quan sát). Trong phân tích thống kê cổ điển người ta không quan tâm tới thứ tự thời gian quan sát diễn ra trong mẫu. Tuy nhiên với mẫu quan sát dưới dạng chuỗi thời gian thì lại không phù hợp vì nó sẽ làm mất đi tính tuần tự của dữ liệu.

Để có thể mô tả tính chất quan trọng này của dữ liệu chuỗi thời gian, giúp tạo điều kiện lựa chọn và đánh giá áp dụng các mô hình phân tích chính xác hơn,

chúng ta coi một chuỗi thời gian là một tập hợp các biến ngẫu nhiên được đánh chỉ số theo thứ tự thời gian. Hay nói cách khác, một **chuỗi thời gian**(1) là một dãy các giá trị quan sát $X = \{x_1, x_2, ..., x_n\}$ được xếp thứ tự theo diễn biến thời gian, x_1 là giá trị quan sát tại thời điểm đầu tiên, x_2 là giá trị tại thời điểm quan sát thứ hai, ... và x_n là giá trị tại thời điểm quan sát thứ n (cũng là thời điểm cuối cùng). Chuỗi thời gian sẽ được biểu diễn bởi một quá trình thông kê ngẫu nhiên $\{x_t\}$ với t có thể liên tiếp $t = 0, \pm1, \pm2, ...$ hoặc đơn giản là một tập các số nguyên có thứ tự. Trên thưc tế, mức tiêu thụ điện theo từng tháng của một hộ gia đình, lượng hành khách hàng ngày trên chuyến tàu Bắc Nam...tất cả đều là thể hiện cụ thể của các chuỗi thời gian. Trong các thí dụ trên, thứ tự thời gian quan sát thực sự đóng vai trò quan trọng, vì thế hầu hết các kỹ thuật thống kê cổ điển ít có tác dụng và do đó cần phải đề xuất những kỹ thuật tính toán mới để bộc lộ được các nét đặc thù của chuỗi thời gian.

Vậy chuỗi thời gian là một chuỗi các giá trị ngẫu nhiên của một đại lượng nào đó được ghi nhận tuần tự theo thời gian.

Tất cả các kỹ thuật dự báo truyền thống theo chuỗi thời gian dựa trên giả định là có một mẫu hình cơ bản tiềm ẩn trong các số liệu nghiên cứu cùng với các yếu tố ngẫu nhiên ảnh hưởng lên hệ thống đang xét. Công việc của phân tích chuỗi thời gian là nghiên cứu kỹ thuật để tách mẫu hình cơ bản này và sử dụng nó như là cơ sở để sản sinh ra dữ liệu dự báo cho tương lai.

1.2 Tính chất đặc trưng của chuỗi thời gian

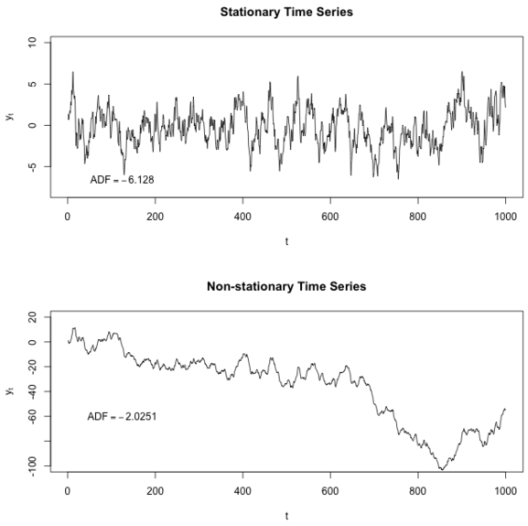
Các tính chất đặc trưng chính (2) của chuỗi thời gian là **tính dừng, tính tuyến tính, tính xu hướng và tính chu kì thời vụ**. Mặc dù, một chuỗi thời gian có thể mang một hoặc nhiều các đặc trưng trên nhưng khi biểu diễn, phân tích hay dự báo giá trị thì mỗi đặc trưng đều có phương pháp đánh giá và xử lý riêng rẽ.

1.2.1 Tính dừng

Trong khái niệm xác suất thì một chuỗi có tính chất dừng khi nó được mô tả bởi một quá trình ngẫu nhiên mà quá trình này nằm trong trạng thái cân bằng xác suất(statistical equilibrium) tức là phân phối xác suất chung của $x(t)$ và $x(t+h)$

chỉ phụ thuộc vào h mà không phụ thuộc vào t (2). Do đó, hoàn toàn có thể xây dựng mô hình ngẫu nhiên có tính chất dừng cho một chuỗi thời gian luôn giữ trạng thái cân bằng quanh giá trị xấp xỉ trung bình.

Phần 1.5 sẽ thảo luận rõ hơn về tính dừng của một chuỗi thời gian vì nó là một trong những tính chất quan trọng được nhắc đến trong hầu hết các mô hình thống kê về chuỗi thời gian. Nhưng có thể quan sát đánh giá một cách tương đối tính dừng của một của một chuỗi thời gian. Một chuỗi dừng thường có đồ thị nhìn khá phẳng, không có xu hướng hay chu kì (Hình 1.1).



Hình 1.1: Chuỗi dừng và không có tính chất dừng -

1.2.2 Tính tuyến tính

Định Nghĩa 1.1 *Một chuỗi thời gian được gọi là tuyến tính(3) nếu nó có thể được biểu diễn bởi công thức:*

$$X_t = \mu + \sum_{i=-\infty}^{\infty} \psi_i Z_{t-i} \tag{1.1}$$

với μ là xấp xỉ trung bình của $\{x_t\}$, $\{\psi_i\}$ là tập tham số thỏa mãn $\sum_{i=-\infty}^{\infty} |\psi_i| < \infty$. Và $|Z_t|$ là chuỗi ngẫu nhiên có xấp xỉ trung bình bằng 0 và phương sai σ^2

Do đó, nếu một chuỗi thời gian là tuyến tính thì nó có thể được biểu diễn bởi một hàm phụ thuộc tuyến tính các giá trị hiện tại và quá khứ.

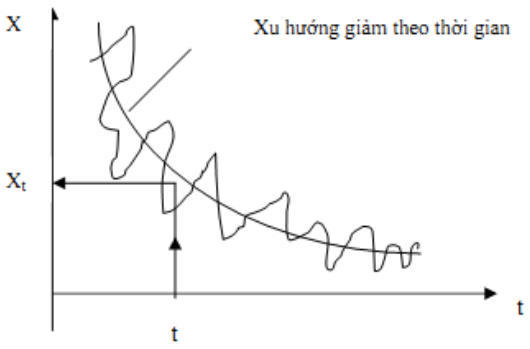
1.2.3 Tính xu hướng

Tính xu hướng của chuỗi dữ liệu thể hiện qua hiện tượng tăng hoặc giảm giá trị xấp xỉ trung bình của các đoạn con một cách liên tục trong phạm vi cục bộ hoặc cả chuỗi. Để xác định những thành phần có tính xu hướng trong chuỗi thời gian, dữ liệu quan sát được thường được ước lượng xấp xỉ vào trong các mô hình hồi quy. Một số mô hình thông thường hay được sử dụng như :

$$x_t = \alpha t + \beta + \varepsilon_t \tag{1.2}$$

$$x_t = exp(\alpha t + \beta + \varepsilon_t) \tag{1.3}$$

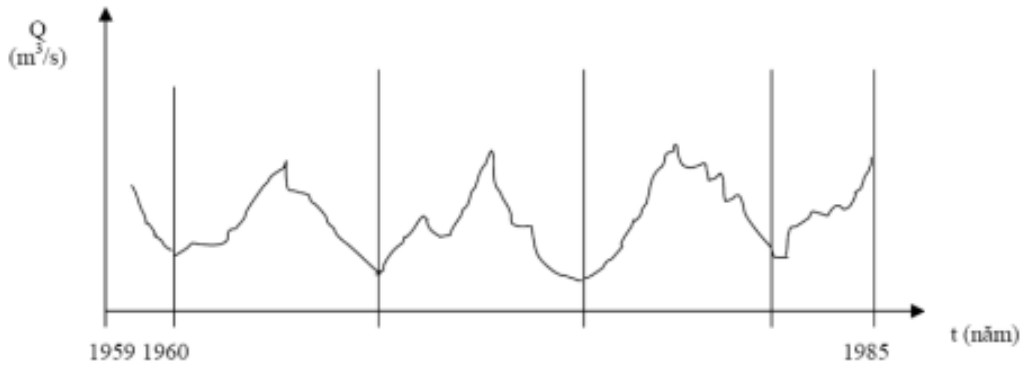
$$x_t = \alpha t + \beta t^2 + \gamma + \varepsilon_t \tag{1.4}$$



Hình 1.2: Xu hướng giảm dần -

1.2.4 Tính chu kỳ thời vụ(seasonality)

Một số chuỗi thời gian xuất hiện các yếu tố dao động theo chu kỳ.Đó chính là đặc trưng mang tính thời vụ rất hay thường gặp trong dữ liệu quan sát thực tế, nhất là trong các dữ liệu về tài chính, kinh tế. Nó thường xuất hiện các đoạn dao động lặp lại giống nhau theo từng giờ, từng ngày, từng tháng, từng quý...Ví dụ như lượng dòng chảy đến hồ chứa Trị An từ năm 1959 đến 1985 (Hình 1.3)



Hình 1.3: Lượng dòng chảy đến hồ chứa Trị An từ năm 1959 đến 1985 -

1.3 Phân loại chuỗi thời gian

Phụ thuộc vào tính chất của dữ liệu mà chuỗi thời gian có thể chia theo một số tiêu chí(2) sau:

- có tính chất dừng và không có tính chất dừng
- có tính chu kỳ thời vụ và không có tính chu kỳ thời vụ
- tuyến tính và không tuyến tính
- đơn chiều và đa chiều
- hỗn loạn

Chuỗi thời gian **đơn chiều** là chuỗi chỉ ghi lại kết quả quan sát theo thời gian của một dạng dữ liệu. Ví dụ như trong một phiên chứng khoán, có rất nhiều giá trị được quan sát như khối lượng giao dịch, giá đóng, giá mở... nhưng chuỗi thời gian đơn chiều chỉ quan tâm đến một trong những giá trị đó. *Khóa luận sẽ tập trung vào phân tích xử lý đối với chuỗi dữ liệu dạng này.*

Ngược lại, chuỗi thời gian **đa chiều** ghi lại kết quả quan sát của hai hay nhiều quá trình đồng thời.Ví dụ, ta có thể ghi lại đồng thời tất cả các giá trị trong phiên chứng khoán ở ví dụ trên. Trong chuỗi đa chiều, dữ liệu của một chuỗi đơn không chỉ phụ thuộc lẫn nhau trong nội bộ chuỗi mà còn phụ thuộc hai chiều với dữ liệu của các chuỗi đơn khác.

1.4 Đo độ phụ thuộc: Tự tương quan và tương quan chéo

Trong trường hợp x_t là biến ngẫu nhiên liên tục có hàm xác suất $F_t(x) = P(x_t \leq x)$ thì hàm mật độ xác suất được tính theo công thức:

$$f_t(x) = \frac{\partial F_t(x)}{\partial x} \tag{1.5}$$

Định Nghĩa 1.2 *Hàm tính xấp xỉ trung bình*

$$\mu_{xt} = E(x_t) = \int_{-\infty}^{\infty} x f_x(x)\, dx \tag{1.6}$$

Định Nghĩa 1.3 *Hàm **tự hiệp phương sai** được xác định bởi công thức:*

$$\gamma_x(s,t) = cov(x_s, x_t) = E[(x_s - \mu_s)(x_t - \mu_t)] \tag{1.7}$$

với mọi thời điểm s, t

Ta luôn có $\gamma_x(s,t) = \gamma_x(t,s)$ với mọi s, t. Hàm tự hiệp phương sai giúp đánh giá độ phụ thuộc tuyến tính giữa hai thời điểm khác nhau trong chuỗi thời gian thực. Trong trường hợp $s = t$ thì ta có giá trị hàm tự hiệp phương sai chính bằng phương sai của x_s hay $\gamma(s,s) = E[(x_s - \mu_s)^2] = var(t)$

Định Nghĩa 1.4 *Hàm tự tương quan (ACF)(4) được xác định bởi:*

$$\rho(s,t) = \frac{\gamma(s,t)}{\sqrt{\gamma(s,s)\gamma(t,t)}} \tag{1.8}$$

Hàm ACF cho biết khả năng dự đoán giá trị x_t mà chỉ dựa vào giá trị của x_s. Dựa vào bất đẳng thức Cosi[1], ta dễ dàng có được $-1 \leq \rho(s,t) \leq 1$. Nếu ta có thể dự đoán giá trị x_s hoàn toàn chỉ dựa vào x_t thông qua hàm tuyến tính $x_t = \beta_0 + \beta_1 x_s$ thì sự tương quan sẽ là $+1$ nếu $\beta_1 > 0$ và là -1 nếu $\beta_1 < 0$. Vậy ta có một công thức đơn giản để đánh giá khả năng dự báo giá trị tại thời điểm t dựa vào giá trị tại thời điểm s trong chuỗi thời gian.

[1]Bất đẳng thức Cosi: $|\gamma(s,t)|^2 \leq \gamma(s,s)\gamma(t,t)$

Ngoài ra, chúng ta còn có thể đo khả năng có thể dự đoán chuỗi y_t từ chuỗi x_t thông qua hàm *ttngquancho*. Giả sử cả hai chuỗi đều có phương sai xác định, ta có các khái niệm sau:

Định Nghĩa 1.5 *Hàm* ***tự hiệp phương sai chéo*** *của hai chuỗi x_t và y_t được cho bởi công thức*

$$\gamma_{xy}(s,t) = cov(x_s, y_t) = E[(x_s - \mu_{xs})(y_t - \mu_{yt})] \tag{1.9}$$

Định Nghĩa 1.6 ***Hàm tự tương quan chéo(CCF)****(4) được cho bởi:*

$$\rho_{xy}(s,t) = \frac{\gamma_{xy}(s,t)}{\sqrt{(\gamma_s(s,s))(\gamma_y(t,t))}} \tag{1.10}$$

1.5 Chuỗi dừng

Chuỗi dừng và không dừng là một trong những khái niệm quan trọng cốt lõi trong việc phân tích chuỗi thời gian.

