

BỘ GIÁO DỤC VÀ ĐÀO TẠO  
TRƯỜNG ĐẠI HỌC SƯ PHẠM KỸ THUẬT  
THÀNH PHỐ HỒ CHÍ MINH



LUẬN VĂN THẠC SĨ  
PHẠM CHÍ CÔNG

DỰ BÁO TRÊN CHUỖI THỜI GIAN SỬ DỤNG  
MÔ HÌNH LAI GHÉP ARIMA VÀ RBFNN

NGÀNH: KHOA HỌC MÁY TÍNH - 8480101



Tp. Hồ Chí Minh, tháng 03/2021

BỘ GIÁO DỤC VÀ ĐÀO TẠO  
TRƯỜNG ĐẠI HỌC SƯ PHẠM KỸ THUẬT  
THÀNH PHỐ HỒ CHÍ MINH

LUẬN VĂN THẠC SĨ  
PHẠM CHÍ CÔNG

DỰ BÁO TRÊN CHUỖI THỜI GIAN SỬ DỤNG MÔ HÌNH  
LAI GHÉP ARIMA VÀ RBFNN

NGÀNH: KHOA HỌC MÁY TÍNH - 8480101

Tp. Hồ Chí Minh, tháng 3 năm 2021

BỘ GIÁO DỤC VÀ ĐÀO TẠO  
TRƯỜNG ĐẠI HỌC SƯ PHẠM KỸ THUẬT  
THÀNH PHỐ HỒ CHÍ MINH

LUẬN VĂN THẠC SĨ  
PHẠM CHÍ CÔNG

DỰ BÁO TRÊN CHUỖI THỜI GIAN SỬ DỤNG MÔ HÌNH  
LAI GHÉP ARIMA VÀ RBFNN

NGÀNH: KHOA HỌC MÁY TÍNH – 8480101

Hướng dẫn khoa học:  
TS. NGUYỄN THÀNH SƠN

Tp. Hồ Chí Minh, tháng 3 năm 2021

# QUYẾT ĐỊNH GIAO ĐỀ TÀI

BỘ GIÁO DỤC VÀ ĐÀO TẠO  
TRƯỜNG ĐẠI HỌC SƯ PHẠM KỸ THUẬT  
THÀNH PHỐ HỒ CHÍ MINH  
Số: 2478/QĐ-DHSPKT

CỘNG HÒA XÃ HỘI CHỦ NGHĨA VIỆT NAM  
Độc lập – Tự do – Hạnh phúc

Tp. Hồ Chí Minh, ngày 15 tháng 9 năm 2020

## QUYẾT ĐỊNH

Về việc giao đề tài luận văn tốt nghiệp và người hướng dẫn năm 2020

### HIỆU TRƯỞNG TRƯỜNG ĐẠI HỌC SƯ PHẠM KỸ THUẬT TP. HỒ CHÍ MINH

Căn cứ Luật Giáo dục đại học ngày 18 tháng 6 năm 2012 và Luật sửa đổi, bổ sung một số điều của Luật Giáo dục đại học ngày 19 tháng 11 năm 2018;

Căn cứ Nghị định 99/2019/NĐ-CP ngày 30 tháng 12 năm 2019 của Chính phủ Quy định chi tiết và hướng dẫn thi hành một số điều của Luật sửa đổi, bổ sung một số điều của Luật giáo dục đại học;

Căn cứ Quyết định số 937/QĐ-TTg ngày 30 tháng 6 năm 2017 của Thủ tướng Chính phủ về việc phê duyệt đề án thí điểm đổi mới cơ chế hoạt động của Trường Đại học Sư phạm Kỹ thuật TP. Hồ Chí Minh;

Căn cứ Nghị quyết số 27/NQ-HĐT ngày 29 tháng 7 năm 2020 của Hội đồng trường ban hành Quy chế tổ chức và hoạt động của trường Đại học Sư phạm Kỹ thuật TP. Hồ Chí Minh;

Căn cứ Thông tư số 15/2014/TT-BGDĐT ngày 15 tháng 5 năm 2014 của Bộ Giáo dục và Đào tạo về việc Ban hành Qui chế đào tạo trình độ thạc sĩ;

Căn cứ vào Biên bản bảo vệ Chuyên đề của ngành Khoa học máy tính vào ngày 26/08/2020;

Theo nhu cầu công tác và khả năng cán bộ;

Theo đề nghị của Trưởng phòng Đào tạo,

## QUYẾT ĐỊNH:

**Điều 1.** Giao đề tài Luận văn tốt nghiệp thạc sĩ và người hướng dẫn Cao học năm 2020 cho:

Học viên : Phạm Chí Công MSHV: 1981301

Ngành : Khoa học máy tính

Tên đề tài : Dự báo trên chuỗi thời gian sử dụng mô hình lai ghép ARIMA và RBFNN

Người hướng dẫn : TS. Nguyễn Thành Sơn

Thời gian thực hiện : Từ ngày 01/09/2020 đến ngày 28/02/2021

**Điều 2.** Giao cho Phòng Đào tạo quản lý, thực hiện theo đúng Qui chế đào tạo trình độ thạc sĩ của Bộ Giáo dục & Đào tạo ban hành.

**Điều 3.** Trưởng các đơn vị, phòng Đào tạo, các Khoa quản ngành cao học và các Ông (Bà) có tên tại Điều 1 chịu trách nhiệm thi hành quyết định này.

Quyết định có hiệu lực kể từ ngày ký./. 

Nơi nhận :

- BGH (để biết);
- Như điều 3;
- Lưu: VT, SDH (3b).

KT. HIỆU TRƯỞNG  
PHÓ HIỆU TRƯỞNG



TS. Lê Hiếu Giang

# BIÊN BẢN CHẤM LUẬN VĂN TỐT NGHIỆP THẠC SĨ

BỘ GIÁO DỤC VÀ ĐÀO TẠO  
TRƯỜNG ĐẠI HỌC SƯ PHẠM KỸ THUẬT  
THÀNH PHỐ HỒ CHÍ MINH

CỘNG HÒA XÃ HỘI CHỦ NGHĨA VIỆT NAM  
Độc lập - Tự do - Hạnh phúc

## BIÊN BẢN CHẤM LUẬN VĂN TỐT NGHIỆP THẠC SĨ\_NĂM 2021 NGÀNH: KHOA HỌC MÁY TÍNH

Hội đồng chấm LVTN theo QĐ số: 1009/QĐ-DHSPKT, ngày 07/04/2021

Có mặt : .....5/5..... Vắng mặt: ...0.....

Chủ tịch Hội đồng : TS. Lê Văn Vinh

Thư ký Hội đồng : TS. Huỳnh Nguyên Chính

Học viên bảo vệ LVTN : Phạm Chí Công

MSHV: 1981301

Giảng viên hướng dẫn : TS. Nguyễn Thành Sơn

Giảng viên phản biện : PGS.TS. Hoàng Văn Dũng

TS. Võ Xuân Thế

Tên đề tài LVTN : *Dự báo trên chuỗi thời gian sử dụng mô hình lai ghép ARIMA và RBFNN*

### I. KẾT QUẢ BẢO VỆ:

STT	Thành viên Hội đồng	Kết quả bảo vệ	Ghi chú
1.	TS. Lê Văn Vinh	7,0	
2.	TS. Huỳnh Nguyên Chính	7,1	
3.	PGS.TS. Hoàng Văn Dũng	7,6	
4.	TS. Võ Xuân Thế	9,0	
5.	PGS.TS. Trần Văn Lăng	8,0	
	Tổng điểm	<u>38,7</u>	
	Điểm trung bình	<u>7,7</u>	

### II. KẾT LUẬN:

(Thư ký hội đồng ghi rõ các ý kiến của thành viên hội đồng về việc chínhしさ, bổ sung những nội dung gì trong LVTN)

.....  
.....  
.....  
.....  
.....

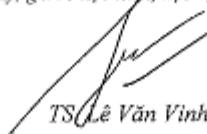
Tp.Hồ Chí Minh, ngày 22 tháng 4 năm 2021

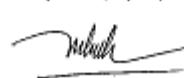
### THƯ KÝ HỘI ĐỒNG

(Ký, ghi rõ họ tên, học hàm, học vị & họ tên)

CHỦ TỊCH HỘI ĐỒNG

(Ký, ghi rõ họ tên, học hàm, học vị & họ tên)

  
TS. Lê Văn Vinh



TS. Huỳnh Nguyên Chính

# NHẬN XÉT CỦA GIẢNG VIÊN PHẢN BIỆN



BỘ GIÁO DỤC VÀ ĐÀO TẠO  
TRƯỜNG ĐẠI HỌC SƯ PHẠM KỸ THUẬT  
THÀNH PHỐ HỒ CHÍ MINH

## PHIẾU NHẬN XÉT LUẬN VĂN THẠC SỸ - HƯỚNG ỨNG DỤNG

(Dành cho giảng viên phản biện)

Tên đề tài luận văn thạc sĩ: *Dự báo trên chuỗi thời gian sử dụng mô hình lai ghép ARIMA và RBFNN*

Tên tác giả: *Phạm Chí Công*

MSHV: 1981301

Ngành: *Khoa học máy tính*

Khóa: 2019-2021

Họ và tên người phản biện: *PGS.TS. Hoàng Văn Dũng*

Chức danh: Phó Giáo sư

Học vị: Tiến Sĩ

Cơ quan công tác: Trường ĐHSPKT TP. HCM

Điện thoại liên hệ: 0913317759

### I. Y KIẾN NHẬN XÉT

#### 1. Về hình thức & kết cấu luận văn.

Kết cấu luận văn được phân chia thành 4 chương chính và phần kết luận hướng phát triển là tương đối phù hợp. Luận văn gồm các nội dung cơ bản Tổng quan, cơ sở mô hình RBFNN, ARIMA, ARIMARBFNN và đề xuất cải tiến lai ghép bằng cách thực hiện song song hai mô hình ARIMA-RBFNN, đánh giá thực nghiệm. Dung lượng giữa các chương, mục được trình bày khá phù hợp. Hình thức trình bày rõ ràng, theo mẫu quy định.

#### 2. Về nội dung

##### 2.1 Nhận xét về tính khoa học, rõ ràng, mạch lạc, khác chiết trong luận văn.

Các nội dung cơ sở, lý thuyết liên quan của luận văn được trình bày khá rõ ràng, đảm bảo tính khoa học và mạch lạc. Tuy nhiên một số công thức và suy diễn cần xem xét để điều chỉnh phù hợp hơn.

##### 2.2 Nhận xét đánh giá việc sử dụng hoặc trích dẫn kết quả NC của người khác có đúng qui định hiện hành của pháp luật sở hữu trí tuệ.

Các tài liệu liên quan, sử dụng trong luận văn được thể hiện ở danh mục tài liệu tham khảo và một số ví dụ trong nội dung có tham chiếu trích dẫn, tuy nhiên chưa thực hiện đầy đủ và chính xác.

Nên chỉ tham khảo những tài liệu xuất bản chính thức, có tính xác thực để viết luận văn và thay vì tham khảo trên các trang web tự do. Các trang web, diễn đàn,... không nên liệt kê ở danh mục tài liệu tham khảo để đảm bảo tính chính xác. Trong một số trường hợp đặc biệt, việc sử dụng minh họa, sao chép từ các trang web cần có chính sửa cho phù hợp với ngữ cảnh nội dung luận văn và nên chỉ rõ tham chiếu đến bằng dạng footnote thay vì ở danh mục tài liệu tham khảo.

##### 2.3 Nhận xét về mục tiêu nghiên cứu, phương pháp nghiên cứu sử dụng trong

### **LVTN.**

Mục tiêu nghiên cứu của luận văn được xác định khá rõ ràng với 4 mục tiêu cơ bản, liên quan đến tìm hiểu kiến thức chung và các thuật toán về phân lớp, luật kết hợp và tối ưu hóa.

Phương pháp nghiên cứu sử dụng thực hiện luận văn liên quan đến nghiên cứu lý thuyết, cải đặt thực nghiệm, phân tích đánh giá là phù hợp với hướng nghiên cứu lý thuyết cơ bản.

### **2.4 Nhận xét Tổng quan của đề tài.**

Học viên đã trình bày tổng quan đề tài khá phù hợp, đánh giá được những vấn đề chung, các nghiên cứu liên quan và hướng phát triển, hướng ứng dụng của các phương pháp dự báo trên chuỗi thời gian bằng cách ghép ARIMA-RBFNN.

### **2.5 Nhận xét đánh giá về nội dung & chất lượng của LVTN.**

Nội dung luận văn tương đối phù hợp với mục tiêu, đối tượng, phạm vi và nhiệm vụ nghiên cứu đã đề ra từ đầu. Luận văn mang tính tìm hiểu các thuật toán và vấn đề trong phân tích chuỗi ứng dụng để dự báo. Đề trình bày được các thuật toán cụ thể về dự báo dựa trên phân tích chuỗi thời gian. Bên cạnh đó, học viên đã thực hiện đánh giá thực nghiệm thuật toán theo hướng ứng dụng kết hợp hai phương pháp dự báo lai ghép song song ARIMA và RBFNN với nhau theo cách tiếp cận của nhóm nghiên cứu L.Zhang trong tài liệu [6].

### **2.6 Nhận xét đánh giá về khả năng ứng dụng, giá trị thực tiễn của đề tài.**

Luận văn đã đạt được những kết quả nhất định trong nghiên cứu lý thuyết. Nội dung luận văn có thể làm tài liệu tham khảo cho những nghiên cứu, ứng dụng sau này liên quan bài toán dự báo kết hợp phân tích chuỗi thời gian và mô hình trí tuệ nhân tạo. Luận văn có ý nghĩa theo hướng nghiên cứu thực nghiệm nhằm làm tiền đề cho các ứng dụng sau này.

### **2.7 Luận văn cần chỉnh sửa, bổ sung những nội dung gì (thiết sót và tồn tại).**

Các biểu đồ trong các hình cần phải có đơn vị đo trong số liệu giữa trực Tung và Hoành, ví dụ Hình 4.1, Hình 4.2, Hình 4.4, Hình 4.6, Hình 4.7, Hình 4.9, Hình 4.11, Hình 4.12, Hình 4.14, Hình 4.16,...

Cần thể hiện cụ thể cách đo sai số từng phần tử trong các độ đo sai số trên tập dữ liệu đánh giá như MAE (*Mean Absolute Error*), RMSE (*Root Mean Squared Error*).

Các công thức cần đánh số thứ tự, có tham chiếu đến trong nội dung luận văn.

Hãy kiểm tra lại các công thức và cách biến đổi ở mục “3.2.4.2. Phương pháp ước lượng bình phương nhỏ nhất” cho rõ ràng và súc tích hơn.

## **II. CÁC VẤN ĐỀ CẦN LÀM RỎ**

(Các câu hỏi của giảng viên phản biện)

1. Ở mục “3.2.4.1. Phương pháp ước lượng tham số Moment”, hãy giải thích công thức *Jule-Walker* trong hệ phương trình trang 31 và cách sử dụng nó trong thực nghiệm luận văn.
2. Kết quả thực nghiệm ở hình “4.20: Thời gian thực thi của các mô hình dự báo khi dùng 128 nút ẩn trên tập dữ liệu City\_temperature”, vì sao thời gian thực hiện

RFBNN bằng 0 (hoặc không đáng kể ~1/6000 các phương pháp khác). Tương tự giải thích kết quả ở Hình 4.13 và Hình 4.15.

### III. ĐÁNH GIÁ

TT	Mục đánh giá	Đánh giá	
		Đạt	Không đạt
1	Tinh khoa học, rõ ràng, mạch lạc, khúc chiết trong luận văn.	x	
2	Đánh giá việc sử dụng hoặc trích dẫn kết quả NC của người khác có đúng qui định hiện hành của pháp luật sở hữu trí tuệ.	x	
3	Mục tiêu nghiên cứu, phương pháp nghiên cứu sử dụng trong LVTN.	x	
4	Tổng quan của đề tài.	x	
5	Đánh giá về nội dung & chất lượng của LVTN.	x	
6	Đánh giá về khả năng ứng dụng, giá trị thực tiễn của đề tài.	x	

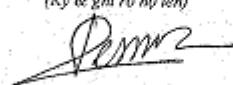
Đánh dấu chéo (x) vào ô muốn Đánh giá

### IV. KẾT LUẬN

(Giảng viên phản biện ghi rõ ý kiến "Tán thành luận văn" hay "Không tán thành luận văn")  
Đồng ý cho bảo vệ

TP Hồ Chí Minh, ngày 20 tháng 4 năm 2021

Người nhận xét  
(Ký & ghi rõ họ tên)



PGS.TS. Hoàng Văn Dũng



BỘ GIÁO DỤC VÀ ĐÀO TẠO  
TRƯỜNG ĐẠI HỌC SƯ PHẠM KỸ THUẬT  
THÀNH PHỐ HỒ CHÍ MINH

HCMUTE

## PHIẾU NHẬN XÉT LUẬN VĂN THẠC SỸ - HƯỚNG ỨNG DỤNG

(Dành cho giảng viên phản biện)

Têu đề tài luận văn thạc sĩ: *Dự báo trên chuỗi thời gian sử dụng mô hình lai ghép ARIMA và RBFNN*

Tên tác giả: *Phạm Chí Công*

MSHV: 1981301

Ngành: *Khoa học máy tính*

Khóa: 2019-2021

Họ và tên người phản biện: *TS. Võ Xuân Thê*

Chức danh: Giảng viên

Học vị: Tiến Sĩ

Cơ quan công tác: Khoa CNTT - ĐH Tài chính Marketing

Điện thoại liên hệ: 0913 660 009 OR 0916 975 888

### I. Ý KIÊN NHẬN XÉT

#### 1. Về hình thức & kết cấu luận văn.

- Cơ bản hình thức và kết cấu luận văn rõ ràng, phù hợp với báo cáo khoa học trình độ sau đại học.
- Có tóm tắt ý chính đầu chương rất rõ ràng.
- Xin có một số góp ý trong phần "chính sửa, bổ sung" (bên dưới).

#### 2. Về nội dung

##### 2.1 Nhận xét về tính khoa học, rõ ràng, mạch lạc, khúc chiết trong luận văn.

- Báo cáo được trình bày mạch lạc, phù hợp với báo cáo khoa học bậc Thạc sỹ, văn phong rõ ràng; tuy nhiên:
- Đề xuất tác giả: Nên diễn đạt lại một số nội dung trong báo cáo cho rõ hơn, tránh bị hiểu lầm (phần "chính sửa, bổ sung" của nhận xét này), nhằm nâng cao giá trị của báo cáo.

##### 2.2 Nhận xét đánh giá việc sử dụng hoặc trích dẫn kết quả NC của người khác có đúng qui định hiện hành của pháp luật sở hữu trí tuệ.

- Việc sử dụng và trích dẫn kết quả nghiên cứu khác cơ bản đúng qui định về sở hữu trí tuệ.
- Nguồn tham khảo chính của báo cáo là [6] L. Zhang, G. X. Zhang, and R. R. Li. "Water Quality Analysis and Prediction Using Hybrid Time Series and Neural Network Models". JAST\_Volume 18\_Issue 4\_Pages 975-983 (2018): Phân tích và dự đoán chất lượng nước bằng cách sử dụng mô hình kết hợp chuỗi thời gian (tuyến tính với phi tuyến) và ANN.

##### 2.3 Nhận xét về mục tiêu nghiên cứu, phương pháp nghiên cứu sử dụng trong LVTN.

- Mục tiêu nghiên cứu của đề tài rõ ràng, phương pháp nghiên cứu tin cậy.

##### 2.4 Nhận xét Tổng quan của đề tài.

- Đề tài có hàm lượng khoa học rõ ràng.
- Có cơ sở LT và thực nghiệm đầy đủ và phù hợp với nội dung đề tài.
- Nội dung chính của báo cáo là 3.3 / cuối tr36 và 3.4 / tr37

### **2.5 Nhận xét đánh giá về nội dung & chất lượng của LVTN.**

- Nội dung luận văn có hàm lượng khoa học rõ ràng, có đầu tư xứng đáng.
- Tác giả có đầu tư nghiêm túc khi thực hiện LVTN với kiến thức và kỹ năng chuyên môn tốt, xứng đáng với trình độ Ths.

### **2.6 Nhận xét đánh giá về khả năng ứng dụng, giá trị thực tiễn của đề tài.**

- Đề tài có giá trị thực tiễn cao: có thể ứng dụng phổ biến trong nhiều lĩnh vực khác nhau một cách đơn giản và hiệu quả.

### **2.7 Luận văn cần chỉnh sửa, bổ sung những nội dung gì (thiết sót và tồn tại).**

NHÀM GIÚP BÁO CÁO CỦA LUẬN VĂN CÓ GIÁ TRỊ CAO HƠN, XIN ĐỀ XUẤT TÁC GIÁ MỘT SỐ CHỈNH SỬA SAU:

[1] Tên Chương 5: nên là : "Kết luận và hướng phát triển" (theo đúng như nội dung bên trong chương). Không nên là "...Kiến nghị" (vì đây chỉ dùng cho báo cáo khoa học tại cơ quan, đơn vị => đã kiến nghị phải có nơi nhận hoặc người nhận: vô lý với báo cáo LV Ths).

[2] Cần rõ danh mục thuật ngữ và từ viết tắt (cần hơn danh mục bảng và hình): để làm rõ một số thuật ngữ và từ viết tắt dùng trong báo cáo.

VD . Mùa vụ 3.2.2. / tr41;

. ETS = không gian trạng thái [cài tiến] mở rộng) đổi mới (đòng 4 tr./ tr37) và nên dẫn nguồn [1]

và một số lời dịch tiếng Việt khác nhau từ thuật ngữ/từ viết tắt gốc.

Ví dụ:

RBF ANN là ....ANN ... dùng hàn cơ sở bán kính [xuyên tâm] .....

ARIMA là, mô hình ..... tự hồi qui .....kết hợp giá trị trung bình động

[3] Cần kèm theo sản phẩm Demo / Python: ArimaRbsnn.py và giải thích rõ QT Designer là công cụ liên quan (hay một phần sản phẩm của đề tài).

[4] Nên diễn đạt rõ hơn mục 2.1.1. / Tr20;.... được ghi nhận thực tế theo [[là] thời gian. Hình 2.1: minh họa Chuỗi thời gian (như đã trình bày rất rõ bên dưới)

[5] mục TLTK [15]/ tr74 cần có lời giải thích kế sau là "nguồn tập DL thực nghiệm"

## **II. CÁC VẤN ĐỀ CẦN LÀM RỎ**

(Các câu hỏi của giảng viên phản biện)

1. Xin tác giả làm rõ về Exponential Smoothing (làm mịn theo cấp số nhân) trong mục 2.3.2 / trang 23 liên quan như thế nào với mô hình lai ghép được dùng trong Đề tài? Có phải là xử lý chuỗi thời gian tuyến tính và phi tuyến trong tiền xử lý số liệu trước khi nạp vào mô hình lai của bài toán đề tài hay không? Và nếu như vậy, thì có thể dùng Tool của Excel được không?

## **III. ĐÁNH GIÁ**

TT	Mục đánh giá	Đánh giá	
		Đạt	Không đạt
1	Tinh khoa học, rõ ràng, mạch lạc, khúc chiết trong luận văn.	x	
2	Đánh giá việc sử dụng hoặc trích dẫn kết quả NC của người khác có đúng qui định hiện hành của pháp luật sở hữu trí tuệ.	x	
3	Mục tiêu nghiên cứu, phương pháp nghiên cứu sử dụng trong LVTN.	x	

4	Tổng quan của đề tài.	X	
5	Đánh giá về nội dung & chất lượng của LVTN.	X	
6	Đánh giá về khả năng ứng dụng, giá trị thực tiễn của đề tài.	X	

Danh dấu chéo (x) vào ô muốn Đánh giá

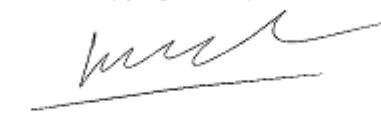
#### IV. KẾT LUẬN

(Giảng viên phản biện ghi rõ ý kiến "Tán thành luận văn" hay "Không tán thành luận văn")

- Xin được đánh giá cao báo cáo của luận văn. - Báo cáo đề tài đạt yêu cầu của một luận văn bậc thạc sĩ. Tuy nhiên đề xuất tác giả nên chỉnh sửa một số nội dung theo góp ý để đề tài có giá trị cao hơn.
- Đồng ý cho bảo vệ.

TP Hồ Chí Minh, ngày 22 tháng 4 năm 2021

**Người nhận xét**  
(Ký & ghi rõ họ tên)



TS.GVC. Võ Xuân Thế

## LÝ LỊCH KHOA HỌC

## I. LÝ LỊCH SƠ LUỐC:

Họ và tên: PHẠM CHÍ CÔNG Giới tính: Nam  
Ngày, tháng, năm sinh: 28/12/1978 Nơi sinh: Đồng Nai  
Quê quán: Ninh Bình Dân tộc: Kinh  
Chỗ ở riêng hoặc địa chỉ liên lạc: 63 Yên Đỗ, Tân Thành, Tân Phú, TPHCM  
Điện thoại cơ quan: 0909821266 Điện thoại nhà riêng: 0283 8121533  
Fax: E-mail: phamchicong78@gmail.com

## **II. QUÁ TRÌNH ĐÀO TẠO:**

## 1. Trung học chuyên nghiệp:

Hệ đào tạo: Thời gian đào tạo từ ...../..... đến ...../  
Nơi học (trường, thành phố):  
Ngành học:

## 2. Đại học:

### **III. QUÁ TRÌNH CÔNG TÁC CHUYÊN MÔN KỂ TỪ KHI TỐT NGHIỆP ĐẠI HỌC:**

Thời gian	Nơi công tác	Công việc đảm nhiệm
2010 đến nay	Viện NCPT Văn Hóa và Giáo Dục Đông Nam Á	P. Viện trưởng

## LỜI CAM ĐOAN

Tôi cam đoan đây là công trình nghiên cứu của tôi.

Các số liệu, kết quả nêu trong luận văn là trung thực và chưa từng được ai công bố trong bất kỳ công trình nào khác

*Tp. Hồ Chí Minh, ngày ... tháng 4 năm 2021*

(Ký tên và ghi rõ họ tên)

Phạm Chí Công

## TÓM TẮT

Hiện nay trong khai phá dữ liệu lớn được quan tâm nhiều nhất trong lĩnh vực khoa học dữ liệu, một trong những vấn đề được đặt lên hàng đầu là những bài toán về dự báo, hầu hết trong các lĩnh vực hoạt động xã hội hiện nay thì vấn đề dự báo đóng góp một phần không nhỏ trong sự tồn tại và phát triển. Người ta đưa ra rất nhiều các kỹ thuật trong khai phá dữ liệu để dự báo, nhưng Bài toán dự báo sử dụng chuỗi số thời gian luôn là một đề tài “nóng” luôn được quan tâm.

Các nhà nghiên cứu đã đưa ra rất nhiều các phương pháp nhằm sử dụng nguồn dữ liệu lớn hiện nay để phục vụ cho các vấn đề về dự báo. Trong luận văn này, chúng tôi cũng đi vào nghiên cứu các phương pháp dự báo sao cho cải thiện được kết quả so với các mô hình dự báo khác. Dựa trên mô hình dự báo ARIMA và RBFNN, L.Zhang và cộng sự [6] đã đưa ra mô hình dự báo lai ARIMA-RBFNN và đã cho kết quả dự báo tốt hơn khi thực hiện từng mô hình. Tuy nhiên, lượng dữ liệu ngày càng lớn nên thời gian thực thi của mô hình sẽ lâu hơn. Do đó, việc cải tiến mô hình ARIMA-RBFNN để thời gian thực thi nhanh hơn là một vấn đề cần quan tâm.

Trong đề tài này, chúng tôi cải tiến mô hình ARIMA-RBFNN đã được L.Zhang và cộng sự giới thiệu, nhằm mục đích cải thiện thời gian thực thi và kết quả của dự báo tốt hơn.

## **ABSTRACT**

Today, big data mining is most interested in data science. One of the top issues is the problem of prediction. Mostly, in the social activities today, the problem of forecasting plays a significant part in the existence and development. Many techniques are given in data mining for prediction, but the forecasting problem using time series is a "hot" topic that is always interested.

Researchers have come up with a variety of methods to use current large data sources to serve the problems of forecasting. In this thesis, we also study prediction methods to improve the results compared to other predictive models. Based on prediction models ARIMA and RBFNN, L.Zhang et al [6] gave the ARIMA-RBFNN hybrid prediction model and gave better predictive results when performing each model. However, the amount of data is increasing, so the execution time of the model will be longer. Therefore, improving the ARIMA-RBFNN model for faster execution time is a matter of concern.

In this study, we refine the ARIMA-RBFNN model introduced by L.Zhang et al in order to improve the execution time and the results of forecasts better.

