

FlinkSQL

Khuong Duy Vu

Department of Informatics

Eötvös Loránd University

A thesis submitted for the degree of Master of Computer Science

05/2012

Acknowledgements

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 !

Abstract

Time series analysis and forecasting has been one of the hot-concerned topics to many data scientists and economists since last few decades. It helps to predict future values based on values both at present and in the past. At first, they utilize statistical methods to figure out the approximate models underlying time series then use that models to predict. Recently, time series forecasting is also solved easily using modified and adaptive machine learning techniques.

According to available data and the efficient of the recent modern techniques, we take the advantages of a traditional statistical method named "ARIMA"(Auto-Agressive Integrated Moving Average Model) based on iterative Box-Jenkins procedure. We also use Artificial Neural Networks accompanying to 2 training methods: Back Propagation and Genetic Algorithm. Both of them are applied to solve univariate time series prediction problem on traffic-forecasting on telecommunication tunnel network, specifically, GTEL telecom network in Vietnam. According to the results, the artificial neural networks produce the better predicted results with smaller Symmetric mean absolute percentage error ($sMAPE = 2.975\% - 3.092\%$) compared to ARIMA's ($sMAPE = 5.34\%$) and its other measures are also more optimistic.

Declaration

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

Contents

List of Figures	ix
List of Tables	xi
1 Tính chất của chuỗi thời gian	1
1.1 Chuỗi thời gian	1
1.2 Tính chất đặc trưng của chuỗi thời gian	2
1.2.1 Tính dừng	2
1.2.2 Tính tuyến tính	3
1.2.3 Tính xu hướng	4
1.2.4 Tính chu kỳ thời vụ(seasonality)	4
1.3 Phân loại chuỗi thời gian	5
1.4 Đo độ phụ thuộc: Hàm tự tương quan và tương quan chéo	6
1.5 Chuỗi dừng	7
1.5.1 Chuỗi dừng và tính chất	7
1.5.2 Kiểm tra chuỗi dừng	9
1.5.2.1 Đồ thị hóa giá trị trung bình và ACF	9
1.5.2.2 Kiểm tra nghiệm đơn vị (unit-root)	9
1.6 Bài toán dự báo chuỗi thời gian	10
1.6.1 Bài toán dự báo	10
1.6.2 Bài toán phân tích chuỗi thời gian và dự báo	12
1.6.2.1 Mô hình thống kê truyền thống	12
1.6.2.2 Mô hình học máy	13

CONTENTS

2	Mô hình ARIMA và Mạng nơ ron nhân tạo	15
2.1	Mô hình tuyến tính ARIMA	15
2.1.1	Nhiều trắng	15
2.1.2	Quá trình tự hồi quy AR	16
2.1.3	Quá trình trung bình trượt MA	18
2.1.3.1	Xác định cấp q trong $MA(q)$	19
2.1.4	Quá trình ARMA	20
2.1.4.1	Xác định mô hình ARMA	20
2.2	Mô hình ARIMA và quy trình Box-Jenkins [1]	23
2.2.1	Mô hình ARIMA cho chuỗi không có tính dừng	23
2.2.2	Xác định mô hình	25
2.2.2.1	Chuẩn bị dữ liệu	25
2.2.2.2	Lựa chọn mô hình:	30
2.2.3	Ước lượng tham số	32
2.2.3.1	Ước lượng bình phương tối thiểu	32
2.2.3.2	Ước lượng hợp lý cực đại	33
2.2.4	Kiểm định mô hình	33
2.2.4.1	Phân tích sai số thặng dư	33
2.2.5	Dự báo	35
2.3	Mô hình mạng nơ ron nhân tạo	36
2.3.1	Kiến trúc mạng nơ ron	36
2.3.1.1	Mô hình nơ ron	36
2.3.1.2	Mạng nơ ron nhân tạo (ANN) cho bài toán dự báo	38
2.3.2	Phương pháp huấn luyện	40
2.3.2.1	Thuật toán lan truyền ngược	40
2.3.2.2	Giải thuật di truyền	41
2.3.3	Dự báo	43
2.3.3.1	Chuẩn bị dữ liệu	44
2.3.3.2	Xác định kiến trúc mạng	46

3	Bài toán dự báo trên dữ liệu Telecom	49
3.1	Dữ liệu telecom và bài toán dự báo lưu lượng	49
3.2	Thực nghiệm với ARIMA	52
3.2.1	Xác định mô hình	52
3.2.2	Ước lượng tham số	57
3.2.3	Kiểm định mô hình	57
3.2.4	Dự báo	58
3.3	Thực nghiệm với mạng nơ ron nhân tạo	59
3.3.0.1	Phương pháp huấn luyện lan truyền ngược	60
3.3.0.2	Sử dụng giải thuật di truyền	62
3.4	So sánh kết quả	63
4	Kết luận và hướng nghiên cứu	67
4.1	Kết quả	67
4.2	Khó khăn	68
4.3	Hướng nghiên cứu tiếp theo	68
	References	69

CONTENTS

List of Figures

1.1	Chuỗi dừng và không có tính chất dừng	3
1.2	Xu hướng giảm dần	4
1.3	Lượng dòng chảy đến hồ chứa Trị An từ năm 1959 đến 1985	5
2.1	Bảng giá trị Hàm tự tương quan mở rộng mẫu	22
2.2	Bảng giá trị SEACF ví dụ cho mô hình ARMA(2,2)	23
2.3	Quy trình Box-Jenkins xác định mô hình ARIMA	24
2.4	Mô hình chuỗi dừng và không có tính chất dừng	26
2.5	Lượng điện sử dụng hàng tháng của Mỹ	27
2.6	Hàm ước lượng hợp lý log dựa vào λ	28
2.7	$\log(\text{Electricity})$	29
2.8	Đồ thị giá trị chuỗi sai phân bậc 1 của $\log(\text{electricity})$	30
2.9	Nơ ron thần kinh	37
2.10	Nơ ron nhân tạo	38
2.11	Phương pháp lan truyền ngược	41
2.12	Mô hình dự báo sử dụng ANN kết hợp giải thuật di truyền	44
2.13	Kỹ thuật cửa sổ trượt	45
2.14	Đồ thị hàm sigmoid	46
3.1	Thống kê lưu lượng theo giờ	51
3.2	Lưu lượng các kênh truyền tính theo ngày	51
3.3	Lưu lượng kênh truyền từ 15/05/2010 đến 14/11/2010 của mạng di động Beeline	52
3.4	Đồ thị hàm hợp lý loga	53
3.5	Đồ thị hàm ACF	53

LIST OF FIGURES

3.6	Đồ thị chuỗi đã được sai phân 1 lần	54
3.7	Đồ thị hàm ACF của chuỗi đã được sai phân	55
3.8	Đồ thị hàm PACF của chuỗi đã được sai phân	55
3.9	Đồ thị hàm ACF của chuỗi $\Delta_7(\Delta(x_t))$	56
3.10	Đồ thị hàm PACF của chuỗi $\Delta_7(\Delta(x_t))$	56
3.11	So sánh kết quả dự đoán 23 ngày tiếp theo sử dụng mô hình $ARIMA(0, 1, 2)(2, 0, 2)_7$	58
3.12	So sánh kết quả dự đoán 23 ngày tiếp theo sử dụng mô hình ANN sử dụng giải thuật lan truyền ngược	61
3.13	So sánh kết quả dự đoán 23 ngày tiếp theo sử dụng mô hình mạng nơ ron sử dụng GA	63
3.14	So sánh kết quả mô hình ANN-GA dự đoán trên tập học	64

List of Tables

2.1	Bảng tham số λ	27
2.2	Bảng dự đoán mô hình dựa vào đồ thị SACF	31
3.1	Độ đo AIC của các mô hình đề xuất	57
3.2	Các độ đo sai số dự đoán của mô hình mạng nơ ron sử dụng lan truyền ngược có 3 nút đầu vào	60
3.3	Trọng số trong mạng nơ ron 3:2:1 lan truyền ngược	61
3.4	Các độ đo sai số dự đoán của mô hình mạng nơ ron sử dụng giải thuật di truyền có 3 nút đầu vào	62
3.5	So sánh độ đo lỗi 3 mô hình	65

LIST OF TABLES

Chapter 1

Tính chất của chuỗi thời gian

1.1 Chuỗi thời gian

Trong quá trình theo dõi sự thay đổi hiện tượng, 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 các 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 $X = \{x_1, x_2, \dots, x_n\}$ (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 phương pháp này 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,

1. TÍNH CHẤT CỦA CHUỖI THỜI GIAN

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**[2] là một dãy các giá trị quan sát $X = \{x_1, x_2, \dots, x_n\}$ 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 $t = 0, \pm 1, \pm 2, \dots$ hoặc đơn giản là một tập các số nguyên có thứ tự. Trên thực 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 [3] của chuỗi thời gian là **tính dừng, tính tuyến tính, tính xu hướng và tính chu kỳ 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 $x(t+h)$

1.2 Tính chất đặc trưng của chuỗi thời gian

chỉ phụ thuộc vào h mà không phụ thuộc vào t [3]. 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 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 chuỗi thời gian. Một chuỗi dừng thường có đồ thị nhìn khá phẳng, không có xu hướng, biến thiên đều. (Hình 1.1).

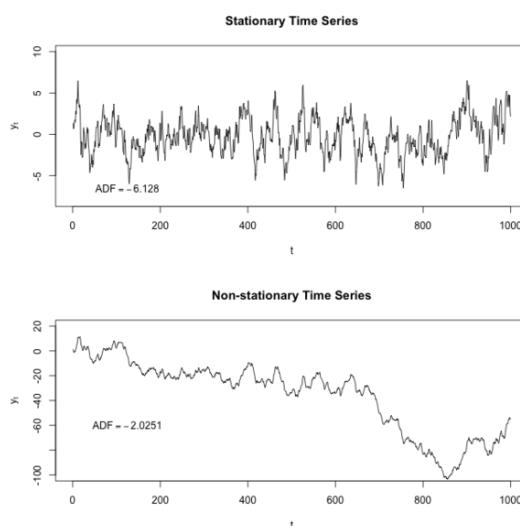


Figure 1.1: Chuỗi dừng và không có tính chất dừng -

1.2.2 Tính tuyến tính

Definition 1.1 Một chuỗi thời gian được gọi là tuyến tính[4] nếu nó có thể được biểu diễn bởi công thức:

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

với μ là giá trị 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 giá trị trung bình bằng 0 và phương sai σ^2

1. TÍNH CHẤT CỦA CHUỖI THỜI GIAN

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 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 :

$$x_t = \alpha t + \beta + \varepsilon_t \quad (1.2)$$

$$x_t = \exp(\alpha t + \beta + \varepsilon_t) \quad (1.3)$$

$$x_t = \alpha t + \beta t^2 + \gamma + \varepsilon_t \quad (1.4)$$

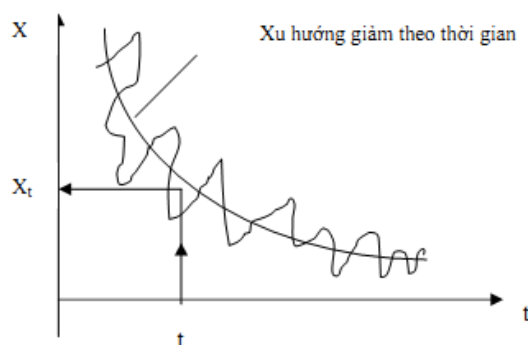


Figure 1.2: Xu hướng giảm dần -

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)

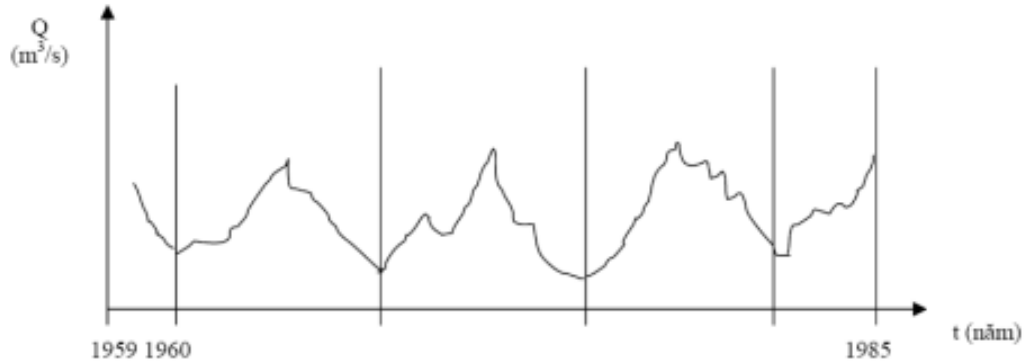


Figure 1.3: Lượng dòng chảy đến hồ chứa Trị An từ năm 1959 đến 1985

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[3] 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. TÍNH CHẤT CỦA CHUỖI THỜI GIAN

1.4 Đo độ phụ thuộc: Hàm tự tương quan và tương quan chéo

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

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

Definition 1.2 *Hàm tính kỳ vọng (giá trị trung bình)*

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

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

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

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 $\gamma(s, s) = E[(x_s - \mu_s)^2] = \text{var}(t)$

Definition 1.4 *Hàm tự tương quan (ACF)[5] được xác định bởi:*

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

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 Cosi¹, 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.