1.5.1 Chuỗi dừng và tính chất

Định Nghĩa 1.7 *Chuỗi x_t được gọi là* ***dừng theo nghĩa chặt****(4) nếu như tại mọi thời điểm $t_1, t_2, \ldots, t_k$ thì ta luôn có $(x_{t_1}, x_{t_2}, \ldots, x_{t_k})$ có cùng phân phối xác suất với $(x_{t_1+h}, x_{t_2+h}, \ldots, x_{t_k+h})$ hay*

$$P(x_{t_1} \leq c_1, x_{t_2} \leq c_2, \ldots, x_{t_k} \leq c_k) = P(x_{t_1+h} \leq c_1, x_{t_2+h} \leq c_2, \ldots, x_{t_k+h} \leq c_k) \tag{1.11}$$

với $k > 0, h > 0$ và mọi số $c_k \geq 0$

Khái niệm chuỗi dừng theo nghĩa chặt khá nghiêm ngặt cho hầu hết các chuỗi dữ liệu. Nó yêu cầu các phân phối đồng thời của x_t bất biến qua phép biến đổi dịch chuyển thời gian. Ví dụ với $k = 2$ ta có

$$P(x_s \leq c_1, x_t \leq c_2) = P(x_{s+h} \leq c_1, x_{t+h} \leq c_2) \tag{1.12}$$

với s, t, h bất kì.Khi đó nếu tồn tại giá trị phương sai của chuỗi thì hàm tự hiệp phương sai của x_t thỏa mãn

$$\gamma(s,t) = \gamma(s+h, t+h) \tag{1.13}$$

Khi đó hàm tự hiệp phương sai chỉ còn phụ thuộc vào khoảng thời gian bước nhảy h bất kì chứ không phải phụ thuộc vào thời điểm s hoặc t. Do đó, đôi khi người ta chỉ đòi hỏi một vài điều kiện lỏng hơn trên một trong hai momment của chuỗi.

Định Nghĩa 1.8 *Chuỗi x_t, có phương sai xác định, được gọi là* ***dừng theo nghĩa rộng*** *(4) nếu*

1. *Hàm xấp xỉ trung bình μ_t là hằng số và không phụ thuộc vào thời gian t*
2. *Hàm hiệp tự phương sai, $\gamma(s,t)$, phụ thuộc vào s và t chỉ thông qua hiệu $|s-t|$*

Đặt $s = t + h$ với h là khoảng cách thời gian thì theo tiêu chí thứ 2 ở trên, ta có

$$\gamma(t+h,t) = cov(x_{t+h}, x_t) = cov(x_h, x_0) = \gamma(h,0) \tag{1.14}$$

Khi đó hàm tự hiệp phương sai sẽ không còn phụ thuộc vào t mà phụ thuộc vào độ trễ.

Trong giới hạn khóa luận, ta dùng thuật ngữ **chuỗi dừng** để ám chỉ định nghĩa **"chuỗi dừng theo nghĩa rộng"**

Áp dụng một số hàm số cho chuỗi dừng

Định Nghĩa 1.9 *Hàm tự hiệp phương sai cho chuỗi dừng:*

$$\gamma(h) = cov(x_{t+h}) = E[(x_{t+h} - \mu)(x_t - \mu)] \tag{1.15}$$

Định Nghĩa 1.10 *Hàm tự tương quan ACF cho chuỗi dừng:*

$$\rho(h) = \frac{\gamma(t+h,t)}{\sqrt{\gamma(t+h,t+h)\gamma(t,t)}} = \frac{\gamma(h)}{\gamma(0)} \tag{1.16}$$

Dựa vào bất đẳng thức Cosi ta cũng có $|\rho(h)| \leq 1$. Do đó, có thể so sánh giá trị với -1 và 1 để đánh giá mối quan hệ của hai thời điểm trong hàm tự tương quan.

Một số tính chất của hàm ACF cho chuỗi dừng:

1. Với $h = 0$ ta có $\gamma(0) = E[(x_t - \mu)^2] = var(x_t)$
2. $\gamma(h) = \gamma(-h)$

1.5.2 Kiểm tra chuỗi dừng

Trong một số nghiên cứu, một số dữ liệu như tỉ lệ lợi tức, tỉ giá tiền đổi hay giá cả thị trường... có thể là dữ liệu không có tính chất dừng. Chúng ta gọi chúng là những chuỗi không dừng nghiệm đơn vị. Dựa vào khái niệm thì một chuỗi dừng sẽ thỏa mãn 2 điều kiện là hàm xấp xỉ trung bình cho luôn giá trị hằng số duy nhất và hàm tự hiệp phương sai, $\gamma(s,t)$ chỉ phụ thuộc vào $h = |s-t|$. Do đó ta có một số phương pháp để kiểm tra một chuỗi có phải là chuỗi dừng không?

1.5.2.1 Mô hình hóa xấp xỉ trung bình và ACF

Khi mô hình hóa chuỗi thời gian, ta nhận thấy chuỗi có xu hướng tăng dần, giảm dần hoặc xấp xỉ trung bình biến đổi thì chuỗi đó chắc chắn không phải là chuỗi dừng theo tính chất thứ nhất trong định nghĩa. Ngoài ra, khi quan sát mô hình trực quan hàm tự tương quan ACF, nếu giá trị đạt cao nhất ở thời điểm đầu tiên, sau đó giảm chậm dần thì đó là một chuỗi không có tính dừng (5).

1.5.2.2 Kiểm tra nghiệm đơn vị (unit-root)

Định Nghĩa 1.11 (Nghiệm đơn vị) *Giả sử một quá trình ngẫu nhiên có thể được viết dưới dạng:*

$$y_t = a_1 y_{t-1} + a_2 y_{t-2} + ... + a_p y_{t-p} + \epsilon_t \tag{1.17}$$

với $a_1, a_2, \ldots$ là các tham số; ϵ_t là quá trình ngẫu nhiên có xấp xỉ bình phương bằng 0 và phương sai không đổi bằng σ^2. Không mất tính tổng quát coi $y_0 = 0$. Khi đó nếu $m = 1$ là nghiệm của phương trình

$$m^p - m^{p-1}a_1 - m^{p-2}a_2 - ... - a_p = 0 \tag{1.18}$$

thì quá trình ngẫu nhiên này có một nghiệm đơn vị m

Khi xảy ra nghiệm đơn vị thì đây sẽ không phải là chuỗi dừng nữa vì mô hình sẽ có thể có xu hướng, xấp xỉ trung bình sẽ biến đổi hay phương sai phụ thuộc thời gian nên không thỏa mãn điều kiện chuỗi dừng. Ví dụ, giả sử đối với một chuỗi x_t đơn giản

$$x_t = \phi_1 x_{t-1} + w_1 \tag{1.19}$$

với tham số ϕ_1 và thành phần nhiễu w_1. Trong trường hợp này, nghiệm đơn vị sẽ xảy ra khi $\phi_1 = 1$. Khi đó ta sẽ có

$$x_t = x_{t_1} + w_1 \tag{1.20}$$

$$x_t = x_0 + \sum_{j=1}^{t} w_j \tag{1.21}$$

$$var(x_t) = \sum_{j=1}^{t} \sigma^2 = t\sigma^2 \tag{1.22}$$

Phương sai phụ thuộc vào thời gian t. Do đó, hàm tự hiệp phương sai cũng phụ thuộc vào t. Đây sẽ không phải là chuỗi dừng. Một số phương pháp kiểm tra nghiệm đơn vị như Dickey–Fuller áp dụng cho mô hình tự hồi quy, mở rộng Augmented Dickey–Fuller cho mô hình ARMA hay phương pháp Phillips–Perron sẽ được nhắc đến sau. Một trong những phương pháp phổ biến để chuyển một chuỗi có tính chất dừng thành chuỗi dừng là phương pháp sai phân

$$c_t = x_t - x_{t-1} \tag{1.23}$$

1.6 Bài toán dự báo chuỗi thời gian

1.6.1 Bài toán dự báo

Trong nền kinh tế thị trường, công tác dự báo là vô cùng quan trọng bởi lẽ nó cung cấp các thông tin cần thiết nhằm phát hiện và bố trí sử dụng các nguồn lực trong tương lai một cách có căn cứ thực tế. Trong môi trường phong phú và có tính trạnh canh cao, thông tin là vô cùng quan trọng. Người nắm được thông tin trước là người đi đầu, chớp được thời cơ. Đó là một phần cực kì quan trọng ảnh hướng đến lợi nhuận, đường lối phát triển của doanh nghiệp. Với những thông tin mà dự báo đưa ra cho phép các nhà hoạch định chính sách có những quyết định về đầu tư, các quyết định về sản xuất, về tiết kiệm và tiêu dùng, các chính sách tài chính, chính sách kinh tế vĩ mô. Dự báo không chỉ tạo cơ sở khoa học cho việc hoạch định chính sách, cho việc xây dựng chiến lược phát triển, cho các quy hoạch tổng thể mà còn cho phép xem xét khả năng thực hiện kế hoạch và hiệu chỉnh kế hoạch.

Trong quản lý vi mô, dự báo là hoạt động gắn liền với công tác hoạch định và chỉ đạo thực hiện chiến lược kinh doanh của doanh nghiệp. Các doanh nghiệp không thể không tổ chức thực hiện tốt công tác dự báo nếu họ muốn đứng vững trong kinh doanh.

Chức năng đầu tiên của quản lý trong doanh nghiệp là xác định mục tiêu của doanh nghiệp trong dài hạn và ngắn hạn. Doanh nghiệp phải lập kế hoạch để thực hiện những mục tiêu đó, tổ chức tốt các nguồn nhân lực và vật tư để thực hiện kế hoạch, điều chỉnh kế hoạch cũng như kiểm soát các hoạt động để tin chắc rằng tất cả diễn ra theo đúng kế hoạch. Phân tích kinh tế và dự báo được tiến hành trong tất cả các bước của quản lý doan nghiệp, nhưng trước hết là trong việc xác định mục tiêu và hoạch định các kế hoạch dài hạn và ngắn hạn.

Trong việc xác định mục tiêu, mỗi doanh nghiệp phải quyết định hàng hóa và dịch vụ nào sẽ được sản xuất và bán ra, mức giá sản phẩm và dịch vụ, vùng tiêu thụ, thị trường tiềm năng về sản phẩm đó. Thị phần mà doanh nghiệp thực tế có thể hi vọng chiếm được, hiệu suất vốn doanh nghiệp có thể kỳ vọng...Những mục tiêu như vậy chỉ có thể trở thành hiện thực nếu doanh nghiệp đã phân tích các xu thế của nền kinh tế, đã dự báo về nhu cầu sản phẩm của mình cả trong dài hạn và ngắn hạn, chi phí các nhân tố sản xuất...Như vậy các dự báo về thị trường, giá cả, tiến bộ khoa học và công nghệ, nguồn nhân lực, sự thay đổi của các nguồn đầu tư vào, đối thủ cạnh tranh,... có tầm quan trọng sống còn đối với doanh nghiệp. Ngoài ra dự báo cung cấp những thông tin cho phép phối hợp hành động giữa các bộ phận trong doanh nghiệp.

Hiện nay có rất nhiều phương pháp dự báo khác nhau về nguồn thông tin được sử dụng, về cơ chế xây dựng dự báo, về độ tin cậy độ xác thực của dự báo. Tuy nhiên, có một số mô hình thông dụng hay được sử dụng (6) như:

Mô hình kinh tế lượng: Là phương pháp dựa trên lý thuyết kinh tế lượng để lượng hoá các quá trình kinh tế xã hội thông qua phương pháp thống kê. ý tưởng chính của phương pháp là mô tả mối quan hệ giữa các đại lượng kinh tế bằng một phương trình hoặc hệ phương trình đồng thời. Với các số liệu quá khứ, tham số của mô hình này được ước lượng bằng phương pháp thông kê. Sử dụng mô hình đã ước lượng này để dự báo bằng kỹ thuật ngoại suy hoặc mô phỏng.

Mô hình I/O: Mô hình I/O là mô hình dựa trên ý tưởng là mối liên hệ liên ngành trong bản đầu vào - đầu ra(Input – Output table) diễn tả mối quan hệ của quá trình sản xuất giữa các yếu tố đầu vào, chi phí trung gian và đầu ra của sản xuất.

Mô hình tôi ưu hoá: Điển hình của mô hình này là bài toán quy hoạch tối ưu, bố trí một nguồn lực nhằn tối ưu hoá một mục tiêu nào đó.

Mô hình chuỗi thời gian: Phương pháp dự báo này được tiến hành trên cơ sở giả định rằng quy luật đã phát hiện trong quá khứ và hiện tại được duy trì sang tương lai trong phạm vi tâm xa dự báo. Các quy luật này được xác định nhờ phân tích chuỗi thời gian và được sử dụng để suy diễn tương lai.

Mô hình nhân tố: Phân tích tương quan giữa các chỉ tiêu(nhân tố) với nhâu và lượng hoá các mối quan hệ này. Việc lượng hoá được thực hiện nhờ phương pháp phân tích hồi quy và dự báo chỉ tiêu kết quả trên cơ sở sự thay đổi của các chỉ tiêu nguyên nhân hay các chỉ tiều giải thích.