# MỤC LỤC

QUYẾT ĐỊNH GIAO ĐỀ TÀI .....	i
LÝ LỊCH KHOA HỌC .....	ix
LỜI CAM ĐOAN .....	x
TÓM TẮT .....	xi
MỤC LỤC.....	xiii
DANH MỤC HÌNH .....	xv
Chương 1: TỔNG QUAN .....	1
1.1. Tính cấp thiết của đề tài .....	1
1.2. Một số các công trình nghiên cứu liên quan .....	1
1.3. Mục đích nghiên cứu, khách thể và đối tượng nghiên cứu của đề tài.....	5
1.4. Nhiệm vụ nghiên cứu và giới hạn .....	6
1.5. Phương pháp nghiên cứu.....	6
1.6. Ý nghĩa thực tiễn của đề tài.....	6
Chương 2: CƠ SỞ LÝ THUYẾT VỀ CHUỖI THỜI GIAN VÀ CÁC MÔ HÌNH DỰ BÁO .....	7
2.1. Chuỗi thời gian ( <i>time series</i> ) và các khái niệm liên quan .....	7
2.1.1. Khái niệm chuỗi thời gian.....	7
2.1.2. Đặc điểm chuỗi thời gian .....	8
2.1.3. Các phương pháp hiển thị chuỗi thời gian .....	10
2.2. Bài toán dự báo chuỗi thời gian và các mô hình dùng trong dự báo chuỗi thời gian.....	12
2.2.1. Các bài toán về dự báo chuỗi thời gian.....	12
2.2.2. Các mô hình dùng trong dự báo chuỗi thời gian.....	13
2.3. Các mô hình lai ghép dùng trong dự báo chuỗi thời gian.....	22
2.3.1. Mô hình ARIMA và ANN .....	22
2.3.2. Mô hình Exponential Smoothing và ANN.....	23
Chương 3: DỰ BÁO TRÊN CHUỖI THỜI GIAN SỬ DỤNG MÔ HÌNH LAI GHÉP ARIMA VÀ RBFNN .....	26

3.1. Mô hình mạng Nơ-ron nhân tạo RBF (Radial Basis Function Neural Network - RBFNN) .....	26
3.2. Mô hình tự hồi quy kết hợp với trung bình di động ARIMA(p,d,q) (AutoRegressive Integrated Moving Average) .....	27
3.2.1. Sai phân I(d).....	27
3.2.2. Mùa vụ (S).....	28
3.2.3. Mô hình ARIMA(p,d,q) .....	28
3.2.4. Phương pháp ước lượng tham số .....	30
3.3. Mô hình lai ghép giữa ARIMA và RBFNN cho bài toán dự báo trên chuỗi thời gian.....	37
3.4. Nghiên cứu cải tiến mô hình lai ghép bằng cách thực hiện song song hai mô hình ARIMA và RBFNN .....	37
<b>Chương 4: ĐÁNH GIÁ THỰC NGHIỆM .....</b>	<b>40</b>
4.1. Môi trường và dữ liệu thực nghiệm .....	40
4.2. Tiêu chí đánh giá.....	41
4.3. Các trường hợp thực nghiệm.....	42
4.3.1. Trường hợp 1: Cố định số nút đầu vào và thay đổi số nút ẩn .....	42
4.3.2. Trường hợp 2: Cố định số nút ẩn và thay đổi số nút đầu vào .....	57
4.4. Nhận xét kết quả thực nghiệm ở các tập dữ liệu.....	69
<b>Chương 5: KẾT LUẬN VÀ HƯỚNG PHÁT TRIỂN .....</b>	<b>71</b>
5.1. Kết quả đạt được .....	71
5.2 Các mặt hạn chế .....	71
5.3. Hướng phát triển .....	71
<b>TÀI LIỆU THAM KHẢO .....</b>	<b>73</b>
<b>PHỤ LỤC .....</b>	<b>75</b>

## DANH MỤC TỪ VIẾT TẮT VÀ THUẬT NGỮ

STT	Từ viết tắt- Thuật ngữ	Tiếng Việt	Tiếng Anh
1	FIR	Mạng đáp ứng xung hữu hạn	Finite Impulse Response
2	AR	Mô hình tự hồi quy	Autoregressive model
3	MA	Mô hình trung bình di động	Moving Average Mode
4	ARIMA	Mô hình tự hồi quy kết hợp Mô hình trung bình di động	
5	ANN	Mô hình mạng Nơ-ron nhân tạo	Artificial Neural Network
6	ETS	Mô hình làm mịn theo cấp số nhân	Exponential Smoothing
7	RBFNN	Mạng Neural nhân tạo truyền thẳng	Radial Basis Function Neural Network
8	LSTM	Bộ nhớ ngắn dài hạn: là một mạng thần kinh hồi quy nhân tạo	Long Short-Term Memory
9	ACF	Hàm tự tương quan	AutoCorrelation Function
10	PACF	Hàm tự tương quan từng phần	Partial AutoCorrelation Function
11	EWMA	Trung bình di động có trọng số theo mũ	Exponentially Weighted Moving Average
12	QT Designer:	Phần mềm thiết kế giao diện cho Demo	

# DANH MỤC HÌNH

Hình 2.1: Dữ liệu chuỗi thời gian .....	10
Hình 2.2: Chuỗi thời gian dừng .....	11
Hình 2.3: Xu hướng tăng theo thời gian .....	12
Hình 2.4: Xu hướng thay đổi theo mùa .....	12
Hình 2.5: Xu hướng thay đổi theo chu kỳ .....	13
Hình 2.6: Đồ thị của $x_t$ theo t .....	14
Hình 2.7: Đồ thị của $x_t$ theo t .....	15
Hình 2.8: Đồ thị $(x_t - x_{t-1})$ theo t .....	15
Hình 2.9: Đồ thị hồi quy đơn giản .....	17
Hình 2.10: Đồ thị hồi quy đơn giản .....	19
Hình 2.11: Kiến trúc của một ANN cho dự báo chuỗi thời gian với 3 ngõ vào, một lớp ẩn hai nơ-ron và một ngõ ra (là giá trị dự báo) .....	24
Hình 2.12: Mô hình lai ARIMA – ANN .....	26
Hình 2.13: Mô hình kết hợp ARIMA – ANN .....	26
Hình 2.14: Mô hình lai ETS – ANN .....	8
Hình 3.1: Mô hình mạng RBF .....	29
Hình 3.2: Sơ đồ mô phỏng mô hình ARIMA .....	33
Hình 3.3: Sơ đồ mô phỏng mô hình lai ARIMA – RBFNN .....	40
Hình 4.1: Biểu đồ thể hiện dữ liệu của chuỗi thời gian AirPassengers .....	41
Hình 4.2: Biểu đồ thể hiện kết quả dự báo của các mô hình khi dùng 64 nút ẩn ....	42
Hình 4.3: Thời gian thực thi của các mô hình khi dùng 64 nút ẩn .....	43
Hình 4.4: Biểu đồ thể hiện kết quả dự báo của các mô hình khi dùng 128 nút ẩn ..	44
Hình 4.5: Thời gian thực thi của các mô hình khi dùng 128 nút ẩn .....	45
Hình 4.6: Biểu đồ thể hiện dữ liệu của chuỗi thời gian Sunspots.csv .....	45
Hình 4.7: Biểu đồ thể hiện kết quả dự báo của các mô hình khi dùng 64 nút ẩn ....	46
Hình 4.8: Thời gian thực thi của các mô hình khi dùng 64 nút ẩn trên tập dữ liệu ..	47
Hình 4.9: Biểu đồ thể hiện kết quả dự báo của các mô hình khi dùng 128 nút ẩn ..	48

Hình 4.10: Thời gian thực thi của các mô hình dự báo khi dùng 128 nút ẩn trên tập Sunspots .....	49
Hình 4.11: Biểu đồ thể hiện dữ liệu của chuỗi thời gian Dentists .....	49
Hình 4.12: Biểu đồ thể hiện kết quả dự báo của các mô hình khi dùng 64 nút ẩn trên chuỗi thời gian Dentists .....	51
Hình 4.13: Thời gian thực thi của các mô hình dự báo khi dùng 64 nút ẩn trên tập dữ liệu Dentists .....	52
Hình 4.14: Biểu đồ thể hiện kết quả dự báo của các mô hình khi dùng 128 nút ẩn trên tập dữ liệu Dentists.....	53
Hình 4.15: Thời gian thực thi của các mô hình dự báo khi dùng 128 nút ẩn trên tập dữ liệu Dentists .....	54
Hình 4.16: Biểu đồ thể hiện dữ liệu của chuỗi thời gian City_temperature .....	54
Hình 4.17: Biểu đồ thể hiện kết quả dự báo của các mô hình khi dùng 64 nút ẩn trên tập dữ liệu City_temperature .....	55
Hình 4.18: Thời gian thực thi của các mô hình dự báo khi dùng 64 nút ẩn trên tập dữ liệu City_temperature .....	56
Hình 4.19: Biểu đồ thể hiện kết quả dự báo của các mô hình khi dùng 128 nút ẩn trên tập dữ liệu City_temperature .....	57
Hình 4.20: Thời gian thực thi của các mô hình dự báo khi dùng 128 nút ẩn trên tập dữ liệu City_temperature .....	58
Hình 4.21: Biểu đồ thể hiện kết quả dự báo của các mô hình khi dùng 32 nút đầu vào trên tập dữ liệu AirPassengers .....	59
Hình 4.22: Thời gian thực thi của các mô hình dự báo khi dùng 32 nút đầu vào trên tập dữ liệu AirPassengers .....	59
Hình 4.23: Biểu đồ thể hiện kết quả dự báo của các mô hình khi dùng 64 nút đầu vào trên tập dữ liệu AirPassengers .....	60
Hình 4.24: Thời gian thực thi của các mô hình dự báo khi dùng 64 nút đầu vào trên tập dữ liệu AirPassengers .....	61
Hình 4.25: Biểu đồ thể hiện kết quả dự báo của các mô hình khi dùng 32 nút đầu vào trên tập dữ liệu Sunspots .....	62

Hình 4.26: Thời gian thực thi của các mô hình dự báo khi dùng 32 nút đầu vào trên tập dữ liệu Sunspots .....	62
Hình 4.27: Biểu đồ thể hiện kết quả dự báo của các mô hình khi dùng 64 nút đầu vào trên tập dữ liệu Sunspots .....	63
Hình 4.28: Thời gian thực thi của các mô hình dự báo khi dùng 64 nút đầu vào trên tập dữ liệu Sunspots .....	63
Hình 4.28: Thời gian thực thi của các mô hình dự báo khi dùng 64 nút đầu vào trên tập dữ liệu Sunspots .....	64
Hình 4.29: Biểu đồ thể hiện kết quả dự báo của các mô hình khi dùng 32 nút đầu vào trên tập dữ liệu Dentists .....	65
Hình 4.30: Thời gian thực thi của các mô hình dự báo khi dùng 32 nút đầu vào trên tập dữ liệu Dentists .....	65
Hình 4.31: Biểu đồ thể hiện kết quả dự báo của các mô hình khi dùng 64 nút đầu vào trên tập dữ liệu Dentists .....	66
Hình 4.32: Thời gian thực thi của các mô hình dự báo khi dùng 64 nút đầu vào trên tập dữ liệu Dentists .....	67
Bảng 4.16: Kết quả thực nghiệm trên tập dữ liệu City_temperature với 32 nút đầu vào .....	67
Hình 4.33: Biểu đồ thể hiện kết quả dự báo của các mô hình khi dùng 32 nút đầu vào trên tập dữ liệu City_temperature .....	68
Hình 4.34: Thời gian thực thi của các mô hình dự báo khi dùng 32 nút đầu vào trên tập dữ liệu City_temperature .....	68
Hình 4.35: Biểu đồ thể hiện kết quả dự báo của các mô hình khi dùng 64 nút đầu vào trên tập dữ liệu City_temperature .....	69
Hình 4.36: Thời gian thực thi của các mô hình dự báo khi dùng 64 nút đầu vào trên tập dữ liệu City_temperature .....	70

## **DANH MỤC BẢNG**

Bảng 2.1: Dữ liệu chuỗi thời gian .....	14
Bảng 2.2: Bảng dữ liệu cân nặng và chiều cao của trẻ .....	17
Bảng 2.3: Bảng dữ liệu cân nặng và chiều cao của trẻ .....	18
Bảng 4.1: Các tập dữ liệu dùng trong thực nghiệm .....	40
Bảng 4.2: Kết quả thực nghiệm trên tập dữ liệu Aripassanger với 64 nút ẩn .....	41
Bảng 4.3: Kết quả thực nghiệm trên tập dữ liệu Aripassanger với 128 nút ẩn .....	43
Bảng 4.3: Kết quả thực nghiệm trên tập dữ liệu Sunspots.csv với 64 nút ẩn .....	46
Bảng 4.5: Kết quả thực nghiệm trên tập dữ liệu Sunspots.csv với 128 nút ẩn .....	47
Bảng 4.6: Kết quả thực nghiệm trên tập dữ liệu Dentists với 64 nút ẩn .....	50
Bảng 4.7: Kết quả thực nghiệm trên tập dữ liệu Dentists với 128 nút ẩn .....	52
Bảng 4.8: Kết quả thực nghiệm trên tập dữ liệu City_temperature với 64 nút ẩn .....	55
Bảng 4.9: Kết quả thực nghiệm trên tập dữ liệu City_temperature với 128 nút ẩn .....	56
Bảng 4.10: Thông kê sử dụng tài nguyên của các mô hình dự báo .....	59
Bảng 4.10: Kết quả thực nghiệm trên tập dữ liệu AirPassengers với 32 nút đầu vào .....	58
Bảng 4.11: Kết quả thực nghiệm trên tập dữ liệu AirPassengers với 64 nút đầu vào .....	60
Bảng 4.12: Kết quả thực nghiệm trên tập dữ liệu Sunspots với 32 nút đầu vào .....	61
Bảng 4.13: Kết quả thực nghiệm trên tập dữ liệu Sunspots với 64 nút đầu vào .....	63
Bảng 4.14: Kết quả thực nghiệm trên tập dữ liệu Dentists với 32 nút đầu vào .....	64
Bảng 4.15: Kết quả thực nghiệm trên tập dữ liệu Dentists với 64 nút đầu vào .....	66
Bảng 4.16: Kết quả thực nghiệm trên tập dữ liệu City_temperature với 32 nút đầu vào .....	67
Bảng 4.17: Kết quả thực nghiệm trên tập dữ liệu City_temperature với 64 nút đầu vào .....	69

# Chương 1

## TỔNG QUAN

### 1.1. Tính cấp thiết của đề tài

Sự phát triển mạnh mẽ của công nghệ và sự bùng nổ của thông tin số trong những năm gần đây, nó đã góp phần không nhỏ vào sự phát triển của xã hội. Với sự đa dạng và lượng dữ liệu khổng lồ là nguồn tài nguyên vô giá nếu chúng ta biết khai thác và sử dụng những thông tin có ích trong đó. Vấn đề đặt ra là khai thác và lưu trữ dữ liệu hiện nay như thế nào, Các phương pháp khai thác dữ liệu truyền thống ngày càng không phù hợp và không đáp ứng được nhu cầu thực tế. Do đó, các công nghệ khai phá dữ liệu mới ra đời đã cho phép chúng ta khai thác được những tri thức hữu dụng bằng cách trích xuất những thông tin có mối quan hệ hoặc có mối tương quan nhất định từ một kho dữ liệu lớn (Big Data), mà bình thường chúng ta không nhận diện và sử dụng được, từ đó chúng ta giải quyết được các bài toán tìm kiếm, dự báo các xu thế, các hành vi trong tương lai, và nhiều tính năng thông minh khác.

Một trong những vấn đề quan trọng nhất hiện nay trong khai phá dữ liệu lớn là những bài toán về dự báo, hầu hết trong các lĩnh vực hoạt động xã hội hiện nay thì vấn đề dự báo đóng góp một phần không nhỏ trong sự tồn tại và phát triển. Hiện nay người ta đưa ra rất nhiều các kỹ thuật trong khai phá dữ liệu để dự báo, nhưng Bài toán dự báo sử dụng chuỗi số thời gian luôn là một đề tài “nóng” luôn được quan tâm.

### 1.2. Một số các công trình nghiên cứu liên quan

- Nguyễn Chí Thành và Hà Gia Sơn (2017) [14]: Nghiên cứu “Kết hợp mạng Neron FIR và mô hình ARIMA theo hình thức động để nâng cao hiệu quả dự báo chuỗi thời gian”, tác giả đưa ra giải pháp kết hợp giữa kết quả dự báo của mạng neron FIR với mô hình ARIMA mà các trọng số sẽ thay đổi để thích nghi với sự biến đổi của chuỗi thời gian, nhằm đạt hiệu quả cao nhất. Đầu tiên, tác giả dùng dữ liệu mẫu để ước lượng các mô hình, sau đó, dự báo các giá trị của biến phụ thuộc, dùng các giá trị này để xây dựng tập các trọng số, tạo các giá trị dự báo ngoài mẫu từ các mô hình riêng biệt và sử dụng các trọng số đã tìm được. Nếu gọi  $Y_t$  là giá trị thực tại thời

điểm t của biến phụ thuộc, và  $f_{t1}, f_{t2}, \dots, f_{tk}$  là các giá trị dự báo được tạo ra bởi k mô hình khác nhau. Phương pháp đương nhiên là tạo ra giá trị trung bình có trọng số của các giá trị dự báo này. Từ đó, tác giả đưa ra giá trị dự báo kết hợp là:

$$f_t = \widehat{\beta}_0 + \widehat{\beta}_1 f_{t1} + \widehat{\beta}_2 f_{t2} + \dots + \widehat{\beta}_k f_{tk}$$

Trong đó:

$\widehat{\beta}_0 + \widehat{\beta}_1 + \widehat{\beta}_2 + \dots + \widehat{\beta}_k$  : là các trọng số cần xác định (Xác định các trọng số bằng phương pháp Uớc lượng tham số của mô hình hồi qui bội), thông thường các trọng số này luôn là một hằng số và sẽ thay đổi theo thời gian, nên tác giả đã sử dụng hàm bậc nhất trong phần ứng dụng để dự báo. Giả định rằng trong mô hình trên  $\beta_i = \alpha_{i0} + \alpha_{i1}t$ , với t thể hiện thời gian từ 1 đến n, và  $i = 0, 1, \dots, k$  ( $k$  là các mô hình phối hợp). Điều này dẫn đến mô hình cải biến như sau:

$$Y_t = \alpha_{00} + \alpha_{01}t + \alpha_{10}f_{t1} + \alpha_{11}(tf_{t1}) + \dots + \alpha_{k0}f_{tk} + \alpha_{k1}(tf_{tk}) + u_t$$

$$\text{Đặt: } \alpha_{00} + \alpha_{01}t = A_0, \alpha_{10} = A_1 \dots \alpha_{1k} = A_k, \alpha_{11} = A_{k+1} \dots \alpha_{k1} = A_{2k}$$

$$f_{t1} = F_1, \dots f_{tk} = F_k, (tf_{t1}) = F_{k+1}, \dots (tf_{tk}) = F_{2k}$$

Ta có phương trình:

$$Y_t = A_0 + A_1F_1 + A_2F_2 + \dots + A_kF_k + A_{k+1}F_{k+1} + \dots + A_{2k}F_{2k}$$

$$\text{Đặt: } n = 2k, \text{ ta sẽ có phương trình } Y_t = A_0 + A_1F_1 + \dots + A_nF_n$$

Đây chính là phương trình hồi qui cơ bản, có thể dùng giải thuật để xác định các hệ số  $A_0, A_1, \dots, A_n$  này.

- Lâm Hoàng Vũ (2012) [13]: “Dự báo chuỗi thời gian sử dụng mô hình ARIMA và giải thuật di truyền”. Tác giả đưa ra một phương pháp để tự động xác định bậc và ước lượng các hệ số của mô hình ARMA, tác giả đề xuất một phương pháp mở rộng không gian tìm kiếm các lời giải của mô hình ARMA dựa trên giải thuật tìm kiếm Tabu trong việc xác định bậc. Kết quả thực nghiệm cho thấy phương pháp mới này đem lại kết quả tốt hơn đối với hầu hết các chuỗi dữ liệu được kiểm tra so với các phương pháp meta-heuristic khác và thời gian chạy dừng ở mức có thể chấp nhận được.

- L. Zhang, G. X. Zhang, and R. R. Li (2018) [6]: “Phân tích và dự đoán chất lượng nước bằng cách sử dụng mô hình lai ARIMA và mạng Nơron RBF”. Nhóm tác giả đã phân tích biến động chất lượng nước ở hồ Chagan trong một khoảng thời gian

bằng cách tập trung vào các cấp độ TN và TP; Phát triển mô hình lai ARIMA và RBFNN để dự đoán dữ liệu chuỗi thời gian chất lượng nước; và đánh giá hiệu suất của các mô hình này bằng cách so sánh dữ liệu được quan sát và dự đoán, từ đó đánh giá hiệu suất dự đoán của mô hình lai ARIMA và RBFNN so với mô hình ARIMA.

Kết quả dự đoán từ các mô hình lai ARIMA - RBFNN được biểu thị như sau:

$$\hat{y}_t = \hat{L}_t + \hat{N}_t$$

Trong đó:

-  $\hat{L}_t$  là kết quả dự báo của mô hình ARIMA

-  $\hat{N}_t$  là kết quả dự báo của mô hình RBFNN

So với mô hình ARIMA, mô hình dự báo lai ARIMA-RBFNN được đề xuất mô tả toàn diện hơn và chính xác, tạo ra các giá trị tương ứng so với các giá trị được tạo bằng các mô hình ARIMA và MAPE. Các mô hình lai cải thiện độ chính xác dự đoán. Do đó, mô hình lai được đề xuất có thể được sử dụng để dự đoán chuỗi thời gian TN và TP cho hồ Chagan.

- Li Wang, Haofei Zou, Jia Su, Ling Li and Sohail Chaudhry (2013) [8]: “Mô hình lai ARIMA-ANN trong dự báo chuỗi thời gian”. Tác giả đề xuất một mô hình lai, đặc biệt trong việc tích hợp các lợi thế của ARIMA và ANN trong việc mô hình hóa các hành vi tuyến tính và phi tuyến trong tập dữ liệu. Vì mô hình ARIMA không xử lý được các thành phần phi tuyến của chuỗi thời gian, mô hình ANN được sử dụng để giải quyết các thành phần phi tuyến của dữ liệu chuỗi thời gian vì chúng có nhiều tế bào nơ ron phi tuyến tương tác trong nhiều lớp. Vì vậy, kết hợp ANN và ARIMA trong dự báo chuỗi thời gian để đối phó với tất cả các thành phần không đồng nhất của các mẫu cơ bản, tác giả sử dụng mô hình cộng ( $L+N$ ) và mô hình nhân ( $L^*N$ ) để kết hợp hai mô hình trong phân tích chuỗi thời gian. Các biểu thức toán học cho hai trường hợp này được thể hiện bằng phương trình:

$$\text{Mô hình cộng: } y_t = L_t + N_t$$

$$\text{Mô hình nhân: } y_t = L_t * N_t$$

Trong đó  $L_t$  đại diện cho thành phần tuyến tính và  $N_t$  là thành phần phi tuyến

Mô hình lai đã được thử nghiệm trên ba bộ dữ liệu thực tế, cụ thể là dữ liệu vết đen mặt trời của Wolf, dữ liệu lynx của Canada và dữ liệu giá cổ phiếu của IBM.

Kết quả cho thấy hiệu quả của mô hình tổ hợp mới trong việc thu được dự báo chính xác hơn so với các mô hình hiện có.

- N. Vijay and G.C. Mishra (2018) [4]: “Thực nghiệm 2 mô hình ARIMA và ANN trong dự báo chuỗi thời gian”. Mục tiêu của nghiên cứu là kiểm tra tính linh hoạt của mô hình mạng nơ ron nhân tạo (ANN) trong dự báo chuỗi thời gian bằng cách so sánh với mô hình ARIMA. Dữ liệu bao gồm diện tích và thời gian sản xuất ngọc trai (bajra) từ 1955-1956 đến 2014-2015 đã được sử dụng trong nghiên cứu để chứng minh tính hiệu quả của mô hình. Các thử nghiệm cho thấy mô hình ANN vượt trội hơn Mô hình ARIMA dựa trên giá trị trung bình gốc (RMSE), MAPE và MSE.

- Haviluddina, Ahmad Jawahirb (2015) [3]: “So sánh mô hình ARIMA và RBFNN trong dự báo ngắn hạn”. Mô hình đề xuất đã được kiểm tra bằng cách sử dụng dữ liệu chuỗi thời gian mô phỏng của khách du lịch đến Indonesia gần đây được xuất bản bởi BPS Indonesia. Kết quả chứng minh rằng mô hình RBFNN tốt hơn mô hình ARIMA được biểu thị bằng giá trị lỗi bình phương trung bình (MSE). Dựa trên kết quả thu được, mô hình RBFNN được khuyến nghị thay thế cho mô hình hiện có vì nó có cấu trúc đơn giản và hợp lý hơn trong dự báo.

- Sibarama Panigrahi, H.S.Behera (2017) [1]: “Sử dụng mô hình lai ETS - ANN để dự báo chuỗi thời gian”. Tác giả đưa phương pháp lai mới bằng cách kết hợp các mô hình làm mịn theo cấp số nhân tuyến tính và phi tuyến (ETS) với ANN, cả hai mô hình ETS và ANN đều có khả năng xử lý dữ liệu tuyến tính cũng như phi tuyến. Tuy nhiên, ANN không thể xử lý các mẫu tuyến tính tốt như các mẫu phi tuyến. Kết quả dự đoán cuối cùng có được bằng cách kết hợp các dự đoán ETS với dự đoán ANN.

Tác giả sử dụng mười sáu bộ dữ liệu chuỗi thời gian để phân tích hiệu suất và so sánh các phương pháp được đề xuất với ARIMA, ETS, MLP và một số mô hình lai ANN ARIMA phiên bản hiện có. Kết quả thử nghiệm cho thấy mô hình lai ETS - ANN được đề xuất cho kết quả tốt hơn về mặt thống kê trên các bộ dữ liệu được sử dụng.

- Sima Siami-Namini, Neda Tavakoli, Akbar Siami Namin (2018) [10]: “So sánh mô hình ARIMA và LSTM trong dự báo chuỗi thời gian”. Tác giả sử dụng các

thuật toán "Long Short-Term Memory (LSTM)" dựa trên học tập sâu mới được phát triển để dự báo dữ liệu chuỗi thời gian, các nghiên cứu thực nghiệm được thực hiện và báo cáo cho thấy các thuật toán dựa trên học tập sâu như LSTM vượt trội hơn các thuật toán dựa trên truyền thống như mô hình ARIMA. Cụ thể hơn, mức giảm trung bình của tỷ lệ lỗi thu được từ LSTM là từ 84 - 87% khi so sánh với ARIMA cho thấy sự vượt trội của LSTM so với ARIMA. Hơn nữa, người ta nhận thấy rằng số lần huấn luyện, được gọi là "era" trong học tập sâu, không ảnh hưởng đến hiệu suất của mô hình dự báo được huấn luyện và được thực hiện ngẫu nhiên.

- Zonghan Wu, Shirui Pan, Guodong Long, Jing Jiang, Xiaojun Chang, Chengqi Zhang (2020) [11]: “Dự báo chuỗi thời gian đa biến với mạng nơ ron đồ thị”. Nhóm tác giả cho rằng dự báo chuỗi thời gian đa biến là các biến của nó phụ thuộc vào nhau, nhưng khi xem xét kỹ, có thể nói rằng các phương thức hiện tại không khai thác triệt để sự phụ thuộc không gian tiềm ẩn giữa các cặp biến, trong khi đó các mạng nơ ron đồ thị (GNNs) đã cho thấy khả năng cao trong việc xử lý các phụ thuộc quan hệ. GNN yêu cầu cấu trúc biểu đồ được xác định rõ để truyền thông tin, có nghĩa là chúng không thể được áp dụng trực tiếp cho chuỗi thời gian đa biến trong đó các phụ thuộc không được biết trước. Tác giả đề xuất một khung mạng nơ ron đồ thị chung được thiết kế dành riêng cho dữ liệu chuỗi thời gian đa biến. Cách tiếp cận của tác giả tự động trích xuất các mối quan hệ đơn hướng giữa các biến thông qua một mô-đun “học” đồ thị, trong đó kiến thức bên ngoài như các thuộc tính biến có thể được tích hợp dễ dàng. Một lớp lan truyền mix-hop mới lạ và lớp khởi động giãn được tiếp tục để xuất để nắm bắt các phụ thuộc không gian và thời gian trong chuỗi thời gian. Các mô-đun học đồ thị, tích chập đồ thị và các mô-đun chập theo thời gian được học cùng nhau trong một khung kết thúc. Kết quả thử nghiệm cho thấy mô hình đề xuất vượt trội so với các phương pháp cơ bản tiên tiến trên 3 trong 4 bộ dữ liệu điểm chuẩn và đạt được hiệu suất ngang bằng với các phương pháp khác trên hai bộ dữ liệu thông tin cấu trúc.

### **1.3. Mục đích nghiên cứu, khách thể và đối tượng nghiên cứu của đề tài**

Nghiên cứu ứng dụng các mô hình lai ghép vào bài toán dự báo trên chỗi thời gian nhằm nâng cao tính hiệu quả trong bài toán dự báo.

## **1.4. Nhiệm vụ nghiên cứu và giới hạn**

### **Nhiệm vụ:**

- Nghiên cứu về chuỗi thời gian và bài toán dự báo trên chuỗi thời gian.
- Nghiên cứu mô hình lai ARIMA và RBFNN ứng dụng trong bài toán dự báo chuỗi thời gian.
  - Nghiên cứu cải tiến mô hình lai ARIMA và RBFNN
  - Đánh giá bằng thực nghiệm mô hình lai ARIMA và RBFNN cải tiến

**Giới hạn:** Chuỗi thời gian và bài toán dự báo trên chuỗi thời gian

## **1.5. Phương pháp nghiên cứu**

Kết hợp các phương pháp: nghiên cứu lý thuyết, lập trình mô phỏng và đánh giá bằng thực nghiệm.