¹Bất 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 tự tương quan chéo*. Giả sử cả hai chuỗi đều có phương sai xác định, ta có các khái niệm sau:

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

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

Definition 1.6 *Hàm tự tương quan chéo (CCF)[5] được cho bởi:*

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

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

Definition 1.7 *Chuỗi x_t được gọi là **dừng theo nghĩa chặt**[5] nếu như tại mọi thời điểm t_1, t_2, \dots, t_k thì ta luôn có $(x_{t_1}, x_{t_2}, \dots, x_{t_k})$ có cùng phân phối xác suất với $(x_{t_1+h}, x_{t_2+h}, \dots, x_{t_k+h})$ hay*

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

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

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

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

$$\gamma(s, t) = \gamma(s + h, t + h) \quad (1.13)$$

1. TÍNH CHẤT CỦA CHUỖI THỜI GIAN

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 moment của chuỗi.

Definition 1.8 Chuỗi x_t , có phương sai xác định, được gọi là **dừng theo nghĩa rộng** [5] nếu

1. Hàm kỳ vọng 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

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

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

Definition 1.9 Hàm tự hiệp phương sai cho chuỗi dừng:

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

Definition 1.10 Hàm tự tương quan ACF cho chuỗi dừng:

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

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 $\gamma(0) = E[(x_t - \mu)^2] = \text{var}(x_t)$
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 tính kỳ vọng luôn cho 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 Đồ thị hóa giá trị trung bình và ACF

Khi mô hình hóa chuỗi thời gian, ta nhận thấy chuỗi có xu hướng tăng dần, giảm dần hoặc giá trị 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 [6].

1.5.2.2 Kiểm tra nghiệm đơn vị (unit-root)

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

$$y_t = a_1 y_{t-1} + a_2 y_{t-2} + \dots + a_p y_{t-p} + \epsilon_t \quad (1.17)$$

với a_1, a_2, \dots 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

$$m^p - m^{p-1}a_1 - m^{p-2}a_2 - \dots - a_p = 0 \quad (1.18)$$

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, giá trị trung bình sẽ biến đổi hay phương sai phụ thuộc thời gian. Do đó, nó không thỏa mãn điều kiện chuỗi dừng. Ví dụ, giả sử đối với một chuỗi x_t đơn giản

$$x_t = \phi_1 x_{t-1} + w_t \quad (1.19)$$

1. TÍNH CHẤT CỦA CHUỖI THỜI GIAN

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

$$x_t = x_{t_1} + w_1 \quad (1.20)$$

$$x_t = x_0 + \sum_{j=1}^t w_j \quad (1.21)$$

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

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 [7] áp dụng cho mô hình tự hồi quy, mở rộng Augmented Dickey–Fuller [8] cho mô hình ARMA hay phương pháp Phillips–Perron [9] 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

$$c_t = x_t - x_{t-1} \quad (1.23)$$

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 cạnh tranh 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 hưởng 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.

1.6 Bài toán dự báo chuỗi thời gian

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 doanh 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 [10] 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 thống kê. Sử dụng mô hình đã ước lượng này để dự báo bằng kỹ thuật ngoại suy hoặc mô phỏng.

1. TÍNH CHẤT CỦA CHUỖI THỜI GIAN

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 tối ưu hoá: Diễn hình của mô hình này là bài toán quy hoạch tối ưu, bố trí một nguồn lực nhằm 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 tầm 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 nhau 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 tiêu 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

Như đã nói ở trên, 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 dựa vào các độ đo, 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, \dots, 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 [3] 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)[2] 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 Mô hình học máy

Khi dự báo chuỗi thời gian sử dụng những kĩ thuật truyền thống 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

1. TÍNH CHẤT CỦA CHUỖI THỜI GIAN

tính, chu kỳ...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 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.

Chapter 2

Mô hình ARIMA và Mạng nơ ron nhân tạo

Chương 2 trình bày 2 mô hình dự báo là mô hình tự hồi quy trung bình trượt và mạng nơ ron nhân tạo sử dụng phương pháp huấn luyện giám sát (lan truyền ngược và giải thuật di truyền).

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 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 kỳ vọng và phương sai xác định hữu hạn. Trong trường hợp đặc biệt, nếu x_t có giá trị 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

2. MÔ HÌNH ARIMA VÀ MẠNG NƠ RON NHÂN TẠO

như chuỗi đó là nhiễu trắng. Một mô hình dự báo chấp nhận được nếu sai số giữa giá trị thực tế và dự báo là một chuỗi nhiễu trắng.

2.1.2 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, $x_{t-1}, x_{t-2}, \dots, x_{t-p}$, 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. Nó có thể được quan sát, xác định bởi đồ thị hàm tự tương quan từng phần (PACF) của chuỗi.

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

$$x_t = \phi_1 x_{t-1} + \phi_2 x_{t-2} + \dots + \phi_p x_{t-p} + w_t \quad (2.1)$$

với x_t có tính dừng, và $\phi_1, \phi_2, \dots, \phi_p$ 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 giá trị trung bình bằng 0 và phương sai σ_w^2 . Nếu giá trị 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:

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

hoặc

$$x_t = \phi_1 x_{t-1} + \phi_2 x_{t-2} + \dots + \phi_p x_{t-p} + w_t \quad (2.3)$$

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

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

Definition 2.2 Toán tử dịch chuyển B được xác định bởi

$$Bx_t = x_{t-1} \quad (2.4)$$

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

$$x_t = B^k x_{t-k} \quad (2.5)$$

2.1 Mô hình tuyến tính ARIMA

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

$$(1 - \phi_1 B - \phi_2 B^2 - \dots - \phi_p B^p)x_t = w_t \quad (2.6)$$

hay

$$\phi(B)x_t = w_t \quad (2.7)$$

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

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

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

$$= \phi^2 \gamma_x(0) \quad (2.10)$$

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 bề gây 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.

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

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.

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

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

và

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

2. MÔ HÌNH ARIMA VÀ MẠNG NƠ RON NHÂN TẠO

với

$$\hat{x}_{t+h} = \beta_1 x_{t+h-1} + \beta_2 x_{t+h-2} + \cdots + \beta_{h-1} x_{t+1} \quad (2.14)$$

$$\hat{x}_t = \beta_1 x_{t+1} + \beta_2 x_{t+2} + \cdots + \beta_{h-1} x_{t+h-1} \quad (2.15)$$

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

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

Khi $h > p$ thì [5] ta tính được

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

Theo đó, khi $h > p$

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

Theo công thức 2.15 thì $x_t - \hat{x}_t$ sẽ chỉ phụ thuộc vào w_i với $i = t+h-1, t+h-2, \dots$ nên chúng sẽ độc lập với w_{h+t} . Khi đó, $\phi_{hh} = 0$ như trên. Cũng theo [5] 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.

2.1.3 Quá trình trung bình trượt MA

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

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

với q là cấp của mô hình trung bình trượt, và $\theta_1, \theta_2, \dots, \theta_q$ 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 giá trị 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 $w_t, w_{t-1}, \dots, w_{t-p}$ trong quá khứ thay vì bằng các giá trị $x_{t-1}, x_{t-2}, \dots, x_{t-p}$. Tương tự, ta cũng có thể biểu diễn 2.19 dưới dạng

$$x_t = \theta(B)w_t \quad (2.20)$$

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

2.1.3.1 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 kỳ vọng và hàm tự hiệp phương sai như sau:

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

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

$$= \begin{cases} \sigma_w^2 \sum_{j=0}^{q-h} \theta_j \theta_{j+h}, & 0 \leq h \leq q \\ 0 & h > q \end{cases} \quad (2.23)$$

Khi đó, ta cũng có

$$\rho(h) = \frac{\gamma(h)}{\gamma(0)} = \begin{cases} \frac{\sum_{j=0}^{q-h} \theta_j \theta_{j+h}}{1 + \theta_1^2 + \dots + \theta_q^2}, & 1 \leq h \leq q \\ 0 & h > q \end{cases} \quad (2.24)$$

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

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

2. MÔ HÌNH ARIMA VÀ MẠNG NƠ RON NHÂN TẠO

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

$$\hat{\rho}_h = \frac{\sum_{t=h+1}^n (x_t - \bar{x})(x_{t-h} - \bar{x})}{\sum_{t=1}^n (x_t - \bar{x})^2} \quad (2.26)$$

với \bar{x} là giá trị trung bình mẫu của x_i . Lúc này ta hoàn toàn có thể tính được ước lượng của $\rho(h)$

2.1.4 Quá trình ARMA

Definition 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

$$x_t = \phi_1 x_{t-1} + \dots + \phi_p x_{t-p} + w_t + \theta_1 w_{t-1} + \dots + \theta_q w_{t-q} \quad (2.27)$$

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 giá trị trung bình $\mu \neq 0$ thì có thể đặt $\alpha = \mu(1 - \phi_1 - \dots - \phi_p)$ và mô hình có thể được viết dưới dạng

$$x_t = \alpha + \phi_1 x_{t-1} + \dots + \phi_p x_{t-p} + w_t + \theta_1 w_{t-1} + \dots + \theta_q w_{t-q} \quad (2.28)$$

Ta vẫn lấy w_t là chuỗi nhiễu trắng Gaussian có giá trị 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 , $MA(q)$

2.1.4.1 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 đề 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

$$x_t = \phi_1 x_{t-1} + \phi_2 x_{t-2} + \dots + \phi_p x_{t-p} + w_t \quad (2.29)$$

Ta sử dụng bình phương tối thiểu để tìm các tham số $\hat{\phi}_i^{(0)}$ với $i = 1, 2, \dots, p$. 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 nhiều trắng

$$\hat{w}_t^{(0)} = x_t - \hat{\phi}_1^{(0)} x_{t-1} - \hat{\phi}_2^{(0)} x_{t-2} - \dots - \hat{\phi}_p^{(0)} x_{t-p} \quad (2.30)$$

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

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

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

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

Ta mới thêm nhiễu $\hat{w}_{t-2}^{(0)}$ để ước lượng mô hình. Nếu thực sự trong thực tế $q = 2$ 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.

2. MÔ HÌNH ARIMA VÀ MẠNG NƠ RON NHÂN TẠO

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 nhiều

$$\hat{w}_t^{(j)} = x_t - \hat{\phi}_1^{(j)} x_{t-1} - \hat{\phi}_2^{(j)} x_{t-2} - \dots - \hat{\phi}_p^{(j)} x_{t-m} \quad (2.33)$$

Ở đây, j là số vòng lặp trong quá trình xấp xỉ mô hình ở trong quá trình ước lượng mô hình ARMA ở Phần 2.1.4.1

AR	MA					
	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	$\hat{\rho}_1^{(4)}$	$\hat{\rho}_2^{(4)}$	$\hat{\rho}_3^{(4)}$	$\hat{\rho}_4^{(4)}$	$\hat{\rho}_5^{(4)}$...
\vdots	\vdots	\vdots	\vdots	\vdots	\vdots	...

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

Theo Tiao và Tsay [11] 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

2.2 Mô hình ARIMA và quy trình Box-Jenkins [1]

tìm kiếm 2 đường tiệm cận "0" trong bảng, điểm giao nhau chính là giá trị của (p, q) . Như trong Hình 2.2, ta tìm được điểm giao là $(2, 2)$, do đó, mô hình dự kiến sẽ là ARMA(2,2)

AR	MA						
	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	...
⋮	⋮	⋮	⋮	⋮	⋮	⋮	...