Khóa luận sẽ tập trung nghiên cứu về việc sử dụng mô hình chuỗi thời gian cho bài toán dự báo nhu cầu đang cực kì cấp thiết hiện nay.

1.6.2 Bài toán phân tích chuỗi thời gian và dự báo

1.6.2.1 Mô hình thống kê truyền thống

Tất cả các kỹ thuật truyền thống dự báo theo chuỗi thời gian dựa trên giả định là có một mẫu hình cơ bản tiềm ẩn trong các số liệu nghiên cứu cùng với các yếu tố ngẫu nhiên ảnh hưởng lên hệ thống đang xét. Công việc của phân tích chuỗi thời gian là nghiên cứu kỹ thuật để tách mẫu hình cơ bản này và sử dụng nó như là cơ sở để sản sinh ra dữ liệu dự báo cho tương lai.

Để làm được điều đó, trước hết ta giả thiết có một mô hình xác suất để biểu diễn dãy số liệu. Sau cùng hy vọng chọn ra một mô hình gần với dãy số liệu, chúng ta tiến hành ước lượng các tham số của mô hình, kiểm tra lại xem mô hình được sử dụng có phù hợp không.

Dự báo dựa trên mô hình chuỗi thời gian là ước lượng các giá trị tương lai x_{t+h} của một biến ngẫu nhiên dựa trên các giá trị quan sát trong quá khứ của nó $x_1, x_2, ..., x_t$. Giá trị dự báo của x_{t+h} thường được ký hiệu là $\hat{x}_t(h)$.

Có 2 hướng tiếp cận chính (2) khi phân tích chuỗi dữ liệu theo phương pháp truyền thống:

Miền thời gian chủ yếu dựa vào việc sử dụng hàm hiệp phương sai của chuỗi thời gian

Miền tần số dựa vào phân tích hàm mật độ phổ và Fourier

Cả hai hướng tiếp cận đều có rất nhiều ứng dụng. Một trong các phương pháp phổ biến theo hướng tiếp cận miền thời gian do George Box và Gwin Jenkins (1970)(1) đề xuất có tên là **Quá trình tự hồi quy trung bình trượt ARMA**(ARMA là AutoRegressive Moving Average. Nó là một quy trình lặp xử lý mô hình chuỗi dừng tuyến tính. Tuy nhiên, nó cũng có thể được mở rộng thành quy trình ARIMA(quy trình tự hồi quy trung bình trượt tích hợp)để xử lý với chuỗi không dừng. Trong quy trình ARIMA, dữ liệu đầu tiên sẽ được sai phân đến khi chuỗi mới biến đổi có tính dừng. Sau đó, ta áp dụng ARMA cho chuỗi mới biến đổi đó.

Ngoài ra, chất lượng của dự báo phụ thuộc vào nhiều yếu tố. Trước hết nó phụ thuộc vào xu hướng phát triển của chuỗi thời gian. Nếu chuỗi thời gian là hàm "đều đặn" theo thời gian thì càng dễ xác định giá trị dự báo. Thí dụ nếu tiến trình phát triển kinh tế không có những biến động đặc biệt thì càng dễ dàng đánh giá tổng sản phẩm quốc nội (GDP) cho những năm tiếp theo. Cho đến nay, các phương pháp dự báo dựa trên chuỗi thời gian chưa cho phép đánh giá ước lượng được các giá trị đột biến. Chất lượng của dự báo dựa vào chuỗi thời gian còn phụ thuộc vào độ xa gần của thời gian. Các giá trị dự báo càng gần hiện tại thì độ chính xác càng cao. Như vậy việc ước lượng GDP cho năm sẽ chính xác hơn việc ước lượng GDP cho 10 năm sau.

1.6.2.2 Kỹ thuật khái phá dữ liệu mới

Khi dự báo chuỗi thời gian sử dụng những kĩ thuật truyền trống như ARIMA, làm trơn, hồi quy hay phân tích xu hướng...ta phải quan tâm đến rất nhiều yếu tố tính chất ảnh hưởng đến việc xác định mô hình xấp xỉ như tính dừng, tuyến

tính, chu kì...Sự phát hiện ra những thành phần mang các yếu tố trên là cực kì quan trọng, bởi mỗi yếu tố sẽ được đánh giá và định hướng ước lượng mô hình khác nhau. Mặt khác, việc phát hiện không phải lúc nào cũng dễ dàng, nếu đánh giá không đúng sẽ dẫn tới việc ước lượng mô hình với sai số khá lớn.

Các kĩ thuật khai phá dữ liệu mới ra đời đã cố gắng giải quyết hoàn toàn vấn đề trên. Một số phương pháp cơ bản như sử dụng mạng nơ ron, hướng tiếp cận logic mờ, tính toán tiến hóa hay có một số kĩ thuật mới hơn như svm, phân lớp mờ...Vai trò của các đặc trưng tính chất trong chuỗi thời gian được giảm đi đáng kể. Một số kĩ thuật còn không quan tâm đến các tính chất đó. Chương 2 sẽ giới thiệu phương pháp *k-người láng giềng gần nhất* và mô hình *mạng nơ ron nhân tạo* để chứng minh điều đó. Hiệu quả của chúng so với phương pháp truyền thống ra sao sẽ được nhắc tới sâu hơn trong khóa luận.

2

Mô hình ARIMA và một số mô hình dự báo hiện đại khác

2.1 Mô hình tuyến tính ARIMA

"ARMA" là tên viết tắt của "Autoregressive Moving Average", có nghĩa là mô hình tự hồi-quy trung bình trượt. Mô hình này được tích hợp từ 2 thành phần: tự hồi quy (AR) và trung bình trượt(MA). Phần này trước tiên sẽ đưa ra những kiến thức cơ bản nhất cần thiết sử dụng cho việc định nghĩa và xây dựng các mô hình AR, MA và ARMA

2.1.1 Hàm tương quan và tự tương quan

2.1.2 Nhiễu trắng

Một chuỗi thời gian x_t được gọi là **nhiễu trắng** nếu x_t là một chuỗi các biến ngẫu nhiên phân phối độc lập với xấp xỉ trung bình và phương sai xác định hữu hạn. Trong trường hợp đặc biệt, nếu x_t có xấp xỉ trung bình bằng 0 và phương sai σ^2 thì gọi là nhiễu trắng Gaussian. Với nhiễu trắng thì tất cả các giá trị hàm ACF đều bằng 0. Do đó, trong thực tế, nếu tất cả các giá trị ACF tiến gần tới 0 thì coi như chuỗi đó là nhiễu trắng

2.1.3 Quá trình tự hồi quy AR

Ý tưởng của mô hình tự hồi quy AR là tính giá trị hiện tại x_t trong chuỗi dựa vào hàm hồi quy của p giá trị, $x_{t-1}, x_{t-2}, \ldots, x_{t-p}$, xảy ra ngay trước nó trong quá khứ. Giá trị p xác định số bước nhảy lùi quá khứ để dự đoán giá trị hiện tại.

Định Nghĩa 2.1 *Một **mô hình tự hồi quy** cấp p, viết tắt là **AR**(p) được xác định bởi công thức*

$$x_t = \phi_1 x_{t-1} + \phi_2 x_{t-2} + \cdots + \phi_p x_{t-p} + w_t \tag{2.1}$$

với x_t có tính dừng, và $\phi_1, \phi_2, \ldots, \phi_p$ là các hằng số $(\phi_p) \neq 0$ mô tả mối quan hệ giữa giá trị hiện tại với các giá trị trước nó. Trong công thức này ta coi w_t là chuỗi nhiễu trắng Gaussian có xấp xỉ trung bình bằng 0 và phương sai σ_w^2. Nếu xấp xỉ trung bình của x_t là $\mu \neq 0$ thì ta sẽ thay x_t bằng $x_t - \mu$ trong công thức 2.2. Ta được công thức sau:

$$x_t - \mu = \phi_1(x_{t-1} - \mu) + \phi_2(x_{t-2} - \mu) + \cdots + \phi_p(x_{t-p} - \mu) + w_t \tag{2.2}$$

hoặc

$$x_t = \phi_1 x_{t-1} + \phi_2 x_{t-2} + \cdots + \phi_p x_{t-p} + w_t \tag{2.3}$$

với $\alpha = \mu(1 - \phi_1 - \phi_2 - \cdots - \phi_p)$

Để dễ dàng biểu diễn mô hình thông qua công thức, ta cần sử dụng thêm khái niệm *toán tử dịch chuyển B*.

Định Nghĩa 2.2 *Toán tử dịch chuyển B được xác định bởi*

$$Bx_t = x_{t-1} \tag{2.4}$$

thực hiện quá trình đệ quy $x_{t-k} = Bx_{t-k+1} = ... = B^k x_t$ hay

$$x_t = B^k x_{t-k} \tag{2.5}$$

Quay lại với mô hình AR, ta có thể biểu diễn nó dưới dạng

$$(1 - \phi_1 B - \phi_2 B^2 - ... - \phi_p B^p)x_t = w_t, \phi(B)x_t = w_t \tag{2.6}$$

hay

$$\phi(B)x_t = w_t \tag{2.7}$$

ϕB ở trên được gọi là toán tử tự hồi quy

Như đã nói ở trên, hàm tự hiệp phương sai có thể đo mức độ phụ thuộc tuyến tính giữa hai thời điểm trong chuỗi thời gian. Nó là sự phụ thuộc chuỗi toàn phần, có nghĩa là x_t phụ thuộc vào x_{t-2} thông qua x_{t-1} rồi x_{t-1} phụ thuộc x_{t-3} thông qua x_{t-2} và cứ như vậy. Ví dụ như mô hình đơn giản $AR(1)$ được cho bởi $x_t = \phi x_{t-1} + w_t$, ta có

$$\gamma_x^2 = cov(x_t, x_{t-2}) = cov(\phi x_{t-1} + w_t, x_{t-2}) \tag{2.8}$$

$$= cov(\phi^2 x_{t-2} + \phi w_{t-1} + w_t, x_{t-2}) \tag{2.9}$$

$$= \phi^2 \gamma_x(0) \tag{2.10}$$

Do giá trị $\gamma_x(2) \neq 0$, nên có thể khẳng định tồn tại tương quan giữa x_t và x_{t-2} thông qua x_{t-1}. Nhưng nếu trong trường hợp ta bỏ đi giá trị x_{t-1} để bẻ gãy chuỗi liên kết liên tục này, mối quan hệ giữa x_t và x_{t-2} sẽ được đo bởi giá trị hàm tự hiệp phương sai giữa $(x_t - \phi x_{t-1})$ và $(x_{t-2} - \phi x_{t-1}$ do thành phần x_{t-1} bị mất đi.

$$cov((x_t - \phi x_{t-1}), (x_{t-2} - \phi x_{t-1}) = cov(w_t, x_{t-2} - \phi x_{t-1})) = 0 \tag{2.11}$$

Do đó, ta cần một hàm tự tương quan từng phần(PACF) có thể dùng để đánh giá mối quan hệ giữa x_s và x_t trong trường hợp trung gian không được tính đến.

Định Nghĩa 2.3 *Hàm **tự tương quan từng phần (PACF)** của một chuỗi dừng x_t, viết tắt là ϕ_{hh} với $h = 1, 2, \ldots$ được xác định :*

$$\phi_{11} = corr(x_{t+1}, x_t) = \rho(1) \tag{2.12}$$

và

$$\phi_{hh} = corr(x_{t+h} - \hat{x}_{t+h}, x_t - \hat{x}_t), \quad h \geq 2 \tag{2.13}$$

với

$$\hat{x}_{t+h} = \beta_1 x_{t+h-1} + \beta_3 x_{t+h-2} + \cdots + \beta_{h-1} x_{t+1} \tag{2.14}$$

$$\hat{x}_t = \beta_1 x_{t+1} + \beta_3 x_{t+2} + \cdots + \beta_{h-1} x_{t+h-1} \tag{2.15}$$

Mô hình AR(p) có thể được cho bởi công thức

$$x_{t+h} = \sum_{j=1}^{p} \phi_j x_{t+h-j} + w_{t+h} \tag{2.16}$$

Khi $h > p$ thì (4) ta tính được

$$\hat{x}_{t+h} = \sum_{j=1}^{p} \phi_j x_{t+h-j} \tag{2.17}$$

Theo đó, khi $h > p$

$$\phi_{hh} = corr(x_{t+h} - \hat{x}_{t+h}, x_t - \hat{x}_t) = corr(w_{t+h}, x_t - \hat{x}_t) = 0 \tag{2.18}$$

Theo công thuwsc 2.15 thì $x_t - \hat{x}_t$ sẽ chỉ phụ thuộc vào w_i với $i = t+h-1, t+h-2, ...$ nên chúng sẽ độc lập với w_{h+t}. Khi đó, $\phi_{hh} = 0$ như trên. Cũng theo (4) thì khi $h < q$ thì $\phi_{hh} \neq 0$. Do đó, trong mô hình $AR(p)$, PACF sẽ bị triệt tiêu sau khi h vượt ngưỡng q.Dựa vào kết quả này, ta có thể dự đoán giá trị q và xây dựng mô hình $AR(q)$ từ chuỗi thời gian thực.