## **1.6. Ý nghĩa thực tiễn của đề tài**

Hiện nay các bài toán về dự báo trên chuỗi thời gian được áp dụng rất rộng, cùng với sự phát triển của công nghệ nên việc khai phá dữ liệu lớn trở nên rất phổ biến, có rất nhiều mô hình đã sử dụng để dự báo trên chuỗi thời gian. Tuy nhiên việc dự báo nhằm mục đích hỗ trợ cho việc ra quyết định và hoạch định cho các hoạt động của các đơn vị có sử dụng chuỗi thời gian, nên việc cải tiến các mô hình dự báo có sẵn với mục đích rút ngắn thời gian và tăng độ chính xác cho các kết quả dự báo là vấn đề cần thiết và luôn được quan tâm thực hiện.

Phần tiếp theo của luận văn sẽ trình bày lý thuyết về chuỗi thời gian và các mô hình thường dùng dự báo chuỗi thời gian trong chương 2; sử dụng mô hình lai ghép ARIMA và RBFNN có cải tiến để dự báo chuỗi thời gian trong chương 3, kết quả đánh giá thực nghiệm sẽ được trình bày trong chương 4.

# Chương 2

## CƠ SỞ LÝ THUYẾT VỀ CHUỖI THỜI GIAN VÀ CÁC MÔ HÌNH DỰ BÁO

Nội dung được trình bày trong chương này gồm: các khái niệm về chuỗi thời gian; giới thiệu một số mô hình dự báo thường sử dụng trên chuỗi thời gian; kết hợp các mô hình để cải thiện kết quả dự báo.

### 2.1. Chuỗi thời gian (*time series*) và các khái niệm liên quan

Trong các bài toán dự báo, dữ liệu thường được biểu diễn dưới dạng chuỗi thời gian. Trong các dạng dữ liệu được phân tích thì dữ liệu chuỗi thời gian luôn được ưu tiên và phổ biến hơn các loại khác.

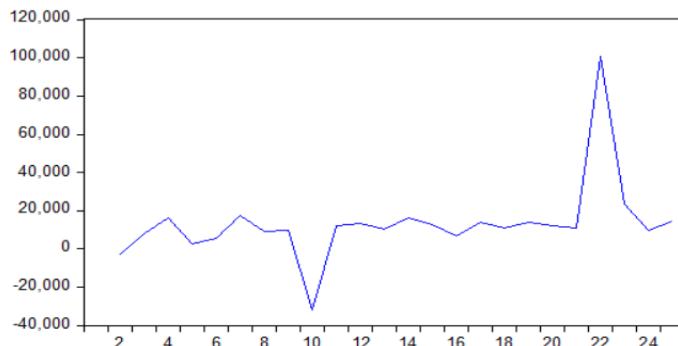
#### 2.1.1. Khái niệm chuỗi thời gian

Theo [2] [10], dữ liệu thời gian thực hay chuỗi thời gian là một chuỗi các giá trị của một đại lượng nào đó được ghi nhận thực tế theo thời gian. Hình 2.1: Minh họa chuỗi thời gian.

Là chuỗi các điểm dữ liệu được đo theo từng khoảng thời gian liền nhau, khoảng cách giữa các lần đo là bằng nhau.

Chuỗi thời gian là dãy các quan sát về một biến số nào đó theo thời gian. Mẫu quan sát có thể xem như một đoạn hữu hạn của một chuỗi vô hạn quan sát.

Ví dụ: Lượng khách hàng tăng, giảm của một công ty trong một khoảng thời gian (2015 - 2020), hay số lượng sản phẩm bán ra của công ty trong khoảng thời gian nào đó. Từ những dữ liệu này người ta dùng để dự báo cho tương lai.



Hình 2.1: Minh họa dữ liệu chuỗi thời gian

**Chuỗi thời gian dừng:** Chuỗi thời gian được coi là dừng nếu như trung bình và phương sai của nó không đổi theo thời gian và giá trị của đồng phương sai giữa hai thời đoạn chỉ phụ thuộc vào khoảng cách và độ trễ về thời gian giữa hai thời đoạn này chứ không phụ thuộc vào thời điểm thực tế mà đồng phương sai được tính. Hình 2.2: Chuỗi thời gian dừng.

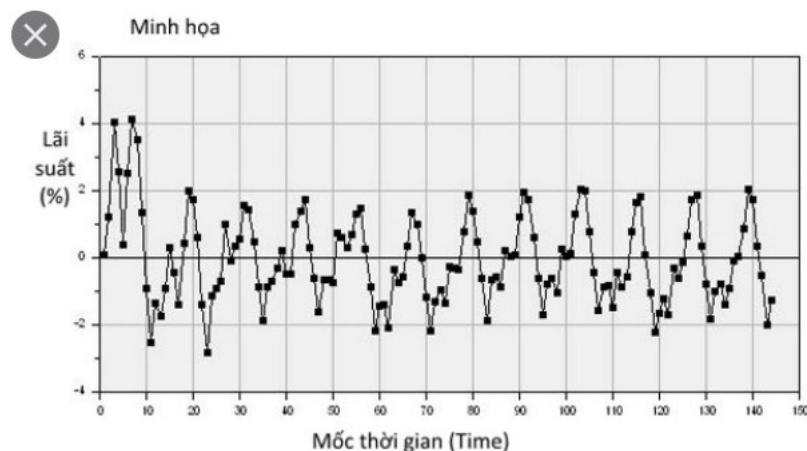
Trung bình:  $E(Y_t) = \mu$

Phương sai:  $Var(Y_t) = E(Y_t - \mu)^2 = \delta^2$

Đồng phương sai:  $\gamma_k = E[(Y_t - \mu)(Y_{t+k} - \mu)]$

Ta dịch chuyển chuỗi  $Y$  ban đầu từ  $Y_t$  đến  $Y_{t+m}$  và nếu  $Y_t$  là dừng, thì trung bình, phương sai và các tự đồng phương sai của  $Y_{t+m}$  phải đúng bằng trung bình, phương sai và các tự đồng phương sai của  $Y_t$ . Tóm lại, nếu một chuỗi thời gian là dừng thì trung bình, phương sai và tự đồng phương sai (tại các độ trễ khác nhau) sẽ giữ nguyên không đổi dù cho chúng được xác định vào thời điểm nào.

Với hầu hết các phương pháp thống kê dự báo, ta đều phải đảm bảo tính dừng của chuỗi dữ liệu vì thế việc kiểm tra tính dừng là rất quan trọng.

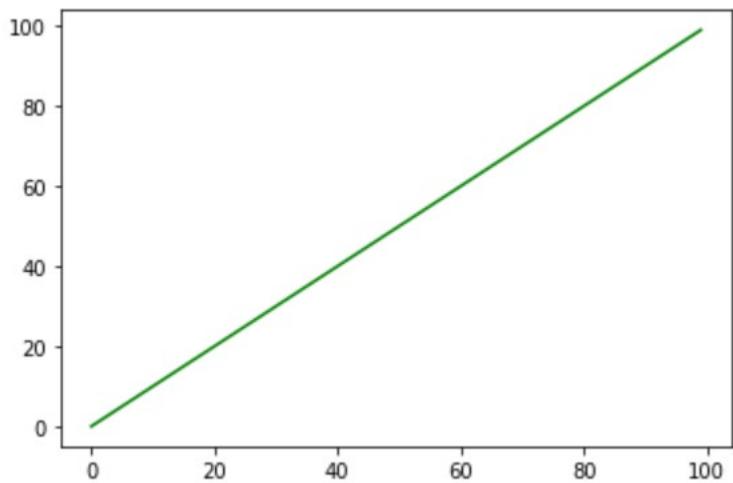


Hình 2.2: Chuỗi thời gian dừng

### 2.1.2. Đặc điểm chuỗi thời gian

#### 2.1.2.1. Xu hướng thay đổi dài hạn

Thành phần này dùng để chỉ xu hướng tăng hay giảm của đại lượng X trong thời gian dài. Về mặt đồ thị thành phần này có thể biểu diễn bởi một đường thẳng hay một đường cong trơn. [5]. Hình 2.3: Xu hướng tăng theo thời gian

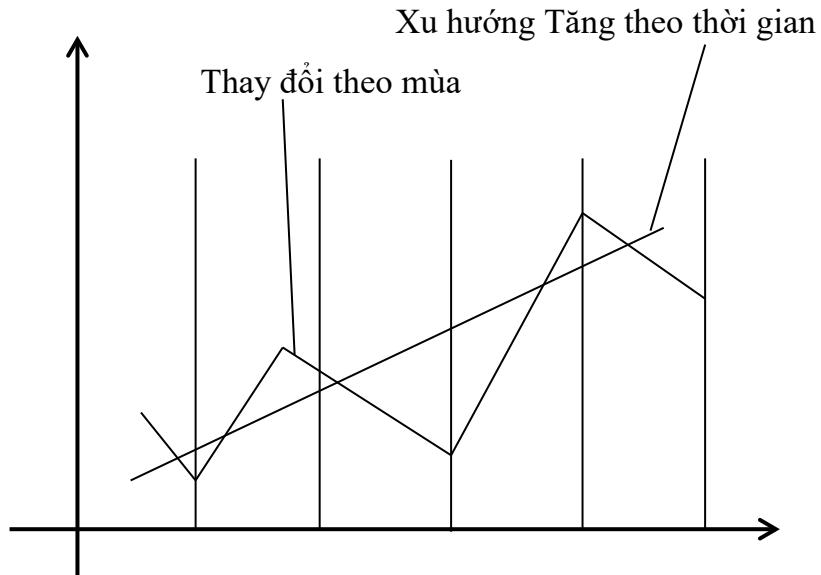


Hình 2.3: Xu hướng tăng theo thời gian

#### 2.1.2.2. Xu hướng thay đổi theo mùa

Thành phần này dùng để chỉ xu hướng tăng hay giảm của đại lượng X tính theo mùa trong năm. Hình 2.2: Xu hướng thay đổi theo mùa

Ví dụ: Giá vé máy bay sẽ tăng vào mùa hè, hoặc sản lượng Bia sẽ tăng vào mùa hè và giảm vào mùa đông.

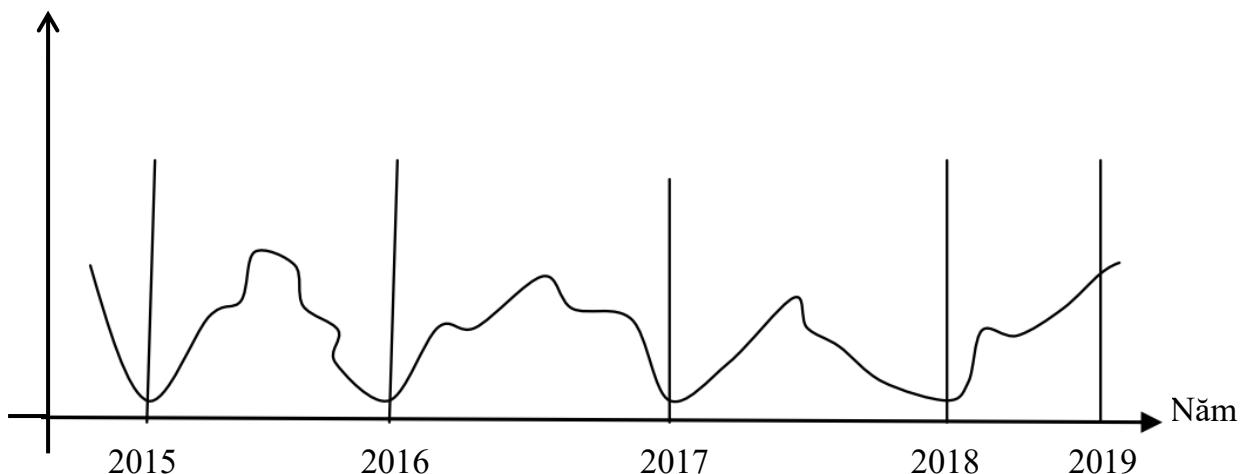


Hình 2.4: Xu hướng thay đổi theo mùa

### 2.1.2.3. Xu hướng thay đổi theo chu kỳ

Xu hướng này chỉ sự thay đổi của đại lượng X theo một chu kỳ. Xu hướng này khác xu hướng theo mùa ở chỗ chu kỳ của đại lượng X kéo dài hơn (thường tính hàng năm). Để đánh giá xu hướng này các giá trị của chuỗi thời gian được quan sát hàng năm. Hình 2.5: Xu hướng thay đổi theo chu kỳ

Ví dụ: Chu kỳ thay đổi thị hiếu sản phẩm của khách hàng



Hình 2.5: Xu hướng thay đổi theo chu kỳ

### 2.1.2.4. Xu hướng thay đổi bất thường

Xu hướng thay đổi này dùng để chỉ sự thay đổi bất thường của các giá trị trong chuỗi thời gian. Sự thay đổi này không thể dự đoán bằng các số liệu kinh nghiệm trong quá khứ, về mặt bản chất nó không có tính chu kỳ.

## 2.1.3. Các phương pháp hiển thị chuỗi thời gian

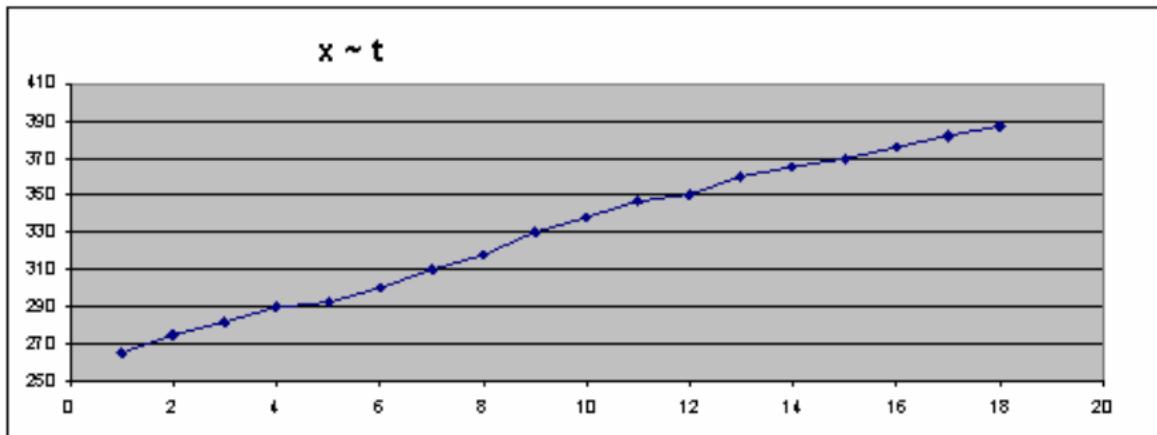
Phân tích chuỗi thời gian bao gồm việc nghiên cứu dạng dữ liệu trong quá khứ và giải thích các đặc điểm chính của nó. Một trong các phương pháp đơn giản và hiệu quả nhất là hiển thị trực quan chuỗi đó. Các đặc điểm không dễ thấy trong bảng dữ liệu thường nổi lên qua các minh họa đồ thị

Bảng 2.1: Dữ liệu chuỗi thời gian

$t$	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18
$x_t$	265	275	282	290	292	300	310	318	330	338	347	350	360	365	370	376	382	387
$x_t/x_{t-1}$		104	103	103	101	103	103	103	104	102	103	101	103	101	101	102	102	101
$x_t - x_{t-1}$		10	7	8	2	8	10	8	12	8	9	3	10	5	5	6	6	5

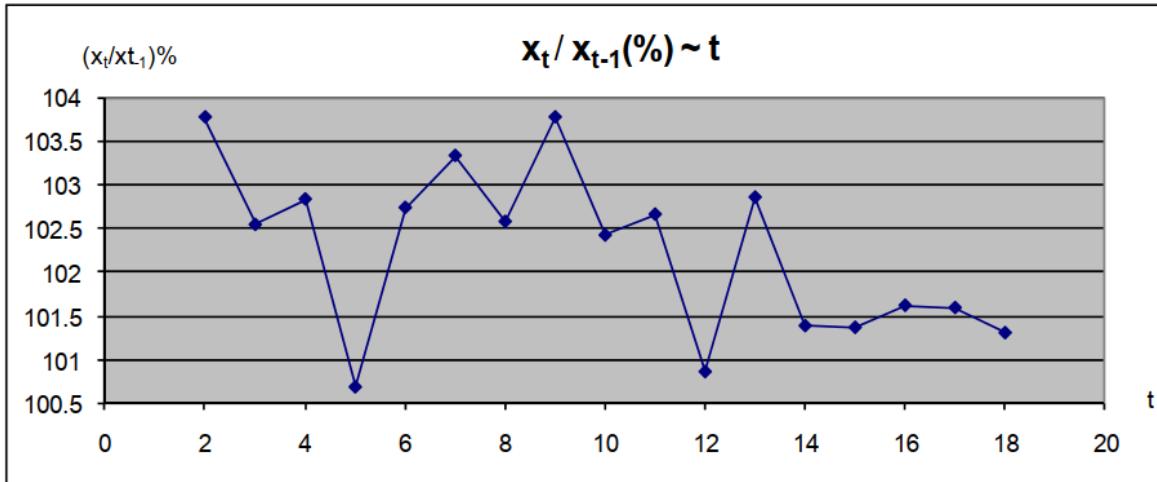
Từ dữ liệu chuỗi thời gian trên, ta có thể minh họa dưới dạng đồ thị như sau:

- Đồ thị của  $x_t$  theo  $t$ : cung cấp lịch sử dữ liệu gốc chưa bị chuyển đổi qua bất cứ phép biến đổi nào, giúp cho việc nghiên cứu xu thế và nhận dạng. Hình 2.6: Đồ thị của  $x_t$  theo  $t$



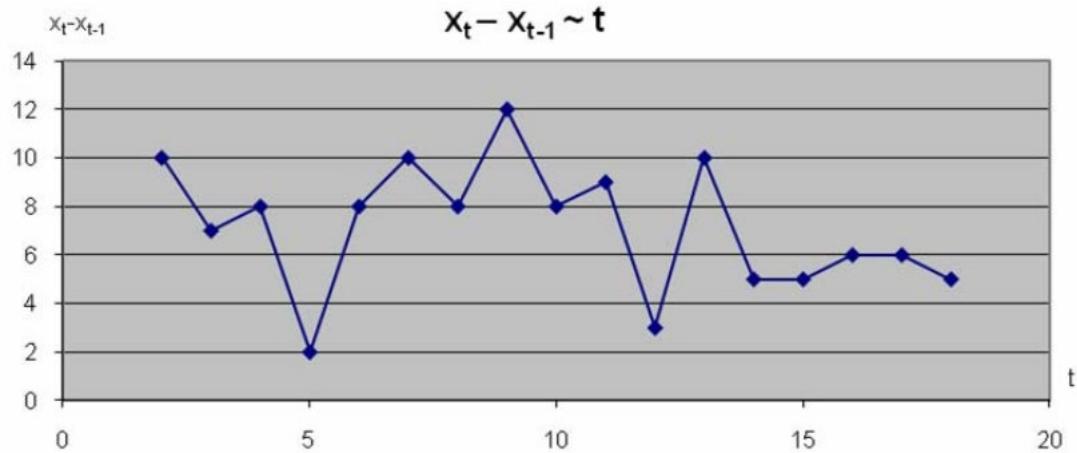
Hình 2.6: Đồ thị của  $x_t$  theo  $t$

- Đồ thị của  $(x_t / x_{t-1}) \times 100$  theo  $t$ : mỗi điểm trên đồ thị này cho biết giá trị hiện thời của chuỗi tăng hay giảm so với giá trị trước đó. Ví dụ giá trị tại thời điểm  $t=2$  là 102,9% chỉ ra rằng chuỗi đã tăng 2,9% từ thời điểm  $t=2$  sang thời điểm  $t=3$ . Nếu mọi giá trị đều lớn hơn 100% nhưng theo xu hướng giảm dần thì đồ thị đó chứng tỏ rằng chuỗi này có xu hướng tăng nhưng tỷ lệ tăng lại giảm dần. Hình 2.7: Đồ thị của  $(x_t / x_{t-1}) \times 100$  theo  $t$



Hình 2.7: Đồ thị của  $x_t$  theo  $t$

- Đồ thị của  $(x_t - x_{t-1})$  theo  $t$ : Đồ thị này biểu diễn sự thay đổi giữa các bước thời gian kế tiếp nhau. Nhìn vào đồ thị ta thấy được khoảng cách giá trị biến đổi giữa các bước kế nhau. Hình 2.8:  $(x_t - x_{t-1})$  theo  $t$



Hình 2.8: Đồ thị  $(x_t - x_{t-1})$  theo  $t$

## 2.2. Bài toán dự báo chuỗi thời gian và các mô hình dùng trong dự báo chuỗi thời gian

### 2.2.1. Các bài toán về dự báo chuỗi thời gian

Chuỗi thời gian được sử dụng để thu thập các dữ liệu quan sát trong rất nhiều lĩnh vực như thống kê, xử lý tín hiệu số, toán tài chính, ... trước khi thực hiện các phân tích thích hợp tùy thuộc vào ứng dụng của mỗi lĩnh vực cụ thể. Phân tích chuỗi thời gian nhằm mục đích đưa ra các thông kê có ý nghĩa, giải quyết vấn đề nhận diện

những đặc trưng cơ bản của chuỗi thời gian cũng như khai phá cấu trúc nội tại của chuỗi thời gian từ dữ liệu quan sát được.

Nghiên cứu khoa học về các đối tượng nào đó (các vấn đề trong kinh tế, vật lý, tự nhiên, ...) thường dựa vào chuỗi thời gian tạo ra từ dữ liệu các mẫu quan sát được theo thời gian, dữ liệu này chính là cơ sở để hiểu được đặc tính cũng như là dự đoán các hành vi tương lai của đối tượng đó. Nếu ta xác định được những phương trình cơ sở thì các đối tượng nghiên cứu này có thể phân tích được và qua đó xác định được các đặc tính của chúng. Tuy nhiên, trong thực tế, ta thường không biết được các phương trình cơ sở của đối tượng nghiên cứu. Trong trường hợp này, những quy tắc quan sát được trong quá khứ sẽ được sử dụng như là những chỉ dẫn để hiểu được đối tượng nghiên cứu và dự đoán hành vi tương lai.

**Bài toán dự báo chuỗi thời gian:** Cho một dãy các dữ liệu quan sát được theo thời gian, một hệ thống dự báo sẽ thực hiện việc ước lượng các giá trị quan sát trong vài chu kỳ thời kế tiếp. Ta có biểu diễn bài toán một cách chi tiết như sau:

Dự báo 1-bước: Cho trước dãy  $x_1, x_2, \dots, x_t$ , dự đoán giá trị của  $x_{t+1}$ .

Bài toán này được tổng quát hóa như sau:

Dự báo n-bước: Tập dữ liệu quan sát được trong quá khứ (Tập huấn luyện) là một tập hợp các chuỗi thời gian tạo ra từ cùng một đối tượng nghiên cứu trên các chu kỳ thời gian khác nhau.

$$TS = \{X_1, X_2, \dots, X_N\}$$

Với  $X_i = x_{ti}, x_{ti+1}, \dots, x_{ti+(l_i-1)}$ , trong đó  $x_t$  là giá trị của chuỗi thời gian tại thời điểm  $t$  và  $l_i$  là độ dài của dãy  $X_i$ . Hệ thống dự báo sẽ được cung cấp tương ứng với tập TS dãy kết quả truy vấn  $Y = y_1, y_2, \dots, y_l$  và ta sẽ cần tìm các giá trị  $y_{l+1}, y_{l+2}, \dots$

Phân tích chuỗi thời gian cho mục đích dự báo là một mảng nghiên cứu lớn với các ứng dụng rộng lớn đa dạng. Rất nhiều lĩnh vực thường dùng dữ liệu chuỗi thời gian để dự báo như: Vật lý, Sinh học, kinh tế, thiên văn học, địa vật lý, ...

### 2.2.2. Các mô hình dùng trong dự báo chuỗi thời gian

#### 2.2.2.1. Mô hình phân tích hồi quy đơn giản

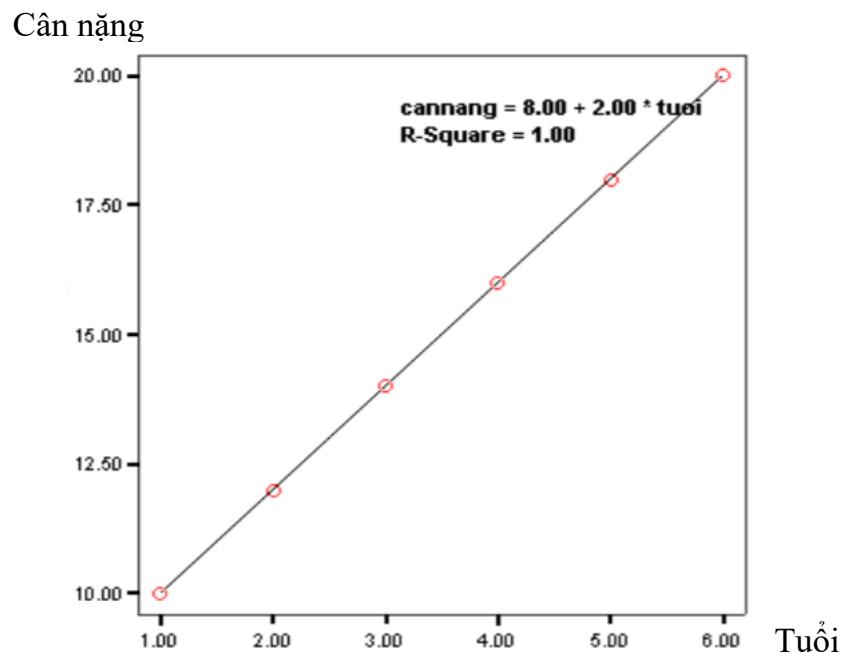
Phân tích hồi qui tuyến tính đơn giản (*Simple Linear Regression Analysis*) là tìm sự liên hệ giữa 2 biến số liên tục: biến độc lập (biến dự đoán) trên trực hoành  $x$

với biến phụ thuộc (biến kết cục) trên trục tung  $y$ . Sau đó vẽ một đường thẳng hồi qui và từ phương trình đường thẳng này ta có thể dự đoán được biến  $y$  (ví dụ: cân nặng) khi đã có  $x$  (ví dụ: tuổi)

Ví dụ: Ta có một mẫu gồm 6 trẻ từ 1 đến 6 tuổi, và có cân nặng như sau: Hình 2.9:  
Đồ thị hồi quy đơn giản

Bảng 2.2: Bảng dữ liệu cân nặng và chiều cao của trẻ

Tuổi	Cân nặng (Kg)
1	10
2	12
3	14
4	16
5	18
6	20



Hình 2.9: Đồ thị hồi quy đơn giản

Nối các cặp  $(x,y)$  này ta thấy có dạng 1 phương trình bậc nhất:  $y=2x+8$  (trong đó 2 là độ dốc và 8 là điểm cắt trên trục tung  $y$  khi  $x=0$ ). Trong thống kê phương trình đường thẳng (bậc nhất) này được viết dưới dạng:

$$y = \beta x + \alpha$$

Đây là phương trình hồi qui tuyến tính, trong đó  $\beta$  gọi là độ dốc (*slope*) và  $\alpha$  là chặn (*intercept*), điểm cắt trên trục tung khi  $x=0$ .

Thực ra phương trình hồi qui tuyến tính này chỉ có trên lý thuyết, nghĩa là các trị số của  $x_i$  ( $i=1,2,3,4,5,6$ ) và  $y_i$  tương ứng, liên hệ với nhau 100% (hoặc hệ số tương quan  $R=1$ )

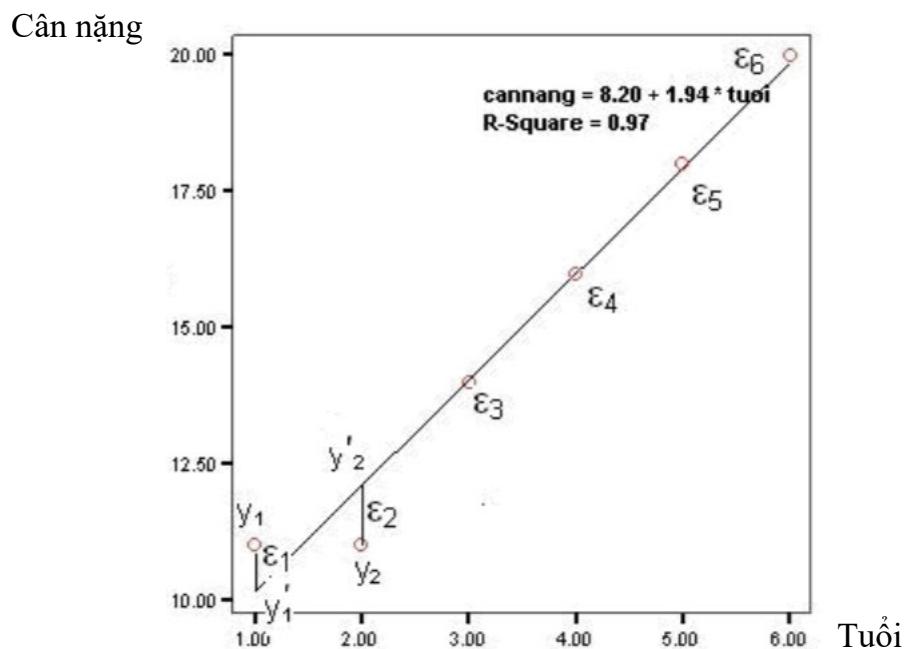
Trong thực tế hiếm khi có sự liên hệ 100% này mà thường có sự sai lệch giữa trị số quan sát  $y_i$  và trị số  $y'_i$  ước đoán nằm trên đường hồi qui. Hình 2.9: Đồ thị hồi quy đơn giản

Ví dụ: Ta có một mẫu gồm 6 trẻ khác có cân nặng như sau:

Bảng 2.3: Bảng dữ liệu cân nặng và chiều cao của trẻ

Tuổi	Cân nặng (Kg)
1	11
2	11
3	14
4	16
5	18
6	20

Hình 2.10: Đồ thị hồi quy đơn giản



Khi vẽ đường thẳng hồi qui, ta thấy các trị số quan sát  $y_3, y_4, y_5, y_6$  nằm trên đường thẳng, còn  $y_1$  và  $y_2$  không nằm trên đường thẳng này và sự liên hệ giữa  $x_i$  và  $y_i$  không còn là 100% mà chỉ còn 97% vì có sự sai lệch tại  $y_1$  và  $y_2$ . Sự sai lệch này trong thống kê gọi là phần dư (*residual*) hoặc *errors*.