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

2.2 Mô hình ARIMA và quy trình Box-Jenkins [1]

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

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

$$\Delta^d x_t = (1 - B)^d x_t \quad (2.34)$$

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

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

2. MÔ HÌNH ARIMA VÀ MẠNG NƠ RON NHÂN TẠO

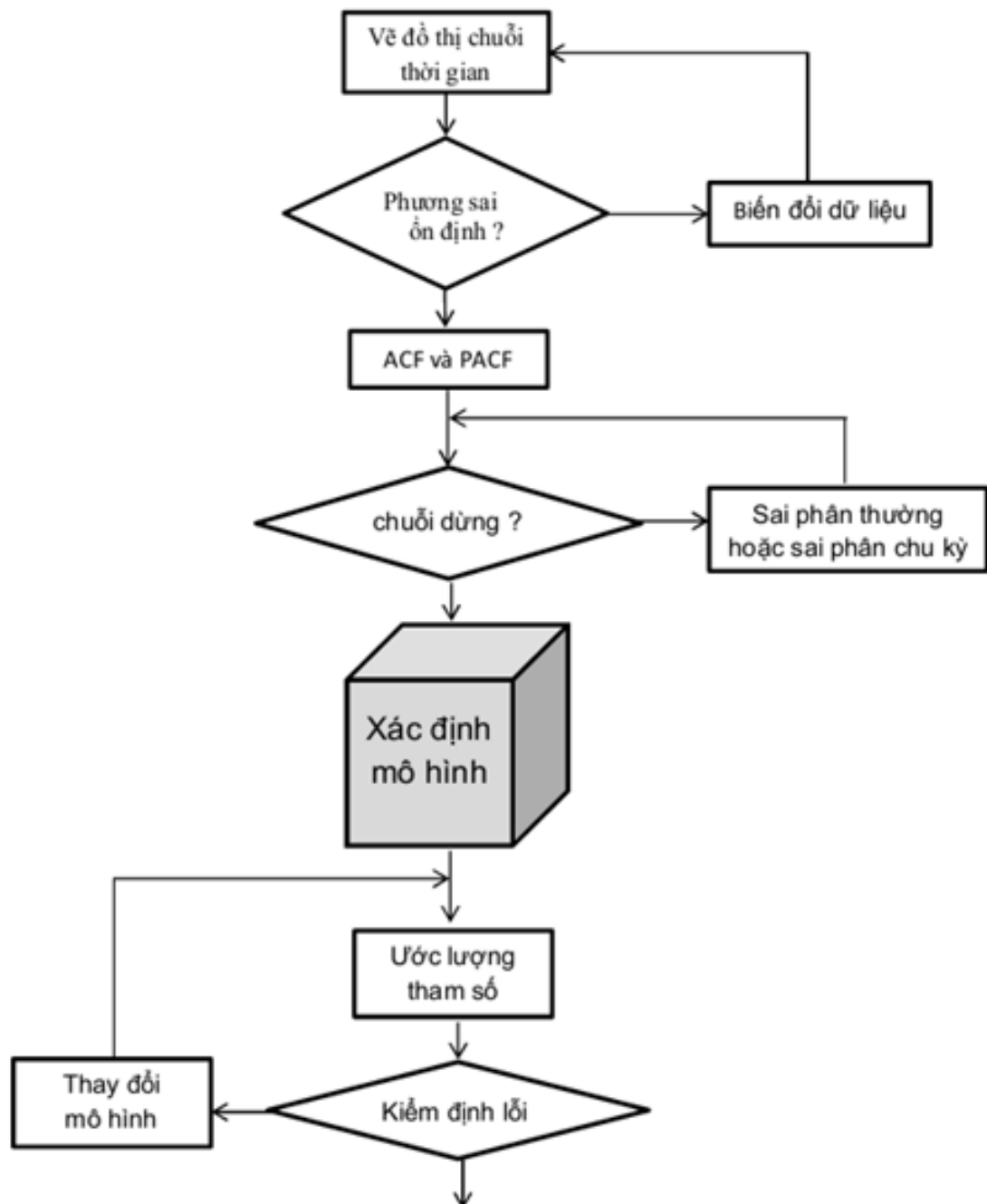


Figure 2.3: Quy trình Box-Jenkins xác định mô hình ARIMA - có cả yếu tố chu kì thời vụ

2.2 Mô hình ARIMA và quy trình Box-Jenkins [1]

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

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**
- **Kiểm định mô hình:** đánh giá độ chính xác mô hình vừa tìm được với dữ liệu quan sát được. Chủ yếu là đánh giá xem lỗi (chênh lệch giá trị dự đoán và giá trị thực tế) có phải là chuỗi có dạng nhiễu trắng không, các độ đo lỗi có trong ngưỡng mong muốn không.

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 tính giá trị trung bình, phương sai và tự tương quan khá ổn định, không thay đổi theo thời gian. 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 kỳ, mùa vụ (Hình 2.4)

2. MÔ HÌNH ARIMA VÀ MẠNG NƠ RON NHÂN TẠO

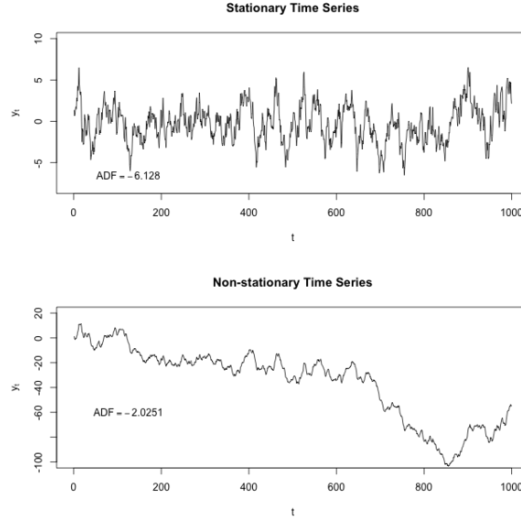


Figure 2.4: Mô hình chuỗi dừng và không có tính chất dừng -

Ta thường gặp một số dữ liệu mà càng về sau thì độ biến thiên của dữ liệu quanh giá trị 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 2.5. 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. Chúng ta có thể sử dụng **phép biến đổi mũ** được giới thiệu bởi Box và Cox [12]. Nếu gọi $T(x_t)$ là hàm ổn định phương sai. Chuỗi sau biến đổi $T(x_t)$ có phương sai không đổi.

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

Tham số λ được cho bởi bảng

Tham số λ có thể được ước lượng thông qua các giá trị loga(log-likelihood) dựa vào các hàm ước lượng cực đại(MLE). Với độ tin tưởng 95%, ta có đồ thị giá trị hợp lý loga như Hình 2.6. Ta thấy có thể chọn được giá trị $\lambda = 0$, sử dụng hàm logarit để biến đổi dữ liệu. Ta có thể quan sát thấy mô hình dữ liệu mới(Hình 2.7)

2.2 Mô hình ARIMA và quy trình Box-Jenkins [1]

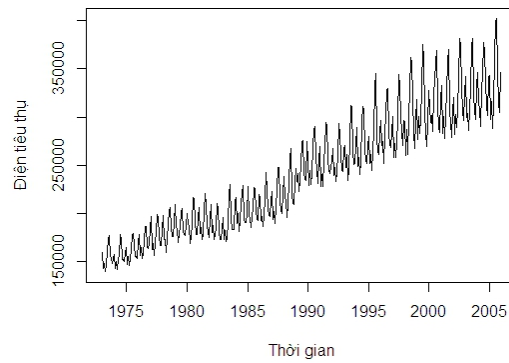


Figure 2.5: Lượng điện sử dụng hàng tháng của Mỹ -

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

Table 2.1: Bảng tham số λ

2. MÔ HÌNH ARIMA VÀ MẠNG NƠ RON NHÂN TẠO

đã có độ biến thiên quanh giá trị giá trị trung bình đều nhau hay nói cách khác, phương sai của chúng đã được giữ ổn định.

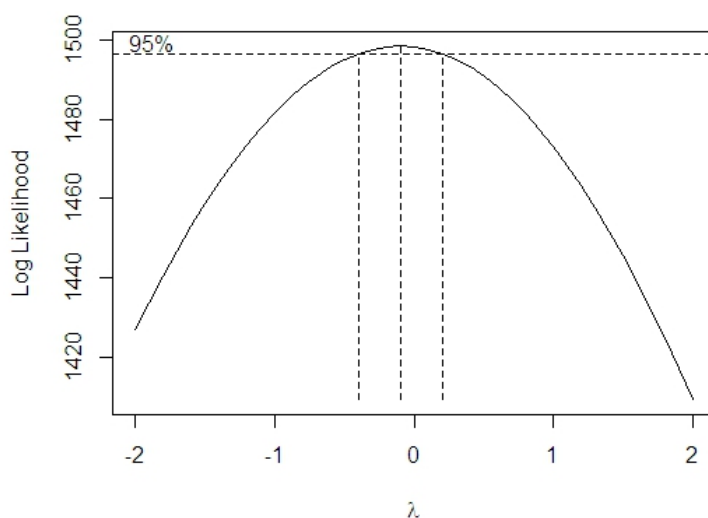


Figure 2.6: Hàm ước lượng hợp lý log dựa vào λ -

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 của chúng không có tính chất trên. Ta nên sử dụng một phương pháp thống 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 Dickey–Fuller(ADF) [7] do Dickey và Fuller tạo ra. Họ đề xuất 3 hàm hồi quy

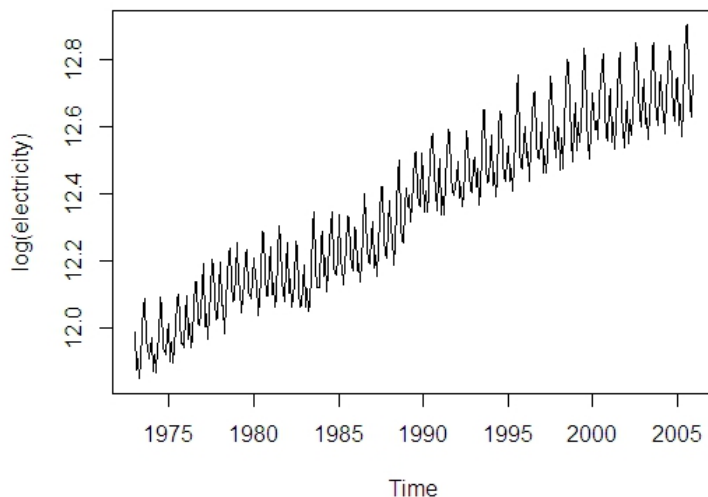


Figure 2.7: $\log(\text{Electricity})$ -

khác nhau để kiểm tra nghiệm đơn vị cho x_t

$$\Delta x_t = ax_{t-1} + \sum_{j=1}^J \varphi_j \Delta x_{t-1} + \beta t + w_t, \quad t = 1, 2, \dots, T \quad (2.37)$$

$$\Delta x_t = ax_{t-1} + \sum_{j=1}^J \varphi_j \Delta x_{t-1} + \mu + w_t, \quad t = 1, 2, \dots, T \quad (2.38)$$

$$\Delta x_t = ax_{t-1} + \sum_{j=1}^J \varphi_j \Delta x_{t-1} + w_t, \quad t = 1, 2, \dots, T \quad (2.39)$$

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 $x_{t-1}, \Delta x_{t-1}, \Delta x_{t-2}, \dots, \Delta x_{t-k}$. 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

2. MÔ HÌNH ARIMA VÀ MẠNG NƠ RON NHÂN TẠO

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 2.8)

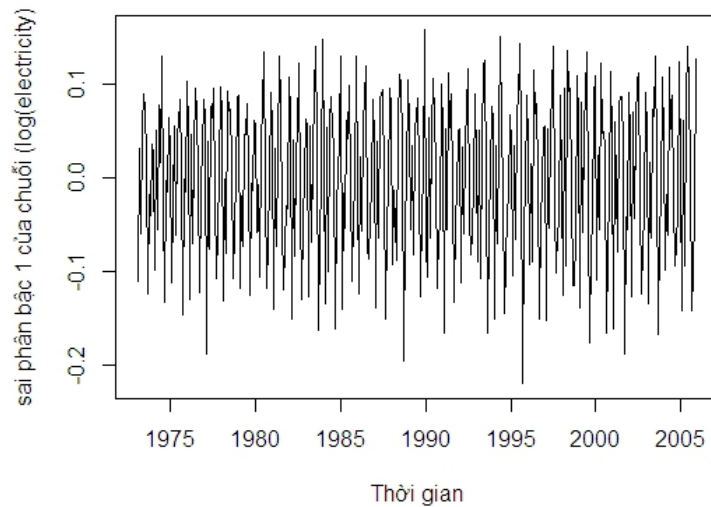


Figure 2.8: Đồ thị giá trị chuỗi sai phân bậc 1 của $\log(electricity)$ -

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

2.2 Mô hình ARIMA và quy trình Box-Jenkins [1]

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 $ARIMA(0, d, q)$ hay $MA(q)$ (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

Table 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. MÔ HÌNH ARIMA VÀ MẠNG NƠ RON NHÂN TẠO

Ngoài ra, một phương pháp khác để lựa chọn mô hình phù hợp nhất là tìm mô hình có độ đo **Akaike's Information Criterion (AIC)** [13] nhỏ nhất.

$$AIC = -2 \ln L + 2k \quad (2.40)$$

với $\ln L$ là logarit tự nhiên của hàm hợp lý cực đại và k là số tham số trong mô hình

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 hình.

2.2.3.1 Ước lượng bình phương tối thiểu

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

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

Ta có,

$$w_t = x_t - \sum_{j=1}^p \phi_j x_{t-j} - \sum_{j=1}^q \theta_j w_{t-j} \quad (2.42)$$

Giả sử chỉ có n mẫu x_1, x_2, \dots, 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

$$S_c(\phi, \theta) = \sum_{t=2}^n w_t^2 \quad (2.43)$$

$$= \sum_{t=2}^n \left[x_t - \sum_{j=1}^p \phi_j x_{t-j} - \sum_{j=1}^q \theta_j w_{t-j} \right]^2 \quad (2.44)$$

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

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

2.2 Mô hình ARIMA và quy trình Box-Jenkins [1]

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

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

2.2.3.2 Ước lượng hợp lý cực đại

Đây 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, \dots, w_n được xác định bởi:

$$f(w_2, w_3, \dots, w_n) = \prod_{t=2}^n f(w_t) \quad (2.45)$$

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

$$L = L(\phi, \theta, \sigma_w^2 | x) = f(x_2, x_3, \dots, x_n | x_1) f(x_1) \quad (2.46)$$

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:

$$\text{Sai số thặng dư tại } t = (\text{giá trị thực tế của mẫu } x_t) - (\text{giá trị của mô hình ước lượng } \hat{x}_t) \quad (2.47)$$

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 Gaussian. Nó gần giống như một chuỗi các biến ngẫu nhiên chuẩn, độc lập và có cùng phân phối xác suất.

2. MÔ HÌNH ARIMA VÀ MẠNG NƠ RON NHÂN TẠO

Ngoài ra, nó có giá trị 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**

$$\hat{e}_t^* = \frac{\hat{e}_t}{\hat{\sigma}_e} \quad (2.48)$$

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 [14] 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 giả 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 giả 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 giả thuyết H_0 sẽ bị loại bỏ.