2.1.4 Quá trình trung bình trượt MA

Định Nghĩa 2.4 *Mô hình **trung bình trượt** cấp q ,viết tắt là mô hình **MA**(q) được xác định bởi công thức:*

$$x_t = w_t + \theta_1 w_{t-1} + \theta_2 w_{t-2} + \cdots + \theta_q w_{t-q} \tag{2.19}$$

với q là cấp của mô hình trung bình trượt, và $\theta_1, \theta_2, \ldots, \theta_q$ là các tham số xác định mối quan hệ phụ thuộc giữa giá trị hiện tại x_t và các giá trị nhiễu sai số trước đó. w_t là chuỗi nhiễu trắng Gaussian có xấp xỉ trung bình bằng 0 và phương sai σ_w^2

Phương trình 2.19 cũng có dạng tuyến tính như phương trình 2.2 nhưng x_t sẽ được tính thông qua các giá trị nhiễu sai số $w_t, w_{t-1}, \ldots, w_{t-p}$ trong quá khứ thay vì bằng các giá trị $x_{t-1}, x_{t-2}, \ldots, x_{t-p}$ Tương tự, ta cũng có thể biểu diễn 2.19 dưới dạng

$$x_t = \theta(B) w_t \tag{2.20}$$

với toán tử trung bình trượt $\theta(B) = 1 + \theta_1 B + \theta_2 B^2 + ... + \theta_q B^q$

2.1.4.1 Xác định cấp q trong $MA(q)$

Vì x_t được tính tuyến tính thông qua các nhiễu trắng với mô hình 2.19, chuỗi này hiển nhiên là chuỗi dừng với xấp xỉ trung bình và hàm tự hiệp phương sai như sau:

$$E(x_t) = \sum_{j=0}^{q} \theta_j E(w_{t-j}) = 0 \tag{2.21}$$

$$\gamma(h) = cov(x_{t+h}, x_t) = cov(\sum_{j=0}^{q} \theta_j w_{t+h-j}, \sum_{k=0}^{q} \theta_k w_{t-k}) \tag{2.22}$$

$$= \begin{cases} \sigma_w^2 \sum\limits_{j=0}^{q-h} \theta_j \theta_{j+h}, & 0 \le h \le q \\ 0 & h > q \end{cases} \tag{2.23}$$

Khi đó,ta cũng có

$$\rho(h) = \frac{\gamma(h)}{\gamma(0)} = \begin{cases} \frac{\sum\limits_{j=0}^{q-h} \theta_j \theta_{j+h}}{1+\theta_1^2+...+\theta_q^2}, & 1 \le h \le q \\ 0 & h > q \end{cases} \tag{2.24}$$

Ta có thể nhận thấy với mô hình $MA(q)$ thì giá trị của $\rho(h)$ sẽ bị triệt tiêu ngay sau cấp q. Đây chính là điểm mấu chốt để xác định q và xây dựng mô hình $MA(q)$ dựa vào việc quan sát đồ thị trực quan của hàm tự tương quan.

Trong thực tế, công thức hàm tự tương quan

$$\rho(h) = \frac{\gamma(h)}{\gamma(0)} \tag{2.25}$$

được gọi là hàm tự tương quan lý thuyết bởi vì nó phụ thuộc vào các tham số thực của x_t. Nhưng trong dữ liệu thực tế, không thể biết được chính xác các tham số này, do đó chúng ta sẽ không thể tính toán chính xác được $\rho(h)$. Tuy nhiên, chúng ta có thể tính toán gần đúng nó dựa vào **hàm tự tương quan mẫu** với độ trễ h được cho bởi công thức

$$\hat{p}_h = \frac{\sum\limits_{t=h+1}^{n} (x_t - \bar{x})(x_{t-h} - \bar{x})}{\sum\limits_{t=1}^{n} (x_t - \bar{x})^2} \tag{2.26}$$

với $\bar{x}$ là xấp xỉ trung bình mẫu của x_i Lúc này tao hoàn toàn có thể tính được ước lượng của $\rho(h)$

2.1.5 Quá trình ARMA

Định Nghĩa 2.5 *Một chuỗi thời gian thực $\{x_t; t = 0, \pm 1, \pm 2, \dots\}$ là một mô hình tự hồi quy - trung bình trượt ARMA(p,q) nếu nó là một chuỗi dừng và được cho bởi công thức*

$$x_t = \phi_1 x_{t-1} + \cdots + \phi_p x_{t-p} + w_t + \theta_1 w_{t-1} + \cdots + \theta_q w_{t-q} \tag{2.27}$$

với $\phi_p \neq 0$, $\theta_q \neq 0$ và $\sigma_w^2 > 0$. Tham số p và q lần lượt được gọi là cấp tự hồi quy và cấp trung bình trượt. Nếu x_t có xấp xỉ trung bình $\mu \neq 0$ thì có thể đặt $\alpha = \mu(1 - \phi_1 - \cdots - \phi_p)$ và mô hình có thể được viết dưới dạng

$$x_t = \alpha + \phi_1 x_{t-1} + \cdots + \phi_p x_{t-p} + w_t + \theta_1 w_{t-1} + \cdots + \theta_q w_{t-q} \tag{2.28}$$

Ta vẫn lấy w_t là chuỗi nhiễu trắng Gaussian có xấp xỉ trung bình bằng 0 và phương sai σ_w^2

Nếu $q = 0$, ta thu được mô hình tự hồi quy cấp p, $AR(p)$, và khi $p = 0$ thì ta lại thu được mô hình trung bình trượt cấp q, $MA(q)$

2.1.5.1 Xác định mô hình ARMA

Hàm tự tương quan ACF có thể giúp xác định cấp q của mô hình $MA(q)$. Ngoài ra, hàm tự tương quan từng phần PACF cũng giúp đưa ra dự đoán cấp p cho mô hình $AR(p)$. Nhưng với mô hình tự hồi quy-trung bình trượt ARMA thì cả ACF và PACF đều không thể giúp xác định bộ (p, q). Do đó, một hàm mở rộng mới được để xuất để khắc phục điều này. Nó có tên là **hàm tự tương quan mở rộng(EACF)**. Tư tưởng chính của EACF gồm 2 bước:

1. Tìm các ước lượng của các tham số tự hồi quy ϕ_i để chuyển x_t sang quá trình trượt trung bình

2. Dựa vào hàm ACF xác định cấp q cho mô hình vừa được biến đổi.

Mô tả quá trình xử lý ước lượng mô hình

1. Đồng nhất mô hình cần xác định với mô hình tự hồi quy

$$x_t = \phi_1 x_{t-1} + \phi_2 x_{t-2} + ... + \phi_p x_{t-p} + w_t \tag{2.29}$$

Ta sử dụng ước lượng bình phương tối thiểu để tìm các ước lượng tham số $\hat{\phi_i}^{(0)}$ với $i = 1, 2, .., p$.Nếu mô hình cần xác định là một mô hình $ARMA(p, q)$ thì các ước lượng tham số trên chưa chính xác nhưng ta có thể thu ước lượng cho nhiễu trắng

$$\hat{w_t}^{(0)} = x_t - \hat{\phi_1}^{(0)} x_{t-1} - \hat{\phi_2}^{(0)} x_{t-2} - ... - \hat{\phi_p}^{(0)} x_{t-p} \tag{2.30}$$

2. Do $\hat{w_t}^{(0)}$ phụ thuộc liên quan đến p theo như công thức trên, ta lại ước lượng mô hình với

$$x_t = \phi_1 x_{t-1}^{(1)} + \phi_2 x_{t-2}^{(1)} + ... + \phi_p x_{t-p} + w_t^{(1)} + \theta_1^{(1)} \hat{w}_{t-1}^{(0)} \tag{2.31}$$

Trong mô hình này ta vừa thêm một nhiễu $\hat{w}_{t-1}^{(0)}$ để ước lượng

- Nếu mô hình cần xác định là mô hình $ARMA(p, q)$ có $q = 1$ thì bộ tham số ϕ_i và θ_1 kia là bộ cần tìm.
- Nếu giá trị thực sự $q > 1$ thì bộ tham số kia vẫn chưa chính xác. Ta chuyển xuống bước 3

3. Trong trường hợp $q > 1$, ta tiếp tục ước lượng với mô hình

$$x_t = \phi_1 x_{t-1}^{(2)} + \phi_2 x_{t-2}^{(2)} + ... + \phi_p x_{t-p} + w_t^{(2)} + \theta_1^{(2)} \hat{w}_{t-1}^{(1)} + \theta_2^{(2)} \hat{w}_{t-2}^{(0)} \tag{2.32}$$

Ta mới thêm nhiễu $\hat{w}_{t-2}^{(0)}$ để ước lượng mô hình. Nếu $q = 2$ thì bộ tham số này là bộ cần tìm. Trường hợp ngược lại, ta tiếp tục thực hiện lặp lại thêm nhiễu và ước lượng đến khi tìm được bộ thỏa mãn.

Đầu ra của EACF là một bảng 2 chiều với chỉ số của dòng tương ứng với cấp p của AR và chỉ số của cột tương ứng với cấp q của MA.Giá trị tại hàng m, cột j là giá trị **hàm tự tương quan mở rộng mẫu(SEACF)**, viết tắt là $\hat{p}_j^{(m)}$. Đó cũng chính là giá trị ước lượng của hàm tự tương quan mẫu SACF cho mô hình $AR(m)$với ước lượng nhiễu

$$\hat{w_t}^{(j)} = x_t - \hat{\phi_1}^{(j)} x_{t-1} - \hat{\phi_2}^{(j)} x_{t-2} - ... - \hat{\phi_p}^{(j)} x_{t-m} \tag{2.33}$$

	MA					
AR	0	1	2	3	4	⋯
0	$\hat{\rho}_1^{(0)}$	$\hat{\rho}_2^{(0)}$	$\hat{\rho}_3^{(0)}$	$\hat{\rho}_4^{(0)}$	$\hat{\rho}_5^{(0)}$	⋯
1	$\hat{\rho}_1^{(1)}$	$\hat{\rho}_2^{(1)}$	$\hat{\rho}_3^{(1)}$	$\hat{\rho}_4^{(1)}$	$\hat{\rho}_5^{(1)}$	⋯
2	$\hat{\rho}_1^{(2)}$	$\hat{\rho}_2^{(2)}$	$\hat{\rho}_3^{(2)}$	$\hat{\rho}_4^{(2)}$	$\hat{\rho}_5^{(2)}$	⋯
3	$\hat{\rho}_1^{(3)}$	$\hat{\rho}_2^{(3)}$	$\hat{\rho}_3^{(3)}$	$\hat{\rho}_4^{(3)}$	$\hat{\rho}_5^{(3)}$	⋯
4	$\hat{\rho}_1^{(4)}$	$\hat{\rho}_2^{(4)}$	$\hat{\rho}_3^{(4)}$	$\hat{\rho}_4^{(4)}$	$\hat{\rho}_5^{(4)}$	⋯
⋮	⋮	⋮	⋮	⋮	⋮	⋯

Hình 2.1: Bảng giá trị Hàm tự tương quan mở rộng mẫu -

Ở đây, j là số vòng lặp trong quá trình xấp xỉ mô hình ở trong quá trình ước lượng mô hình 2.1.5.1 ở trên

Theo Tiao và Tsay (7) thì trong mô hình ARMA(p,1) thì

$$\hat{p}_j^{(m)} \longrightarrow \begin{cases} c, & j = q + m - p \\ 0, & j > q + m - p \end{cases}$$

với $|c| < 1$

Công thức trên giúp xác định mô hình ARMA(p,q) trong thực tế. Bảng giá trị hàm SEACF (Hình 2.1) có chỉ số của hàng biểu diễn cho cấp p của $AR(p)$ còn chỉ số của cột biểu diễn cho cấp q của $MA(q)$. Trong dữ liệu mẫu thực tế, khó để $\hat{p}_j^{(m)}$ đạt giá trị 0 một cách chính xác mà ta phải xấp xỉ rồi biểu diễn. Trong mô hình, "0" biểu diễn cho giới hạn 0 của $\hat{p}_j^{(m)}$, "x" biểu diễn giá trị $\hat{p}_j^{(m)} \neq 0$ hay đúng hơn là có giá trị lớn hơn 2 lần giá trị của ước lượng sai số tiêu chuẩn. Ta tìm kiếm 2 đường tiệm cận "0" trong bảng, điểm giao nhau chính là giá trị của (p, q). Như trong Hình 2.1.5.1, ta tìm được điểm giao là $(2, 2)$, do đó, mô hình dự kiến sẽ là ARMA(2,2)

	MA						
AR	0	1	2	3	4	5	⋯
0	x	x	x	x	x	x	⋯
1	x	x	x	x	x	x	⋯
2	x	x	0	0	0	0	⋯
3	x	x	x	0	0	0	⋯
4	x	x	x	x	0	0	⋯
5	x	x	x	x	x	0	⋯
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋯

Hình 2.2: Bảng giá trị SEACF ví dụ cho mô hình ARMA(2,2) -

2.2 Mô hình ARIMA và quy trình Box-jenkins

2.2.1 Mô hình ARIMA cho chuỗi không có tính dừng

Mô hình ARIMA(p,a,q) là sự kết hợp của mô hình ARMA(p,d) và quá trình sai phân dữ liệu(**"I"**). Nhờ quá trình sai phân d lần dữ liệu mà chuỗi không có tính chất dừng chuyển được về chuỗi mới có tính chất dừng. Sau đó ta sẽ xử lý chuỗi mới với mô hình ARMA

Định Nghĩa 2.6 *Một quá trình ngẫu nhiên x_t được gọi là ARIMA(p,d,q) nếu sau quá trình sai phân cấp d*

$$\Delta^d x_t = (1-B)^d x_t \tag{2.34}$$

ta được ARMA(p,d). x_t có thể được viết dưới dạng

$$\phi(B)(1-B)^d x_t = \delta + \theta(B) w_t \tag{2.35}$$

với $\delta = \mu(1-\phi_1 - ... - \phi_p)$, $E(\Delta^d x_t) = \mu$

Quy trình Box-Jenkins gồm 3 bước chính:

- **Xác định mô hình:** xác định xem mô hình đó thuộc loại nào AR,MA,ARMA hay ARIMA

- **Ước lượng tham số mô hình**
- **Đánh giá mô hình:** đánh giá độ chính xác mô hình vừa tìm được với dữ liệu quan sát được