Gọi  $y_1, y_2, y_3, y_4, y_5, y_6$  là trị số quan sát và  $y'_1, y'_2, y'_3, y'_4, y'_5, y'_6$  là trị số ước đoán nằm trên đường hồi qui;  $\epsilon_1, \epsilon_2, \epsilon_3, \epsilon_4, \epsilon_5, \epsilon_6$  là phần dư.

Như vậy:

$$\epsilon_1 = y_1 - y'_1$$

$$\epsilon_2 = y_2 - y'_2$$

$$\epsilon_3 = y_3 - y'_3$$

$$\epsilon_4 = y_4 - y'_4$$

$$\epsilon_5 = y_5 - y'_5$$

$$\epsilon_6 = y_6 - y'_6$$

Khi đó phương trình hồi qui tuyến tính được viết dưới dạng tổng quát như sau:

$$y' = \beta x_i + \alpha_i + \epsilon_i$$

Như vậy nếu phần dư  $\epsilon_i$  càng nhỏ sự liên hệ giữa  $x, y$  càng lớn và ngược lại. Phần liên hệ còn được gọi là phần hồi qui. Mô hình hồi qui tuyến tính được mô tả như sau:

$$\text{Dữ liệu} = \text{Hồi quy (Regression)} + \text{Phần dư (Residual)}$$

#### 2.2.2.2. Hàm tự tương quan ACF

Hàm tự tương quan đo lường phụ thuộc tuyến tính giữa các cặp quan sát  $y(t)$  và  $y(t+k)$ , ứng với thời đoạn  $k = 1, 2, \dots$  ( $k$  còn gọi là độ trễ). Với mỗi độ trễ  $k$ , hàm tự tương quan tại độ trễ  $k$  được xác định qua độ lệch giữa các biến ngẫu nhiên  $Y_t, Y_{t+k}$  so với các giá trị trung bình, và được chuẩn hóa qua phương sai.

Giả thiết rằng các biến ngẫu nhiên trong chuỗi dùng thay đổi quanh giá trị trung bình  $\mu$  với phương sai hằng số  $\sigma^2$ . Hàm tự tương quan tại các độ trễ khác nhau sẽ có giá trị khác nhau.

Trong thực tế, ta có thể ước lượng hàm tự tương quan tại độ trễ thứ  $k$  qua phép biến đổi trung bình của tất cả các cặp quan sát, phân biệt bằng các độ trễ  $k$ , với giá trị

trung bình mẫu là  $\mu$ , được chuẩn hóa bởi phuơng sai  $\sigma^2$ . Chẳng hạn, cho mỗi chuỗi  $N$  điểm, giá trị  $r_k$  của hàm tự tương quan tại độ trễ thứ  $k$  được tính như sau:

$$r_k = \frac{\frac{1}{N} \sum_{t=1}^{N-k} (y_t - \mu)(y_{t+k} - \mu)}{\delta^2}$$

$$\text{Với: } \mu = \frac{1}{N} \sum_{t=1}^N (y_t); \quad \delta^2 = \frac{1}{N} \sum_{t=1}^N (y_t - \mu)^2$$

Trong đó:

$y_t$ : chuỗi thời gian dừng tại thời điểm  $t$

$y_{t+k}$ : chuỗi thời gian dừng tại thời điểm  $t+k$

$\mu$ : giá trị trung bình của chuỗi dừng

$r_k$ : giá trị tương quan giữa  $y_t$  và  $y_{t+k}$  tại độ trễ  $k$

$r_k = 0$  thì không có hiện tượng tự tương quan

Về mặt lý thuyết, chuỗi dừng khi tất cả các  $r_k = 0$  hay chỉ vài  $r_k$  khác không.

Do chúng ta xem xét hàm tự tương quan mẫu, do đó sai số mẫu sẽ xuất hiện, vì vậy, hiện tượng tự tương quan khi  $r_k = 0$  theo ý nghĩa thống kê.

Khi hàm tự tương quan ACF giảm đột ngột, có nghĩa  $r_k$  rất lớn ở độ trễ 1, 2 và có ý nghĩa thống kê ( $|t| > 2$ ). Những  $r_k$  này được xem là những “đỉnh” và ta nói rằng hàm tự tương quan ACF giảm đột ngột sau độ trễ  $k$  nếu không có những “đỉnh” ở độ trễ  $k$  lớn hơn  $k$ . Hầu hết hàm tự tương quan ACF sẽ giảm đột ngột sau độ trễ 1, 2.

Nếu hàm tự tương quan ACF của chuỗi thời gian không dừng, không giảm đột ngột mà trái lại giảm nhanh nhưng đều: không có đỉnh, ta gọi chiều hướng này là “tắt dần”.

#### 2.2.2.3. Hàm tự tương quan từng phần PACF

Song song với việc xác định hàm tự tương quan giữa các cặp  $y(t)$  và  $y(t+k)$ , ta xác định hàm tự tương quan từng phần cũng có hiệu lực trong việc can thiệp đến các quan sát  $y(t+1), \dots, y(t+k-1)$ . Hàm tự tương quan từng phần tại độ trễ  $k$ ,  $C_{kk}$  được ước lượng bằng hệ số liên hệ  $y(t)$  trong mỗi kết hợp tuyến tính bên dưới. Sự kết hợp được tính dựa trên tầm ảnh hưởng của  $y(t)$  và các giá trị trung gian  $y(t+k)$ .

$$y(t+k) = C_{k1}y(t+k-1) + C_{k2}y(t+k-2) + \dots + C_{kk-1}y(t+1) + C_{kk}y(t) + e(t)$$

Giải phương trình hồi quy dựa trên bình phương tối thiểu vì hệ số hồi quy  $C_{kj}$  phải được tính ở mỗi độ trễ  $k$ , với  $j$  chạy từ 1 đến  $k$ .

Giải pháp ít tốn kém hơn do Durbin phát triển dùng để xấp xỉ độ quy hệ số hồi quy cho mô hình ARIMA chuỗi dừng, sử dụng giá trị hàm tự tương quan tại độ trễ  $k$ ,  $r_k$  và hệ số hồi quy của độ trễ trước.

Tổng quan, hàm tự tương quan từng phần được tính theo phương pháp Durbin:

$$C_{kk} = \frac{r_k - \sum(C_{k-1,j})r_{k-j}}{1 - \sum(C_{k-1,j})r_j}$$

Trong đó :

$r_k$ : Hàm tự tương quan tại độ trễ  $k$

$v$ : Phương sai

$C_{kj}$ : Hàm tự tương quan từng phần cho độ trễ  $k$ , loại bỏ những ảnh hưởng của các độ trễ can thiệp.

Tóm lại, hàm tự tương quan ACF và hàm tự tương quan từng phần PACF của chuỗi thời gian có các đặc tính khác nhau. Hàm tự tương quan ACF đo mức độ phụ thuộc tuyến tính giữa các cặp quan sát. Hàm tự tương quan từng phần PACF đo mức độ phụ thuộc tuyến tính từng phần. ARIMA khai thác những điểm khác biệt này để xác định cấu trúc mô hình cho chuỗi thời gian.

#### 2.2.2.4. Mô hình tự hồi quy (AR(p) – Autoregressive model)

Mô hình tự hồi qui là giá trị ước tính tương lai của mô hình phân tích chuỗi thời gian chỉ phụ thuộc vào giá trị trong quá khứ.

Theo [11] mô hình AR(p) là hồi quy trên chính số liệu quá khứ ở những chuk trước.

$$Y_t = a_0 + a_1y_{t-1} + a_2y_{t-2} + \dots + a_py_{t-p} + e_t$$

Trong đó :

$y_t$ : quan sát dừng hiện tại

$y_{t-1}, y_{t-2}, \dots$ : quan sát dừng quá khứ (thường sử dụng không quá 2 biến này)

$a_0, a_1, a_2, \dots$  : các tham số phân tích hồi quy.

$e_t$  : sai số dự báo ngẫu nhiên của giai đoạn hiện tại. Giá trị trung bình được mong đợi bằng 0.

$y_t$  là một hàm tuyến tính của những quan sát dùng quá khứ  $y_{t-1}, y_{t-2}, \dots$ . Nói cách khác khi sử dụng phân tích hồi quy  $y_t$  theo các giá trị chuỗi thời gian dùng có độ trễ, chúng ta sẽ được mô hình AR (yếu tố xu thế đã được tách khỏi yếu tố thời gian, chúng ta sẽ mô hình hóa những yếu tố còn lại – đó là sai số).

Số quan sát dùng quá khứ sử dụng trong mô hình hàm tự tương quan là bậc  $p$  của mô hình AR. Nếu ta sử dụng hai quan sát dùng quá khứ, ta có mô hình tương quan bậc hai AR(2).

Điều kiện dừng là tổng các tham số phân tích hồi quy nhỏ hơn 1:

$$a_0 + a_1 + \dots + a_p < 1$$

Mô hình AR(1):  $y_t = a_0 + a_1 y_{t-1} + e_t$

Mô hình AR(2):  $y_t = a_0 + a_1 y_{t-1} + a_2 y_{t-2} + e_t$

#### 2.2.2.5. Mô hình trung bình di động (MA(q) – Moving Average Model)

Mô hình trung bình di động thuộc về lớp các mô hình thường dùng trong dự báo chuỗi thời gian. Giả sử ta cần dự báo chuỗi thời gian được thu thập theo từng tháng trong năm, có thể ta phải dùng đến mô hình sau:

$$f(t) = \frac{1}{12} (y_{t-1} + y_{t-2} + \dots + y_{t-12})$$

Như vậy, giá trị dự báo 1 - bước ứng với mô hình này là:

$$\hat{y}_{T+1} = \frac{1}{12} (y_T + y_{T-1} + \dots + y_{T-11})$$

Mô hình trung bình di động sẽ hữu dụng nếu ta tin rằng giá trị mong đợi ở tháng kế tiếp của chuỗi thời gian chỉ đơn thuần là giá trị trung bình của 12 tháng trước đó. Điều này có vẻ không thực tế, tuy nhiên, giá trị dự báo tốt có thể đạt được từ việc lấy trung bình đơn giản như vậy. Để hợp lý hơn, ta có thể cho rằng các quan sát gần nhất (với thời điểm dự báo) có vai trò quan trọng hơn là các quan sát trước đó nữa. Trong trường hợp này ta sẽ gán cho các quan sát một hệ số để thể hiện vai trò của nó, quan sát gần nhất sẽ nhận hệ số lớn nhất. Mô hình trung bình di động hoàn thiện theo cách này còn được gọi là trung bình di động có trọng số theo mũ (EWMA):

$$\begin{aligned}\hat{y}_{T+1} &= \alpha y_T + \alpha(1-\alpha)y_{T-1} + \alpha(1-\alpha)^2 y_{T-2} + \dots \\ &= \alpha \sum_{\tau=0}^{\infty} (1-\alpha)^{\tau} y_{T-\tau}, \quad 0 < \alpha \leq 1\end{aligned}$$

Với  $\alpha = 1$ , ta bỏ qua bất kỳ quan sát nào xuất hiện trước  $y_T$  và giá trị dự báo trở thành:

$$\hat{y}_{T+1} = y_T$$

Khi  $\alpha$  bé thì mô hình cho thấy các giá trị quan sát càng xa so với thời điểm dự báo càng có vai trò lớn hơn. Chú ý rằng phương trình trên biểu diễn mức trung bình vì

$$\alpha \sum_{\tau=0}^{\infty} (1-\alpha)^\tau = \frac{\alpha}{1-(1-\alpha)} = 1$$

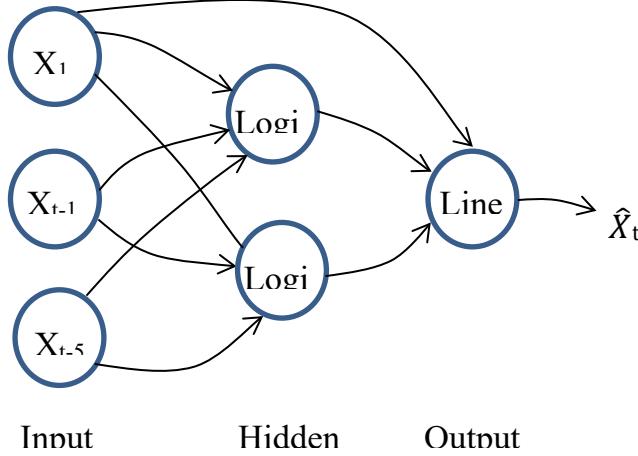
Nếu chuỗi thời gian có xu hướng tăng hoặc giảm thì mô hình *EWMA* sẽ đưa ra giá trị dự báo tương ứng ở mức thấp hơn hoặc cao hơn giá trị tương lai (trường hợp này thực sự có thể xảy ra vì mô hình này lấy trung bình các giá trị trong quá khứ để đưa ra giá trị dự báo, nếu chuỗi thời gian tăng đều đặn thì *EWMA* sẽ có giá trị dự báo bé hơn so với các giá trị của chuỗi gần thời điểm dự báo). Do đó, một kỹ thuật thường thấy trong vấn đề dự báo (không chỉ đối với mô hình *EWMA*) được áp dụng là loại bỏ các yếu tố xu hướng khỏi dữ liệu chuỗi thời gian trước khi dùng đến mô hình *EWMA*. Mỗi khi giá trị dự báo của chuỗi đã loại bỏ yếu tố xu hướng được tạo ra thì một số hạng biểu diễn xu hướng sẽ được cộng thêm vào để đạt được giá trị dự báo cuối cùng.

Nếu ta sử dụng mô hình *EWMA* thực hiện dự báo hơn một bước  $\hat{y}_{T+1}$ , ta sẽ hiệu chỉnh lại phương trình để mở rộng mô hình *EWMA* như sau:

$$\begin{aligned}\hat{y}_{T+1} &= \alpha \hat{y}_{T+1-1} + \alpha(1-\alpha) \hat{y}_{T+1-2} + \dots + \alpha(1-\alpha)^{l-2} \hat{y}_{T+1} \\ &\quad + \alpha(1-\alpha)^{l-1} y_T + \alpha(1-\alpha)^l y_{T-1} + \alpha(1-\alpha)^{l+1} y_{T-2} + \dots\end{aligned}$$

#### 2.2.2.6. Mô hình mạng Nơ-ron nhân tạo ANN (*Artificial Neural Network*)

Mạng nơ-ron nhân tạo là một lĩnh vực nghiên cứu rất lớn trong lĩnh vực trí tuệ nhân tạo, ANN được xem như một hệ thống kết nối tập hợp các ngõ vào (*inputs*) đến tập hợp các ngõ ra (*outputs*) qua một hay nhiều lớp nơ-ron, các lớp này được gọi là các lớp ẩn. Việc xác định có bao nhiêu ngõ vào, ngõ ra, số lớp ẩn cũng như là số lượng nơ-ron của mỗi lớp tạo thành kiến trúc của mạng.



Hình 2.11: Kiến trúc của một ANN cho dự báo chuỗi thời gian với 3 ngõ vào, một lớp ẩn hai nơ-ron và một ngõ ra (là giá trị dự báo)

Trong ngữ cảnh chuỗi thời gian, ngõ ra là giá trị của chuỗi thời gian được dự báo, ngõ vào có thể là có giá trị quan sát trước thời điểm dự báo (xác định bởi độ trễ) của chuỗi thời gian và các biến giải thích khác.[12]

Đối với các ANN một lớp ẩn có  $H$  nơ-ron, phương trình tổng quát để tính giá trị dự báo  $\hat{x}_t$  (ngõ ra) sử dụng đến các mẫu quan sát quá khứ  $x_{t-j1}, x_{t-j2}, \dots, x_{t-jk}$  làm ngõ vào được viết dưới dạng sau:

$$\hat{x}_t = \phi_0 (w_{c0} + \sum_{h=1}^H w_{h0} \phi_h (w_{ch} + \sum_{i=1}^k w_{ih} x_{t-ji}))$$

Trong đó:

- $\{w_{ch}\}_{h=1,2, \dots, H}$  biểu thị các trọng số cho kết nối giữa hằng số ngõ vào và các nơ-ron lớp ẩn.

- $w_{c0}$  là trọng số kết nối trực tiếp giữa ngõ vào hằng số và ngõ ra.

- $\{w_{ih}\}$  và  $\{w_{h0}\}$  là các trọng số của các kết nối khác giữa các ngõ vào và các nơ-ron lớp ẩn, giữa các nơ-ron lớp ẩn với ngõ ra.

- $\phi_0$  và  $\phi_h$  là hai hàm kích hoạt lần lượt được sử dụng tại ngõ ra và tại các nơ-ron lớp ẩn.

Lớp vào sẽ là nơi nhận các tín hiệu đầu vào. Các tín hiệu này có thể là một hằng số, dữ liệu thô hoặc cũng có thể là đầu ra của một mạng nơ-ron khác. Các giá trị này sẽ tác động đến các nơ-ron lớp ẩn thông qua bộ trọng số  $w_c$ . Tại lớp ẩn, tín

hiệu của lớp vào sẽ được xử lý bằng một hàm kích hoạt (activate function)  $\phi_0$  và  $\phi_h$ , sau đó tín hiệu sẽ được truyền qua lớp ra thông qua bộ trọng số  $w_h$ .

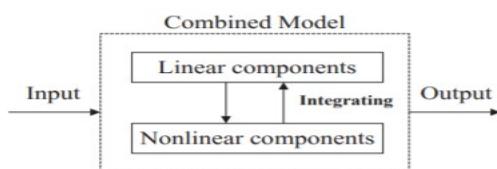
Tương tự như các mô hình hồi quy tuyến tính, mô hình ANN xác lập mối quan hệ giữa một tập hợp các biến đầu vào  $X_i$  ( $i = \overline{1, m}$ ) với một hoặc nhiều biến đầu ra  $\hat{X}_k$  ( $k = \overline{1, n}$ ) dựa vào dữ liệu trong quá khứ. Điều khác biệt là sự tồn tại của các “lớp ẩn” (*hidden layer*), các lớp này liên kết giữa lớp vào và lớp ra của mạng nơ-ron. Chính các lớp ẩn này đã giúp cho mô hình mạng thần kinh có khả năng mô phỏng mối tương quan phi tuyến tốt hơn so với mô hình truyền thống.[7]

Mục tiêu của mô hình ANN là tính toán và dự báo giá trị của biến đầu ra với một tập hợp các thông tin của biến đầu vào được cho trước. Mô hình ANN sẽ được “huấn luyện” để có thể “học” từ những thông tin quá khứ. Từ đó, mạng có thể đưa ra kết quả dự báo dựa trên những gì đã được học. Quá trình này sẽ được tiến hành bằng các thuật toán huấn luyện mạng, phổ biến là thuật toán lan truyền ngược (*back-propagation algorithm*) và thuật toán di truyền (*genetic algorithm*).

### 2.3. Các mô hình lai ghép dùng trong dự báo chuỗi thời gian.

#### 2.3.1. Mô hình ARIMA và ANN

Mô hình lai ARIMA và ANN. Như đã trình bày phần trên, mỗi chuỗi thời gian được chia thành 2 thành phần một thành phần tuyến tính và một thành phần phi tuyến bằng các kỹ thuật phân rã (ví dụ: phân tách Fourier và wavelet phân hủy). Mô hình dự báo ARIMA thực hiện thành công trên các thành phần tuyến tính của chuỗi thời gian, còn các thành phần phi tuyến của thời gian thì không thành công. Do đó phải dùng một mô hình khác để giải quyết cho các thành phần phi tuyến của chuỗi thời gian, mô hình ANN có khả năng mô hình hóa đặc biệt hiện tượng khá phức tạp vì chúng có nhiều tế bào thần kinh phi tuyến tương tác trong nhiều lớp, có thể được sử dụng để giải quyết thành phần phi tuyến của dữ liệu chuỗi thời gian.



Hình 2.12: Mô hình lai ARIMA - ANN

Chúng ta có thể xem xét hai mô hình để phân tích chuỗi thời gian như trên là mô hình cộng ( $L + N$ ) và mô hình nhân ( $L * N$ ) theo công thức sau:

$$y_t = L_t + N_t$$

$$y_t = L_t * N_t$$

Trong đó:

-  $L_t$ : là thành phần tuyến tính

-  $N_t$ : là thành phần phi tuyến tính

Dữ liệu sử dụng trong hai thành phần này phải được xử lý trước.

Mô hình ARIMA (Box et al., 1994) đã được sử dụng để dự đoán  $y_t$  và để cho  $\hat{L}_t$  biểu thị kết quả dự đoán. Các  $e_t$  là phần dư giữa các chuỗi của mô hình ARIMA.

$$e_t = y_t - \hat{L}_t \text{ hoặc}$$

$$e_t = y_t / \hat{L}_t$$

+  $e_t$  được coi là đầu vào của mô hình ANN, sau đó mô hình ANN có thể được biểu thị như sau:

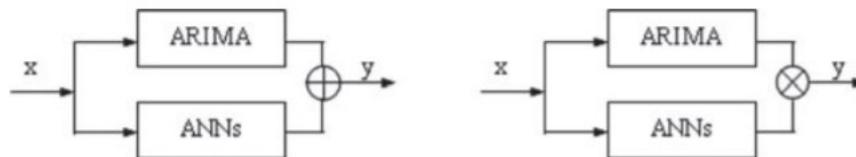
$$e_t = f(e_{t-1}, e_{t-2}, \dots, e_{t-n}) + \varepsilon_t$$

Trong đó,  $f$  là hàm phi tuyến được xác định bởi mạng nơ ron và  $\varepsilon_t$  là lỗi ngẫu nhiên.

Kết quả đầu ra của ANN được định nghĩa là  $\hat{N}_t$ .

+ Hai mô hình được kết hợp để dự báo và kết quả dự đoán từ các mô hình lai ARIMA - ANN được biểu thị như sau:

$$y_t = \hat{L}_t + \hat{N}_t \quad \text{Hoặc } y_t = \hat{L}_t * \hat{N}_t$$

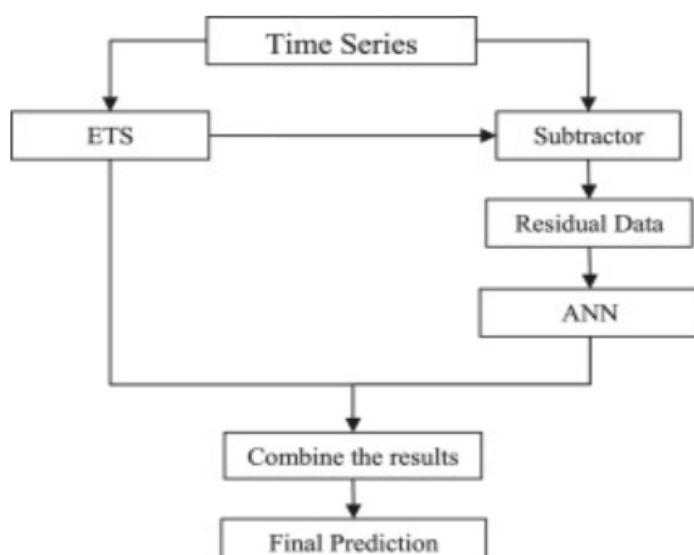


Hình 2.13: Mô hình kết hợp ARIMA - ANN

### 2.3.2. Mô hình Exponential Smoothing và ANN

Các mô hình lai bằng cách kết hợp các mô hình tuyến tính như trung bình di chuyển tích hợp tự động (ARIMA) với các mô hình phi tuyến như mạng nơ ron nhân tạo (ANN) đã trở nên phổ biến do hiệu suất vượt trội so với các mô hình riêng lẻ. Các mô hình này giả định chuỗi thời gian là tổng của thành phần tuyến tính và phi tuyến.

Tuy nhiên, một chuỗi thời gian trong thế giới thực có thể hoàn toàn tuyến tính hoặc hoàn toàn phi tuyến hoặc thường chứa sự kết hợp của các mẫu tuyến tính và phi tuyến. Một phương pháp mới được phát triển bằng cách kết hợp các mô hình làm mịn theo cấp số nhân tuyến tính và phi tuyến từ không gian trạng thái đổi mới (ETS) với ANN. Mô hình lai ETS - ANN kết hợp sự khác nhau của các mẫu tuyến tính và phi tuyến trong chuỗi thời gian. Điều này là do cả hai mô hình ETS và ANN đều có khả năng xử lý tuyến tính cũng như phi tuyến. Tuy nhiên, ANN không thể xử lý các mẫu tuyến tính tốt như các mẫu phi tuyến. Do đó, trong phương pháp đề xuất, ETS đầu tiên được áp dụng cho chuỗi thời gian nhất định và dự đoán thu được. Điều này giúp tăng khả năng nắm bắt tốt các mẫu tuyến tính hiện có (nếu có) bằng cách sử dụng các mô hình ETS tuyến tính. Sau đó, chuỗi lỗi dư được tính bằng cách trừ các dự đoán ETS khỏi chuỗi ban đầu. Chuỗi lỗi dư được xử lý bằng mô hình ANN. Sau đó, dự đoán cuối cùng có được bằng cách kết hợp các dự đoán ETS với dự đoán ANN. Thủ nghiệm được sử dụng trên nhiều bộ dữ liệu chuỗi thời gian để phân tích hiệu suất so sánh của phương pháp được đề xuất với ARIMA, ETS, perceptron đa lớp (MLP) và một số mô hình ANN lai ARIMA phiên bản hiện có. Kết quả thử nghiệm cho thấy mô hình lai được đề xuất cho thấy kết quả tốt hơn về mặt thống kê cho các bộ dữ liệu được sử dụng.



Hình 2.14: Mô hình lai ETS – ANN

Trong chương này chúng tôi đã trình bày khái quát về chuỗi thời gian, một số mô hình dự báo trên chuỗi thời gian, trong chương 3 tiếp theo của luận văn, chúng tôi sẽ giới thiệu một số mô hình lai của các nhà khoa học đã thực hiện, để từ đó đề xuất và thực nghiệm một mô hình lai mới trong chương 4.

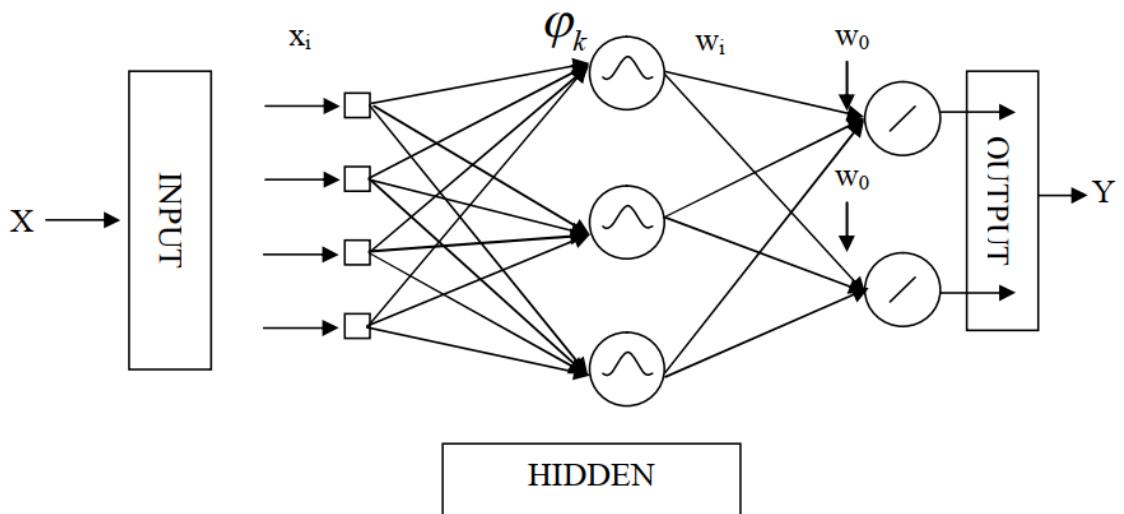
## Chương 3

# DỰ BÁO TRÊN CHUỖI THỜI GIAN SỬ DỤNG MÔ HÌNH LAI GHÉP ARIMA VÀ RBFNN

Nội dung trong chương này: Cơ sở của các mô hình RBFNN, ARIMA, ARIMA-RBFNN sẽ được trình bày và đề xuất mô hình ARIMA-RBFNN cải tiến.