Ngoài ra, Ljung và Box [15] đã phát triển một hàm kiểm tra dựa vào sự 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 thức:

$$Q_* = n(n+2) \sum_{k=1}^K \frac{\hat{r}_k^2}{n-k} \quad (2.49)$$

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, \dots, x_t , chúng ta cần dự đoán các kết quả tương lai x_{t+1}, x_{t+2}, \dots . 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, \dots, x_t . Ta gọi nó là Hàm dự đoán $MMSE$ (minimum mean squared error forecast) được cho bởi công thức sau:

$$\hat{x}_t(l) = E(x_{t+l} | x_1, x_2, \dots, x_t) \quad (2.50)$$

Một số tính chất của kỳ vọng có điều kiện:

- $E(c | x_1, x_2, \dots, x_t) = c$ với c là hằng số
- $E(\phi_i x_i | x_1, x_2, \dots, x_t) = \phi_i x_i$ với $i = 1, 2, \dots, t$
- $E(w_{t+1} | x_1, x_2, \dots, x_t) = E(w_{t+1}) = 0$ với w_t là nhiễu trắng, bởi vì w_{t+1} độc lập với x_1, x_2, \dots, x_t

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

$$x_t = \phi_1 x_{t-1} + \dots + \phi_p x_{t-p} + w_t + \theta_1 w_{t-1} + \dots + \theta_q w_{t-q} \quad (2.51)$$

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

$$\hat{x}_t(l) = \sum_{j=1}^p \phi_j \hat{x}_t(l-j) + w_t + \sum_{j=1}^q \theta_j E(w_{t+l-j} | x_1, x_2, \dots, x_t) \quad (2.52)$$

Trong đó,

$$E(w_{t+l-j} | x_1, x_2, \dots, x_t) = \begin{cases} 0, & l-j > 0 \\ w_{t+l-j}, & l-j \leq 0 \end{cases} \quad (2.53)$$

$$\hat{x}_t(l-j) = E(x_{t+l-j} | x_1, x_2, \dots, x_t) \quad j = 1, 2, \dots, p \quad (2.54)$$

2. MÔ HÌNH ARIMA VÀ MẠNG NƠ RON NHÂN TẠO

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

$$x_t = \phi x_{t-1} + w_t + \theta w_{t-1} \quad (2.55)$$

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

$$\hat{x}_t(1) = \phi x_t - \theta w_t \quad (2.56)$$

$$\hat{x}_t(l) = \phi \hat{x}_t(l-1) \quad (2.57)$$

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 (dendrite – nhận tín hiệu từ các tế bào não khác) và sợi trục (axon – giúp truyền tín hiệu đến các tế bào não khác)(Hình 2.9). Khi tua gai của tế bào thần kinh nhận kích thích, một tín hiệu điện (xung điện) sẽ truyền qua thân 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

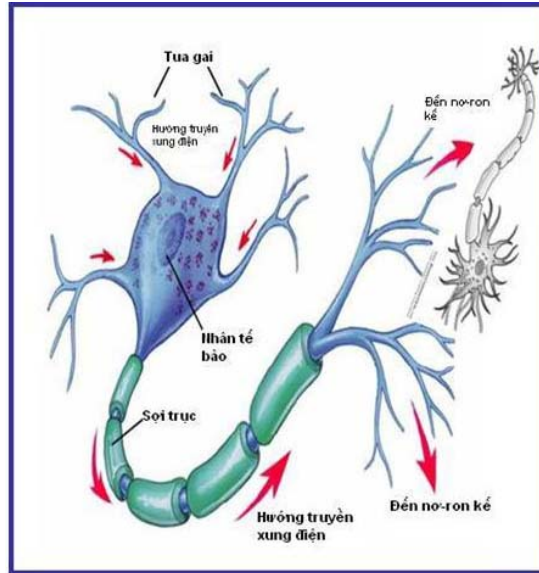


Figure 2.9: Nơ ron thần kinh -

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 [16] được mô tả trong Hình 2.10

Nơ ron nhân tạo gồm 3 thành phần chính:

- Các tín hiệu đầu vào x_i với $i = 1, 2, \dots, n$ 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)

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

với điều kiện

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

2. MÔ HÌNH ARIMA VÀ MẠNG NƠ RON NHÂN TẠO

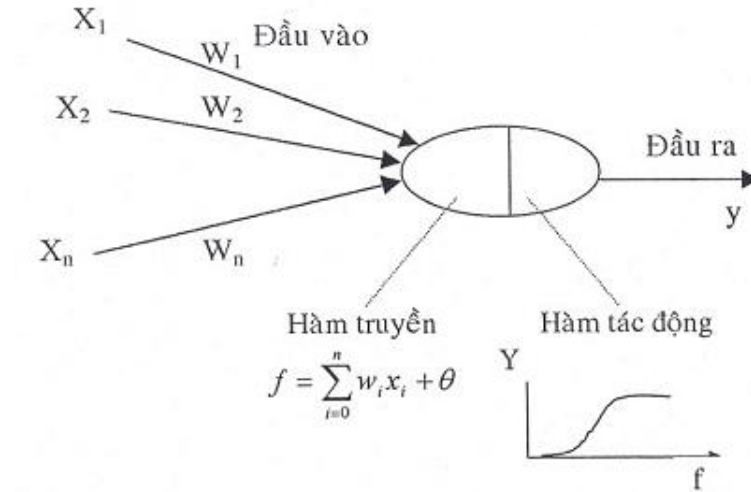


Figure 2.10: Nơ ron nhân tạo -

Một số hàm kích hoạt hay được sử dụng như:

Hàm bước nhảy

$$f(x) = \begin{cases} 1 & \text{khi } f \geq 0 \\ 0 & \text{khi } f < 0 \end{cases} \quad (2.58)$$

Hàm dấu

$$f(x) = \text{sgn}(x) \begin{cases} 1 & \text{khi } f \geq 0 \\ -1 & \text{khi } f < 0 \end{cases} \quad (2.59)$$

Hàm sigmoid

$$f(x) = \frac{1}{1 + e^{-x}} \quad (2.60)$$

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 RBF(Radial Basic Function)
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

$$y = f_o(\sum w_h f_h(\sum f_i(w_i^T x))) \quad (2.61)$$

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 $f : [0, 1]^n \rightarrow R$ nào cũng có thể được viết dưới dạng:

$$f(x_1, x_2, \dots, x_n) = \sum_{i=1}^{2n+1} \psi_i(\sum_{j=1}^n \varphi_{ij}(x_j)) \quad (2.62)$$

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

Mặt khác, theo Hecht-Nielsen [17] 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 [18] và không bao giờ cần đến quá 4 lớp mạng để giải quyết các bài toán phức tạp nhất [19]

2. MÔ HÌNH ARIMA VÀ MẠNG NƠ RON NHÂN TẠO

Rumelhart và cộng sự [20] đề xuất phương pháp học lan truyền ngược (back-propagation) 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 của mạng.

2.3.2 Phương pháp huấn luyện

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 [20] 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. Chênh lệch được tính toán sau đó quay ngược lại thay đổi trọng số ở lớp trước nó (Hình 2.11). Việc tối ưu hóa trọng số dựa vào luật **giảm 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 nơ ron thứ j lớp đầu ra, u_j là đầu vào của nơ ron này.

$$u_j = \sum_{i=1}^n w_i x_i + \theta_j \quad (2.63)$$

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

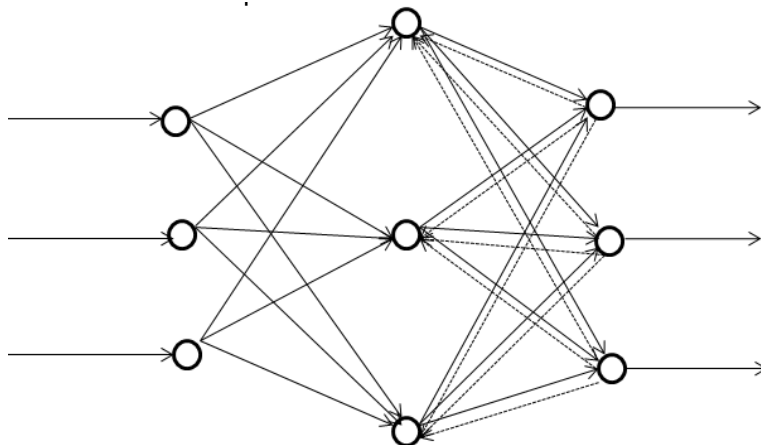


Figure 2.11: Phương pháp lan truyền ngược -

quan sát được tại nơ ron đầu ra thứ j thì hàm sai số được cho bởi

$$S_j = 0.5(d_j - y_j)^2 = 0.5(e_j)^2 \quad (2.64)$$

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:

$$\Delta w_i = -\eta \frac{\nabla S_j}{\nabla w_i} \quad (2.65)$$

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

$$\Delta w_i = \eta \delta_j . x_i \quad (2.66)$$

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 dựa 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.

2. MÔ HÌNH ARIMA VÀ MẠNG NƠ RON NHÂN TẠO

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 trong 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 bit 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 dừng 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

2.3 Mô hình mạng nơ ron nhân tạo

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" tại <http://code.google.com/p/cudaann/>. 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. MÔ HÌNH ARIMA VÀ MẠNG NƠ RON NHÂN TẠO

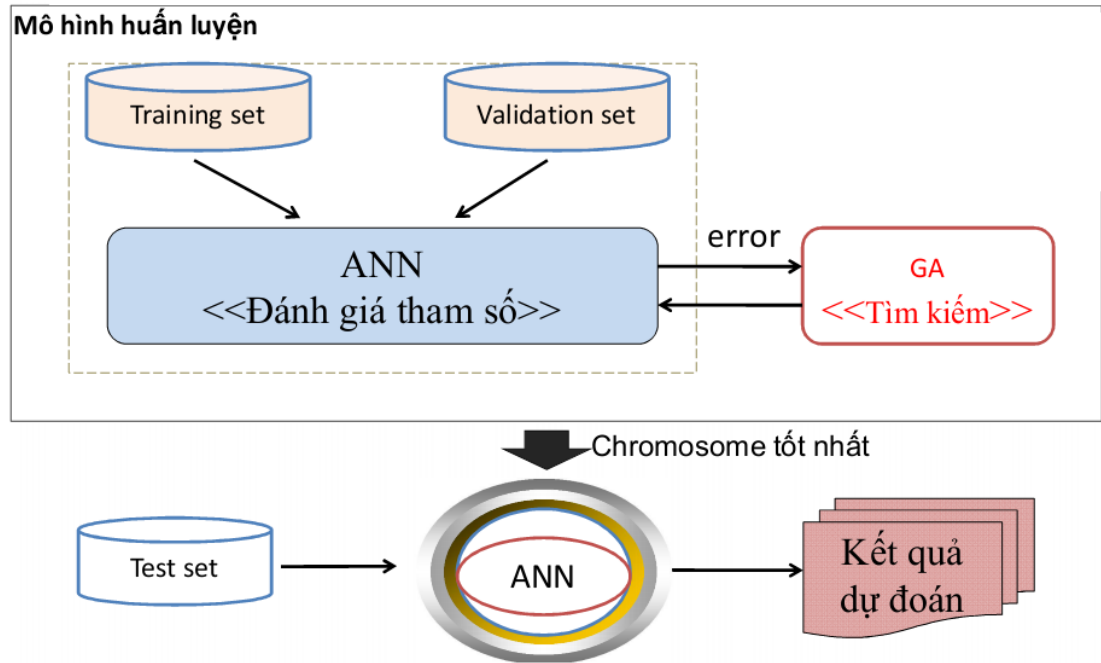


Figure 2.12: Mô hình dự báo sử dụng ANN kết hợp giải thuật di truyền

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 **kỹ thuật cửa sổ trượt** (sliding-window) (Hình 2.13). MLP nhận vector $\{x_{t-1}, x_{t-2}, \dots, x_{t-p}\}$ 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.