2.2.2 Xác định mô hình

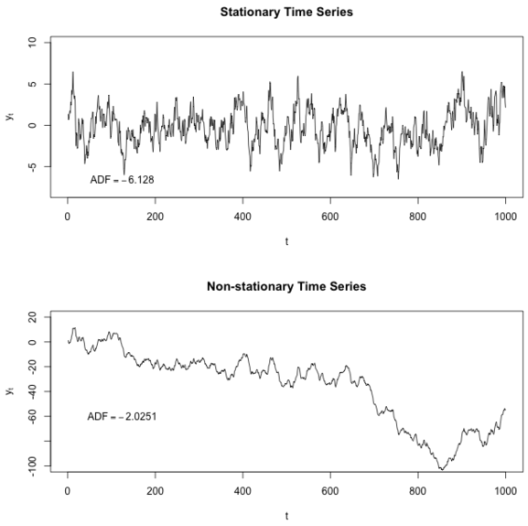
Xác định mô hình gồm 2 bước chính:

- **Chuẩn bị dữ liệu:** trước tiên ta cần xác định xem chuỗi dữ liệu có phải chuỗi dừng không(quan sát đồ thị hoặc qua hàm kiểm tra nghiệm đơn vị) và có tồn tại yếu tố mùa vụ nào trong mô hình không. Ta cần sử dụng sai phân và sai phân mùa vụ để khử các yếu tố này. Ngoài ra, với từng loại dữ liệu cụ thể chúng ta cần thực hiện chuyển đổi dữ liệu (ví dụ: logarit...) nếu dữ liệu xuất hiện xu hướng, giá trị tăng theo cấp số mũ, phương sai thay đổi...Việc này cần thực hiện trước các hàm sai phân.
- **Lựa chọn mô hình**Dữ liệu sau khi được chuyển đổi và sai phân, ta dựa vào đồ thị hàm SACF, SPACF và SEACF để xác định ước lượng cấp p và q

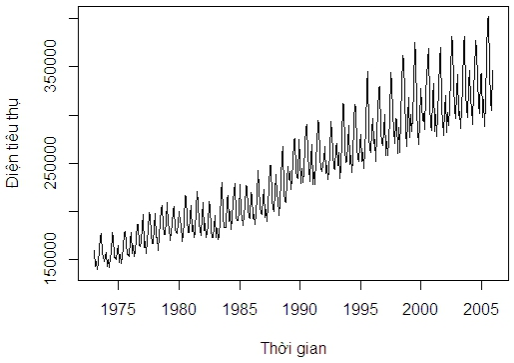
2.2.2.1 Chuẩn bị dữ liệu

Tất cả các mô hình chúng ta nhắc tới MA, AR, $ARMA$ đều ngầm định rằng dữ liệu chuỗi đầu vào là chuỗi dừng. Một chuỗi dừng thì phải có tính chất cấu trúc đồ thị hàm xấp xỉ trung bình, phương sai và tự tương quan khá ổn định, không thay đổi theo thời gian. link. Dựa vào đồ thị, chuỗi dừng nhìn phẳng,không có xu hướng, giá trị phương sai và tự tương quan ổn định, không thay đổi theo thời gian và cũng không có yếu tố chu kì, mùa vụ (Hình 2.3)

Ta thường gặp một số dữ liệu mà càng về sau thì độ biến thiên của dữ liệu quanh xấp xỉ trung bình càng lớn (hoặc càng nhỏ). Những dữ liệu như thế thì phương sai của chúng sẽ thay đổi theo thời gian. Vì vậy, nó không phải là chuỗi dừng. Ví dụ như biểu đồ lượng điện sử dụng hàng tháng của Mỹ từ tháng 1/1973 đến 12/2005 trong Hình 2.4.Từ đồ thị chúng ta có thể thấy rằng phương sai tăng dần theo thời gian, càng về cuối thì biên độ giao động càng lớn. Trước tiên chúng ta cần phải biến đổi để chuyển dữ liệu thành dữ liệu mới có phương sai ổn định.



Hình 2.3: Mô hình chuỗi dừng và không có tính chất dừng -



Hình 2.4: Lượng điện sử dụng hàng tháng của Mỹ -

Chúng ta có thể sử dụng **phép biến đổi mũ** được giới thiệu bởi Box và Cox (1964). Nếu gọi $T(x_t)$ là hàm ổn định phương sai. Chuỗi sau biến đổi $T(x_t)$ có phương sai không đổi.

$$T(x_t) = \begin{cases} \frac{x_t^\lambda - 1}{\lambda}, & \lambda \neq 0 \\ \ln(x_t), & \lambda = 0 \end{cases} \tag{2.36}$$

Tham số λ được cho bởi bảng

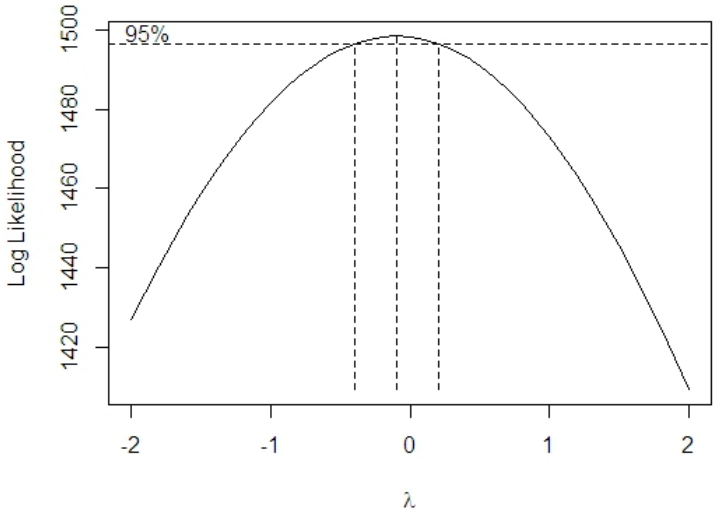
λ	$T(x_t)$
-2.0	$1/x_t^2$
-1.0	$1/x_t$
-0.5	$1/\sqrt{x_t}$
0.0	$\ln(x_t)$
0.5	$\sqrt{x_t}$
1.0	x_t
2.0	x_t^2

Bảng 2.1: Bảng tham số λ

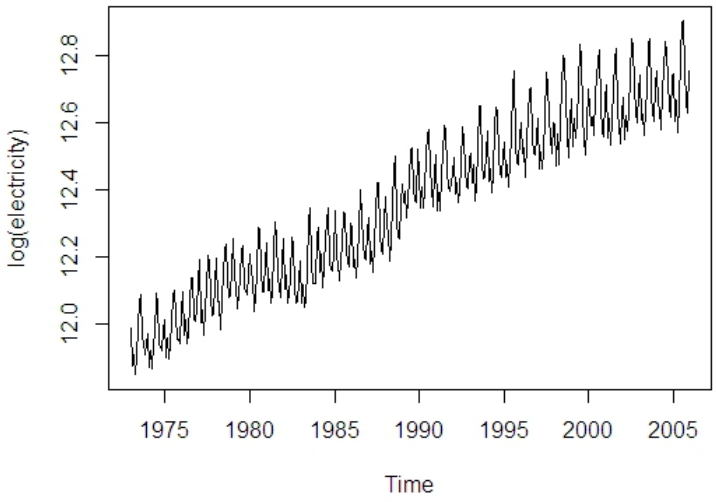
Tham số λ có thể được ước lượng thông qua các giá trị hợp lý loga(log-likehood) dựa vào các hàm ước lượng hợp lý cực đại(MLE). Với độ tin tưởng 95%, ta có đồ thị giá trị hợp lý loga như Hình 2.5.Ta thấy có thể chọn được giá trị $\lambda = 0$, sử dụng biến đổi logarit để biến đổi dữ liệu. Ta có thể quan sát thấy mô hình dữ liệu mới(Hình 2.6) đã có độ ổn định phương sai

Sau khi khử đi sự biến thiên mạnh của phương sai, chúng ta kiểm tra lần cuối tính dừng của chuỗi. Như đã mô tả trong Phần 1.5.2, ta có thể quan sát biểu đồ hàm tự tương quan mẫu SACF để dự đoán tính dừng của chuỗi. Nếu nó đạt giá trị ban đầu lớn nhưng giảm từ từ thì chuỗi đó là chuỗi không có tính chất dừng.

Ngoài ra, một số chuỗi dữ liệu không có tính chất dừng nhưng biểu đồ SACF không có tính chất trên. Ta nên sử dụng phương pháp một phương pháp thông kê khác kiểm tra một lần nữa. Cũng ở trong Phần 1.5.2, ta chỉ ra rằng nếu một mô hình có nghiệm đơn vị thì nó không phải là chuỗi dừng. Một trong những phương pháp phổ biến kiểm tra mô hình ARIMA là phương pháp Augmented



Hình 2.5: Hàm ước lượng hợp lý log dựa vào λ -



Hình 2.6: log(Electricity) -

Dickey–Fuller(ADF) (c2 trang 13) do Dickey và Fuller (1979 và 1981) tạo ra. Họ nghiên cứu 3 hàm hồi quy khác nhau để kiểm tra nghiệm đơn vị cho x_t

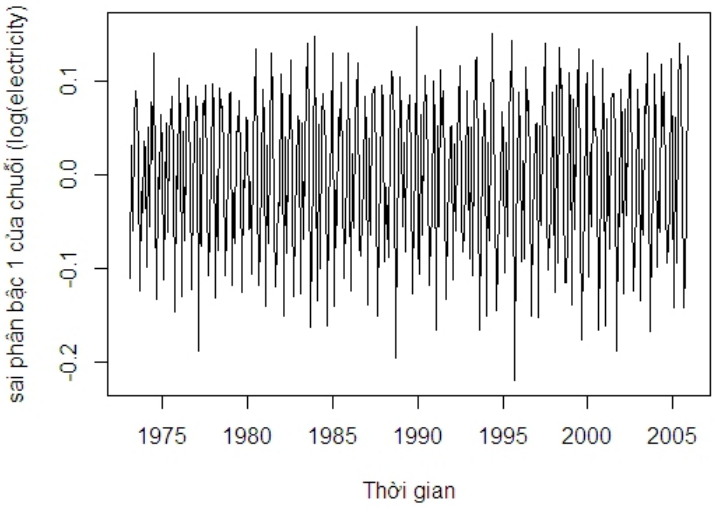
$$\Delta x_t = ax_{t-1} + \sum_{j=1}^{J} \varphi_j \Delta x_{t-1} + \beta t + w_t, \qquad t = 1, 2, ..., T \tag{2.37}$$

$$\Delta x_t = ax_{t-1} + \sum_{j=1}^{J} \varphi_j \Delta x_{t-1} + \mu + w_t, \qquad t = 1, 2, ..., T \tag{2.38}$$

$$\Delta x_t = ax_{t-1} + \sum_{j=1}^{J} \varphi_j \Delta x_{t-1} + w_t, \qquad t = 1, 2, ..., T \tag{2.39}$$

với Δx_0 cố định. Biến phụ thuộc là Δx_t Họ đã chứng minh rằng x_t có nghiệm đơn vị nếu $a = 0$. ADF sẽ kiểm tra giả thiết xuôi $H_0 : a = 0$ (không có tính chất dừng) và giả thiết ngược $H_1 : a < 0$(có tính chất dừng) Chúng ta sẽ kiểm tra bằng cách hồi quy Δx_t trên x_{t-1}, Δx_{t-1}, Δx_{t-2},...,Δx_{t-k}. Ta ước lượng a bằng phương pháp bình phương tối thiểu. Nếu giá trị ước lượng a nhỏ hơn 0 đáng kể thì kết luận được đây là chuỗi dừng còn ngược lại thì nó là chuỗi không có tính chất dừng. Vậy thế nào là "đáng kể"?. Ta sẽ dựa vào giá trị xác suất **p-value** để đánh giá giả thuyết. p-value chính là ước tính xác suất tồn tại các tham số nêu giả thuyết null là đúng. Nếu giá trị này nhỏ hơn ngưỡng quy định (thường là 0.1 hoặc 0.05) thì giả thuyết null bị bác bỏ giá trị ước lượng được chấp nhận còn ngược lại thì kết quả chưa tin tưởng được. Ví dụ như với dữ liệu nhiệt độ trái đất (không khí và nước) hàng năm tính từ giữa thế kỉ 20, ta thu được kết quả kiểm tra ADF như sau: $(a \ -0.245, p - value = 0.1)$. Do $p - value > 0.1$ nên dù $a < 0$ thì vẫn không tin tưởng được vào ước lượng. Nhưng nếu giả sử $p - value = 0.01 < 0.1$ thì do $a < 0$ nên ta có thể kết luận chuỗi này có tính chất dừng.