### 3.1. Mô hình mạng Nơ-ron nhân tạo RBF (Radial Basis Function Neural Network - RBFNN)

RBFNN là một loại mạng Neural nhân tạo truyền thẳng gồm có ba lớp. Nó bao gồm  $n$  nút của lớp đầu vào cho vector đầu vào  $x \in \mathbb{R}^n$ ,  $N$  neuron ẩn (giá trị của neuron ẩn thứ  $k$  chính là giá trị trả về của hàm cơ sở bán kính  $\varphi_k$ ) và  $m$  neuron đầu ra. Hình 3.1: Mô hình mạng RBF



Hình 3.1: Mô hình mạng RBF

Mạng RBF có thể biểu diễn bằng công thức toán học sau:

$$\varphi(x^j) = \sum_{k=1}^N (w_k \varphi_k(x^j) + w_{0k}) = \sum_{k=1}^N \left( w_k e^{-\frac{\|x^j - v^k\|^2}{\sigma^2_k}} + w_{0k} \right) = y^j$$

Với tầng ẩn thì thường dùng hàm tổng là hàm  $S = \|x - w\|^2$ , còn hàm chuyển là hàm Gauss  $\varphi(v) = e^{-v}$

Tầng ra thì dùng hàm tổng là hàm  $S = \sum_{i=1}^N w_i x_i$ , hàm chuyển là hàm tuyến tính  $\varphi(v) = av$

Có nhiều cách huấn luyện mạng RBF. Có thể tách riêng một pha để xác định các tham số độ rộng  $\sigma_k$  của mỗi hàm bán kính và sau đó tìm các tham số  $w_k$  (phương pháp 2 pha) hoặc huấn luyện 1 lần nhờ tìm cực tiểu sai số tổng các bình phương.

### **3.2. Mô hình tự hồi quy kết hợp với trung bình di động ARIMA(p,d,q) (AutoRegressive Integrated Moving Average)**

#### **3.2.1. Sai phân I(d)**

Trong thực tế, rất hiếm khi gặp một chuỗi thời gian dừng bởi sự dao động lên xuống của thị trường. Do đó, trước khi áp dụng mô hình ARIMA vào dữ liệu chuỗi thời gian phải chuyển đổi chuỗi thời gian thành chuỗi dừng mới.

Để xác định tính dừng của chuỗi thời gian:

- Dựa vào biểu đồ tự tương quan ACF

- + Nếu hàm tự tương quan ACF của chuỗi thời gian hoặc giảm thật nhanh hoặc giảm dần khá nhanh thì giá trị của chuỗi thời gian được xem là dừng.

- + Nếu hàm tự tương quan ACF của chuỗi thời gian giảm dần thật chậm thì chuỗi thời gian được xem là không dừng.

- Dựa trên đồ thị  $Y(t) = f(t)$ , một cách trực quan chuỗi  $Y(t)$  có tính dừng nếu như đồ thị cho thấy trung bình và phương sai của quá trình  $Y_t$  không thay đổi theo thời gian.

Sai phân chỉ sự khác nhau giữa giá trị hiện tại và giá trị trước đó. Phân tích sai phân nhằm làm cho ổn định giá trị trung bình của chuỗi dữ liệu, giúp cho việc chuyển đổi chuỗi thành một chuỗi dừng.

$$\text{Sai phân lần 1 (I(1)) : } z(t) = y(t) - y(t-1)$$

$$\text{Sai phân lần 2 (I(2)) : } h(t) = z(t) - z(t-1)$$

Ví dụ: Xét chuỗi dữ liệu, cột thời gian tính bằng mini giây

70.3	100.5	130.2	160.7	190.5	220.2	250.4
------	-------	-------	-------	-------	-------	-------

Sai phân bậc một cung cấp một chuỗi dừng dao động quanh giá trị trung bình 30 của chuỗi:

70.3	100.5	130.2	160.7	190.5	220.2
30.2	29.7	30.5	29.8	29.7	30.2

Trong ví dụ này, ta chỉ thực hiện một lần chuyển đổi sai phân cho toàn bộ dữ liệu nhằm làm ổn định giá trị trung bình. Tuy nhiên, trong thực tế, có rất nhiều chuỗi dữ liệu cần thực hiện sai phân nhiều hơn để có thể đạt được tính dừng.

### 3.2.2. Mùa vụ ( $S$ )

Hiện tượng có thành phần mùa vụ trong dữ liệu chuỗi thời gian cũng là một hiện tượng khá phổ biến khi sử dụng mô hình ARIMA với dữ liệu trong một khoảng thời gian dài.

Vì lý do đó, để có thể áp dụng được mô hình ARIMA vào dữ liệu chuỗi thời gian cần khử tính mùa vụ trước. Nếu  $Y(t)$  có tính mùa vụ, với chu kỳ  $s$ , thì để khử tính mùa vụ ta lấy sai phân thứ  $s$  :  $Z(t) = Y(t) - Y(t-s)$ . Và sử dụng chuỗi dữ liệu mới  $Z(t)$  sau khi đã khử tính mùa vụ vào mô hình ARIMA.

Thông thường, tính mùa vụ của chuỗi dữ liệu vào khoảng

- 4 mùa trong một năm  $S(4) : z(t) = y(t) - y(t-4)$
- 12 tháng trong một năm  $S(12) : z(t) = y(t) - y(t-12)$

**Mô hình ARMA( $p,q$ ):** là mô hình hỗn hợp của AR và MA. Hàm tuyến tính sẽ bao gồm những quan sát dừng quá khứ và những sai số dự báo quá khứ và hiện tại:

$$Y(t) = a_0 + a_1y(t-1) + a_2y(t-2) + \dots + a_py(t-p) + e(t) + b_1e(t-1) + b_2e(t-2) + \dots + b_qe(t-q)$$

Trong đó :

$y(t)$  : quan sát dừng hiện tại

$y(t-p),$  và  $e(t-q)$ : quan sát dừng và sai số dự báo quá khứ.

$a_0, a_1, a_2, \dots, b_1, b_2, \dots$  : các hệ số phân tích hồi quy

Ví dụ: ARMA(1,2) là mô hình hỗn hợp của AR(1) và MA(2)

Đối với mô hình hỗn hợp thì dạng  $(p,q) = (1,1)$  là phổ biến. Tuy nhiên, giá trị  $p$  và  $q$  được xem là những độ trễ cho ACF và PACF quan trọng sau cùng. Cả hai điều kiện bình quân di động và điều kiện dừng phải được thỏa mãn trong mô hình hỗn hợp ARMA.

### 3.2.3. Mô hình ARIMA( $p,d,q$ )

Mô hình tự hồi quy tích hợp với trung bình di động(ARIMA) là một mô hình tuyến tính có khả năng biểu diễn cả chuỗi thời gian tĩnh lẫn không tĩnh. Mô hình ARIMA dựa vào các mẫu tự tương quan trong bản thân của chuỗi thời gian để sinh ra dự đoán. Hệ thống các phương pháp dùng để xác định, kiểm tra và cải tiến mô hình ARIMA có sự đóng góp rất lớn của hai nhà thống kê, G.E.P.Box và G.M.Jenkins. Do đó việc mô hình và dự đoán dựa trên mô hình ARIMA còn được gọi là phương pháp luận Box-Jenkins [1]

Các mô hình chỉ mô tả chuỗi dừng hoặc những chuỗi đã sai phân hóa, nên mô hình ARIMA(p,d,q) thể hiện những chuỗi dữ liệu không dừng, đã được sai phân (ở đây, d chỉ mức độ sai phân).

Khi chuỗi thời gian dừng được lựa chọn (hàm tự tương quan ACF giảm đột ngột hoặc giảm đều nhanh), chúng ta có thể chỉ ra một mô hình dự định bằng cách nghiên cứu xu hướng của hàm tự tương quan ACF và hàm tự tương quan từng phần PACF.

Theo lý thuyết, nếu hàm tự tương quan ACF giảm đột biến và hàm tự tương quan từng phần PACF giảm mạnh thì chúng ta có mô hình tự tương quan. Nếu hàm tự tương quan ACF và hàm tự tương quan từng phần PACF đều giảm đột ngột thì chúng ta có mô hình hỗn hợp.

Về mặt lý thuyết, không có trường hợp hàm tự tương quan ACF và hàm tự tương quan từng phần cùng giảm đột ngột. Trong thực tế, hàm tự tương quan ACF và hàm tự tương quan từng phần PACF giảm đột biến khá nhanh. Trong trường hợp này, chúng ta nên phân biệt hàm nào giảm đột biến nhanh hơn, hàm còn lại được xem là giảm đều. Do đôi lúc sẽ có trường hợp giảm đột biến đồng thời khi quan sát biểu đồ hàm tự tương quan ACF và hàm tự tương quan từng phần PACF, biện pháp khắc phục là tìm vài dạng hàm dự định khác nhau cho chuỗi thời gian dừng. Sau đó, kiểm tra độ chính xác mô hình tốt nhất. Hình 3.2: Sơ đồ mô phỏng mô hình ARIMA

Mô hình ARIMA (1, 1, 1) :  $y(t) - y(t-1) = a_0 + a_1(y(t-1) - y(t-2)) + e(t) + b_1e(t-1)$ )

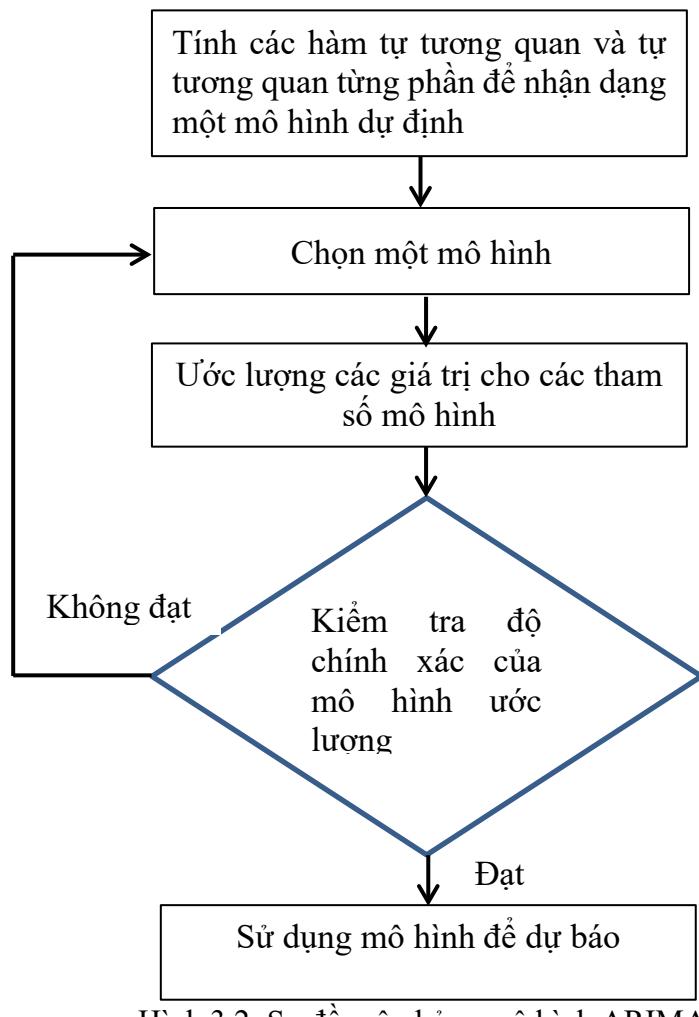
$$\text{Hoặc } z(t) = a_0 + a_1z(t-1) + e(t) + b_1e(t-1),$$

Với  $z(t) = y(t) - y(t-1)$  ở sai phân đầu tiên:  $d = 1$ .

Tương tự ARIMA(1,2,1):  $h(t) = a_0 + a_1z(t-1) + e(t) + b_1e(t-1)$ ,

Với  $h(t) = z(t) - z(t-1)$  ở sai phân thứ hai:  $d = 2$ .

Tuy nhiên, trong thực hành  $d$  lớn hơn 2 rất ít được sử dụng.



Hình 3.2: Sơ đồ mô phỏng mô hình ARIMA

### 3.2.4. Phương pháp ước lượng tham số

#### 3.2.4.1. Phương pháp ước lượng tham số Moment

Phương pháp ước lượng *moment* thường là một trong những phương pháp dễ nhất, nếu không phải là hiệu quả nhất để có được các ước tính tham số. Phương pháp này bao gồm việc cân bằng *moment* mẫu với các *moment* lý thuyết tương ứng và giải các phương trình cân bằng để có được ước tính của bất kỳ tham số nào chưa biết. Ví

dụ đơn giản nhất của phương pháp là ước tính trung bình của một quá trình dừng theo trung bình mẫu.[9]

- Áp dụng phương pháp ước lượng tham số Moment vào mô hình tự hồi quy (AR(p) - *Autoregressive Models*)

Trước tiên hãy xem xét trường hợp *AR (1)*. Đối với quá trình này, chúng ta có mối quan hệ đơn giản  $p_1 = \emptyset$ . Trong phương pháp này,  $p_1$  tương đương với  $r_1$ , với độ trễ thứ nhất. Ta có thể ước lượng  $\emptyset$

$$\hat{\emptyset} = r_1$$

Xét trường hợp AR(2). Mỗi quan hệ giữa các tham số  $\emptyset_1$  và  $\emptyset_2$  và các thời điểm khác nhau được đưa ra bởi các phương trình Yule-Walker

$$p_1 = \emptyset_1 + p_1\emptyset_2 \text{ và } p_2 = p_1\emptyset_1 + \emptyset_2 \quad (1)$$

Theo phương pháp ước lượng Moment, thay đổi  $p_1$  thành  $r_1$  và  $p_2$  thành  $r_2$  ta được công thức

$$r_1 = \emptyset_1 + r_1\emptyset_2 \text{ và } r_2 = r_1\emptyset_1 + \emptyset_2 \quad (2)$$

Sau đó tham số sẽ được tính:

$$\hat{\emptyset}_1 = \frac{r_1(1 - r_2)}{1 - r^2_1} \text{ và } \hat{\emptyset}_2 = \frac{r_2 - r^2_1}{1 - r^2_1} \quad (3)$$

Tương tự cho trường hợp *AP(p)*, thay thế  $p_k$  thành  $r_k$ , thông qua phương trình *Yule-Walker*, ta có:

$$\left. \begin{array}{l} \emptyset_1 + r_1\emptyset_2 + r_2\emptyset_3 + \dots + r_{p-1}\emptyset_p = r_1 \\ r_1\emptyset_1 + \emptyset_2 + r_1\emptyset_3 + \dots + r_{p-2}\emptyset_p = r_2 \\ \dots \\ r_{p-1}\emptyset_1 + r_{p-2}\emptyset_2 + r_{p-3}\emptyset_3 + \dots + \emptyset_p = r_p \end{array} \right\} \quad (4)$$

Sau đó giải các phương trình tuyến tính này. Để đơn giản có thể dùng phương pháp đệ quy của phương trình Durbin-Levinson, nhưng không tốt về vấn đề làm tròn nếu giá trị làm tròn gần với giá trị ước lượng. Các ước tính thu được theo cách này cũng được gọi là ước tính Yule-Walker.

- Áp dụng phương pháp ước lượng tham số Moment vào mô hình trung bình di động (MA(q) - *Moving average mode*)

Phương pháp ước lượng *Moments* này không hỗ trợ tốt cho mô hình trung bình di động. Hãy xem xét trường hợp  $MA(1)$  đơn giản:

$$p_1 = -\frac{\theta}{1 + \theta^2} \quad (5)$$

Tương đương  $p_1$  với  $r_1$ , giải phương trình bậc 2 theo  $\theta$ . Nếu  $|r_1| < 0.5$ , sau đó hai giá trị gốc được đưa ra bởi

$$-\frac{1}{2r_1} \pm \sqrt{\frac{1}{4r_1^2} - 1}$$

Hai giá trị luôn bằng 1, do đó chỉ có một giá trị thỏa mãn điều kiện  $|\theta| < 1$ .

Sau khi thao tác đại số hơn nữa, chúng ta thấy rằng giải pháp khả nghịch có thể được viết là

$$\hat{\theta} = \frac{-1 + \sqrt{1 - 4r_1^2}}{2r_1} \quad (6)$$

Nếu  $r_1 = \pm 0.5$ , duy nhất các giải pháp tồn tại, cụ thể là  $\mp 1$ , nhưng không thể đảo ngược. Nếu  $|r_1| > 0.5$  (điều này chắc chắn là có thể mặc dù  $|p_1| < 0.5$ ), không có giải pháp thực sự nào tồn tại, và vì vậy phương pháp ước lượng *Moments* không mang lại một ước lượng của  $\theta$ . Dĩ nhiên, nếu  $|r_1| > 0.5$ , đặc điểm kỹ thuật của mô hình  $MA(1)$  sẽ bị nghi ngờ đáng kể.

Đối với các mô hình  $MA$  bậc cao, phương pháp ước lượng *Moments* rất phức tạp. Phương pháp ước lượng *Moments* ít dùng cho mô hình  $MA$ , vì các giá trị ước lượng tạo ra không tốt.

- Áp dụng phương pháp ước lượng tham số Moment vào mô hình kết hợp  $ARMA(p,q)$

Xem xét trường hợp  $ARMA(1,1)$

$$p_k = \frac{(1 - \theta\phi)(\phi - \theta)}{1 - 2\theta\phi + \theta^2} \phi^{k-1} \quad \text{for } k \geq 1 \quad (7)$$

Lưu ý  $p_2 / p_1 = \phi$ , ta có thể ước lượng  $\phi$

$$\hat{\phi} = \frac{r_2}{r_1}$$

Như vậy ta có thể sử dụng

$$r_1 = \frac{(1 - \theta\hat{\phi})(\hat{\phi} - \theta)}{1 - 2\theta\hat{\phi} + \theta^2} \quad (8)$$

Để giải quyết  $\hat{\phi}$ . Lưu ý khi giải phương trình bậc 2 và thành phần khả nghịch nếu có phải giữ lại

### 3.2.4.2. Phương pháp ước lượng bình phương nhỏ nhất (*Least Squares Estimation*)

- Áp dụng phương pháp ước lượng tham số bình phương nhỏ nhất vào mô hình tự hồi quy (AR(p) - *Autoregressive Models*)

Theo [9], xem xét trường hợp sau:

$$Y_t - \mu = \phi(Y_{t-1} - \mu) + e_t \quad (9)$$

Chúng ta có thể xem đây là mô hình hồi quy với biến dự đoán  $Y_{t-1}$  và biến trả lời  $Y_t$ . Ước tính bình phương nhỏ nhất sau đó tiến hành bằng cách tối thiểu hóa tổng bình phương của sự khác biệt.

$$(Y_t - \mu) - \phi(Y_{t-1} - \mu)$$

Vì chỉ quan sát thấy  $Y_1, Y_2, \dots, Y_n$ , nên chỉ có thể tính tổng từ  $t = 2$  đến  $t = n$ .

Để cho

$$S_c(\phi, \mu) = \sum_{t=2}^n [(Y_t - \mu) - \phi(Y_{t-1} - \mu)]^2 \quad (10)$$

Điều này thường được gọi là hàm tổng bình phương có điều kiện. Theo nguyên tắc bình phương nhỏ nhất, ta ước tính  $\phi$  và  $\mu$  bằng các giá trị tương ứng làm giảm thiểu  $S_c(\phi, \mu)$  cho các giá trị quan sát của  $Y_1, Y_2, \dots, Y_n$ .

Xét phương trình  $\partial S_c / \partial \mu = 0$ , ta có

$$\frac{\partial S_c}{\partial \mu} = \sum_{t=2}^n 2[(Y_t - \mu) - \phi(Y_{t-1} - \mu)](-1 + \phi) = 0 \quad (11)$$

Hoặc đơn giản hóa và tính theo  $\mu$

$$\mu = \frac{1}{(n-1)(1-\phi)} \left[ \sum_{t=2}^n Y_t - \phi \sum_{t=2}^n Y_{t-1} \right] \quad (12)$$

Khi  $n$  lớn:

$$\frac{1}{(n-1)} \sum_{t=2}^n Y_t \approx \frac{1}{(n-1)} \sum_{t=2}^n Y_{t-1} \approx \bar{Y}$$

Do đó, bát kề giá trị của  $\emptyset$ , phương trình sẽ là:

$$\hat{\mu} \approx \frac{1}{(1-\emptyset)} (\bar{Y} - \emptyset \bar{Y}) = \bar{Y} \quad (13)$$

Đôi khi,  $\hat{\mu} = \bar{Y}$

Bây giờ hãy xem xét việc giảm thiểu  $S_c(\emptyset, \bar{Y})$  đối với  $\emptyset$ . Ta có

$$\frac{\partial S_c(\emptyset, \bar{Y})}{\partial \emptyset} = \sum_{t=2}^n 2[(Y_t - \bar{Y}) - \emptyset(Y_{t-1} - \bar{Y})](Y_{t-1} - \bar{Y}) \quad (14)$$

Đặt giá trị này =0 và  $\emptyset$  được tính như sau:

$$\hat{\emptyset} = \frac{\sum_{t=2}^n (Y_t - \bar{Y}) - (Y_{t-1} - \bar{Y})}{\sum_{t=2}^n (Y_{t-1} - \bar{Y})^2} \quad (15)$$

Ngoại trừ một thuật ngữ bị thiếu trong mẫu số, cụ thể  $(Y_{t-1} - \bar{Y})^2$ , điều này giống như  $r_1$ . Trong trường hợp quá trình dừng, thì phương pháp ước lượng bình phương nhỏ nhất và phương pháp ước lượng *Moments* gần như giống hệt nhau, đặc biệt là đối với mẫu lớn.

Đối với mô hình *AR(p)*, phương pháp cũng được áp dụng, cụ thể  $\hat{\mu} = \bar{Y}$

Để khái quát hóa việc ước tính  $\emptyset$ , ta xem xét mô hình bậc hai, thay thế bằng hàm tổng bình phương có điều kiện, vì vậy

$$S_c(\emptyset_1, \emptyset_2, \bar{Y}) = \sum_{t=3}^n [(Y_t - \bar{Y}) - \emptyset_1(Y_{t-1} - \bar{Y}) - \emptyset_2(Y_{t-2} - \bar{Y})]^2 \quad (16)$$

Đặt  $\partial S_c / \partial \emptyset_1 = 0$ , ta có

$$-2 \sum_{t=3}^n [(Y_t - \bar{Y}) - \emptyset_1(Y_{t-1} - \bar{Y}) - \emptyset_2(Y_{t-2} - \bar{Y})] (Y_{t-1} - \bar{Y}) = 0$$

Ta có thể viết lại như sau

$$\sum_{t=3}^n (Y_t - \bar{Y})(Y_{t-1} - \bar{Y}) = \left( \sum_{t=3}^n (Y_{t-1} - \bar{Y})^2 \right) \emptyset_1 + \left( \sum_{t=3}^n (Y_{t-1} - \bar{Y})(Y_{t-2} - \bar{Y}) \right) \emptyset_2 \quad (17)$$

Tổng các thành phần lõi  $\sum_{t=3}^n (Y_t - \bar{Y})(Y_{t-1} - \bar{Y})$  thì gần bằng tử số của  $r_1$  – ta thiếu một thành phần  $(Y_2 - \bar{Y})(Y_1 - \bar{Y})$ , nhưng ta đang thiếu  $(Y_n - \bar{Y})(Y_{n-1} - \bar{Y})$ .

$\bar{Y}$ ). Nếu ta chia cả hai phương trình cho  $\sum_{t=3}^n (Y_t - \bar{Y})^2$ , sau đó loại trừ các ảnh hưởng không đáng kể, ta có:

$$r_1 = \phi_1 + r_1 \phi_2 \quad (18)$$

Sắp sỉ tương tự với phương trình  $\partial S_c / \partial \phi_2 = 0$ , dẫn đến

$$r_2 = r_1 \phi_1 + \phi_2$$

Nhưng phương trình trên chỉ là phương trình mẫu Yule-Walker cho mô hình AR(2).

Kết quả hoàn toàn tương tự cho trường hợp AR(p): Đến một xấp xỉ xuất sắc, ước tính bình phương tối thiểu có điều kiện của  $\phi$  được tính bằng cách giải các phương trình Yule-Walker mẫu.

- Áp dụng phương pháp ước lượng tham số bình phương nhỏ nhất vào mô hình trung bình di động (MA(q) - *Moving average mode*)

Theo [9], hãy xem xét ước lượng bình phương nhỏ nhất của  $\theta$  trong mô hình MA(1):

$$Y_t = e_t - \theta e_{t-1}$$

Phương trình của MA(1) được biểu diễn là:

$$Y_t = -\theta Y_{t-1} - \theta^2 Y_{t-2} - \theta^3 Y_{t-3} - \dots + e_t \quad (19)$$

Một mô hình hồi quy vô hạn. Do đó, bình phương tối thiểu có thể được thực hiện bằng cách chọn giá trị  $\theta$  nhỏ nhất.

$$S_c(\theta) = \sum (e_t)^2 = \sum [Y_t + \theta Y_{t-1} + \theta^2 Y_{t-2} + \theta^3 Y_{t-3} + \dots]^2 \quad (20)$$

Trong đó, ngầm định,  $e_t = e_t(\theta)$  là một hàm của chuỗi quan sát và tham số  $\theta$  chưa biết.

Để giải vấn đề này, hãy xem xét đánh giá  $S_c(\theta)$  cho một giá trị đã cho là  $\theta$ . Với  $Y$ , ta có các quan sát  $Y_1, Y_2, \dots, Y_n$ , phương trình được viết lại như sau

$$e_t = Y_t + \theta e_{t-1}$$

Sử dụng phương trình đệ quy để tính toán  $e_1, e_2, \dots, e_n$  nếu ta có giá trị ban đầu  $e_0$ . Giá trị sắp sỉ mong đợi khi đặt  $e_0 = 0$ , điều kiện trên  $e_0 = 0$ . Ta có

$$\left. \begin{array}{l} e_1 = Y_1 \\ e_2 = Y_2 + \theta e_1 \\ e_3 = Y_3 + \theta e_2 \\ \vdots \\ e_n = Y_n + \theta e_{n-1} \end{array} \right\} \quad (21)$$

Và tính toán  $S_c(\emptyset) = \sum_{t=2}^n (e_t)^2$ , với điều kiện  $e_0 = 0$ , và một giá trị  $\theta$  đã cho.

Đối với trường hợp đơn giản, ta có thể chọn giá trị bình phương nhỏ nhất cho tham số  $\theta$  trong khoảng (-1,1). Đối với trường hợp tổng quát MA(q), ta cần sử dụng các thuật toán tối ưu như, Gauss-Newton hoặc NelderMead.

Tương tự cho các mô hình có bậc cao hơn. Ta tính  $e_t = e_t(\theta_1, \theta_2, \dots, \theta_q)$ , từ:

$$e_t = Y_t + \theta_1 e_{t-1} + \theta_2 e_{t-2} + \dots + \theta_q e_{t-q} \quad (22)$$

Với  $e_0 = e_{-1} = \dots = e_{-q} = 0$ . Tổng bình phương nhỏ nhất trong  $\theta_1, \theta_2, \dots, \theta_q$  được sử dụng phương pháp đa biến.

- Áp dụng phương pháp ước lượng tham số bình phương nhỏ nhất vào mô hình kết hợp ARMA(p,q)

Theo [9], xét trường hợp ARMA(1,1)

$$Y_t = \emptyset Y_{t-1} + e_t - \emptyset e_{t-1} \quad (23)$$

Trong trường hợp MA thuần túy, xét  $e_t = e_t(\emptyset, \theta)$ , và  $S_c(\emptyset, \theta) = \sum e_t^2$  nhỏ nhất. ta có thể viết lại phương trình:

$$e_t = Y_t - \emptyset Y_{t-1} + \theta e_{t-1}$$

Để có được  $e_1$ , đặt  $Y_0 = 0$  hoặc  $\bar{Y}$  có giá trị trung bình khác. Tuy nhiên, ta có thể để quy từ giá trị  $t = 2$ , để tránh  $Y_0$ , và giảm xuống nhỏ nhất.

$$S_c(\emptyset, \theta) = \sum_{t=2}^n e_t^2 \quad (24)$$

Mô hình ARMA(p,q) được tính như sau:

$$e_t = Y_t - \emptyset_1 Y_{t-1} - \emptyset_2 Y_{t-2} - \dots - \emptyset_p Y_{t-p} + \theta_1 e_{t-1} + \theta_2 e_{t-2} + \dots + \theta_q e_{t-q} \quad (25)$$

Với  $e_p = e_{p-1} = \dots = e_{p+1-q} = 0$ , và khi  $S_c(\emptyset_1, \emptyset_2, \dots, \emptyset_p, \theta_1, \theta_2, \dots, \theta_q)$ , thì các tham số được ước lượng bằng phương pháp bình phương nhỏ nhất.

Đối với các tham số  $\theta_1, \theta_2, \dots, \theta_q$  tương ứng với các mô hình, các giá trị  $e_p, e_{p+1}, \dots, e_{p+1-q}$  rất ít ảnh hưởng đến các ước tính cuối cùng của các thông số cho mẫu lớn.