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 sigmoid (Hình 2.14), ta thấy đồ thị tiến về tiệm cận khi $x \rightarrow \infty$. 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

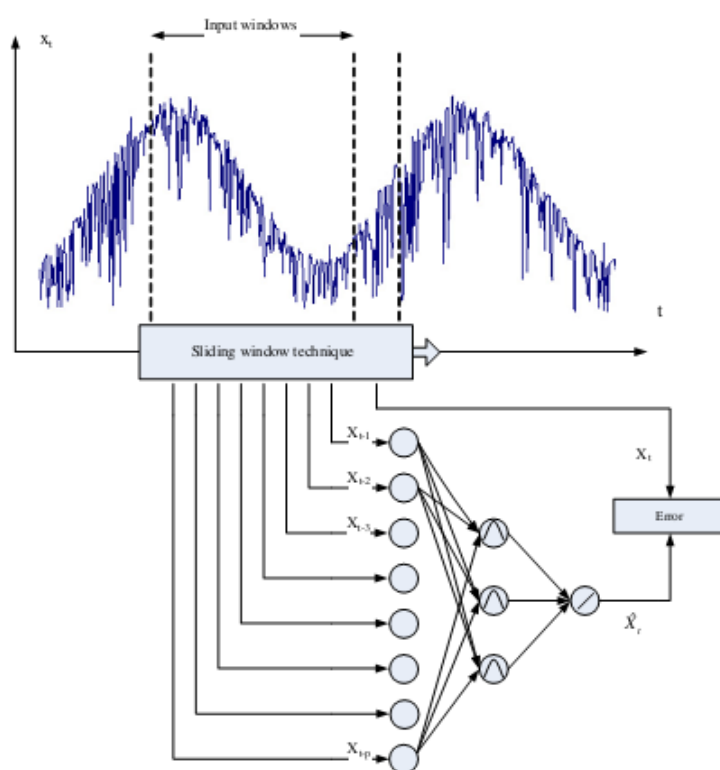


Figure 2.13: Kỹ thuật cửa sổ trượt -

2. MÔ HÌNH ARIMA VÀ MẠNG NƠI NHẬN TẠO

thức chuẩn hóa tuyến tính hay được sử dụng hơn

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

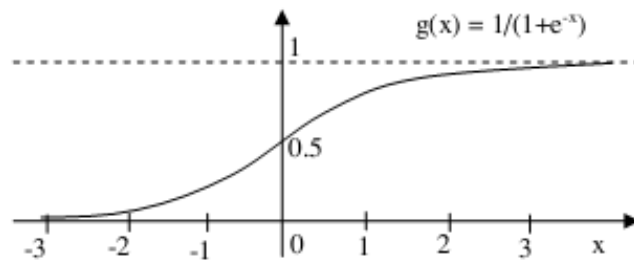


Figure 2.14: Đồ thị hàm sigmoid -

Ngoài ra theo Zhang (2005)[21], đối với dữ liệu có thành phần mang tính chu kỳ thời vụ hay là mang tính xu hướng, ta nên thực hiện khử xu hướng và khử yếu tố chu kỳ. Nó giúp việc huấn luyện và giảm thiểu lỗi tốt hơn nhiều so với dữ liệu thô chưa khử xu hướng hay chu kỳ.

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, tùy 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.

2.3 Mô hình mạng nơ ron nhân tạo

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

$$N_h = \alpha \sqrt{N_i * N_o} \quad (2.67)$$

trong đó N_o, N_h, N_i lần lượt là số lượng nơron ở lớp đầu vào, lớp ẩn và lớp đầu ra. α là tham số trong khoảng $[0.5, 2]$. Baum and Haussler [22] thì đề xuất công thức

$$N_h \leq \frac{N_{tr} * E_{tol}}{N_{dp} + N_o} \quad (2.68)$$

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. MÔ HÌNH ARIMA VÀ MẠNG NƠI NHÂN TẠO

Chapter 3

Bài toán dự báo trên dữ liệu Telecom

3.1 Dữ liệu telecom và bài toán dự báo lưu lượng

Trong lĩnh vực viễn thông, **lưu lượng** là thông tin (người sử dụng/báo hiệu) mang trên các kênh truyền dẫn trên mạng viễn thông. Chất lượng dịch vụ mạng phụ thuộc rất lớn vào việc lưu lượng thông tin trên mạng có được thông suốt hay không. Trong bối cảnh, cơ sở vật chất một số mạng chưa theo kịp với tốc độ phát triển bùng nổ các số thuê bao ở một số khu vực. Tình trạng này có thể dẫn đến việc xảy ra nghẽn mạng, dịch vụ mạng chập chờn, kết nối kém...đặc biệt là trong giờ cao điểm (hay còn gọi là giờ bận). Đó một phần do không dự báo được trước chính xác cường độ lưu lượng trên kênh truyền tại thời điểm đó để phân bổ và có phương án hợp lý. Một số khu vực thì kênh truyền khá rỗi rãi và không hoạt động hết công suất, trong khi một số khu vực khác thì nghẽn do lưu lượng quá lớn. Dự báo được thời điểm, lưu lượng tại một khu vực sẽ giúp nhà mạng phân bổ lại cơ sở hạ tầng và chia sẻ kênh truyền giúp cải thiện dịch vụ tốt hơn. Bài toán đặt ra ở đây là từ dữ liệu các cuộc gọi, tin nhắn và các dịch vụ giá trị gia tăng khác (call detail record) làm thế nào để ta có thể xác định tương đối lưu lượng kênh truyền thời điểm hiện tại và dự đoán lưu lượng trọng tương lai.

Ta tập trung vào phân tích lưu lượng tại các **giờ bận**. Đó là giờ có lưu lượng trao đổi qua các kênh là lớn nhất trong ngày. Lưu lượng trong giờ bận được xác

3. BÀI TOÁN DỰ BÁO TRÊN DỮ LIỆU TELECOM

định theo công thức Erlangs [23]:

$$A = C * t/T \quad (3.1)$$

với C là số cuộc gọi trong giờ bận, t là thời gian trung bình của một cuộc gọi và T là thời gian khảo sát. Trong thực tế, không chỉ có các cuộc gọi được thực hiện mà còn có các tin nhắn và các dịch vụ giá trị gia tăng khác như tải nhạc chuông, nghe nhạc... Công thức 3.1 có thể được mở rộng cho phù hợp với dữ liệu như sau:

$$A' = (\sum(n_{sms}) + \sum callDuration)/T \quad (3.2)$$

hay nói cách khác, lưu lượng trong giờ bận được tính bằng tổng số tin nhắn và tổng thời gian gọi hoặc sử dụng dịch vụ giá trị gia tăng.

Dữ liệu phân tích

Thực nghiệm sử dụng dữ liệu thu thập thông tin các cuộc gọi, tin nhắn, dịch vụ của người dùng mạng di động BeeLine từ 15/05/2010 đến 14/11/2010. Bản ghi chi tiết cuộc gọi gồm rất nhiều trường dữ liệu nhưng ta quan tâm đến một số trường quan trọng:

Rate – DT: thời điểm bắt đầu cuộc gọi, tin nhắn, dịch vụ

Trans – DT: thời điểm kết thúc cuộc gọi, dịch vụ

PRIMARY – UNIT: đo mức độ sử dụng dịch vụ, cuộc gọi... Với tin nhắn thì

PRIMARY – UNIT = 1, đối với đại lượng này tính bằng thời gian đàm thoại, đối với một số dịch vụ giá tăng thì có thể tính bằng lượng tải về Kb, Mb, Gb...

Sau khi thống kê xác định các giờ bận trong ngày thì ta có kết quả như Hình 3.1 Dựa vào biểu đồ, ta chọn khoảng thời gian từ 8h sáng đến 12h đêm là khoảng giờ bận. Tính tổng lưu lượng của các giờ bận trong ngày ta có kết quả như Hình Tập dữ liệu là chuỗi thời gian gồm 184 phần tử. Để tiến hành thực nghiệm độ hiệu quả của mô hình ta chia tập theo tỉ lệ *training : test* = 160 : 24. Trong quá trình thực nghiệm, ta xây dựng mô hình dựa trên dữ liệu 160 ngày đầu tiên và dự đoán kết quả của 24 ngày tiếp theo. Trong hình ??, đồ thị bên trái đường nét đứt mô tả tập học còn phần bên phải đường nét đứt mô tả tập kiểm tra mà ta sẽ sử dụng để đối chiếu với kết quả dự báo. Mô hình được ứng dụng thử nghiệm là mô hình ARIMA và Mạng nơ ron nhân tạo được nói tới trong chương 2.

3.1 Dữ liệu telecom và bài toán dự báo lưu lượng

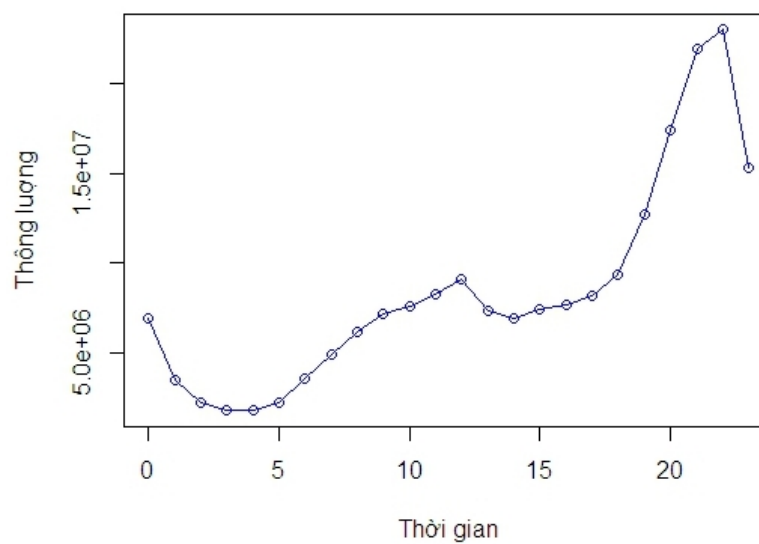


Figure 3.1: Thống kê lưu lượng theo giờ -

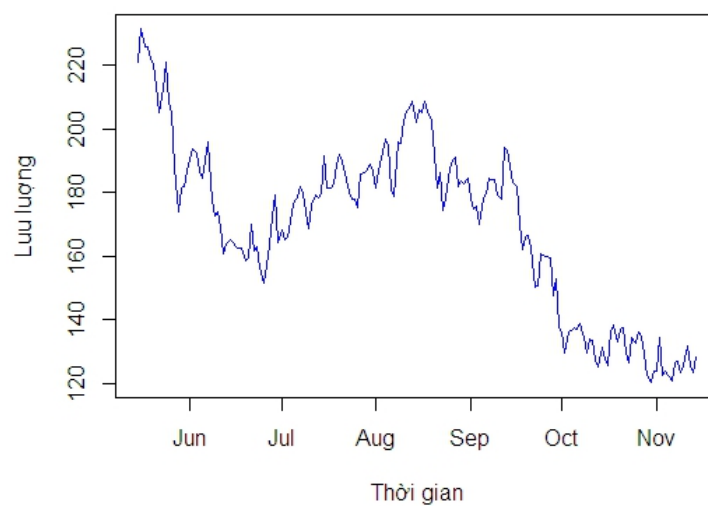


Figure 3.2: Lưu lượng các kênh truyền tính theo ngày - Đơn vị: 1 triệu Elangs

3.2 Thực nghiệm với ARIMA

3.2.1 Xác định mô hình

Bước đầu tiên là quan sát đồ thị của chuỗi thời gian (Hình 3.3). Ta nhận xét, chuỗi này đang có xu hướng giảm dần và có thể xuất hiện tính chu kỳ thời vụ với chu kỳ là 7 (ứng với 7 ngày trong một tuần lễ).

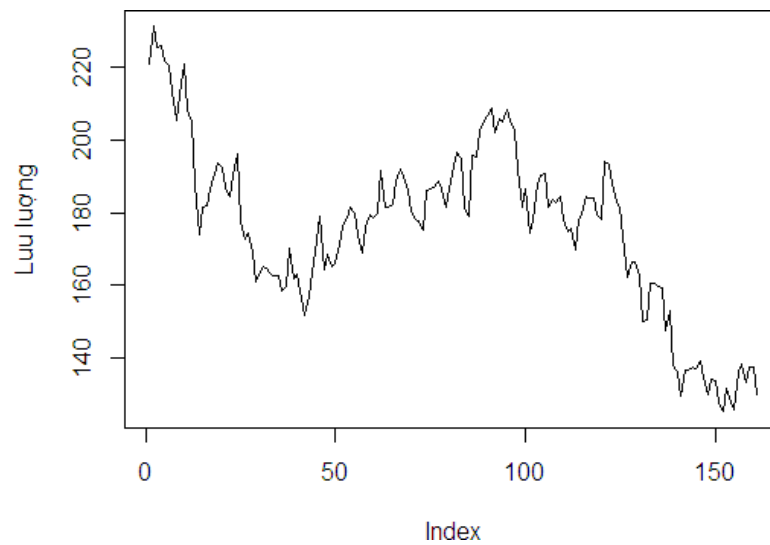


Figure 3.3: Lưu lượng kênh truyền từ 15/05/2010 đến 14/11/2010 của mạng di động Beeline -

Bước tiếp theo là kiểm tra độ ổn định của phương sai hay nói cách khác là kiểm tra sự thay đổi mức độ biến thiên quanh giá trị trung bình của chuỗi theo thời gian. Trong trường hợp, phương sai không ổn định, ta cần phải biến đổi dữ liệu. Đối với dữ liệu hiện tại, dựa vào hình 3.4, với độ chính xác 95%, giá trị hợp lý logarit cực đại có thể đạt được nếu $\lambda \in [0, 1]$. Khi $\lambda = 1$, theo Phần 2.2.2.1, dữ liệu không cần phải biến đổi. Ta sẽ chọn phương án không biến đổi dữ liệu, nếu mô hình đạt được cho kết quả dự đoán không tốt, việc biến đổi dữ liệu có thể thực hiện ở các vòng lặp sau.