Sau khi đã xác định được rằng chuỗi dữ liệu không có tính chất dừng, ta cần làm thực hiện sai phân d lần cho đến khi thu được chuỗi mới có tính chất dừng. Thông thường trong thực tế, chỉ cần thực hiện tối đa 2 lần là thu được chuỗi dừng. Trong ví dụ về mức tiêu thụ điện ở trên, sau khi sai phân chuỗi $log(electricity)$ ta thu được chuỗi mới có đồ thị(Hình 2.7)



Hình 2.7: Đồ thị giá trị chuỗi sai phân bậc 1 của $log(electricity)$ -

2.2.2.2 Lựa chọn mô hình:

Sau khi xác định được tham số d trong mô hình ARIMA, ta cần ước lượng 2 tham số p và d còn lại. Phương pháp chính để ước lượng dự đoán 2 tham số này là sử dụng đồ thị của hàm tự tương quan của mẫu(SACF), hàm tự tương quan từng phần của mẫu(SPACF) và hàm tự tương quan mở rộng của mẫu ((SEACF) Theo như Phần 2.1.3 thì giá trị hàm tự tương quan từng phần của mô hình $AR(p)$ có giá trị xấp xỉ bằng 0 từ trễ $p+1$ trở đi.Do đó, ta sẽ quan sát đồ thị SPACF giá trị p mà sau giá trị đó đồ thị tiến sát giá trị 0. Thông thường, trong đồ thị SPACF có 2 đường giới hạn giá trị sai số là $y = \pm 2/\sqrt{N}$ với N là cỡ của mẫu dữ liệu. Nếu quan sát được giá trị p phù hợp như vậy thì một ước lượng của mô hình ta cần tìm là $ARIMA(p, d, 0)$ hay $AR(p)$(sau khi sai phân d lần)

Tương tự, theo Phần 2.1.4, giá trị hàm tự tương quan của mô hình $MA(q)$ sẽ giảm xuống 0 từ giá trị $p+1$ trở đi. Mô hình SACF cũng có 2 đường giới hạn sai số $\pm 2/\sqrt{N}$ với N là cỡ của dữ liệu mẫu. Nếu quan sát được giá trị q mà sau nó, đồ thị giao động nhẹ quanh giá trị 0 thì có thể kết luận rằng một trong những ước lượng mô hình phù hợp là $ARIMA(0, d, q)$ hày $MA(q)$(sau khi sai phân d lần).

Dưới đây là một số dự đoán mô hình dựa vào đồ thị SACF:

Hình dạng	**Mô hình ước lượng**
Giá trị lớn, triệt tiêu dần	Mô hình RA (dùng thêm đồ thị SACF để xác định)
Cả giá trị âm và dương, triệt tiêu dần	Mô hình RA
Một hoặc một số đỉnh trội , còn lại xấp xỉ 0	Mô hình MA, tìm điểm q mà đồ thị bắt đầu về 0
Giảm dần, tăng sau một số khoảng trễ	Mô hình ARMA
Tất cả bằng hoặc xấp xỉ gần 0	dữ liệu ngẫu nhiên
Cao tại một số điểm nhất định	chứa yếu tố chu kì
Không giảm tới 0	không phải chuỗi dừng

Bảng 2.2: Bảng dự đoán mô hình dựa vào đồ thị SACF

Nếu không thể dự đoán được bằng đồ thị SACF và SPACF, đó có thể là mô hình trộn ARMA. Theo phần 2.1.5, ta cần sử dụng hàm tự tương quan mở rộng để ước lượng tham số p và q

2.2.3 Ước lượng tham số

Sau khi xác định p, d, q ta thu được mô hình. Bước tiếp theo là xác định các tham số cho công thức mô mình.

2.2.3.1 Ước lượng bình phương tối thiểu

Theo công thức 2.27, mô hình được cho bởi công thức:

$$x_t = \sum_{j=1}^{p} \phi_j x_{t-j} + w_t + \sum_{j=1}^{q} \theta_j w_{t-j} \tag{2.40}$$

Ta có,

$$w_t = x_t - \sum_{j=1}^{p} \phi_j x_{t-j} - \sum_{j=1}^{q} \theta_j w_{t-j} \tag{2.41}$$

Giả sử chỉ có n mẫu $x_1, x_2, ..., x_n$ được quan sát. Khi đó ta chỉ cần tính toán từ $t = 2$ đến $t = n$. Gọi hàm $S_c(\phi, \theta)$ là hàm tổng bình phương có điều kiện, được cho bởi

$$S_c(\phi, \theta) = \sum_{t=2}^{n} w_t^2 \tag{2.42}$$

$$= \sum_{t=2}^{n} [x_t - \sum_{j=1}^{p} \phi_j x_{t-j} - \sum_{j=1}^{q} \theta_j w_{t-j}]^2 \tag{2.43}$$

Mục tiêu là ước lượng các giá trị ϕ, θ sao cho giá trị hàm $S_c(\phi, \theta)$ đạt giá trị cực tiểu. Giá trị nhiễu trắng w_t sẽ được ước lượng với sai số phương sai

AR(p): $\hat{\sigma}_w^2 = (1 - \sum_{j=1}^{p} \hat{\phi}_j \hat{\rho}_j)$

MA(q): $\hat{\sigma}_w^2 = \frac{S^2}{1+\hat{\theta_1^2}+\hat{\theta_2^2}+\cdots+\hat{\theta_q^2}}$

ARMA(1,1) $\hat{\sigma}_w^2 = S^2 . \frac{1-\hat{\phi^2}}{1-2\hat{\phi}\hat{\theta}+\hat{\theta^2}}$

Ví dụ(trang 200 f11notes)

2.2.3.2 Ước lượng hợp lý cực đại

Dây là phương pháp hay được sử dụng nhất để ước lượng các tham số chưa biết. Gọi hàm mật độ xác suất của nhiễu w_t là $f(w_t)$. Hàm xác suất đồng thời của $w_2, w_3, ..., w_n$ được xác định bởi:

$$f(w_2, w_3, ..., w_n) = \prod_{t=2}^{n} f(w_t) \tag{2.44}$$

Giá trị x_1 cố định.**Hàm hợp lý** sẽ được cho bởi công thức:

$$L = L(\phi, \theta, \sigma_w^2 | x) = f(x_2, x_3, ..., x_n | x_1) f(x_1) \tag{2.45}$$

Khi đó, ước lượng hợp lý cực đại của ϕ, θ, σ_w^2 sẽ là những giá trị mà làm cho $L(\phi, \theta, \sigma_w^2 | x)$ đạt giá trị cực đại.

2.2.4 Kiểm định mô hình

Bước cuối cùng để xác định hoàn toàn công thức của mô hình là kiểm định lại các giá trị đã được ước lượng trong phần trước.

2.2.4.1 Phân tích sai số thặng dư

Sai số thặng dư(residuals) là lượng chênh lệch giữa mẫu quan sát và giá trị của hàm ước lượng cho tập mẫu ấy tại một thời điểm xác định. Nó được xác định bởi công thức:

Sai số thặng dư tại t = (giá trị thực tế của mẫu x_t) - (giá trị của mô hình ước lượng x_t) (2.46)

Nếu mô hình được ước lượng tốt thì chuỗi sai số thặng dư của nó thường sẽ có một số tính chất gần giống như tính chất của nhiễu trắng Gassian. Nó gần giống như một chuỗi các biến ngẫu nhiên chuẩn, độc lập và có cùng phân phối xác suất. Ngoài ra, nó có xấp xỉ trung bình $\mu_w = 0$ và có độ lệch tiêu chuẩn ổn định. Về cơ bản, sai số thặng dư được tính bởi công thức $\hat{e}_t = x_t - \hat{x_t}$ nhưng chúng ta thường xử lý trên **sai số thặng dư chuẩn hóa**

$$\hat{e_t}^* = \frac{\hat{e_t}}{\hat{\sigma_e}} \tag{2.47}$$

với $\hat{\sigma_e^2}$ là ước lượng của phương sai nhiễu trắng σ_w^2 Ta cần kiểm tra mức độ chuẩn tắc và độc lập của chuỗi sai số thặng dư chuẩn hóa. Biểu đồ và đồ thị qq của sai số thặng dư có thể dùng để đánh giá trực quan về tính chuẩn tắc của nó. Đồ thị chuỗi của sai số thặng dư có thể giúp phát hiện các mẫu mà vi phạm đến tính độc lập của các sai số thặng dư. Ngoài ra chúng ta áp dụng kiểm tra giả thuyết về chuẩn tắc của Shapiro-Wilk và tính độc lập (runs test) đối với sai số thặng dư chuẩn hóa. Từ phân phối chuẩn chuẩn hóa, ta biết rằng hầu hết các giá trị $\hat{e_t^*}$ nằm trong khoảng -3 đến 3 theo tiêu chuẩn Bon-ferroni outlier với $\alpha = 5\%$ và $n = 241$

- Phương pháp **Shapiro-Wilk** kiểm tra 2 giải thuyết:
  H_0: chuỗi sai số thặng dư chuẩn hóa có phân phối chuẩn.
  H_1: chuỗi sai số thặng dư chuẩn hóa không có phân phối chuẩn

- Phương pháp **runs test** kiểm tra 2 giải thuyết:
  H_0: chuỗi sai số thặng dư chuẩn hóa có tính độc lập.
  H_1: chuỗi sai số thặng dư chuẩn hóa không có tính độc lập.

- Trong cả 2 phương pháp, nếu giá trị xác suất $p - value$ quá nhỏ thì giải thuyết H_0 sẽ bị loại bỏ.

Ví dụ trang 217 f11notes

Ngoài ra, Ljung và Box (1978) đã phát triển một hàm kiểm tra dựa vào sự tương tương quan mẫu của sai số thặng dư để kiểm tra liệu một mô hình $ARMA(p,q)$ có phù hợp hay không. Thật vậy, công thức biến đổi $Ljung - Box$ được cho bởi công thực:

$$Q_* = n(n+2)\sum_{k=1}^{K}\frac{\hat{r_k^2}}{n-k} \tag{2.48}$$

với $\hat{r_k}$ là các giá trị tự tương quan mẫu của sai số thặng dư được tính dựa vào mô hình ARMA(p,q). được sử dụng để kiểm tra giả thuyết xuôi H_0(mô hình ARMA phù hợp) và giả thiết ngược H_1(mô hình ARMA không phù hợp) Với K xác định cho trước, chọn một mức α để quyết định giả thuyết nào đúng. Nếu giá trị của Q_* vượt quá phân phối χ với điểm sai phân α và độ tự do $K-p-q$ hay nói cách khác nếu $Q_* > \chi^2_{K-p-q}$ thì mô hình $ARMA(p,q)$ này không phù hợp.

2.2.5 Dự báo

Từ dữ liệu mẫu có sẵn $x_1, x_2, ..., x_t$, chúng ta cần dự đoán các kết quả tương lai $x_{t+1}, x_{t+2}, ...$Giá trị dự đoán tại thời điểm $t+l$, kí hiệu là $\hat{x}_t(l)$, chính là giá trị kì vọng của x_{t+l} với điều kiện đã có t phần tử mẫu quan sát $x_1, x_2, ..., x_t$. Ta gọi nó là $HmdØonMMSE$(minimum mean squared error forecast) được cho bởi công thức sau:

$$\hat{x}_t(l) = E(x_{t+l}|x_1, x_2, ..., x_t) \tag{2.49}$$

Một số tính chất của kỳ vọng có điều kiện:

- $E(c|x_1, x_2, ..., x_t) = c$ với c là hằng số

- $E(\phi_i x_i|x_1, x_2, ..., x_t) = \phi_i x_i$ với $i = 1, 2, ..., t$

- $E(w_{t+1}|x_1, x_2, ..., x_t) = E(w_{t+1}) = 0$ với w_t là nhiễu trắng, bởi vì w_{t+1} độc lập với $x_1, x_2, ..., x_t$

Nếu mô hình ước lượng là ARMA(p,q) được cho bởi công thức

$$x_t = \phi_1 x_{t-1} + \cdots + \phi_p x_{t-p} + w_t + \theta_1 w_{t-1} + \cdots + \theta_q w_{t-q} \tag{2.50}$$

Ta tính được giá trị tương lai tại $t + l$ là

$$\hat{x}_t(l) = \sum_{j=1}^{p} \phi_j \hat{x}_t(l-j) + w_t + \sum_{j=1}^{p} \phi_j E(w_{t+l-j}|x_1, x_2, ..., x_t) \tag{2.51}$$

Trong đó,

$$E(w_{t+l-j}|x_1, x_2, ..., x_t) = \begin{cases} 0, & l-j > 0 \\ w_{t+l-j}, & l-k \leq 0 \end{cases} \tag{2.52}$$

$$\hat{x}_t(l-j) = E(x_{t+l-j}|x_1, x_2, ..., x_t) \qquad j = 1, 2, .., p \tag{2.53}$$

Ta thực hiện đệ quy nhiều vòng để tính được công thức chung cho $\hat{x}_t(l)$ Ví dụ với mô hình $ARMA(1,1)$ được cho bởi

$$x_t = \phi x_{t-1} + w_t + \theta w_{t-1} \tag{2.54}$$

Sử dụng một số tính chất kỳ vọng có điều kiện và tính toán đệ quy, ta thu được công thức

$$\hat{x}_t(1) = \phi x_t - \theta w_t \tag{2.55}$$

$$\hat{x}_t(l) = \phi \hat{x}_t(l-1) \tag{2.56}$$

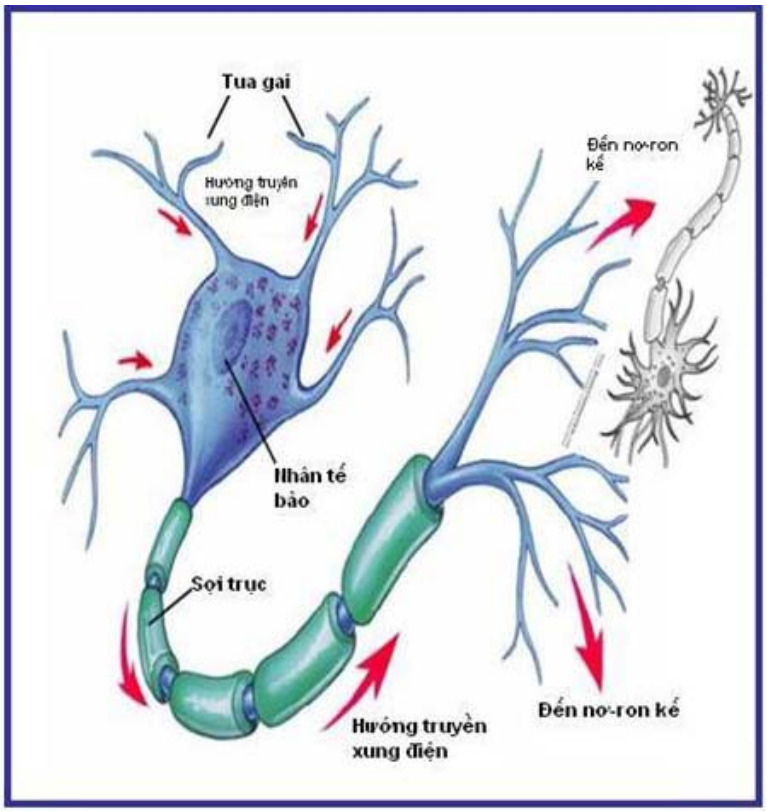
2.3 Mô hình mạng nơ ron nhân tạo

Mạng nơ ron nhân tạo là một cấu trúc tính toán mô phỏng theo hoạt động của bộ não người. Mạng nơ ron nhân tạo giải quyết tốt các bài toán tìm mẫu và phân lớp, xấp xỉ hàm, tối ưu hóa, lượng giá vectơ và thu gộp dữ liệu. Mạng nơ ron nhân tạo còn là một hệ thống bao gồm nhiều phần tử xử lý đơn giản hoạt động song song. Tính năng của hệ thống này tùy thuộc vào cấu trúc của hệ, các trọng số liên kết nơ ron và quá trình tính toán tại các nơ ron đơn lẻ. Mạng nơ ron có thể học từ dữ liệu mẫu và tổng quát hóa dựa trên các dữ liệu mẫu học đó.