### **3.3. Mô hình lai ghép giữa ARIMA và RBFNN cho bài toán dự báo trên chuỗi thời gian**

Theo [6] tác giả *L. Zhang và cộng sự*, xem xét một chuỗi thời gian ( $y_t$ ) được cấu thành từ cấu trúc tự tương quan tuyến tính ( $L_t$ ) và thành phần phi tuyến ( $N_t$ ). Đó là:

$$y_t = L_t + N_t$$

Tác giả dự đoán chuỗi thời gian bằng mô hình lai ARIMA và RBFNN như sau:

+ Mô hình ARIMA (*Box et al., 1994*) đã được sử dụng để dự đoán  $y_t$  và để cho  $\hat{L}_t$  biểu thị kết quả dự đoán. Các  $e_t$  là phần dư giữa các chuỗi của mô hình ARIMA.

$$e_t = y_t - \hat{L}_t$$

+  $e_t$  được coi là đầu vào của mô hình RBFNN (*Moody and Darken, 1989*), sau đó mô hình RBFNN có thể được biểu thị như sau:

$$e_t = f(e_{t-1}, e_{t-2}, \dots, e_{t-n}) + \varepsilon_t$$

Trong đó,  $f$  là hàm phi tuyến được xác định bởi mạng nơ ron và  $\varepsilon_t$  là lỗi ngẫu nhiên.

Kết quả đầu ra của RBFNN được định nghĩa là  $\hat{N}_t$ .

+ Hai mô hình được kết hợp để dự báo và kết quả dự đoán từ các mô hình lai ARIMA RBFNN được biểu thị như sau:

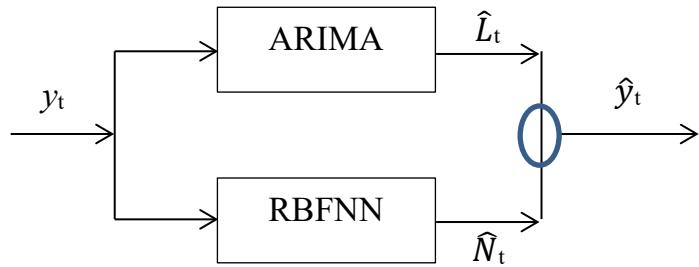
$$\hat{y}_t = \hat{L}_t + \hat{N}_t$$

Vì vậy, kết quả dự đoán được tạo ra thông qua mô hình lai ARIMA-RBFNN đã thu được thông qua sự kết hợp giữa dự đoán tuyến tính của ARIMA và kết quả dự đoán phi tuyến được dự đoán bởi mô hình RBFNN (through qua phần lỗi của mô hình ARIMA).

### **3.4. Nghiên cứu cải tiến mô hình lai ghép bằng cách thực hiện song song hai mô hình ARIMA và RBFNN**

Trong các mô hình lai ARIMA-RBFNN hiện nay, đa số các tác giả thường phân tích chuỗi thời gian thành 2 thành phần, tuyến tính và phi tuyến tính, sau đó sử dụng mô hình ARIMA để dự báo trên chuỗi thời gian, kết quả của mô hình ARIMA sẽ gồm 2 phần, phần kết quả dự báo và phần lỗi (Thành phần phi tuyến), phần lỗi này tiếp tục được sử dụng để dự báo bằng mô hình RBFNN. Kết quả cuối cùng các tác giả sử dụng phết cộng hoặc nhân hai kết quả của hai mô hình.

Tuy nhiên, hiện tại các tác giả đang thực hiện tuần tự từng mô hình, sau đó gộp kết quả lại. Để kiểm nghiệm về thời gian cũng như kết quả dự báo, chúng ta sử dụng mô hình lai cải tiến bằng cách thực hiện song song hai mô hình, từ kết quả đạt được, chúng ta sẽ xem xét để đề xuất mô hình tốt hơn.



Hình 3.3: Sơ đồ mô phỏng mô hình lai ARIMA - RBFNN

Gọi  $\hat{L}_t$  là giá trị dự báo của mô hình ARIMA,  $\hat{N}_t$  là giá trị dự báo của mô hình RBFNN, giá trị dự báo của  $y$  được tính như sau:

$$\hat{y} = \alpha \hat{L}_t + (1 - \alpha) \hat{N}_t \quad \alpha \in (0,1)$$

Để xác định tham số trọng số  $\alpha$ , chúng ta sẽ tìm giá trị của  $\alpha$  để hệ số dự báo lỗi MSE là nhỏ nhất.

$$MSE = \sum_{i=1}^n (Y_i - Y_{hybrid,i})^2 = \sum_{i=1}^n (Y_i - [\alpha Y_{NN,i} + (1 - \alpha) Y_{DTW,i}])^2$$

Trong đó  $Y_i$  là giá trị thực tế tại thời điểm  $i$ ,  $Y_{NN,i}$  là giá trị dự báo tại thời điểm  $i$  được tạo bởi ANN và  $Y_{DTW,i}$  là giá trị dự báo tại thời điểm  $i$  được tạo bởi khớp mẫu trong  $DTW$ . Đây là một hàm bậc hai, do đó chúng ta có thể rút ra giá trị của  $\alpha$  làm cho lỗi dự báo  $MSE$  nhỏ nhất như sau:

$$\alpha = \frac{\sum_{i=1}^n (Y_{NN,i} - Y_{DTW,i})(Y_i - Y_{DTW,i})}{\sum_{i=1}^n (Y_{NN,i} - Y_{DTW,i})^2}$$

Vì  $\alpha$  nằm trong phạm vi  $[0, 1]$ , nếu giá trị tính toán của  $\alpha$  là âm, chúng ta có thể chọn giá trị của nó là 0 và nếu giá trị tính toán của lớn hơn 1, chúng ta có thể chọn giá trị của nó là 1.

Để thấy được hiệu quả của bài toán dự báo trên chuỗi thời gian, các khái niệm về chuỗi thời gian, các bài toán về dự báo trên chuỗi thời gian, cũng như các mô hình thường dùng trong dự báo chuỗi thời gian đã được giới thiệu trong chương này. Tổng quan các nghiên cứu về bài toán dự báo của một số tác giả cũng được nêu trong chương 2. Tuy nhiên, mỗi tác giả đều đưa ra các mô hình dự báo khác nhau và chủ yếu tập trung vào từng dữ liệu chuyên biệt, ít có tác giả nghiên cứu chung cho nhiều dữ liệu chuỗi thời gian, do đó vẫn đề so sánh kết quả giữa các mô hình còn hạn chế. Trong chương 4, chúng tôi sẽ thực nghiệm mô hình lai cải tiến ARIMA và RBFNN, nhằm mục đích cải thiện thời gian và độ chính xác của dự báo và đưa ra các đề xuất cho mô hình lai này.

## Chương 4

# ĐÁNH GIÁ THỰC NGHIỆM

*Mô hình tự hồi quy kết hợp với trung bình di động ARIMA (AutoRegressive Integrated Moving Average) đã được áp dụng thành công như một mô hình tuyến tính phổ biến để dự báo chuỗi thời gian. Ngoài ra, trong những năm gần đây, mạng nơ-ron nhân tạo truyền thống (RBFNN) đã được sử dụng để nắm bắt phức hợp các mối quan hệ với nhiều kiểu mẫu khác nhau vì chúng đóng vai trò như một công cụ tính toán mạnh mẽ và linh hoạt. Để cải thiện kết quả tốt hơn, L. Zhang và các cộng sự đã tích hợp các ưu điểm của ARIMA và RBFNN trong việc mô hình hóa các hành vi tuyến tính và phi tuyến tính trong tập dữ liệu bằng cách thực hiện tuần tự hai mô hình sau đó thực hiện phép cộng 2 kết quả dự báo. Kết quả cho thấy mô hình kết hợp này hiệu quả hơn kết quả của từng mô hình. Tuy nhiên, chúng tôi muốn cải tiến mô hình này bằng cách thực hiện song song, nhằm mục đích cải thiện thời gian thực hiện và kết quả của dự báo.*

*Mô hình kết hợp cải tiến này được thực nghiệm trên bốn bộ dữ liệu thực tế. Qua phần thực nghiệm, tính toán chúng tôi sẽ so sánh kết quả dự báo của mô hình kết hợp cải tiến so với các mô hình hiện có.*

### 4.1. Môi trường và dữ liệu thực nghiệm

Thực hiện đánh giá thực nghiệm trên máy tính Dell Inspiron 15, Inter® core™ i5-5300U CPU @ 2.3 GHz, 16 GB RAM, hệ điều hành Windows 10. Sử dụng phần mềm **QT Designer** để tạo giao diện chương trình và dùng ngôn ngữ lập trình Python 3.77 để thực nghiệm chương trình.

Tập dữ liệu sử dụng để thực nghiệm gồm: Dữ liệu *AirPassengers*, *Sunspots*, *Dentists*, *City\_temperature*. Các bộ dữ liệu này được cộng đồng mạng về khai phá dữ liệu công bố [15].

- Tập dữ liệu *AirPassengers.csv*: Tập dữ liệu này cung cấp tổng số hành khách hàng tháng của một hãng hàng không Hoa Kỳ từ năm 1949 đến năm 1960, tập dữ liệu này được lấy từ một tập dữ liệu có sẵn của R và được dùng cho mục đích nghiên cứu.

- Tập dữ liệu *Sunspots.csv*: Tổng số vết đen Mặt trời trung bình hàng tháng, từ 01/01/1749 đến 31/08/2017. Cơ sở dữ liệu từ SIDC - Trung tâm Phân tích Dữ liệu

Ảnh hưởng Mặt trời - bộ phận nghiên cứu vật lý mặt trời của Đài quan sát Hoàng gia Bỉ. Trang web SIDC

- Tập dữ liệu *Dentists.csv*: Số lượng các nha sĩ có sẵn trên 10.000 dân số. Tập dữ liệu này nằm trong tập dữ liệu bao gồm các thông kê y tế trên thế giới cập nhật gần đây nhất (gồm các quốc gia được WHO công nhận).

- Tập dữ liệu *City\_temperature.csv*: Tập dữ liệu về nhiệt độ trung bình hàng ngày của các thành phố lớn trên thế giới, tập dữ liệu này được cung cấp bởi trường Đại học Dayton, và chỉ dùng cho mục đích nghiên cứu và phi thương mại.

#### 4.2. Tiêu chí đánh giá.

Một mô hình dự báo được đánh giá tốt khi sai số dự báo nhỏ. Ngoài ra tính ngẫu nhiên của sai số cũng là một tham số quan trọng để đánh giá độ chính xác của dự báo.

Khi tiến hành dự báo người ta thường giả định dữ liệu ban đầu ngẫu nhiên; các tính toán, đánh giá, kiểm định cũng đều dựa trên giả định này (ngẫu nhiên, phân phối chuẩn) nên nếu mô hình đúng thì sai số cũng phải không theo một chiều hướng nào cả.

Các tiêu chí đánh giá thường được sử dụng trong thực tế dự báo như sau:

- Thời gian thực thi sẽ tính từ thời điểm bắt đầu thực hiện mô hình và kết thúc khi tìm được dự báo. Sẽ bỏ qua thời gian tiền xử lý dữ liệu như load dữ liệu hay kiểm tra dữ liệu.

- Sai số tuyệt đối trung bình MAE (*Mean Absolute Error*): MAE đo độ lớn trung bình của các lỗi trong một tập hợp các dự đoán mà không cần xem xét hướng của chúng. Đó là giá trị trung bình trên mẫu thử nghiệm về sự khác biệt tuyệt đối giữa dự đoán và quan sát thực tế, trong đó tất cả các khác biệt riêng lẻ có trọng số bằng nhau.

- Sai số bình phương trung bình gốc RMSE (*Root Mean Squared Error*): RMSE là một quy tắc tính điểm bậc hai cũng đo độ lớn trung bình của lỗi. Đó là căn bậc hai của trung bình của sự khác biệt bình phương giữa dự đoán và quan sát thực tế.

### 4.3. Các trường hợp thực nghiệm

Trong phần thực nghiệm này, em chia tập dữ liệu thành 2 phần, 66% tập dữ liệu dùng cho training và 34% của tập dữ liệu dùng cho testing. Tuy nhiên tỷ lệ này có thể thay đổi cho các tập dữ liệu khác nhau. (trong demo cho phép tự chọn tỷ lệ cho hai phần training và testing).

Để chọn tham số cho mô hình ARIMA, trong thực nghiệm em dùng phương pháp bình phương nhỏ nhất để ước lượng các tham số. Trong ngôn ngữ *Python* em dùng một vòng lặp và hàm *fit()* để tính chỉ số *RMSE* sau đó so sánh các chỉ số đã tìm được để tìm ra chỉ số *RMSE* nhỏ nhất, và các tham số được dùng trong mô hình ARIMA khi chỉ số *RMSE* nhỏ nhất.

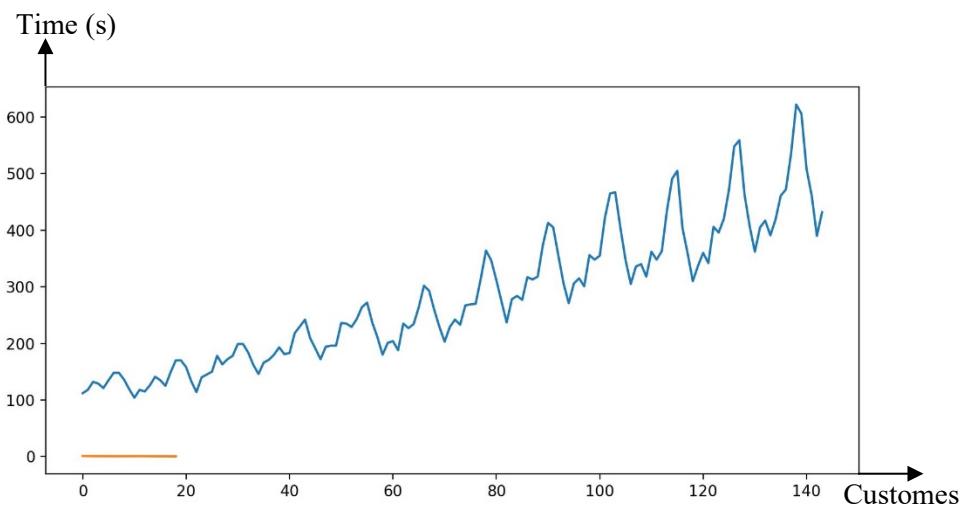
Đối với mô hình RBFNN, số lượng các nút trong lớp đầu vào và lớp ẩn có ảnh hưởng sâu sắc đến hiệu suất của RBFNN. Theo [16] nếu sử dụng quá ít nút thì mối quan hệ trong dữ liệu chưa được xác định một cách chính xác nhất và số lượng nút quá lớn sẽ làm tăng độ phức tạp của mạng và thời gian trong quá trình thực thi, người ta thấy rằng số lượng nút tối ưu trong một lớp thường được tính bằng phương pháp thử và sai (*Train and error*) [16,17]. Trong đề tài luận văn em sẽ thực nghiệm hai trường hợp, trường hợp thứ nhất: cố định số nút đầu vào và thay đổi số nút ẩn; trường hợp 2: Cố định số nút ẩn và thay đổi số nút đầu vào (trong thực nghiệm em chọn hai giá trị để thay đổi là 64 và 128).

Bảng 4.1: Các tập dữ liệu dùng trong thực nghiệm

STT	Tập dữ liệu	Độ dài	Chú thích
1	AirPassengers	145	
2	Sunspots	150	
3	Dentists	2135	
4	City temperature	5328	

#### 4.3.1. Trường hợp 1: Cố định số nút đầu vào và thay đổi số nút ẩn

##### 4.3.1.1. Thực nghiệm trên tập dữ liệu AirPassengers.csv



Hình 4.1: Biểu đồ thể hiện dữ liệu của chuỗi thời gian *AirPassengers*

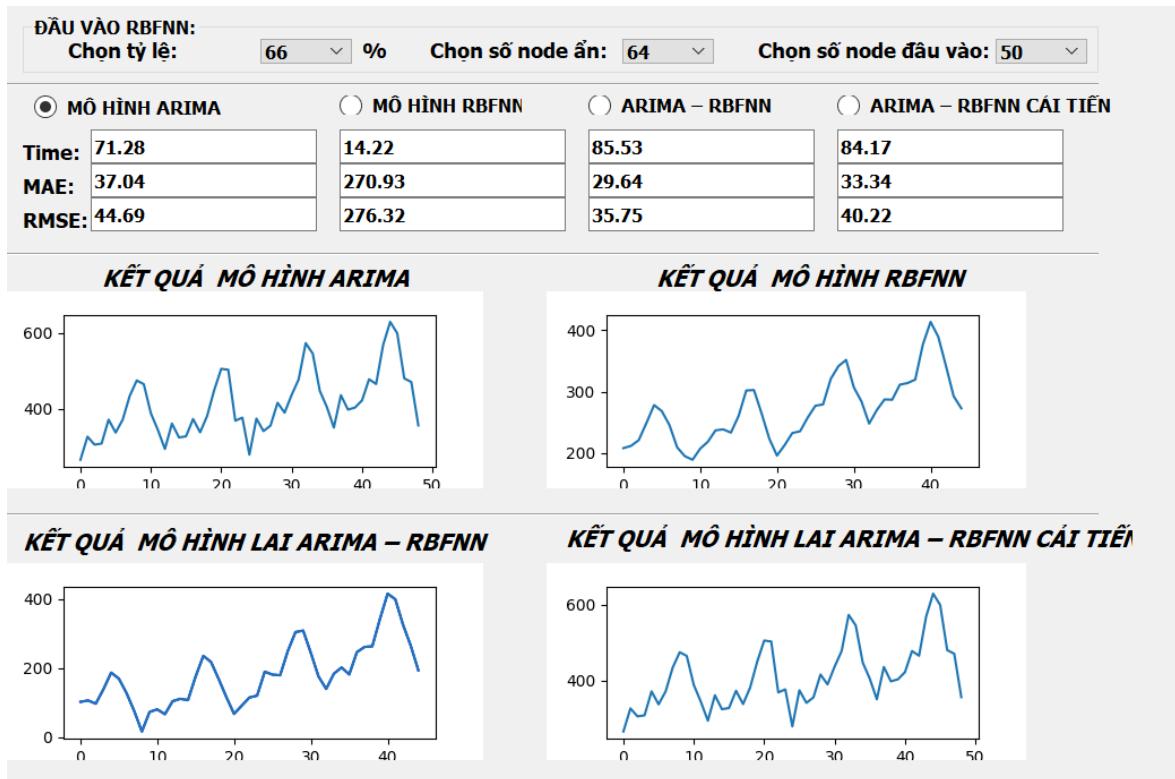
**Trường hợp cố định số nút đầu vào, số nút ẩn trong mô hình RBFNN là 64.**

Qua thực nghiệm trên tập dữ liệu *AirPassenger.csv* kết quả cho trong bảng 4.2 thể hiện thời gian và các chỉ số đánh giá mô hình như sau:

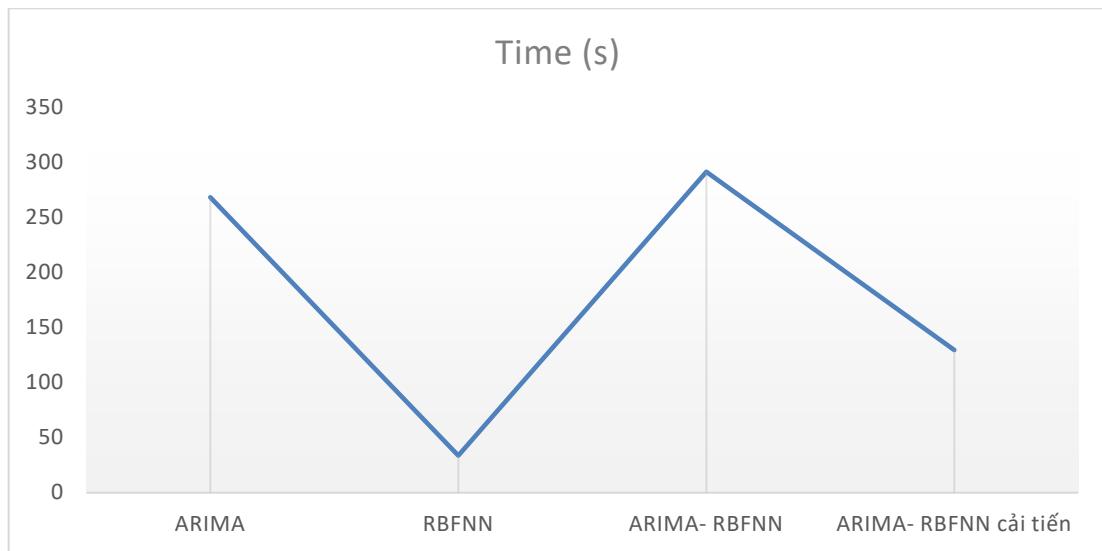
Bảng 4.2: Kết quả thực nghiệm trên tập dữ liệu Aripassanger với 64 nút ẩn

Mô hình	AirPassengers 64		
	Time (s)	RMSE	MAE
ARIMA	268.16	44.81	37.3
RBFNN	33.74	152.07	116.09
ARIMA- RBFNN	291.32	35.85	29.84
ARIMA- RBFNN cài tiến	129.62	40.33	32.54
Tỷ lệ ARIMA- RBFNN cài tiến so với ARIMA- RBFNN	44.49%	112.50%	109.05%
Chênh lệch tỷ số giữa ARIMA- RBFNN cài tiến so với ARIMA- RBFNN	55.51%	-12.50%	-9.05%

Hình 4.2 là kết quả dự báo của bốn mô hình được thể hiện chi tiết qua biểu đồ



Hình 4.2: Biểu đồ thể hiện kết quả dự báo của các mô hình khi dùng 64 nút ẩn  
 Biểu đồ so sánh thời gian thực thi của các mô hình được thể hiện trong hình 4.3



Hình 4.3: Thời gian thực thi của các mô hình khi dùng 64 nút ẩn  
**Nhận xét:** Với kết quả thực nghiệm chạy trên tập dữ liệu *AirPassengers* với số nút ẩn của mô hình RBFNN là 64 cho thấy thời gian thực thi của ARIMA-RBFNN cải

tiến tốt hơn mô hình ARIMA-RBFNN, tuy nhiên xét về mặt chính xác thì chưa được cải thiện cụ thể như trong bảng 4.2 cho thấy, thời gian cải thiện được 55.51%, tuy nhiên độ chính xác lại giảm 12.5% (Bảng 4.2).

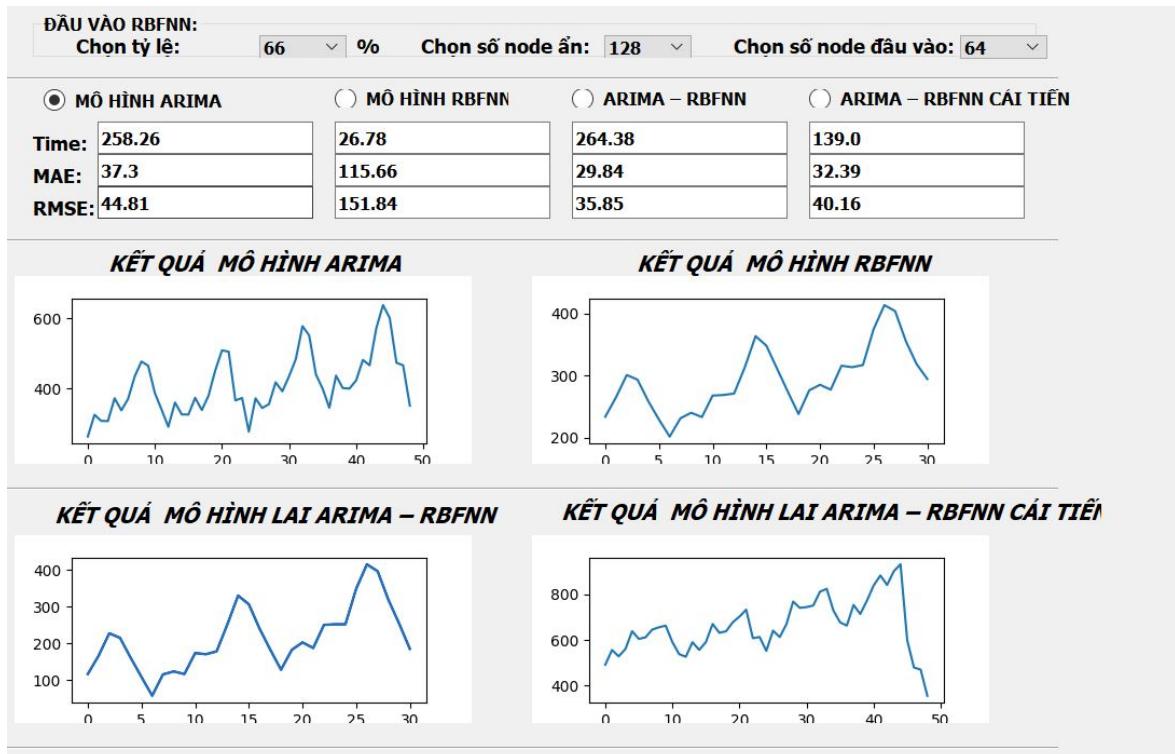
**Trường hợp có định số nút đầu vào, số nút ẩn trong mô hình RBFNN là 128.**

Kết quả thực nghiệm trên tập dữ liệu *AirPassengers.csv* cho trong bảng 4.3 thể hiện thời gian và các chỉ số đánh giá mô hình như sau:

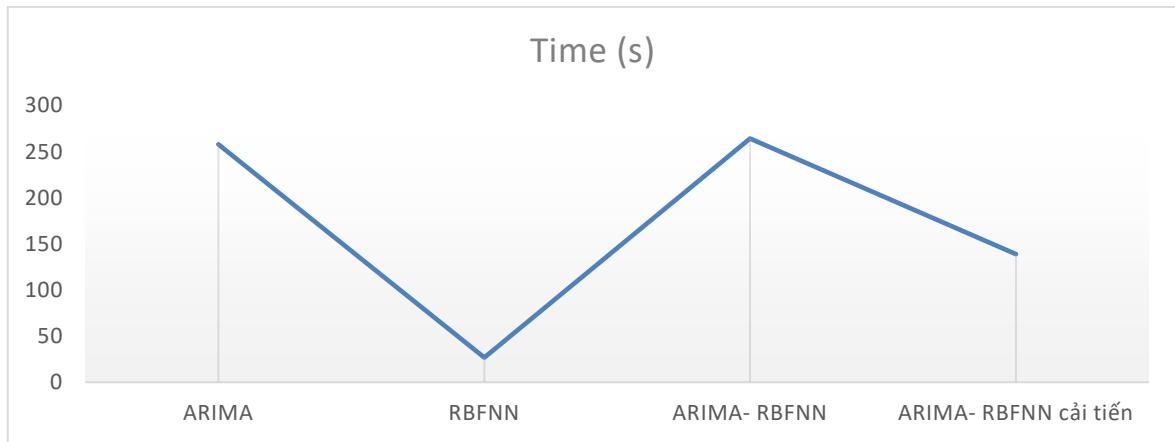
Bảng 4.3: Kết quả thực nghiệm trên tập dữ liệu Aripassanger với 128 nút ẩn

Mô hình	AirPassengers 128		
	Time(s)	RMSE	MAE
ARIMA	258.26	44.81	37.3
RBFNN	26.78	151.84	115.66
ARIMA- RBFNN	264.38	35.85	29.84
ARIMA- RBFNN cải tiến	139.0	40.16	32.39
Tỷ lệ ARIMA- RBFNN cải tiến so với ARIMA- RBFNN	52.58%	112.02%	108.55%
Chênh lệch tỷ số giữa ARIMA- RBFNN cải tiến so với ARIMA- RBFNN	47.42%	-12.02%	-8.55%

Hình 4.4 là kết quả dự báo của bốn mô hình được thể hiện chi tiết qua các biểu đồ sau:



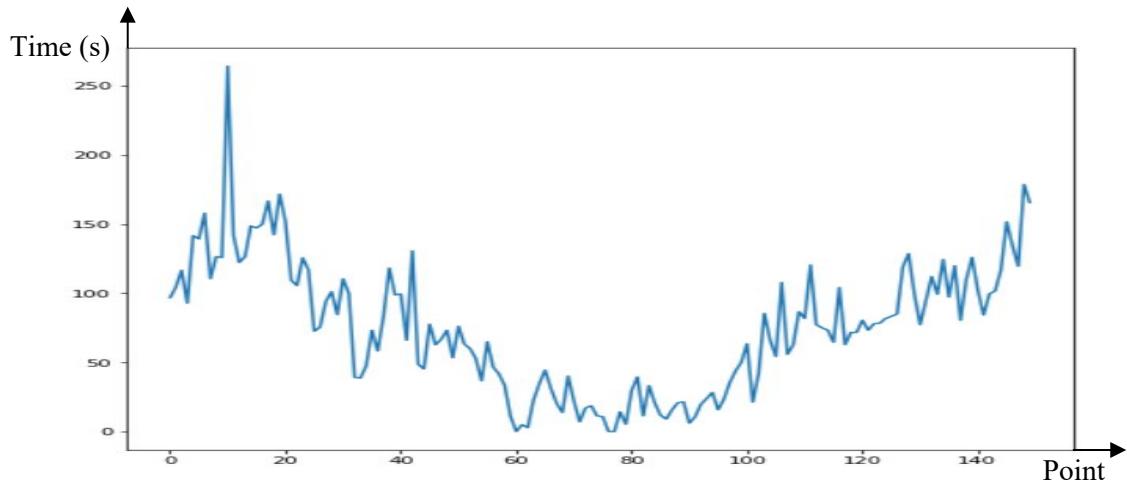
Hình 4.4: Biểu đồ thể hiện kết quả dự báo của các mô hình khi dùng 128 nút ẩn  
 Biểu đồ so sánh thời gian thực thi của các mô hình được thể hiện trong hình 4.4



Hình 4.5: Thời gian thực thi của các mô hình khi dùng 128 nút ẩn

**Nhận xét:** Khi tăng số nút ẩn lên 128 thì mô hình ARIMA không bị ảnh hưởng, trong khi đó mô hình RBFNN được cải thiện không nhiều về kết quả dự báo nhưng thời gian tăng lên đáng kể, hai mô hình ARIMA-RBFNN và mô hình ARIMA-RBFNN cải tiến không thay đổi về kết quả dự báo, nhưng về mặt thời gian thì lại giảm, nguyên nhân có thể do tập dữ liệu nhỏ nên không ảnh hưởng nhiều về kết quả dự báo khi thay đổi số nút ẩn.