Bước thứ 3 trong pha xác định mô hình là kiểm tra tính dừng của chuỗi thời gian. Ta nhận thấy đồ thị hàm ACF (Hình 3.5) của chuỗi thời gian lúc đầu có giá trị lớn ở đầu rồi sau đó giảm dần đều. Chứng tỏ đây có thể là chuỗi không có tính chất dừng. Để khẳng định chắc chắn, ta sử dụng phương pháp kiểm tra

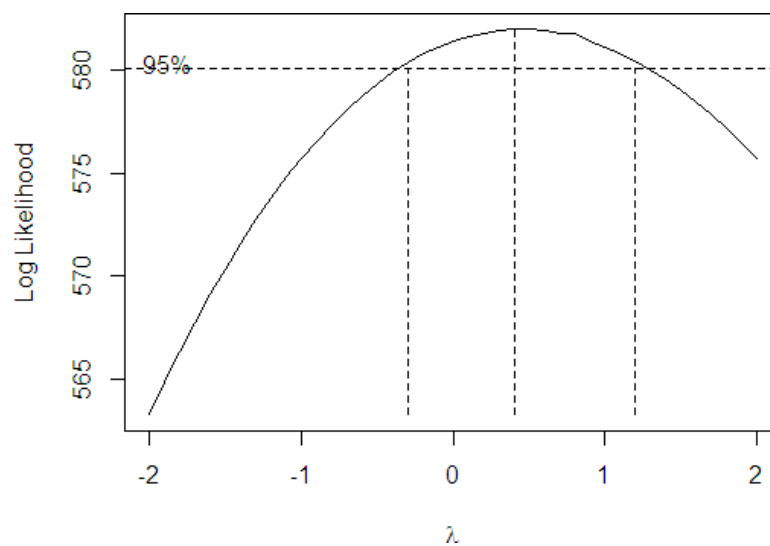


Figure 3.4: Đồ thị hàm hợp lý loga -

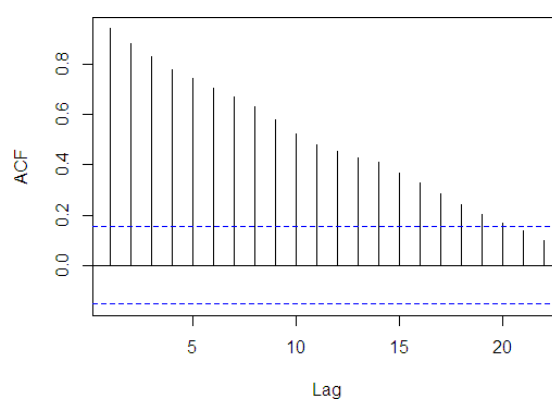


Figure 3.5: Đồ thị hàm ACF -

3. BÀI TOÁN DỰ BÁO TRÊN DỮ LIỆU TELECOM

Augmented Dicky-Fuller trên dữ liệu và thu được kết quả:

- giá trị thống kê $statistic = -2.0188$
- $p - value = 0.3068$

Giá trị $p - value > 0.05$ khá lớn, do đó ta chấp nhận giả thuyết H_0 cho rằng chuỗi thời gian đang kiểm tra không phải là chuỗi dừng. Ta cần phải thực hiện sai phân cấp 1 trên chuỗi thời gian.

Kiểm tra *Augmented Dicky-Fuller* với chuỗi mới thu được sau khi sai phân, ta có kết quả

- giá trị thống kê $statistic = -9.7984$
- $p - value = 0.01$

Giá trị $p - value < 0.05$, ta có thể bác bỏ giả thuyết H_0 . Do đó, chuỗi mới là chuỗi có tính chất dừng (Hình 3.6).

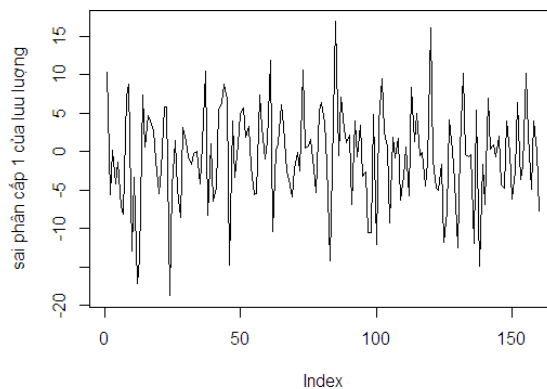


Figure 3.6: Đồ thị chuỗi đã được sai phân 1 lần -

Quan sát đồ thị hàm ACF (Hình 3.7) và PACF (Hình 3.8) của chuỗi thời gian sau khi đã sai phân ta thấy tại các độ trễ 7, 14, 21... giá trị tăng cao hơn so với các giá trị xung quanh. Dự đoán chắc chắn hơn chu kỳ sẽ là $lags = 7$.

Tại $lag = 14$ (là lags trội trong chu kỳ thứ 2) có giá trị vượt trội và quá ngưỡng sai số cho phép, dự đoán thành phần P, Q của tính chu kỳ mùa vụ có giá trị $P_7 = Q_7 = 2$. Tại $lag = 2$ cả giá trị acf và pacf đều có giá trị trội hơn xung

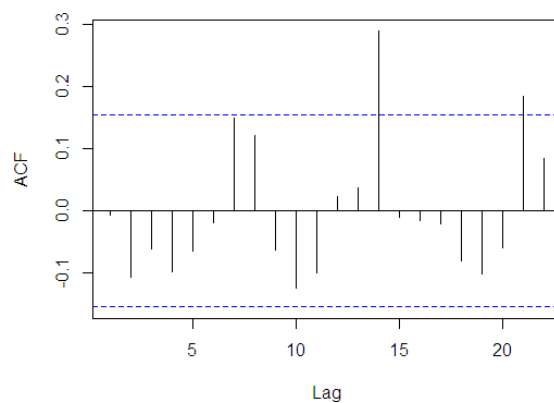


Figure 3.7: Đồ thị hàm ACF của chuỗi đã được sai phân -

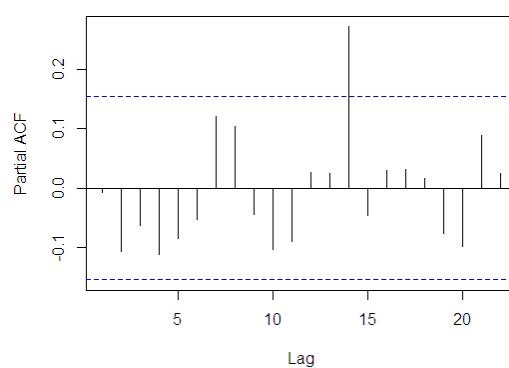


Figure 3.8: Đồ thị hàm PACF của chuỗi đã được sai phân -

3. BÀI TOÁN DỰ BÁO TRÊN DỮ LIỆU TELECOM

quanh, dự đoán đây có thể là mô hình $ARIMA(2, 1, 0)$ với thành phần chu kỳ mùa $(2, 0, 2)_7$ hoặc $ARIMA(0, 1, 2)$ với thành phần chu kỳ mùa $(2, 0, 2)_7$.

Nếu áp dụng khử *sai phân chu kỳ* với chu kỳ $k = 7$ cho chuỗi Δx_t

$$\Delta_7(\Delta(x_t)) = (1 - B^7)\Delta x_t = \Delta x_t - \Delta x_{t-7} = x_t - x_{t-1} - x_{t-7} + x_{t-8} \quad (3.3)$$

Ta thu được đồ thị hàm ACF và hàm PACF của chuỗi mới $(\Delta_7(\Delta(x_t)))$ như sau

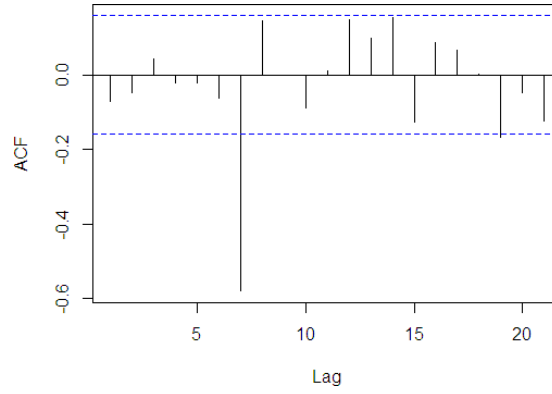


Figure 3.9: Đồ thị hàm ACF của chuỗi $\Delta_7(\Delta(x_t))$ -

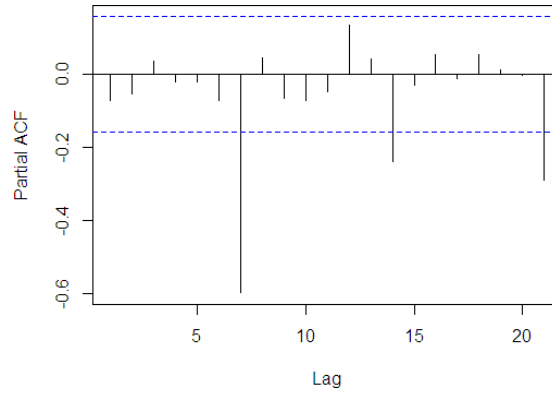


Figure 3.10: Đồ thị hàm PACF của chuỗi $\Delta_7(\Delta(x_t))$ -

Quan sát Hình 3.9 và 3.10, ta nhận thấy giá trị ACF và PACF của chuỗi $\Delta_7(\Delta(x_t))$ có giá trị rất trội tại điểm trễ $lag = 7$. Do vậy, ta đề xuất thêm mô hình có thành phần chu kỳ thời vụ $(1, 1, 1)_7$, $(0, 1, 1)_7$, $(1, 1, 0)_7$

Table 3.1: Độ đo AIC của các mô hình đề xuất

	(0,1,2)	(2,1,0)
$(2, 0, 0)_7$	1029.12	1029.37
$(0, 0, 2)_7$	1036.33	1036.72
$(2, 0, 2)_7$	1021.29	1021.48
$(1, 1, 0)_7$	1011.63	1011.58
$(0, 1, 1)_7$	990.39	990.41
$(1, 1, 1)_7$	989.01	989.06

Do số lượng mô hình đề xuất khá lớn, ta sử dụng chỉ số sai số AIC(the Akaike Information Criterion)(Bảng 3.1) để chọn mô hình có AIC nhỏ hơn.

Nhận thấy mô hình $ARIMA(0, 1, 2)(1, 1, 1)_7$ có chỉ số AIC nhỏ nhất. Đầu tiên, ta tiến hành tiếp tục ước lượng tham số và kiểm định mô hình này. Nếu trong trường hợp mô hình này không thỏa mãn, thử nghiệm sẽ được tiếp tục với các mô hình còn lại theo độ ưu tiên mô hình có chỉ số AIC nhỏ hơn. Kết quả thực nghiệm cho thấy, mô hình $ARIMA(0, 1, 2)(2, 0, 2)_7$ cho kết quả về tính chất nhiễu và sai số mô hình tốt nhất. Quá trình ước lượng và kiểm định tiếp theo sẽ được thực hiện với mô hình này.

3.2.2 Ước lượng tham số

Sử dụng ước lượng tham số cực đại, ta thu được các tham số

- Thành phần trung bình trượt MA(2): $ma1 = -0.0510$, $ma2 = -0.1344$
- Thành phần tự hồi quy chu kỳ SAR(1): $sar1 = -0.0795$, $sar2 = 0.9177$
- Thành phần trung bình trượt chu kỳ SMA(2): $sma1 = 0.2494$, $sma2 = -0.7136$

3.2.3 Kiểm định mô hình

1. Kiểm tra *Shapiro-Wilk* cho sai số chuẩn hóa: Kết quả đo được độ chuẩn tắc của sai số là $W = 0.991$ với $p = 0.4 > 0.05$. Do giá trị $p - value$ khá lớn,

3. BÀI TOÁN DỰ BÁO TRÊN DỮ LIỆU TELECOM

nên ta chấp nhận giả thuyết H_0 cho rằng sai số chuẩn hóa có tính độc lập cao.

2. Kiểm tra *runs test* cho sai số chuẩn hóa: Kết quả thu được $p\text{-value} = 0.997$. Tương tự, ta kết luận được chuỗi sai số chuẩn hóa có tính độc lập cao
3. Kiểm tra **Ljung-Box**: Kết quả thu được $p\text{-value} = 0.8463$ rất cao nên chấp nhận giả thuyết H_0 cho rằng chuỗi sai số là chuỗi nhiễu trắng Gaussian. Ngoài ra $Q_* = 4.8607, \chi_{0.95,9}^2 = 16.91798$ nên $Q_* < \chi_{0.95,9}^2$, ta chấp nhận giả thuyết cho rằng đây là mô hình ARIMA phù hợp