2.3.1 Kiến trúc mạng nơ ron

2.3.1.1 Mô hình nơ ron

Nơ ron là một phần tử đơn giản nhất trong một mạng lưới phức tạp khoảng 100 tỷ nơ ron thần kinh. Mỗi nơ-ron có 3 phần chính: thân tế bào (soma), tua gai

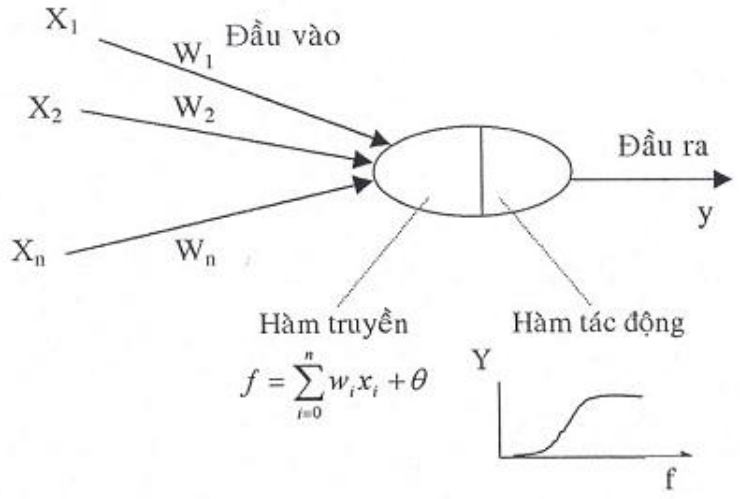


Hình 2.8: Nơ ron thần kinh - link

(dendrite – nhận tín hiệu từ các tế bào não khác) và sợi trục (axon – giúp truyền tín hiệu đến các tế bào não khác). Khi tua gai của tế bào thần kinh nhận kích thích, một tín hiệu điện (xung điện) sẽ truyền qua nhân tế bào và dọc theo sợi trục. Ở cuối sợi trục có một khe nhỏ giữa hai nơ-ron (khe xi-náp). Tín hiệu được truyền đi từ đầu cuối sợi trục của tế bào thần kinh này đến tua gai của một tế bào khác thông qua các chất dẫn truyền thần kinh (các chất hóa học ở cuối đầu sợi trục). Chất dẫn truyền thần kinh sẽ băng qua khe xi-náp và kết dính vào các thụ thể (receptors) ở tua gai của một tế bào khác, rồi lại kích thích nó tạo ra

xung điện truyền đến tế bào kế tiếp.

Dựa vào mô hình nơ ron thần kinh, Widrow và Hoff đề xuất mô hình nơ ron nhân tạo (8) được mô tả trong Hình 2.9



Hình 2.9: Nơ ron nhân tạo -

Nơ ron nhân tạo gồm 3 thành phần chính:

- Các tín hiệu đầu vào x_i với $i = 1, 2, ..., n$ kết nối tới lõi. Mỗi tín hiệu x_i sẽ có một trọng số tương ứng w_i
- Lõi(perceptron) nhận tín hiệu đầu vào là các x_i và một giá trị ngưỡng θ. Hàm truyền vào lõi được tính bởi công thức

$$g = \sum_{i=1}^{n}(w_i x_i + \theta)$$

- Tín hiệu đầu ra y được tính bởi công thức hàm kích hoạt (tác động)

$$y = f(\sum_{i=1}^{n}(w_i x_i + \theta))$$

với điều kiện

$$\sum_{i=1}^{n}(w_i x_i + \theta) \geq 0$$

Một số hàm kích hoạt hay được sử dụng như:

Hàm bước nhảy

$$f(x) = \begin{cases} 1 & khif \geq 0 \\ 0 & khif < 0 \end{cases} \tag{2.57}$$

Hàm dấu

$$f(x) = sgn(x) \begin{cases} 1 & khif \geq 0 \\ -1 & khif < 0 \end{cases} \tag{2.58}$$

Hàm sigmoid

$$f(x) = \frac{1}{1 + e^{-x}} \tag{2.59}$$

2.3.1.2 Mạng nơ ron nhân tạo (ANN) cho bài toán dự báo

Mạng nơ ron nhân tạo được sử dụng trong để giải quyết rất nhiều bài toán thuộc nhiều lĩnh vực khác nhau, giải quyết hiệu quả một số bài toán như: bài toán phân lớp, bài toán điều khiển và tối ưu hóa,bài toán dự báo... Trong bài toán dự báo, mạng nơ ron nhân tạo được sử dụng để xây dựng mô hình dự báo dựa vào các dữ liệu trong quá khứ. Sau đó, ta có thể sử dụng mô hình đó để dự báo kết quả trong tương lai. Một số mô hình mạng nơ ron nhân tạo như :

1. Mạng nơ ron nhiều lớp MLP
2. Mạng bán kính-tâm(Radial Basic Function RBF)
3. Mạng hồi quy
4. Mạng chống lan truyền
5. Mạng nơ ron xác suất

Mạng nơ ron nhiều lớp (Multiplayer Perceptron Network), viết tắt là **MLP**, hay còn gọi là mạng nơ ron truyền thẳng nhiều lớp gồm có 3 tầng chính:Tầng thứ nhất gồm một **lớp đầu vào**, tầng thứ hai trung gian gồm một hoặc một vài **lớp ẩn**, và tầng cuối cùng gồm một lớp đầu ra. Các lan truyền thông tin được xử lý thẳng từ đầu vào cho đến đầu ra theo một hướng. Do kiến trúc phân tầng nên MLP có thể biểu diễn một quan hệ phụ thuộc phi tuyến tính

giữa đầu vào và đầu ra. Ví dụ như mô hình mạng nơ ron 3 lớp (chỉ có một lớp ẩn), ta có công thức liên hệ giữa đầu ra và đầu vào

$$y = f_o(\sum w_h f_h(\sum f_i(w_i^T x))) \tag{2.60}$$

với f_i, f_h, f_o lần lượt là hàm kích hoạt của lớp đầu vào, ẩn và đầu ra. w là trọng số

Trong hầu hết các ứng dụng dự báo thông thường, ta chỉ cần sử dụng một lớp ẩn cho tầng trung gian. Thật vậy,theo **Định lý chồng Kolmogorov**, bất kỳ hàm liên tục đa chiều $f : [0, 1]^n \to R$ nào cũng có thể được viết dưới dạng:

$$f(x_1, x_2, ..., x_n) = \sum_{i=1}^{2n+1} \psi_i(\sum_{j=1}^{n} \varphi_{ij}(x_j)) \tag{2.61}$$

với ψ_i và φ_{ij} là các hàm liên tục một biến.

Mặt khác, theo Hecht-Nielsen (9) thì công thức trong định lý của Kolmogorov có thể được biểu diễn hoàn toàn bởi một mạng nơ ron truyền thẳng 3 lớp trong đó lớp đầu vào có n phần tử , lớp đầu ra có m phần tử còn lớp trung gian có $2n + 1$ phần tử. Do vậy, một mô hình nơ ron 3 lớp như vậy có thể biểu diễn mọi hàm liên tục nhiều chiều. Hay nói cách khác, một mô hình mạng truyền thẳng 3 lớp đủ để giải quyết các bài toán dạng xác định kiến trúc mô hình mạng nơ ron truyền thẳng (Lippmann 197)và không bao giờ cần đến quá 4 lớp mạng để giải quyết các bài toán phức tạp nhất(Cybenko 1989)

Rumelhart và cộng sự (10) đề xuất phương pháp học lan truyền ngược(backpropagation) cho MLP. Và ngày nay, nó trở thành một phương pháp được áp dụng rộng rãi cho mạng nơ ron nhiều lớp. Ý tưởng chính của phương pháp này là đánh giá sai số giữa dữ liệu đầu ra tính toán trong mô hình và đầu ra quan sát được. Sai số sẽ được lan truyền ngược lại lớp trước để điều chỉnh tối ưu các trọng số w_i của mạng.

Ngoài ra, ta cũng có thể sử dụng giải thuật di truyền để tối ưu các trọng số.

2.3.2 Phương pháp huấn luyên

Phần này ta tập trung vào một số phương pháp huấn luyện cho mạng nơ ron truyền thẳng nhiều lớp.

2.3.2.1 Thuật toán lan truyền ngược

Thuật toán lan truyền ngược(backpropagation) là phương pháp học có giám sát được Paul Werbos phát triển năm 1971 nhưng nó chỉ thực sự trở lên phổ biến và được sử dụng rộng rãi sau khi được Rumelhard (10) chỉ ra rõ ràng năm 1986.

Trong quá trình huấn luyện, mạng nơ ron nhiều lớp học thông qua quá trình điều chỉnh dần các trọng số sao cho chênh lệch giữa đầu ra theo tính toán và đầu ra thực tế quan sát được là nhỏ nhất. Việc tối ưu hóa trọng số dựa vào luật **giảm đạo hàm** (descent gradient),tức là lấy đạo hàm của hàm chi phí(hàm giá) theo trọng số, thay đổi các trọng số đó tiến dần tới cực trị địa phương của hàm chi phí.

Giả sử hàm $y_j = f(u_j)$ là giá trị hàm kích hoạt tại noron thứ j lớp đầu ra, u_j là đầu vào của nơ ron này.

$$u_j = \sum_{i=1}^{n} w_i x_i + \theta_j \tag{2.62}$$

với n là số lượng nơ ron lớp ẩn cuối cùng trực tiếp nối đến nơ ron lớp đầu ra. x_i,w_i là đầu vào và trọng số tương ứng thứ i nối tới nơ ron. Nếu d_j là giá trị thực quan sát được tại nơ ron đầu ra thứ j thì hàm sai số được cho bởi

$$S_j = 0.5(d_j - y_j)^2 = 0.5(e_j)^2 \tag{2.63}$$

Mục tiêu của giải thuật học là đưa S_j tiến về giá trị tối thiểu, bằng cách thay đổi các trọng số w_j. Sau mỗi vòng lặp, Δw_i được tính thông qua công thức:

$$\Delta w_i = -\eta \frac{\nabla S_j}{\nabla w_i} \tag{2.64}$$

với $\eta > 0$ là tham số xác định tốc độ hội tụ về cực tiểu. Ta tính được

$$\Delta w_i = \eta \delta_j . x_i \tag{2.65}$$

với $\delta_j = e_j f'(u_j)$ Cuối cùng, ta thực hiện cập nhật thay đổi lại cho từng trọng số w_i. Lặp lại đến khi $S_j < S_{max}$ là ngưỡng sai số lớn nhất cho phép.

2.3.2.2 Giải thuật di truyền

Phương pháp lan truyền ngược trình bày ở trên điều chỉnh các trọng số để đưa về mục tiêu cuối cùng là tối thiểu hóa hàm sai số S_j. Nó dự trên cơ chế giảm đạo hàm(giảm gradient) để đưa hàm giá (trong trường hợp này là hàm sai số) về cực trị. Đôi khi rất khó khăn để đưa về cực trị đối với một số hàm sai số. Một trong những hướng giải quyết khác là sử dụng giải thuật di truyền dựa (vào chọn lọc tự nhiên đưa) giúp tối ưu trọng số để đưa hàm giá về cực trị toàn cục.

Mỗi vector đầu vào được coi là một nhiễm sắc thể.

Giải thuật di truyền đơn giản gồm có 3 toán tử:

- **Tái tạo**(reproduction) là quá trình trong đó các chuỗi nhiễm sắc thể được sao chép lại để thực hiện thao tác sinh sản. Xác suất được chọn cho quá trình sinh sản phụ thuộc với giá trị hàm mục tiêu (hàm thích nghi). Nếu nhiễm sắc thể có độ thích nghi cao thì khả năng nó được chọn để thao tác sinh sản cũng lớn hơn.

- **Lai ghép**(Crossover) Khi mỗi chuỗi được chọn để sinh sản thì một bản sao chính xác của nó sẽ được cho vào bể ghép. Hai nhiễm sắc thể sẽ được lai ghép ngẫu nhiên tron bể. Quá trình lai ghép có thể đơn giản là chỉ đổi các gen có cùng vị trí hay số thứ tự giống nhau. Ví dụ như trao đổi bít thứ 2 trong chuỗi cho nhau.