#### 4.3.1.2. Thực nghiệm trên tập dữ liệu Sunspots.csv



Hình 4.6: Biểu đồ thể hiện dữ liệu của chuỗi thời gian Sunspots.csv

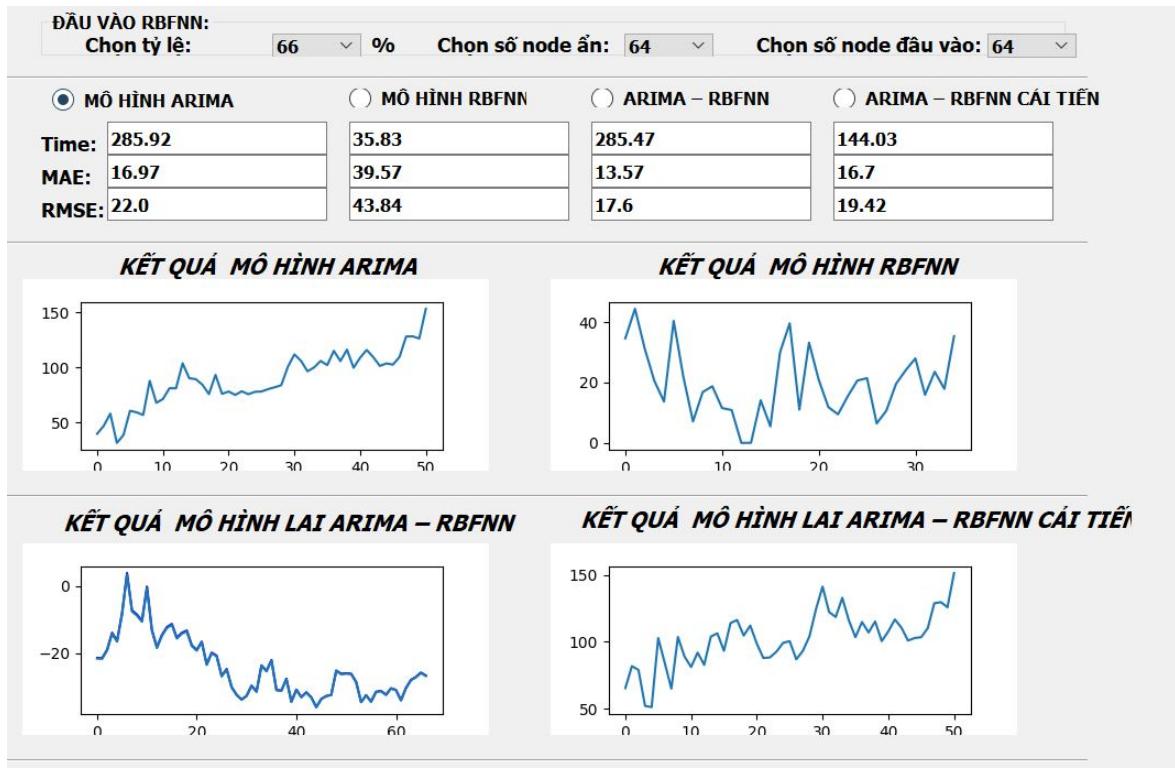
**Thực nghiệm trên tập dữ liệu Sunspots.csv khi sử dụng 64 nút ẩn trong mô hình RBFNN.**

Kết quả thực nghiệm trên tập dữ liệu Sunspots.csv cho trong bảng 4.4 thể hiện thời gian và các chỉ số đánh giá mô hình như sau:

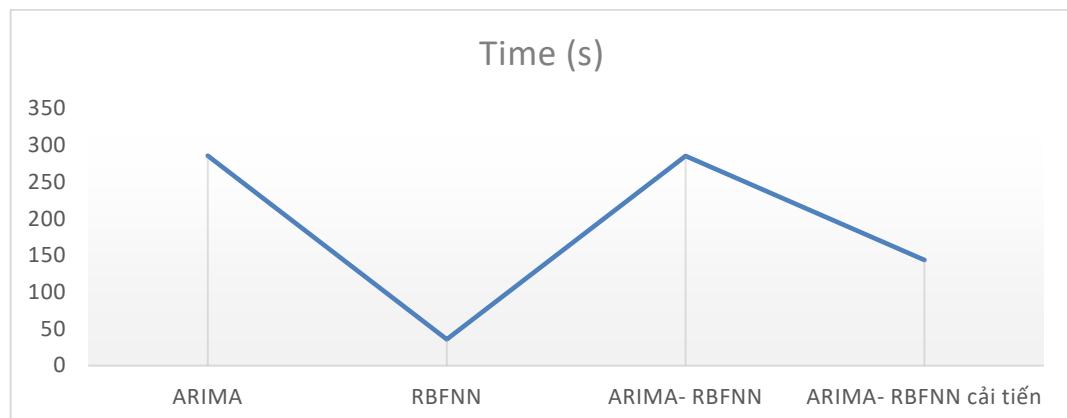
Bảng 4.3: Kết quả thực nghiệm trên tập dữ liệu Sunspots.csv với 64 nút ẩn

Mô hình	Sunspots 64		
	Time (s)	RMSE	MAE
ARIMA	285.92	22.0	16.97
RBFNN	35.83	43.84	39.57
ARIMA- RBFNN	285.47	17.6	13.57
ARIMA- RBFNN cải tiến	144.03	19.42	16.7

Hình 4.7 là kết quả dự báo của bốn mô hình được thể hiện chi tiết qua các biểu đồ sau:



Hình 4.7: Biểu đồ thể hiện kết quả dự báo của các mô hình khi dùng 64 nút ẩn  
Biểu đồ so sánh thời gian thực thi của các mô hình được thể hiện trong hình 4.8



Hình 4.8: Thời gian thực thi của các mô hình khi dùng 64 nút ẩn trên tập dữ liệu  
*Sunspots*

**Nhận xét:** Với kết quả thực nghiệm chạy trên tập dữ liệu *Sunspots* với số nút ẩn của mô hình RBFNN là 64 cho thấy mô hình ARIMA-RBFNN hiệu quả nhất, trong khi đó mô hình ARIMA-RBFNN cải tiến lại cho kết quả không tốt về thời gian thực thi vẫn tốt hơn các mô hình khác, nguyên nhân có thể do tập dữ liệu không theo một xu hướng nhất định.

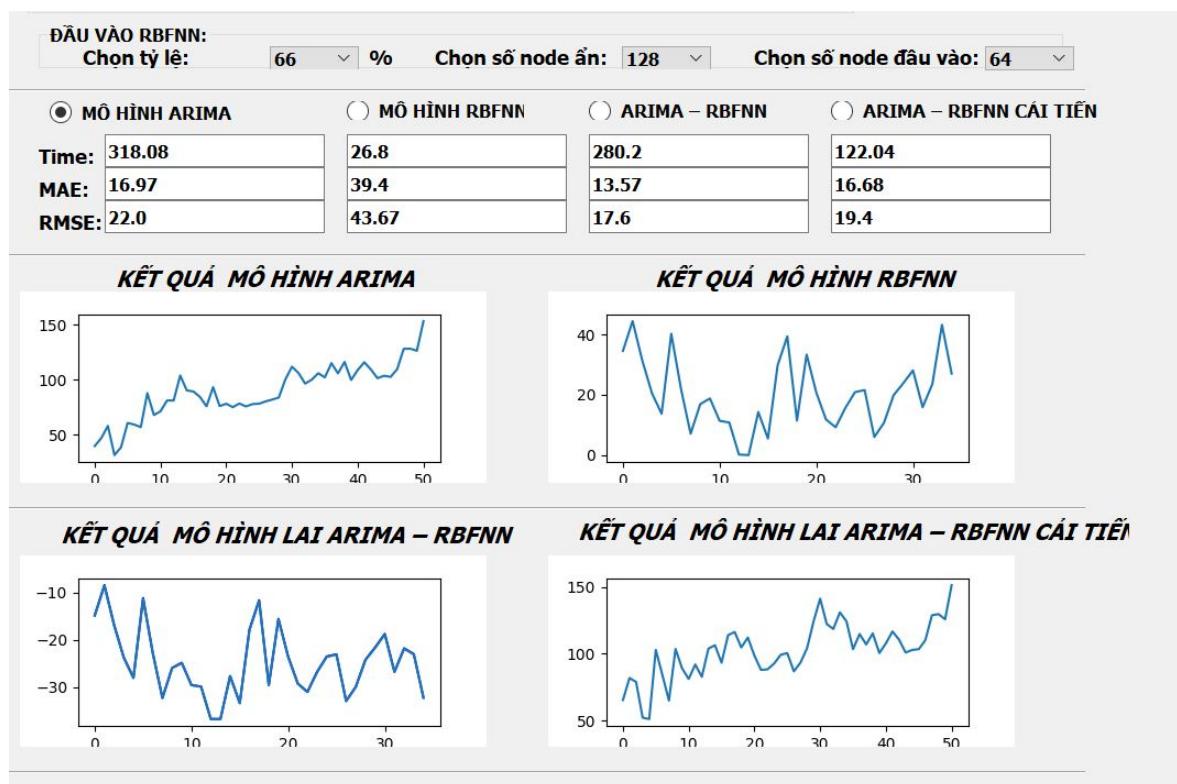
**Thực nghiệm trên tập dữ liệu Sunspots.csv khi sử dụng 128 nút ẩn trong mô hình RBFNN.**

Kết quả thực nghiệm trên tập dữ liệu *Sunspots.csv* cho trong bảng 4.5 thể hiện thời gian và các chỉ số đánh giá mô hình như sau:

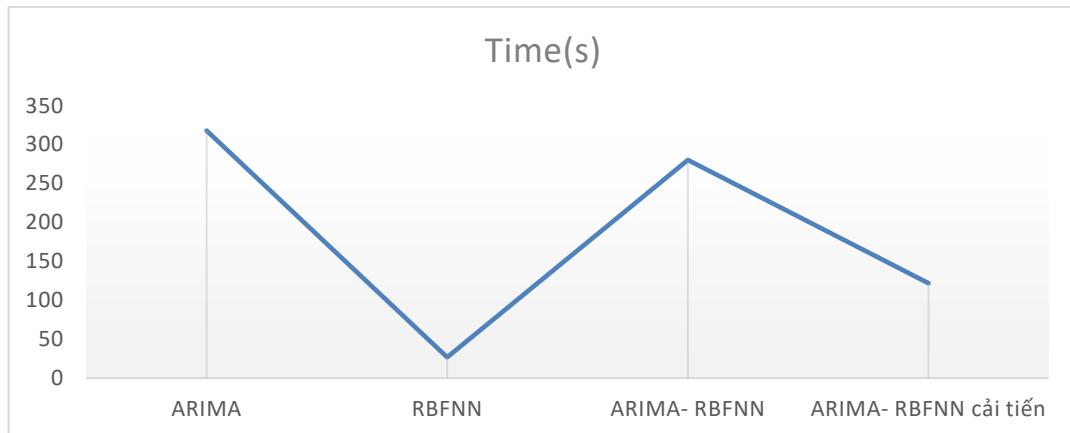
Bảng 4.5: Kết quả thực nghiệm trên tập dữ liệu *Sunspots.csv* với 128 nút ẩn

Mô hình	Sunspots 128		
	Time(s)	RMSE	MAE
ARIMA	318.08	22.0	16.97
RBFNN	26.8	43.67	39.4
ARIMA- RBFNN	280.2	17.6	13.57
ARIMA- RBFNN cải tiến	122.04	19.4	16.68

Hình 4.9 là kết quả dự báo của bốn mô hình dự báo khi sử dụng 128 nút ẩn được thể hiện chi tiết qua các biểu đồ sau:



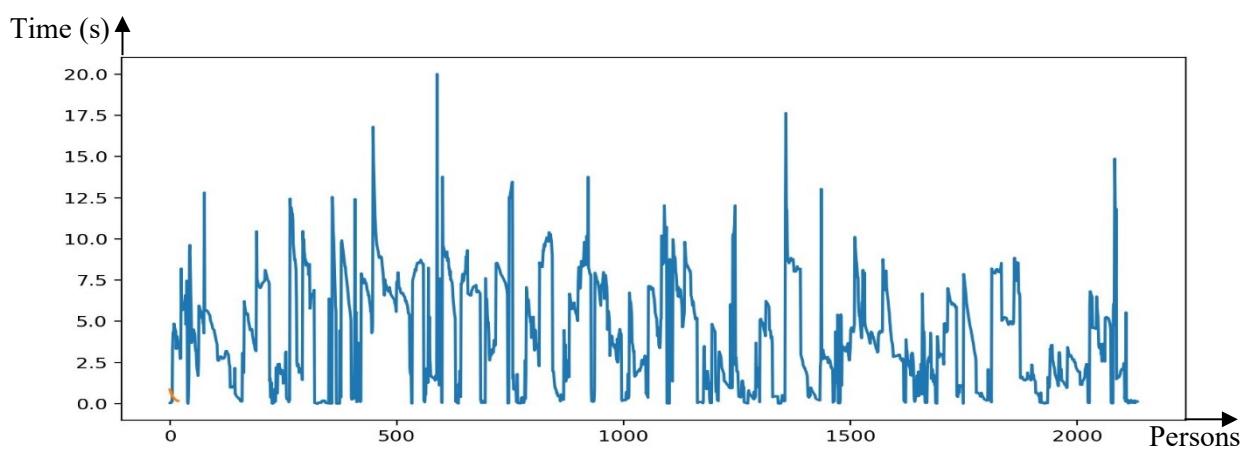
Hình 4.9: Biểu đồ thể hiện kết quả dự báo của các mô hình khi dùng 128 nút ẩn  
Biểu đồ so sánh thời gian thực thi của các mô hình khi sử dụng 128 nút ẩn  
được thể hiện trong hình 4.9



Hình 4.10: Thời gian thực thi của các mô hình dự báo khi dùng 128 nút ẩn trên tập dữ liệu *Sunspots*

**Nhận xét:** Khi sử dụng số nút ẩn là 128 thì mô hình RBFNN cho kết quả dự báo tốt hơn khi dùng 64 nút ẩn, tuy nhiên thời gian lại tăng lên đáng kể, hai mô hình ARIMA-RBFNN và ARIMA-RBFNN cải tiến thì ít có sự thay đổi cả về thời gian và độ chính xác về dự báo của mô hình do tập dữ liệu *Sunspots* nhỏ nên khi thực hiện các kết quả dự báo thay đổi không đáng kể.

#### 4.3.1.3. Thực nghiệm trên tập dữ liệu Dentists.csv



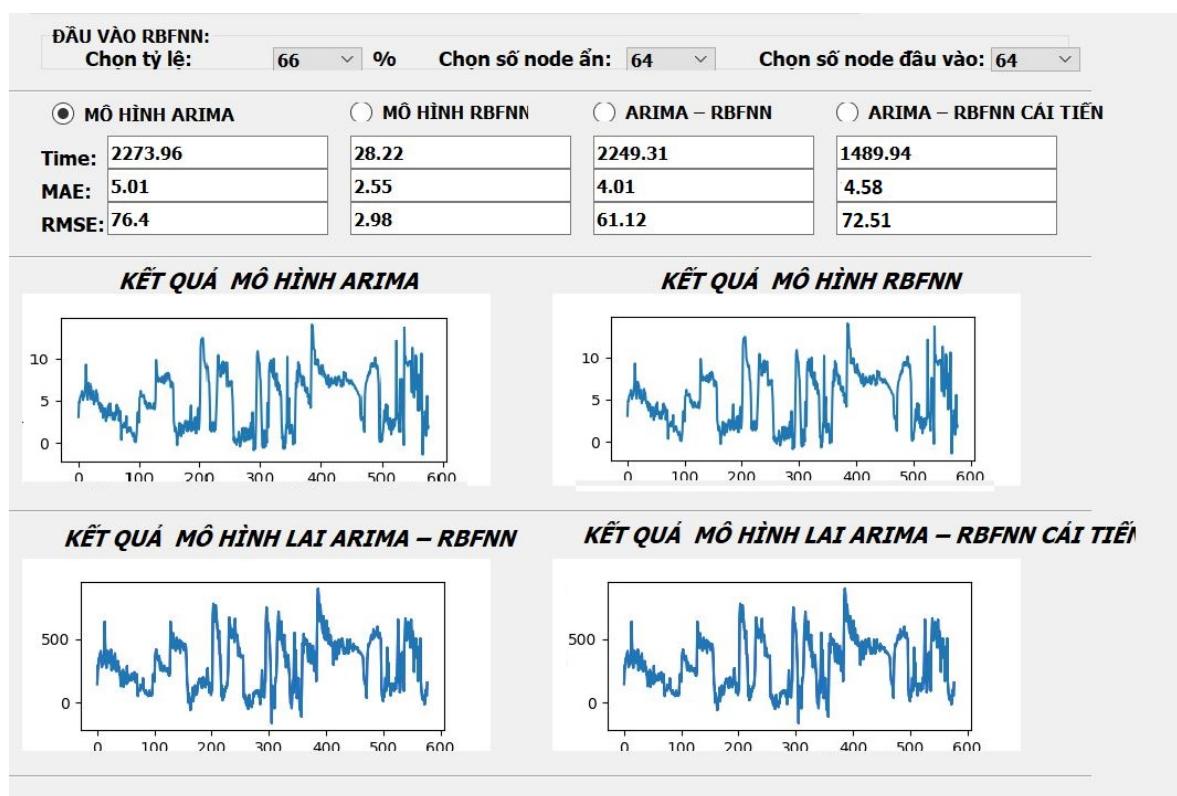
Hình 4.11: Biểu đồ thể hiện dữ liệu của chuỗi thời gian *Dentists*  
**Thực nghiệm trên tập dữ liệu Dentists.csv khi sử dụng 64 nút ẩn trong mô hình RBFNN.**

Kết quả thực nghiệm trên tập dữ liệu *Dentists* cho trong bảng 4.6 thể hiện thời gian và các chỉ số đánh giá mô hình như sau:

Bảng 4.6: Kết quả thực nghiệm trên tập dữ liệu *Dentists* với 64 nút ẩn

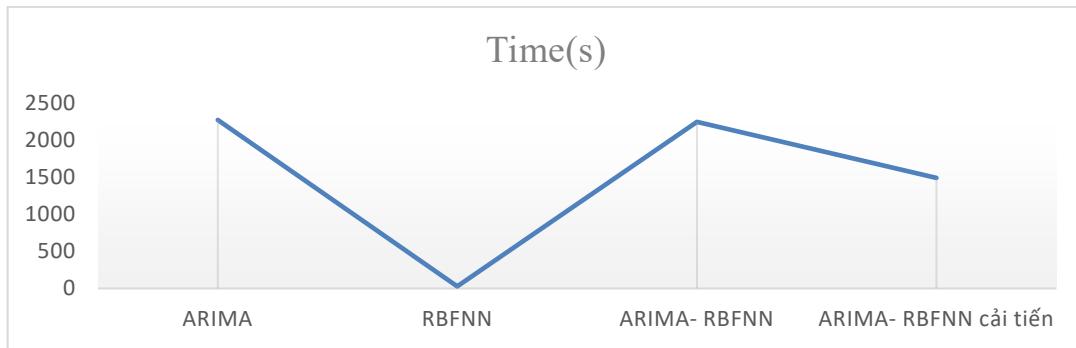
Mô hình	Dentists 64		
	Time(s)	RMSE	MAE
ARIMA	2273.96	76.4	5.01
RBFNN	28.22	2.98	2.55
ARIMA- RBFNN	2249.31	61.12	4.01
ARIMA- RBFNN cải tiến	1489.94	72.51	4.58

Hình 4.12 là kết quả dự báo của bốn mô hình dự báo khi sử dụng 64 nút ẩn được thể hiện chi tiết qua các biểu đồ sau:



Hình 4.12: Biểu đồ thể hiện kết quả dự báo của các mô hình khi dùng 64 nút ẩn trên tập dữ liệu *Dentists.csv*

Biểu đồ so sánh thời gian thực thi của các mô hình khi sử dụng 128 nút ẩn được thể hiện trong hình 4.13



Hình 4.13: Thời gian thực thi của các mô hình dự báo khi dùng 64 nút ẩn trên tập dữ liệu *Dentists*

**Nhận xét:** khi thực nghiệm trên tập dữ liệu được tăng lên về kích thước (độ dài chuỗi 2135), thì các mô hình dự báo cho kết quả tốt hơn rất nhiều so với các tập dữ liệu nhỏ hơn (độ dài chuỗi 150). Xét về thời gian thực thi thì mô hình ARIMA-RBFNN cải tiến nhanh hơn mô hình ARIMA-RBFNN, trong khi đó thì kết quả dự báo của mô hình ARIMA-RBFNN cải tiến lại thấp hơn ARIMA-RBFNN.

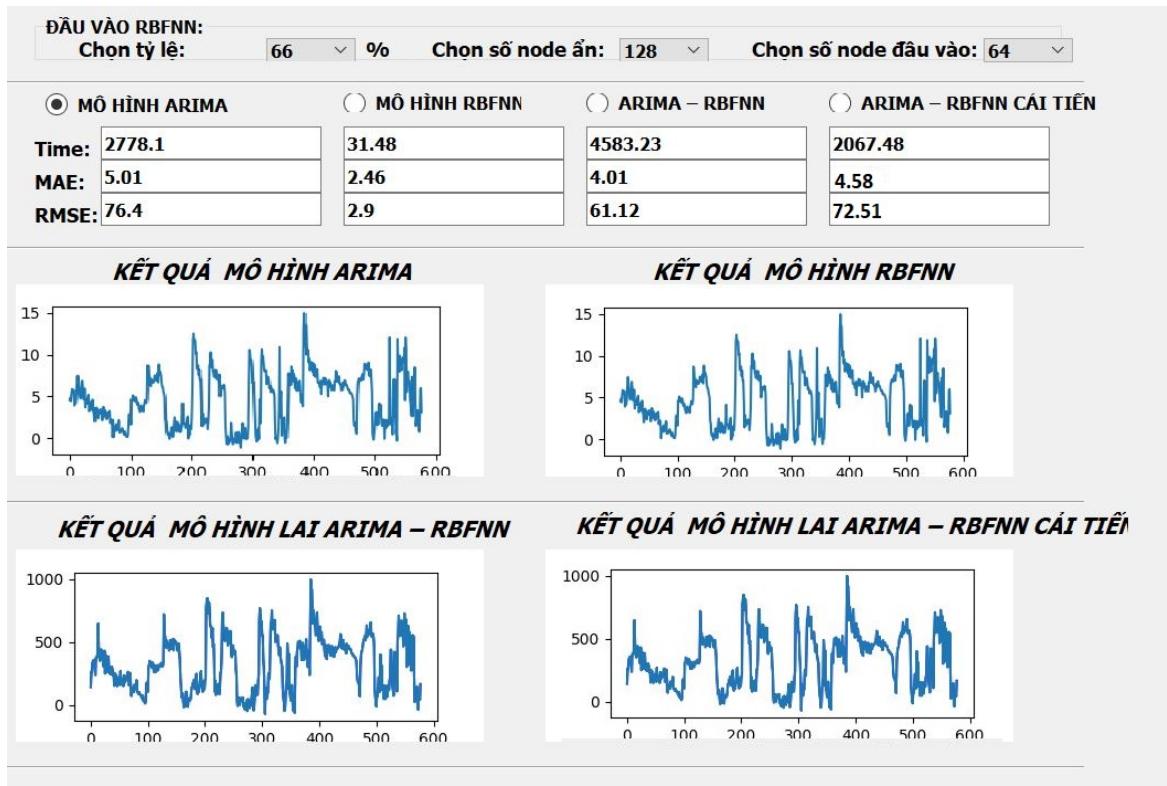
**Thực nghiệm trên tập dữ liệu *Dentists.csv* khi sử dụng 128 nút ẩn trong mô hình RBFNN.**

Kết quả thực nghiệm trên tập dữ liệu *Dentists* cho trong bảng 4.7 thể hiện thời gian và các chỉ số đánh giá mô hình như sau:

Bảng 4.7: Kết quả thực nghiệm trên tập dữ liệu *Dentists* với 128 nút ẩn

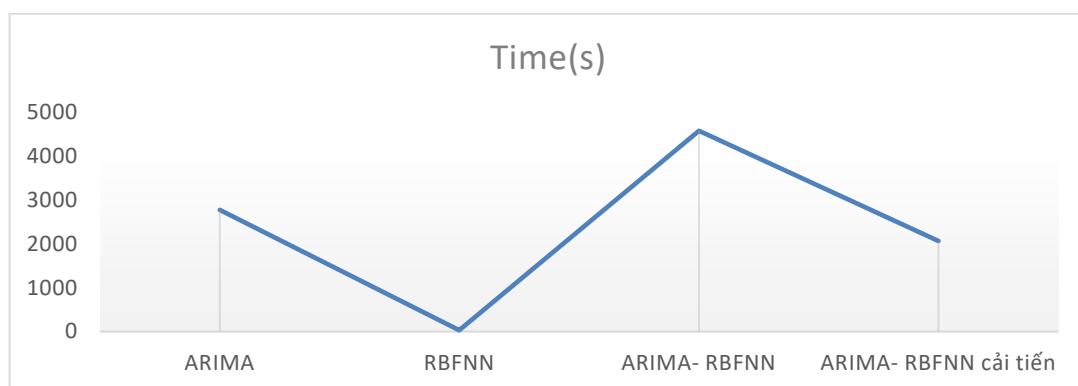
Mô hình	Dentists 128		
	Time(s)	RMSE	MAE
ARIMA	2778.1	76.4	5.01
RBFNN	31.48	2.9	2.46
ARIMA- RBFNN	4583.23	61.12	4.01
ARIMA- RBFNN cải tiến	2067.48	72.51	4.58

Hình 4.14 là kết quả dự báo của bốn mô hình dự báo khi sử dụng 128 nút ẩn được thể hiện chi tiết qua các biểu đồ sau:



Hình 4.14: Biểu đồ thể hiện kết quả dự báo của các mô hình khi dùng 128 nút ẩn trên tập dữ liệu *Dentists.csv*

Biểu đồ so sánh thời gian thực thi của các mô hình khi sử dụng 128 nút ẩn được thể hiện trong hình 4.14

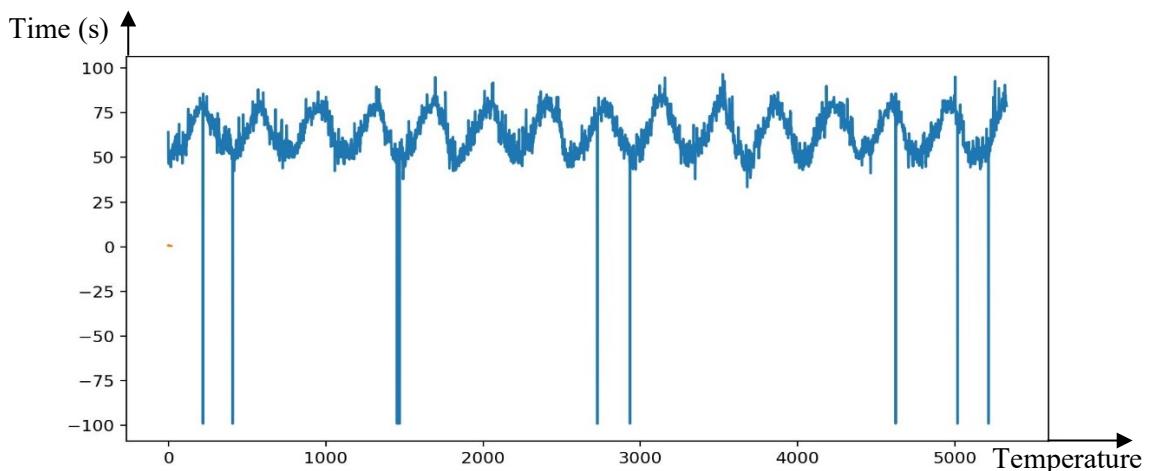


Hình 4.15: Thời gian thực thi của các mô hình dự báo khi dùng 128 nút ẩn trên tập dữ liệu *Dentists*

**Nhận xét:** Khi tăng số nút ẩn lên 128, thời gian thực hiện của mô hình RBFNN tăng nhưng kết quả dự báo lại tăng không đáng kể, riêng mô hình ARIMA-RBFNN thời gian thực thi lại giảm, hai mô hình còn lại là ARIMA và ARIMA-RBFNN cải

tiến thay đổi không nhiều, nguyên nhân do khi thực thi tập dữ liệu lớn mà tài nguyên hệ thống chưa đáp ứng được nên gây hạn chế về kết quả thực thi của các mô hình.