3.2.4 Dự báo

Kết quả dự báo cho 23 ngày tiếp theo được mô tả ở Hình 3.11 Mô hình dự báo

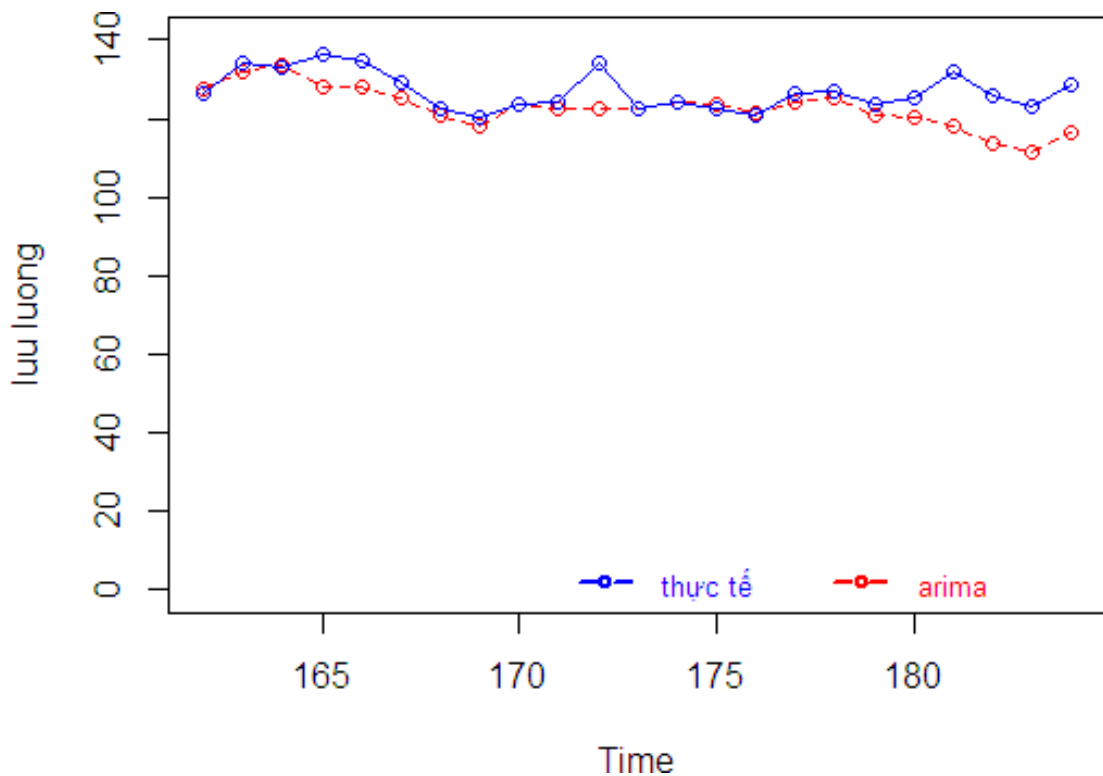


Figure 3.11: So sánh kết quả dự đoán 23 ngày tiếp theo sử dụng mô hình $ARIMA(0, 1, 2)(2, 0, 2)_7$ - (đường dự báo arima - nét đứt, đường thực tế - nét liền)

khá chính xác tại một số điểm như ngày thứ 3, 9, 10, 14, 15. Ngoài ra, do tính

chất không ổn định của dữ liệu nên dự đoán ở một số ngày vẫn có sai số đáng kể. Theo như xu hướng thì chuỗi có xu hướng đi xuống nên dự báo cũng chính xác theo hướng như vậy. Ta thấy ngoài một số điểm bất thường thì ARIMA dự báo khá chính xác mục tiêu ngắn hạn nhưng để đạt đến mức độ hoàn thiện hơn thì cần quay trở lại xác định lại mô hình. Việc này đòi hỏi cần kinh nghiệm và tốn rất nhiều công sức.

3.3 Thực nghiệm với mạng nơ ron nhân tạo

Trong mô hình mạng nơ ron nhân tạo với cả hai kĩ thuật huấn luyện lan truyền ngược và giải thuật di truyền, ta tiến hành thực nghiệm với mô hình đơn giản có 3 phần tử lớp đầu vào, phần tử lớp ẩn nằm trong khoảng từ 2 đến 9 và 1 phần tử lớp đầu ra. Như vậy, giá trị x_t sẽ được dự báo thông qua đầu vào là giá trị của 3 phần tử liền trước nó trong chuỗi thời gian $x_{t-1}, x_{t-2}, x_{t-3}$. Dữ liệu trước khi đưa vào học sẽ được chuyển sang dạng chuẩn hóa trong $[0, 1]$ theo công thức

$$x'_i = \frac{x_i - \min(X)}{\max(X) - \min(X)} \quad (3.4)$$

với x_i trong tập học.

Một số độ đo được lựa chọn để tìm ra mạng nơ ron nhân tạo có kiến trúc và trọng số phù hợp nhất như RMSE (căn quân phương của trung bình bình phương lỗi), MAE (trung bình của sai số tuyệt đối), E (chỉ số hiệu quả), R^2 (chỉ số xác định), sMAPE (trung bình đối xứng của tỉ lệ lỗi) ... để xác định độ chính xác của dự đoán

$$RMSE = \sqrt{\frac{1}{n} \sum_{i=1}^n (x_i - \hat{x}_i)^2} \quad (3.5)$$

$$MAE = \frac{1}{n} \sum_{i=1}^n |x_i - \hat{x}_i| \quad (3.6)$$

$$E = 1 - \frac{\sum (x_i - \hat{x}_i)^2}{\sum (x_i - \bar{x})^2} \quad (3.7)$$

$$R^2 = \frac{\sum (x_i - \bar{x})(\bar{x}_i - \hat{\hat{x}})}{\sqrt{\sum ((x_i - \bar{x})^2 \sum ((\bar{x}_i - \hat{\hat{x}})^2)}} \quad (3.8)$$

3. BÀI TOÁN DỰ BÁO TRÊN DỮ LIỆU TELECOM

$$sMAPE = \frac{1}{n} \sum \frac{|x_i - \hat{x}_i|}{\frac{|x_i| + |\hat{x}_i|}{2}} * 100 \quad (3.9)$$

Trong đó, độ đo $sMAPE$ (Symmetric mean absolute percentage error) là độ đo chính bởi nó không chỉ dựa vào chỉ số lỗi mà còn dựa vào độ lớn dữ liệu được dự báo. Trong một số trường hợp với dữ liệu có giá trị lớn thì lỗi được tính bằng hiệu giá trị dự báo và giá trị thực tế không phản ánh được mức độ sai số bằng tỉ lệ độ lớn chênh lệch đó so với độ lớn của giá trị dự báo và giá trị thực tế.

3.3.0.1 Phương pháp huấn luyện lan truyền ngược

Với mô hình sử dụng phương pháp huấn luyện lan truyền ngược, ta sử dụng thêm kết nối truyền thẳng (skip-layer) từ mỗi điểm ở lớp đầu vào tới điểm ở lớp đầu ra. Với mô hình mạng nơ ron có 3 nút đầu vào và một nút đầu ra, số nơ ron lớp ẩn được thực nghiệm từ 2 đến 9. Dựa vào kết quả kiểm tra các độ đo sai số ở Bảng 3.2, ta nhận thấy mô hình 3 : 2 : 1 với 2 nơ ron lớp ẩn cho kết quả các độ đo sai số nhỏ nhất.

Table 3.2: Các độ đo sai số dự đoán của mô hình mạng nơ ron sử dụng lan truyền ngược có 3 nút đầu vào

hid-neurons	rmse	mae	e	R2	sMAPE
2	4.831967	3.934214	1.01993	0.436791	3.09247
3	5.608044	4.277518	1.37387	0.303458	3.350934
4	5.774985	4.661212	1.456882	0.320368	3.642291
5	10.81326	8.214711	5.107821	-0.21396	6.25575
6	6.632455	5.0617	1.921637	0.23636	3.9351
7	6.188345	4.542465	1.672906	0.265164	3.544303
8	5.356544	4.273932	1.253407	0.340559	3.355468
9	6.462559	4.962741	1.824448	0.213298	3.884248

Giá trị trọng số các liên kết được mô tả trong Bảng 3.3 với i là nút đầu vào, h là nút ẩn, o là nút đầu ra, b là nút trực tiếp cung cấp giá trị ngưỡng.

Kết quả dự báo được mô tả ở Hình 3.12 Quan sát mô hình có thể thấy mô

3.3 Thực nghiệm với mạng nơ ron nhân tạo

Table 3.3: Trọng số trong mạng nơ ron 3:2:1 lan truyền ngược

b->h1	i1->h1	i2->h1	i3->h1
5.13	-2.01	-9.06	-3.44
b->h2	i1->h2	i2->h2	i3->h2
0.15	4.35	-4.44	-0.52
b->o	i1->o	i2->o	i3->o
1.43	-0.06	-2.52	3.59
	-2.84	-0.25	

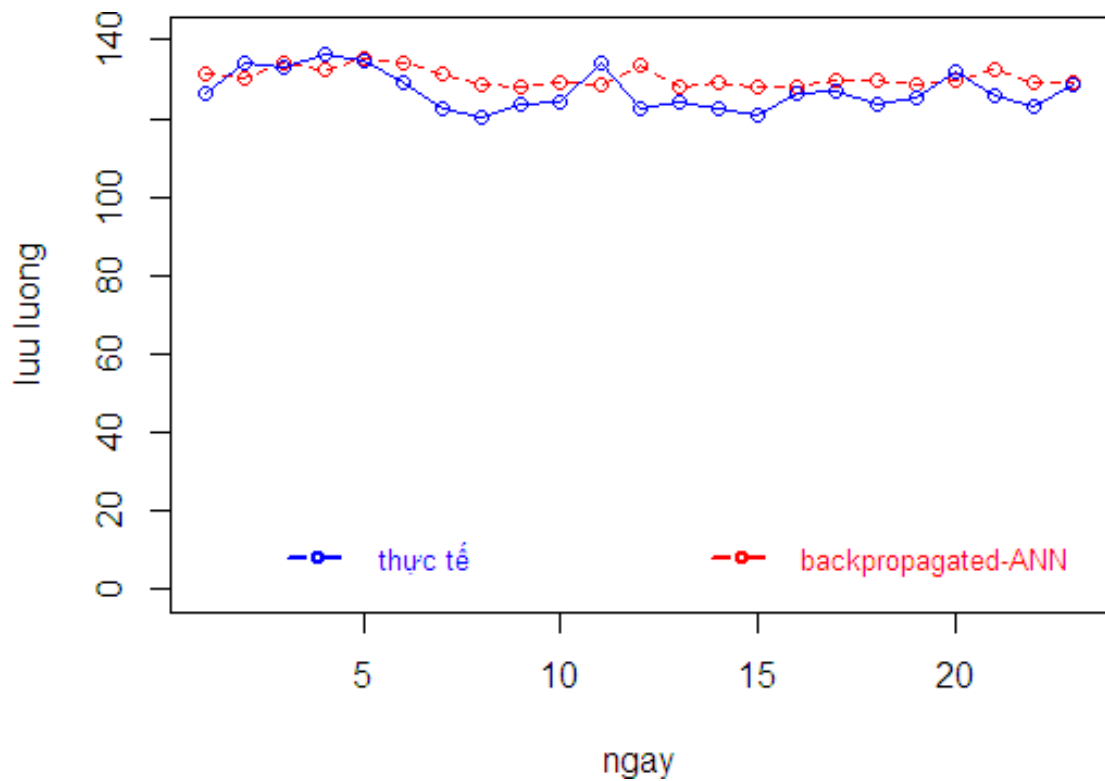


Figure 3.12: So sánh kết quả dự đoán 23 ngày tiếp theo sử dụng mô hình ANN sử dụng giải thuật lan truyền ngược - đường dự báo - nét đứt, đường thực tế - nét liền

3. BÀI TOÁN DỰ BÁO TRÊN DỮ LIỆU TELECOM

hình dự đoán có hình dạng gần giống chuỗi thời gian thực tế nhưng bị trễ một nhịp.

3.3.0.2 Sử dụng giải thuật di truyền

Thực nghiệm sử dụng mô hình mạng nơ ron áp dụng giải thuật di truyền có thêm một số tham số như

- tỉ lệ lai ghép *crossover - rate* = 0.6
- tỉ lệ biến dạng *mutation - rate* = 0.2
- Số vòng lặp (tương ứng với số thế hệ được tạo ra) *maxGen* = 1500
- Số cá thể trong quần thể là 500
- Tỉ lệ lỗi cho phép 0.001

Tương tự như mạng nơ ron sử dụng phương pháp lan truyền ngược, phần thực nghiệm mạng nơ ron với giải thuật di truyền cũng kiểm tra sai số đối với các mô hình có 3 nút đầu vào, 1 nút đầu ra và số nút ẩn từ 2 đến 9. Kết quả sai số thu được như Bảng 3.4:

Table 3.4: Các độ đo sai số dự đoán của mô hình mạng nơ ron sử dụng giải thuật di truyền có 3 nút đầu vào

hid_neurons	rmse	mae	e	R2	sMAPE
2	4.655608	3.794555	0.946837	0.433741	2.975085
3	6.163781	5.435273	1.659652	0.451339	4.248716
4	4.908285	4.242395	1.052403	0.44639	3.327021
5	5.331018	4.765645	1.241489	0.450103	3.733707
6	4.814318	4.005334	1.012493	0.438789	3.140415
7	7.126121	6.270289	2.218344	0.43476	4.872308
8	7.233547	6.353208	2.285731	0.442229	4.935695
9	5.457268	4.710235	1.300988	0.439929	3.684482

Quan sát Bảng 3.4 thì với các chỉ số sai số nhỏ nhất, mạng 3:2:1 vẫn thích hợp hơn cả. Sai số của nó so với mạng nơ ron sử dụng giải thuật lan truyền ngược nhỏ hơn nhưng không đáng kể.