- **Đột biến**(Mutation) Việc đột biến là cần thiết bởi vì đôi khi chỉ với 2 quá trình trên có thể làm mất đi một vài gen có ích nào đó. Sự đột biến diễn ra ngẫu nhiên và xác suất nhỏ đối với một số vị trí gen trong chuỗi. Ví dụ như đơn giản chỉ tự động đổi giá trị của một vị trí gen nào đó. Nhưng đôi khi nó lại giúp bảo vệ duy trì một số gen quan trọng trong một số trường hợp.

Sơ đồ của một giải thuật di truyền đơn giản:

1. Khởi tạo quần thể ban đầu của chuỗi nhiễm sắc thể

2. Xác định hàm giá trị mục tiêu cho mỗi chuỗi nhiễm sắc thể

3. Tạo các chuỗi nhiễm sắc thể mới bằng sinh sản từ các chuỗi nhiễm sắc thể hiện tại, có thể dùng đến ghép chéo hoặc đột biến nếu cần

4. Xác định hàm mục tiêu cho các chuỗi nhiễm sắc thể mới và đưa nó vào trong một quần thể mới.

5. Nếu điều kiện dừng đã thỏa mãn thì dừ ng lại và trả về chuỗi nhiễm sắc thể tốt nhất cùng với giá trị hàm mục tiêu của nó, nếu không thì quay về bước 3

Ta cài đặt giải thuật di truyền để tối ưu giá trị trọng số trong mô hình mạng MLP. Phương pháp này được Emre Gaglar đề xuất trong dự án "CUDAANN r6". Ta có một số tùy biến sau:

- **Hàm mục tiêu** sử dụng hàm trung bình của bình phương lỗi (MSE). Lỗi ở đây chính là chênh lệch giữa giá trị đầu ra tính toán được và giá trị đầu ra quan sát thực tế.
$$MSE = \frac{1}{N}\sum(y_i - d_i)^2$$

- **Quần thể ban đầu** Quần thể ban đầu được chọn ngẫu nhiên.

- **Đột biến** Đột biến giá trị trọng số được tính theo công thức
$$w_{ji}^{new} = w_{ki} + \alpha(w_{nj} - w_{mi})$$
với α là tỉ lệ đột biến, w_{ji} là trọng số thứ i của nhiễm sắc thể thứ j và bộ $\{k, j, i\}$ đôi một khác nhau, được chọn ngẫu nhiên từ quần thể.

- **Lai ghép** khi 2 nhiễm sắc thể lai ghép với nhau, chúng trao đổi lần lượt từng trọng số tương ứng với nhau với xác suất là *tỉ lệ lai ghép* cho trước.

2.3.3 Dự báo

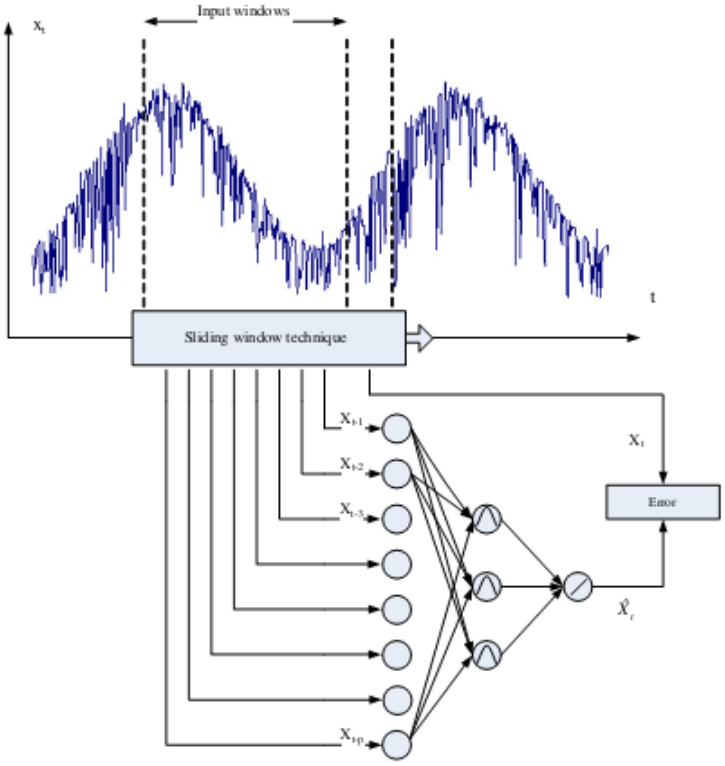
Dự báo chuỗi thời gian đơn chiều(univariate) sử dụng mạng nơ ron truyền thẳng nhiều lớp, ta xử lý qua 4 bước:

- Chuẩn bị dữ liệu: sử dụng kĩ thuật cửa sổ trượt để chia thành các tập dữ liệu(bộ học, kiểm tra, đánh giá), chuẩn hóa...

- Xác định kiến trúc mạng nơ ron: kiểu mạng, số lớp, số nơ ron mỗi lớp, hàm kích hoạt...
- Huấn luyện mạng: chọn giải thuật huấn luyện, tối ưu thuật toán...
- Đánh giá kết quả dự báo

2.3.3.1 Chuẩn bị dữ liệu

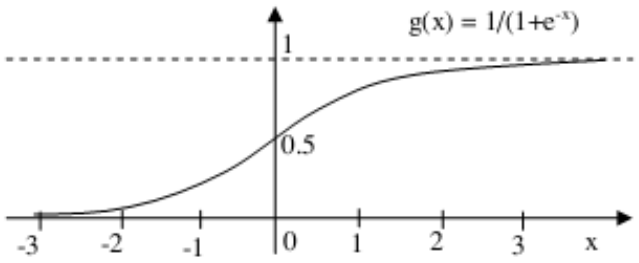
Trong bài toán dự báo chuỗi thời gian, để dự đoán 1 thời điểm trong tương lai, mạng MLP nhận đầu vào p giá trị trong quá khứ trước nó và đầu ra là giá trị thời điểm cần dự báo. Do tính chất của chuỗi thời gian đơn biến chỉ gồm một chuỗi đơn, trong khi đầu vào mạng là một vector nên ta sử dụng **kĩ thuật cửa sổ trượt**(sliding-window) (Hình 2.10). MLP nhận vector $\{x_{t-1}, x_{t-2}, ...x_{t-p}\}$ đầu vào và giá trị x_t là đầu ra. $\hat{x}_t$ là giá trị đầu ra dựa vào tính toán của mô hình.



Hình 2.10: Kĩ thuật cửa sổ trượt -

Để tăng chất lượng cho MLP sử dụng hàm kích hoạt sigmoid, dữ liệu đầu vào nên được **chuẩn hóa** về miền $[0, 1]$. Thật vậy, quan sát đồ thị hàm sigmod (Hình 2.11), ta thấy đồ thị tiến về tiệm cận khi $x \to \infty$. Do đó, khi $|x|$ không nằm trong $[-1, 1]$ thì giá trị hàm sigmoid tiến gần tới giá trị bão hòa. Ta cần phải chuẩn hóa các giá trị đầu vào lớn , nếu không thì các nơ ron ở ngay lớp ẩn đầu tiên đã đạt tới giá trị bão hòa, quá trình học không được chất lượng như mong muốn. Trong thực tế, ta sử dụng chuẩn hóa đơn giản $x_{ni} = x_i/x_{max}$ hoặc công thức chuẩn hóa tuyến tính hay được sử dụng hơn

$$x_{ni} = \frac{x_i - x_{min}}{x_{max} - x_{min}}$$



Hình 2.11: Đồ thị hàm sigmoid -

2.3.3.2 Xác định kiến trúc mạng

Để xác định kiến trúc mạng, ta quan tâm đến một số yếu tố sau đây

- Xác định số nút lớp đầu vào: số nút đầu vào phần lớn chủ yếu phụ thuộc vào số lượng biến độc lập trong tập dữ liệu. Mỗi biến độc lập nên được cho tương ứng với một nút, tuy từng nhu cầu. Trong trường hợp dữ liệu đầu vào cho dự đoán thì số lượng nút đầu vào chính bằng độ dài vector các giá trị quá khứ dùng để dự đoán tương lai.

- Xác định số nút lớp đầu ra: tùy vào yêu cầu bài toán.

- Số lớp ẩn trong tầng trung gian. Như đã trình bày ở Mục 2.3.1.2 thì đối với các bài toán thông thường, mô hình MLP với chỉ một lớp ẩn là đủ để truyền đạt hết thông tin của dữ liệu. Trong một số trường hợp bài toán khó, ta có thể dùng đến 2 lớp ẩn.

- Xác định số noron ở lớp ẩn. Không có một phương pháp chính thức nào để xác định số noron trong lớp ẩn. Dựa vào kinh nghiệm, có một số quy tắc để thử như: số nơ ron lớp sau bằng khoảng 75% số nơ ron của lớp liền trước hoặc tỉ lệ này nằm trong khoảng 0.5-3,...Theo **luật hình tháp** thì đề xuất

$$N_h = \alpha\sqrt{N_i * N_o} \tag{2.66}$$

  trong đó N_o, N_h, N_i lần lượt là số lượng nơ ron ở lớp đầu vào, lớp ẩn và lớp đầu ra. α là tham số trong khoảng $[0.5, 2]$. Baum and Haussler (1989) thì đề xuất công thức

$$N_h \neq \frac{N_{tr} * E_{tol}}{N_{dp} + N_o} \tag{2.67}$$

  với N_{tr} là số mẫu trong bộ học, E_{tol} là sai số tối đa cho phép, N_{dp} là số phần tử dữ liệu trong một mẫu học.

- Chọn hàm kích hoạt(chuyển): thường sử dụng hàm sigmoid, hàm tiếp tuyến hyperbolic...

2.4 Một số mô hình và phương pháp khác(có thời gian sẽ giới thiệu ngắn gọi khoảng 1 trang)

2.4.1 Phương pháp k-người láng giềng gần nhất

2.4.2 Chuỗi Markov

3

Bài toán dự báo trên dữ liệu Telecom

3.1 Dữ liệu telecom và bài toán dự báo nhu cầu

3.2 Thực nghiệm với ARIMA

3.2.1 Xác định mô hình

3.2.2 Ước lượng tham số mô hình

3.2.3 Đánh giá mô hình

3.2.4 Dự báo

3.3 Thực nghiệm với k-người láng giềng gần nhất

3.4 Thực nghiệm với mạng nơ ron nhân tạo

3.5 So sánh kết quả

3. BÀI TOÁN DỰ BÁO TRÊN DỮ LIỆU TELECOM

Tham khảo

[1] Ross Ihaka. **Time Series Analysis**. Lecture Notes for 475.726. University of Auckland, 2005. 2, 13

[2] Ajoy K. Palit; Dobrivoje Popovic. *Computational Intelligence in Time Series Forecasting: Theory and Engineering Applications*. Springer, Berlin, 1st edition, 2005. 2, 3, 5, 13

[3] Ruey S. Tsay. *Analysis of Financial Time Series*. Wiley, 3rd edition, 2010. 3

[4] và David S. Stoffer Robert H. Shumway. *Time Series Analysis and Its Applications: With R Examples*. Springer, Berlin, 3rd edition, 2010. 6, 7, 8, 18

[5] B. L.; O'Connell Bowerman. *Time Series and Forecasting*. Duxbury Press, North Scituate, Massachusetts., 1st edition, 1979. 9

[6] **Liên quan tới khái niệm về dự báo**. *eFinance*, **21**, 2005. 11

[7] G. C. Tsay, R. S.; Tiao. *Consistent estimates of autoregressive parameters and extended sample autocorrelation function for stationary and nonstationary ARMA models*. J. Am. Stat. Assoc, 1984. 22

[8] Widrow B ; Hoff ME. **Adaptive Switching Circuits**. *Anderson J and Rosenfeld E. (eds.) Neurocomputing*, pages 126–134, 1960. 36

[9] Hecht-Nielsen R. **Kolmogorov's Mapping Neural Network Existence Theorem**. **III**, pages 11–14, San Diego, CA, 1987. 38

[10] Rumelhart D. E; McClelland. *Parallel Distributed Processing: Explorations in the Microstructure of Cognition.* MIT Press, Cambridge, MA, 1986. 38, 39

Tham khảo

[1] Ross Ihaka. **Time Series Analysis**. Lecture Notes for 475.726. University of Auckland, 2005. 2, 13

[2] Ajoy K. Palit; Dobrivoje Popovic. *Computational Intelligence in Time Series Forecasting: Theory and Engineering Applications*. Springer, Berlin, 1st edition, 2005. 2, 3, 5, 13

[3] Ruey S. Tsay. *Analysis of Financial Time Series*. Wiley, 3rd edition, 2010. 3

[4] và David S. Stoffer Robert H. Shumway. *Time Series Analysis and Its Applications: With R Examples*. Springer, Berlin, 3rd edition, 2010. 6, 7, 8, 18

[5] B. L.; O'Connell Bowerman. *Time Series and Forecasting*. Duxbury Press, North Scituate, Massachusetts., 1st edition, 1979. 9

[6] **Liên quan tới khái niệm về dự báo**. *eFinance*, **21**, 2005. 11

[7] G. C. Tsay, R. S.; Tiao. *Consistent estimates of autoregressive parameters and extended sample autocorrelation function for stationary and nonstationary ARMA models*. J. Am. Stat. Assoc, 1984. 22

[8] Widrow B ; Hoff ME. **Adaptive Switching Circuits**. *Anderson J and Rosenfeld E. (eds.) Neurocomputing*, pages 126–134, 1960. 36

[9] Hecht-Nielsen R. **Kolmogorov's Mapping Neural Network Existence Theorem**. **III**, pages 11–14, San Diego, CA, 1987. 38

[10] Rumelhart D. E; McClelland. *Parallel Distributed Processing: Explorations in the Microstructure of Cognition.* MIT Press, Cambridge, MA, 1986. 38, 39