3.4 So sánh kết quả

Ta có kết quả dự đoán Hình 3.13. Độ fitness tốt nhất của quần thể : 0.00298121. Nhiễm sắc thể tốt nhất tìm được tương ứng với tập trọng số tốt nhất là $-5.43/0.62/-0.53/0.90/10.40/-0.91/-0.53/-8.34/-4.94/2.56/0.70/$

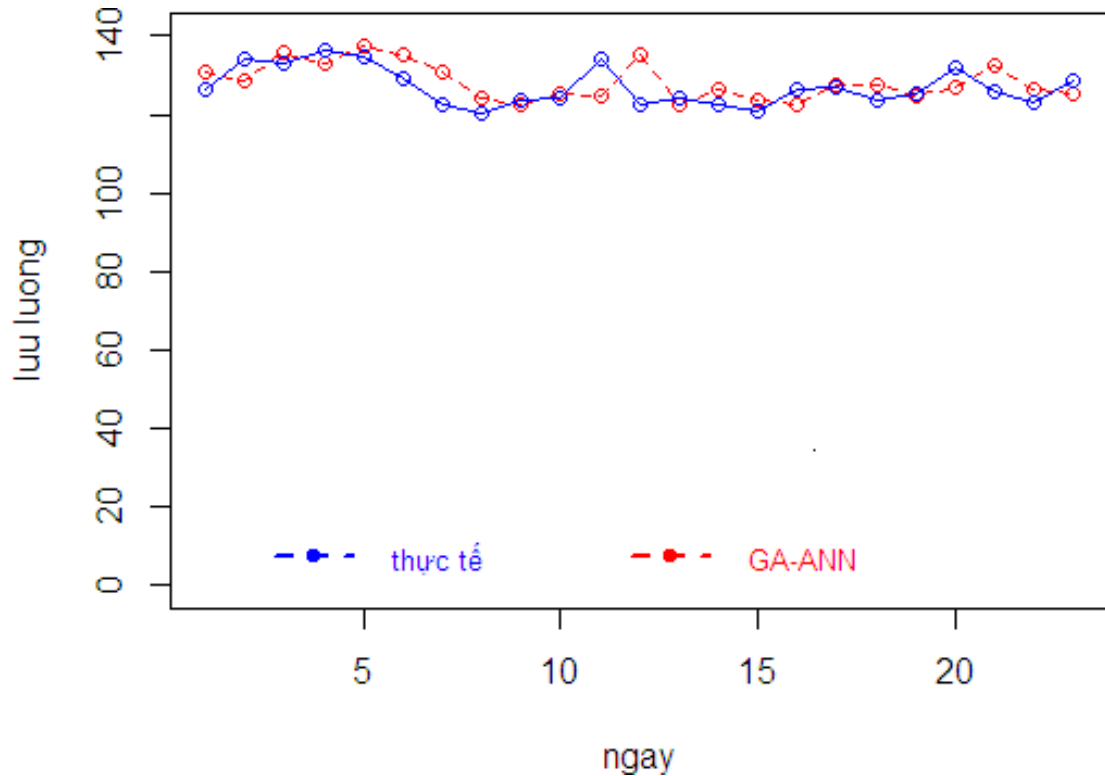


Figure 3.13: So sánh kết quả dự đoán 23 ngày tiếp theo sử dụng mô hình mạng nơ ron sử dụng GA -

Mô hình chuỗi thời gian dự đoán bởi mạng nơ ron mới xây dựng với đầu vào là tập học(160 ngày) có hình dáng khá giống với chuỗi thời gian thực tế(Hình 3.14). Ta hoàn toàn có thể tin tưởng được.

3.4 So sánh kết quả

So sánh độ đo lỗi của đại diện 2 phương pháp: ARIMA và mạng nơ ron nhân tạo ta có được kết quả như Bảng 3.5. Kết quả so sánh các mô hình đã thử nghiệm cho thấy 2 mạng nơ ron sử dụng 2 phương pháp học khác nhau cho kết quả gần

3. BÀI TOÁN DỰ BÁO TRÊN DỮ LIỆU TELECOM

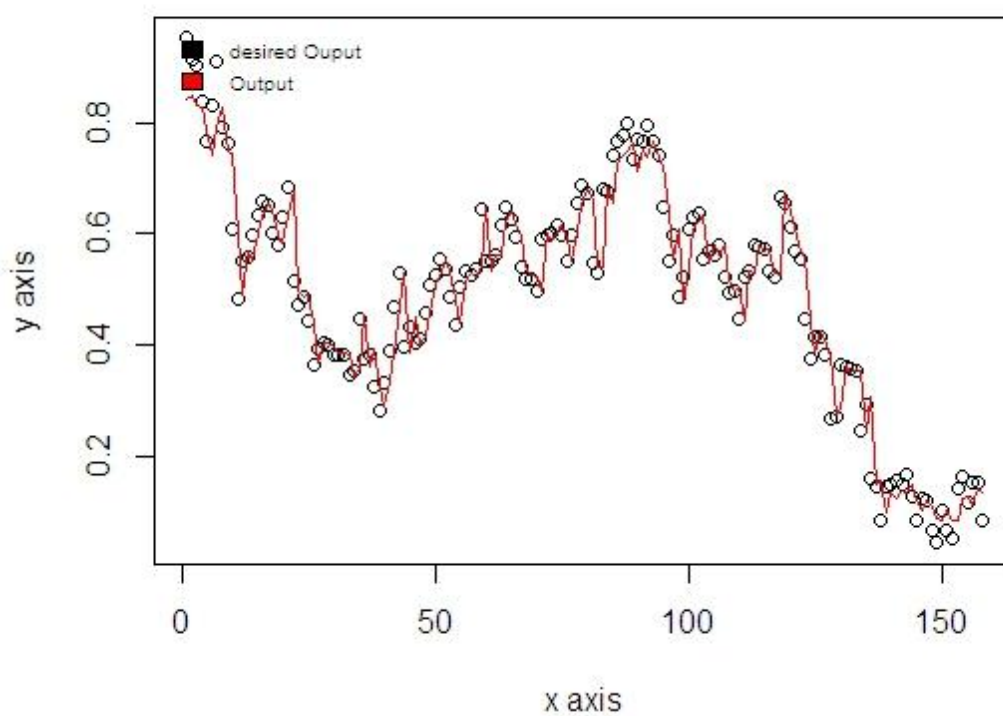


Figure 3.14: So sánh kết quả mô hình ANN-GA dự đoán trên tập học -
(Đường dự đoán - nét liền)

3.4 So sánh kết quả

bằng nhau và tốt hơn so với mô hình ARIMA. Trung bình tỉ lệ lỗi w_t so với giá trị của chuỗi thời gian x_t của mô hình mạng nơ ron là khoảng 2.975%-3.092%.

Table 3.5: So sánh độ đo lỗi 3 mô hình

	rmse	mae	e	R2	sMAPE
GA-ANN(3:2:1)	4.655608	3.794555	0.946837	0.433741	2.975085
ANN(3:2:1) lan truyền ngược	4.831967	3.934214	1.01993	0.436791	3.09247
ARIMA	8.21929	6.573086	2.95115	0.506937	5.349483

3. BÀI TOÁN DỰ BÁO TRÊN DỮ LIỆU TELECOM

Chapter 4

Kết luận và hướng nghiên cứu

4.1 Kết quả

Qua quá trình tìm hiểu bài toán dự báo với việc sử dụng mô hình thống kê ARIMA và mô hình mạng nơ ron nhân tạo (sử dụng giải thuật di truyền và lan truyền ngược), khóa luận đạt được một số kết quả sau đây:

- Trình bày mô hình thống kê ARIMA và các mô hình con, sử dụng thủ tục Box-Jenkins áp dụng bài toán dự báo chuỗi thời gian đơn chiều tuyến tính.
- Trình bày mô hình mạng nơ ron nhân tạo sử dụng phương pháp học bán giám sát là lan truyền ngược và giải thuật di truyền
- Đề xuất phương pháp tính lưu lượng kênh truyền vào giờ bận mở rộng không chỉ với các cuộc gọi mà còn tính toán cả với các tin nhắn và các dịch vụ giá trị gia tăng
- Thử nghiệm các mô hình vừa khảo sát với bài toán dự báo lưu lượng kênh truyền. Với mô hình tốt nhất đã thử nghiệm (mạng nơ ron nhân tạo có 3 nút đầu vào, 2 nút ẩn và 1 nút đầu ra sử dụng giải thuật di truyền tối ưu bộ trọng số) cho trung bình tỉ lệ lỗi so với giá trị của chuỗi chỉ khoảng $sMAPE = 2.975\%$. Với dữ liệu có giá trị trung bình $\mu = 192.9621$ thì tỉ lệ lỗi như vậy chấp nhận được. Mô hình đề xuất hoàn toàn có thể sử dụng được. Ta cũng đạt được kết quả với mô hình $ARIMA(0, 1, 2)(2, 0, 2)_7$ cho $sMAPE = 5.349\%$

4. KẾT LUẬN VÀ HƯỚNG NGHIÊN CỨU

4.2 Khó khăn

Trong quá trình thực hiện khóa luận, tồn tại một số khó khăn sau:

- Dữ liệu rất lớn (350GB) nhưng đa số các trường thuộc tính mập mờ, không có đặc tả dữ liệu hoàn chỉnh. Điều đó tạo rất nhiều khó khăn trong việc xác định phương pháp khai phá và các mô hình phù hợp cho dữ liệu.
- Mô hình ARIMA thực hiện rất nhiều bước trong thủ tục với rất nhiều tiêu chuẩn khác nhau. Cần chọn ra những tiêu chuẩn phù hợp nhất để áp dụng cho mô hình. Ngoài ra, việc dự đoán mô hình thường phải trải qua rất nhiều bước lặp lại nhưng không có một phương pháp cụ thể nào xác định được mô hình tối ưu nhất.

4.3 Hướng nghiên cứu tiếp theo

Bài toán dự đoán mở ra rất nhiều hướng nghiên cứu mới với khả năng áp dụng nhiều mô hình biến đổi khác nhau. Nghiên cứu tiếp theo là làm thế nào để xác định tối ưu các tham số trong cả 2 mô hình với số bước thử lặp đi lặp lại ít nhất. Ngoài ra, chúng tôi đang tiếp tục thử nghiệm một số mô hình dự báo khác như mô hình Markov ẩn, mô hình Bayesian, phương pháp k-người láng giềng gần nhất để áp dụng cho bài toán dự báo.

References

- [1] BOX G.;JENKINS G.M.; REINSEL G. *Time series analysis: Forecasting and control*. Prentice Hall, 3rd edition, 1994. vi, 23, 25, 27, 29, 31, 33, 35
- [2] ROSS IHAKA. **Time Series Analysis**. 2005. 2, 13
- [3] AJOY K. PALIT; DOBRIVOJE POPOVIC. *Computational Intelligence in Time Series Forecasting: Theory and Engineering Applications*. Springer, Berlin, 1st edition, 2005. 2, 3, 5, 13
- [4] RUEY S. TSAY. *Analysis of Financial Time Series*. Wiley, 3rd edition, 2010. 3
- [5] ROBERT H. SHUMWAY; DAVID S. STOFFER. *Time Series Analysis and Its Applications: With R Examples*. Springer, Berlin, 3rd edition, 2010. 6, 7, 8, 18
- [6] B. L.; O'CONNELL BOWERMAN. *Time Series and Forecasting*. Duxbury Press, North Scituate, Massachusetts., 1st edition, 1979. 9
- [7] D.A.; W.A. FULLER DICKEY. **Distribution of the Estimators for Autoregressive Time Series with a Unit Root**. *Journal of the American Statistical Association*, **74**:427–431, 1979. 10, 28
- [8] T. J.; J. H. STOCK ELLIOTT, G.; ROTHENBERG. **Efficient Tests for an Autoregressive Unit Root**. *Econometrica*, **64**:813–836, 1996. 10
- [9] P.C.B; P. PERRON PHILLIPS. **Testing for a Unit Root in Time Series Regression**. *Biometrika*, **75**:335–346, 1988. 10

REFERENCES

- [10] **Liên quan tới khái niệm về dự báo.** *eFinance*, **21**, 2005. 11
- [11] 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
- [12] D. R. BOX, GEORGE E. P.; COX. **An analysis of transformations.** *Journal of the Royal Statistical Society*, pages 211–252, 1964. 26
- [13] H. AKAIKE. **Maximum likelihood identification of Gaussian autoregressive moving-average models.** *Biometrika*, pages 255–266, 1973. 32
- [14] M. B. SHAPIRO, S. S.; WILK. **An analysis of variance test for normality (complete samples).** *Biometrika*, page 591–611, 1965. 34
- [15] G. M. LJUNG; G. E. P. BOX. **On a Measure of a Lack of Fit in Time Series Models.** *Biometrika*, **2**:297–303, 1978. 34
- [16] WIDROW B ; HOFF ME. **Adaptive Switching Circuits.** *Anderson J and Rosenfeld E. (eds.) Neurocomputing*, pages 126–134, 1960. 37
- [17] HECHT-NIELSEN R. **Kolmogorov’s Mapping Neural Network Existence Theorem. III,** pages 11–14, San Diego, CA, 1987. 39
- [18] LIPPMANN R.P. **An introduction to computing with neural nets.** *IEEE ASSP Magazine*, page 4–22, 1987. 39
- [19] CYBENKO G. **Continuous valued neural networks with two hidden layers are sufficient.** Technical report, Taft University, 1988. 39
- [20] RUMELHART D. E; MCCLELLAND. *Parallel Distributed Processing: Explorations in the Microstructure of Cognition.* MIT Press, Cambridge, MA, 1986. 40
- [21] G.PETER ZHANG AND MIN QI. **Neural network forecasting for seasonal and trend time series.** *European Journal of Operational Research*, **160**(2):501 – 514, 2005. 46

REFERENCES

- [22] BAUM EB; HAUSSLER D. **What Size Net Gives Valid Generalisation?**
Neural Computation, page 151–160, 1989. 47
- [23] J.E. FLOOD. *Telecommunication Switching Traffic and Networks*. Prentice Hall, 1995. 50