#### 4.3.1.4. Thực nghiệm trên tập dữ liệu City\_temperature.csv



Hình 4.16: Biểu đồ thể hiện dữ liệu của chuỗi thời gian *City\_temperature*

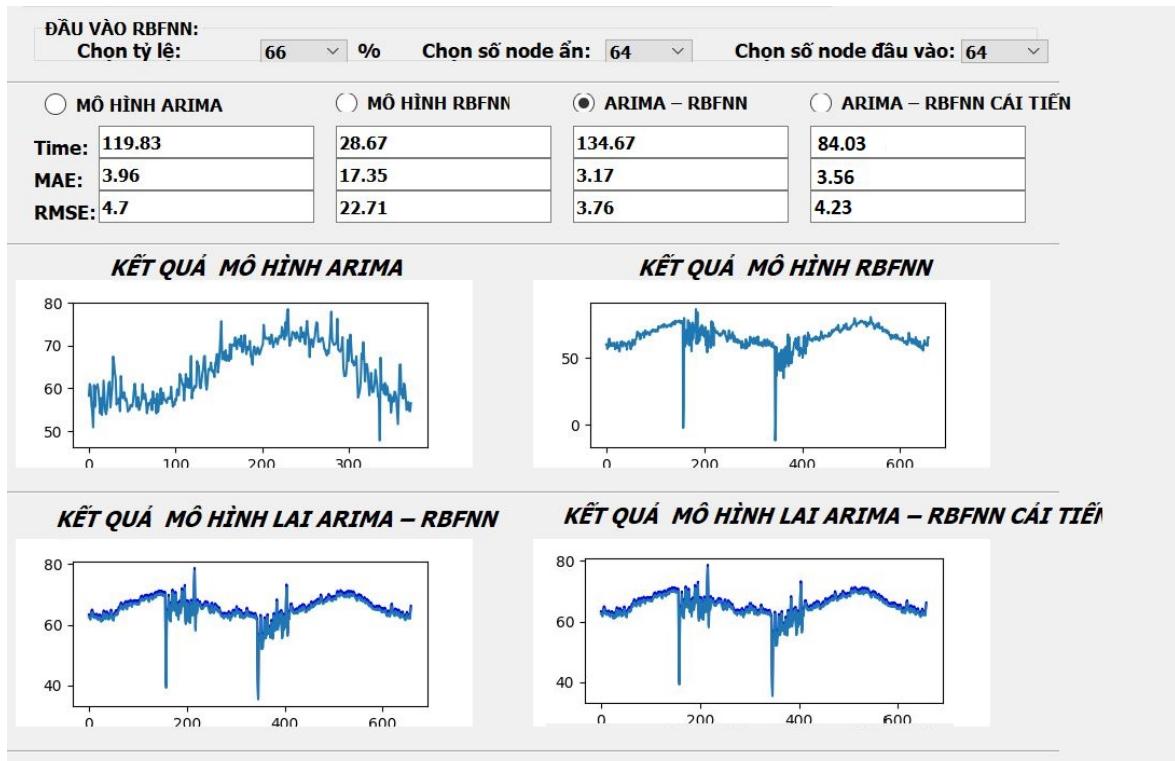
**Thực nghiệm trên tập dữ liệu *City\_temperature.csv* khi sử dụng 64 nút ẩn trong mô hình RBFNN.**

Kết quả thực nghiệm trên tập dữ liệu *City\_temperature* cho trong bảng 4.8 thể hiện thời gian và các chỉ số đánh giá mô hình như sau:

Bảng 4.8: Kết quả thực nghiệm trên tập dữ liệu *City\_temperature* với 64 nút ẩn

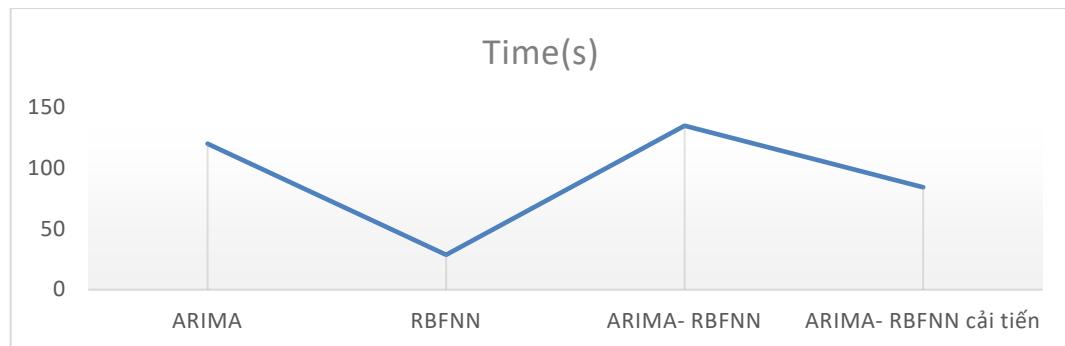
Mô hình	City_temperature 64		
	Time(s)	RMSE	MAE
ARIMA	119.83	4.7	3.96
RBFNN	28.67	22.71	17.35
ARIMA- RBFNN	134.67	3.76	3.17
ARIMA- RBFNN cải tiến	84.03	4.23	3.56

Hình 4.17 là kết quả dự báo của bốn mô hình dự báo khi sử dụng 64 nút ẩn được thể hiện chi tiết qua các biểu đồ sau:



Hình 4.17: Biểu đồ thể hiện kết quả dự báo của các mô hình khi dùng 64 nút ẩn trên tập dữ liệu *City\_temperature*

Biểu đồ so sánh thời gian thực thi của các mô hình khi sử dụng 64 nút ẩn được thể hiện trong hình 4.18



Hình 4.18: Thời gian thực thi của các mô hình dự báo khi dùng 64 nút ẩn trên tập dữ liệu *City\_temperature*

**Nhận xét:** Khi tập dữ liệu đủ lớn thì thời gian thực thi của mô hình ARIMA-RBFNN cải tiến là tốt nhất, nhanh gấp gần 3 lần so với mô hình ARIMA-RBFNN tuy nhiên về độ chính xác lại không bằng ARIMA-RBFNN điều đó cho thấy vấn đề tiền xử lý dữ liệu là điều rất quan trọng và nó ảnh hưởng rất nhiều tới kết quả dự báo của các mô hình.

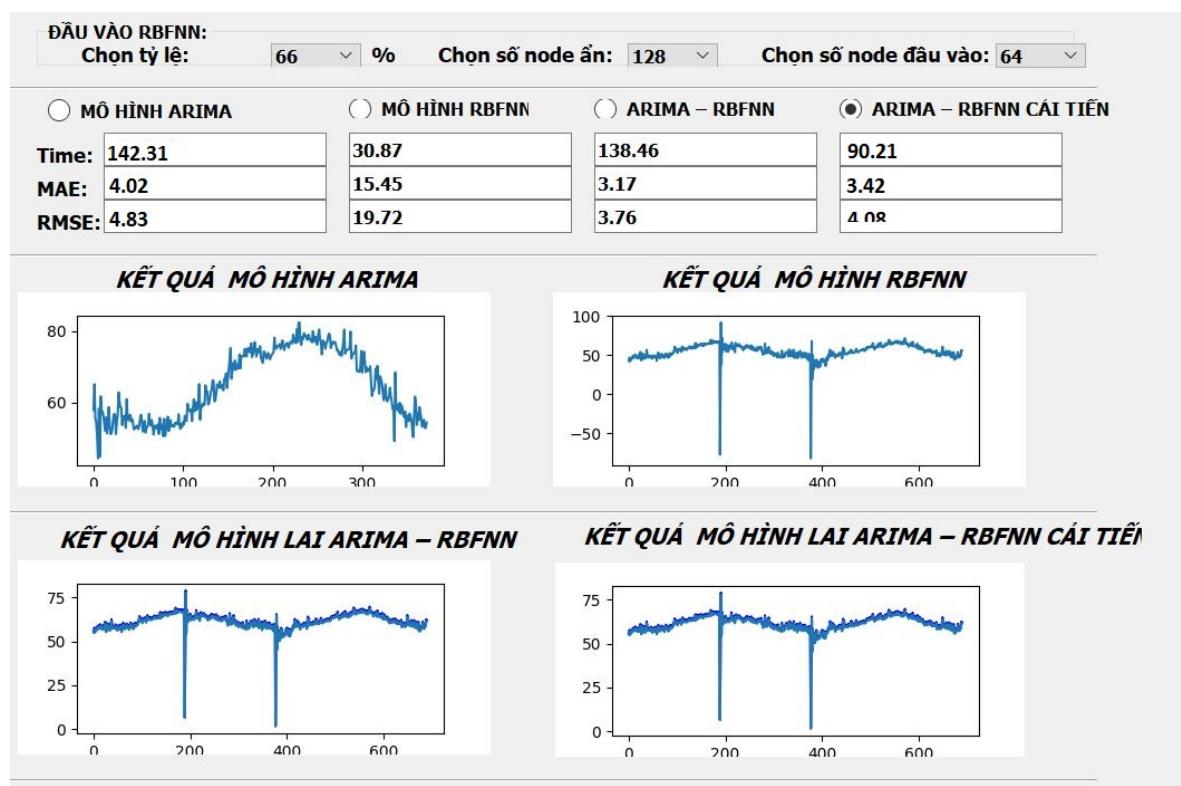
**Thực nghiệm trên tập dữ liệu City\_temperature.csv khi sử dụng 128 nút ẩn trong mô hình RBFNN.**

Kết quả thực nghiệm trên tập dữ liệu City\_temperature cho trong bảng 4.9 thể hiện thời gian và các chỉ số đánh giá mô hình như sau:

Bảng 4.9: Kết quả thực nghiệm trên tập dữ liệu City\_temperature với 128 nút ẩn

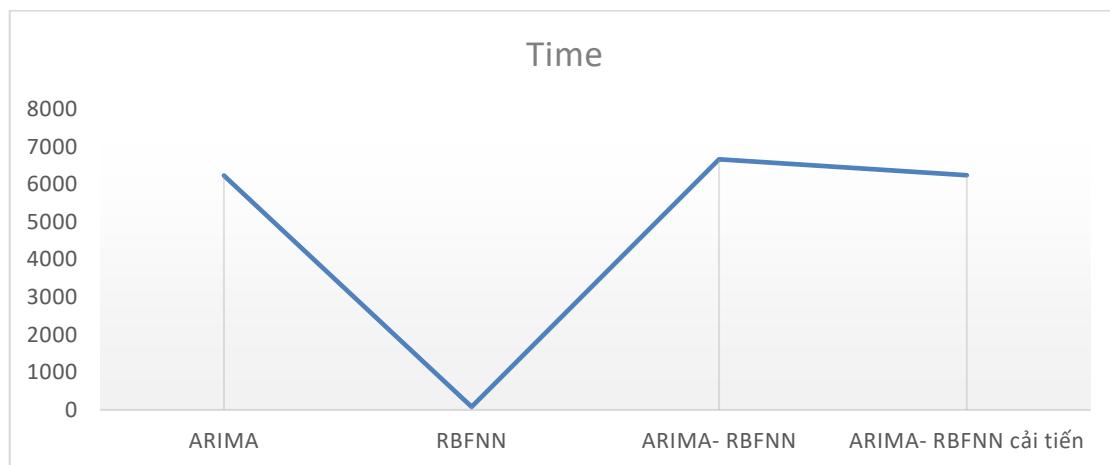
Mô hình	City_temperature 128		
	Time(s)	RMSE	MAE
ARIMA	142.31	4.83	4.02
RBFNN	30.87	19.72	15.45
ARIMA- RBFNN	138.46	3.76	3.17
ARIMA- RBFNN cải tiến	90.21	3.42	4.08

Hình 4.19 là kết quả dự báo của bốn mô hình dự báo khi sử dụng 128 nút ẩn được thể hiện chi tiết qua các biểu đồ sau:



Hình 4.19: Biểu đồ thể hiện kết quả dự báo của các mô hình khi dùng 128 nút ẩn trên tập dữ liệu City\_temperature

Biểu đồ so sánh thời gian thực thi của các mô hình khi sử dụng 64 nút ẩn được thể hiện trong hình 4.19



Hình 4.20: Thời gian thực thi của các mô hình dự báo khi dùng 128 nút ẩn trên tập dữ liệu *City\_temperature*

**Nhận xét:** Khi thay đổi số nút ẩn từ 64 lên 128 trên chuỗi dữ liệu lớn mà tài nguyên hệ thống không tối ưu dẫn tới hiệu quả của các mô hình dự báo ảnh hưởng theo. Tuy nhiên xét về mặt thời gian thực thi thì mô hình ARIMA-RBFNN cải tiến vẫn nhanh hơn mô hình ARIMA-RBFNN, nhưng về mặt kết quả của mô hình thì chưa thấy cải tiến so với trường hợp sử dụng 64 nút ẩn.

#### 4.3.2. Trường hợp 2: Cố định số nút ẩn và thay đổi số nút đầu vào

4.3.2.1. Thực nghiệm trên tập dữ liệu *AirPassengers.csv*

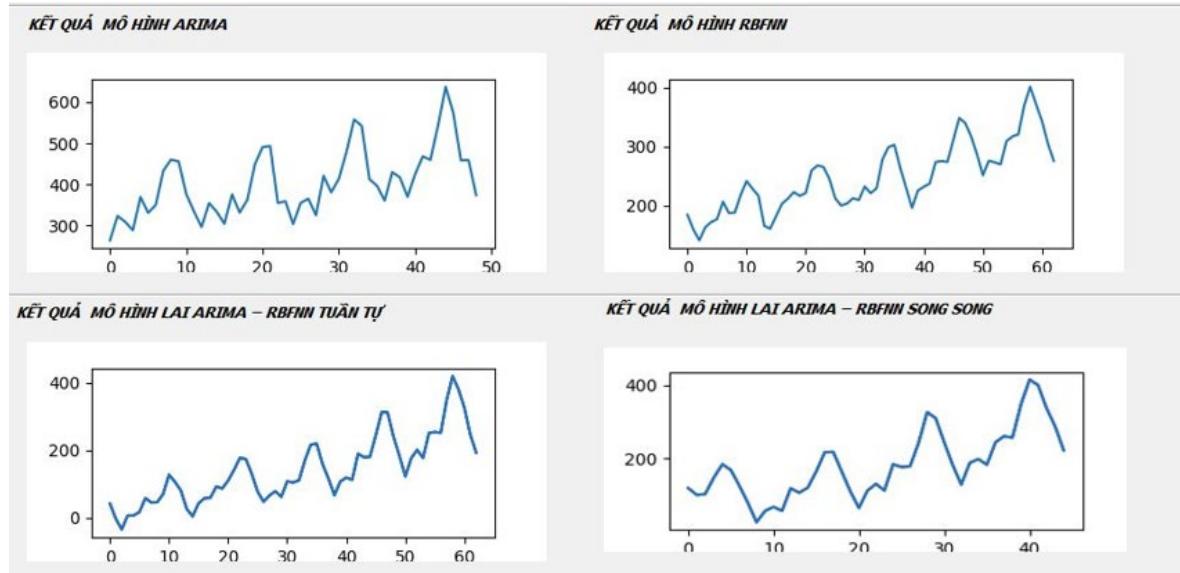
**Thực nghiệm trên tập dữ liệu AirPassengers khi sử dụng 32 nút đầu vào trong mô hình RBFNN.**

Kết quả thực nghiệm trên tập dữ liệu *AirPassengers* cho trong bảng 4.10 thể hiện thời gian và các chỉ số đánh giá mô hình như sau:

Bảng 4.10: Kết quả thực nghiệm trên tập dữ liệu *AirPassengers* với 32 nút đầu vào

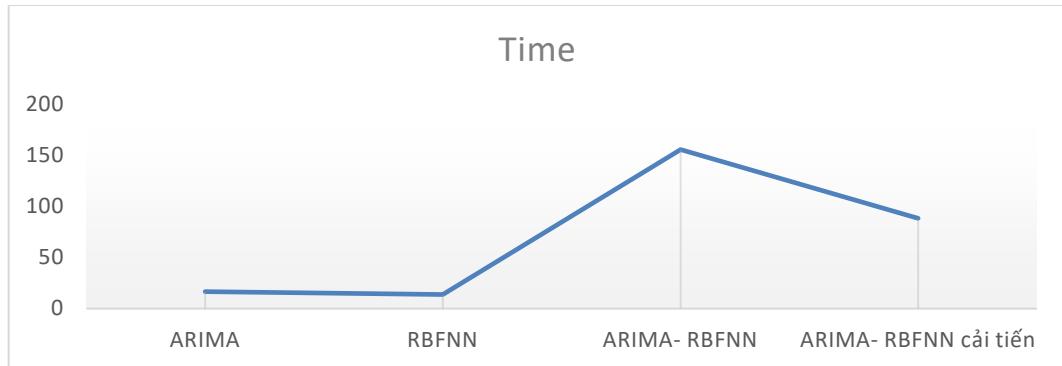
Mô hình	AirPassengers 32		
	Time (s)	RMSE	MAE
ARIMA	16.47	45.25	37.85
RBFNN	13.8	228.31	211.13
ARIMA- RBFNN	155.45	44.81	37.3
ARIMA- RBFNN cải tiến	88.14	45.03	37.575

Hình 4.21 là kết quả dự báo của bốn mô hình dự báo khi sử dụng 32 nút đầu vào được thể hiện chi tiết qua các biểu đồ sau:



Hình 4.21: Biểu đồ thể hiện kết quả dự báo của các mô hình khi dùng 32 nút đầu vào trên tập dữ liệu *AirPassengers*

Biểu đồ so sánh thời gian thực thi của các mô hình khi sử dụng 32 nút đầu vào được thể hiện trong hình 4.22



Hình 4.22: Thời gian thực thi của các mô hình dự báo khi dùng 32 nút đầu vào trên tập dữ liệu *AirPassengers*

**Nhận xét:** Khi sử dụng 32 nút đầu vào cho mô hình RBFNN, thì mô hình RBFNN có thời gian thực thi ít nhất, nhưng kết quả lại không bằng các mô hình khác, trong trường hợp này thì mô hình ARIMA cho kết quả phù hợp nhất với thời gian thực thi khá tốt so với hai mô hình ARIMA-RBFNN và ARIMA-RBFNN cải tiến.

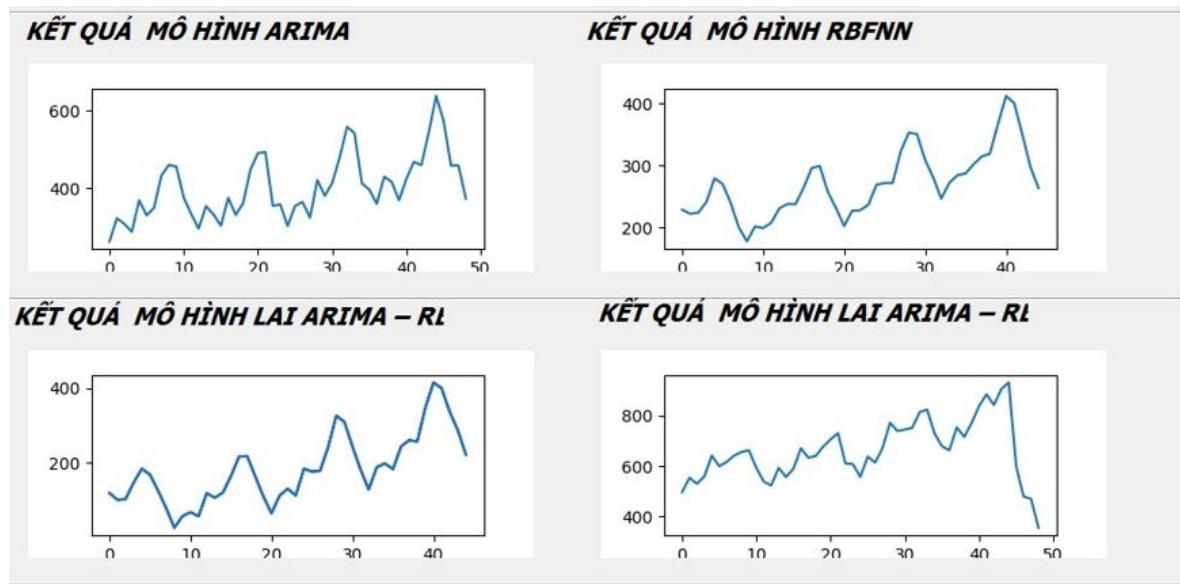
**Thực nghiệm trên tập dữ liệu AirPassengers khi sử dụng 64 nút đầu vào trong mô hình RBFNN.**

Kết quả thực nghiệm trên tập dữ liệu *AirPassengers* cho trong bảng 4.11 thể hiện thời gian và các chỉ số đánh giá mô hình như sau:

Bảng 4.11: Kết quả thực nghiệm trên tập dữ liệu *AirPassengers* với 64 nút đầu vào

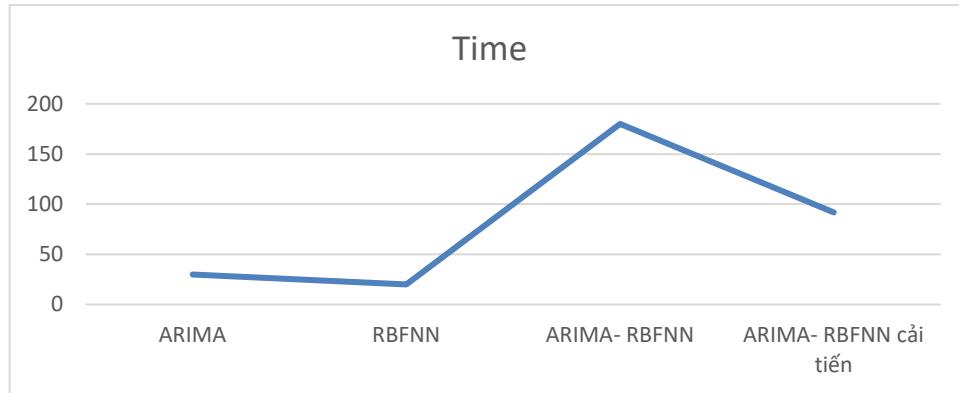
Mô hình	AirPassengers 64		
	Time (s)	RMSE	MAE
ARIMA	16.37	45.25	37.85
RBFNN	14.73	198.92	161.22
ARIMA- RBFNN	153.59	44.81	37.3
ARIMA- RBFNN cải tiến	91.66	45.03	37.575

Hình 4.23 là kết quả dự báo của bốn mô hình dự báo khi sử dụng 64 nút đầu vào được thể hiện chi tiết qua các biểu đồ sau:



Hình 4.23: Biểu đồ thể hiện kết quả dự báo của các mô hình khi dùng 64 nút đầu vào trên tập dữ liệu *AirPassengers*

Biểu đồ so sánh thời gian thực thi của các mô hình khi sử dụng 64 nút đầu vào được thể hiện trong hình 4.24



Hình 4.24: Thời gian thực thi của các mô hình dự báo khi dùng 64 nút đầu vào trên tập dữ liệu *AirPassengers*

**Nhận xét:** Khi thay đổi số nút đầu vào từ 32 lên 64 thì mô hình ARIMA không bị ảnh hưởng, tuy nhiên hai mô hình ARIMA-RBFNN và ARIMA-RBFNN cũng không có sự thay đổi về kết quả dự báo nhưng thời gian lại cao hơn khi sử dụng 32 nút đầu vào cho tập dữ liệu *AirPassengers*

#### 4.3.2.2. Thực nghiệm trên tập dữ liệu Sunspots.csv

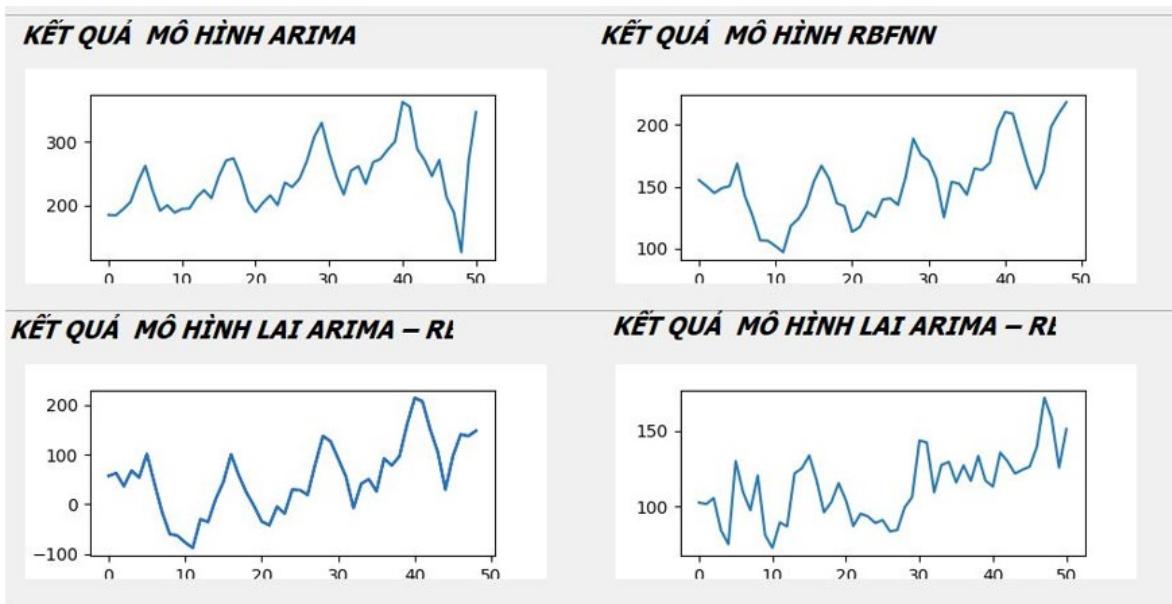
**Thực nghiệm trên tập dữ liệu Sunspots.csv khi sử dụng 32 nút đầu vào trong mô hình RBFNN.**

Kết quả thực nghiệm trên tập dữ liệu *Sunspots* cho trong bảng 4.12 thể hiện thời gian và các chỉ số đánh giá mô hình như sau:

Bảng 4.12: Kết quả thực nghiệm trên tập dữ liệu *Sunspots* với 32 nút đầu vào

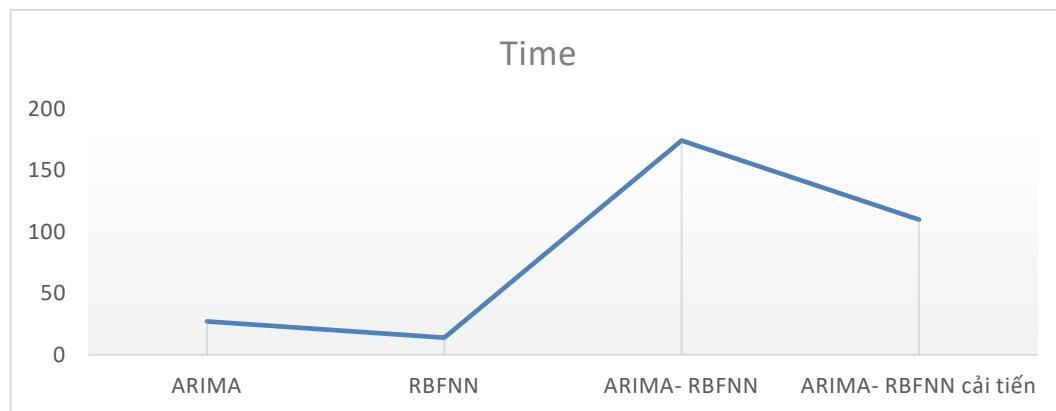
Mô hình	Sunspots 32		
	Time (s)	RMSE	MAE
ARIMA	27.09	22.28	16.44
RBFNN	13.94	61.1	46.34
ARIMA- RBFNN	174.2	22	16.97
ARIMA- RBFNN cải tiến	109.94	22.14	16.705

Hình 4.25 là kết quả dự báo của bốn mô hình dự báo khi sử dụng 32 nút đầu vào được thể hiện chi tiết qua các biểu đồ sau:



Hình 4.25: Biểu đồ thể hiện kết quả dự báo của các mô hình khi dùng 32 nút đầu vào trên tập dữ liệu *Sunspots*

Biểu đồ so sánh thời gian thực thi của các mô hình khi sử dụng 32 nút đầu vào được thể hiện trong hình 4.26



Hình 4.26: Thời gian thực thi của các mô hình dự báo khi dùng 32 nút đầu vào trên tập dữ liệu *Sunspots*

**Nhận xét:** hai mô hình ARIMA-RBFNN và ARIMA-RBFNN cải tiến đều cho kết quả không tốt ở tập dữ liệu này, thời gian thực thi của 2 mô hình này khá nhiều nhưng kết quả lại tương đương với mô hình ARIMA, điều này cho thấy kết quả của các mô hình phụ thuộc vào tập dữ liệu rất nhiều.

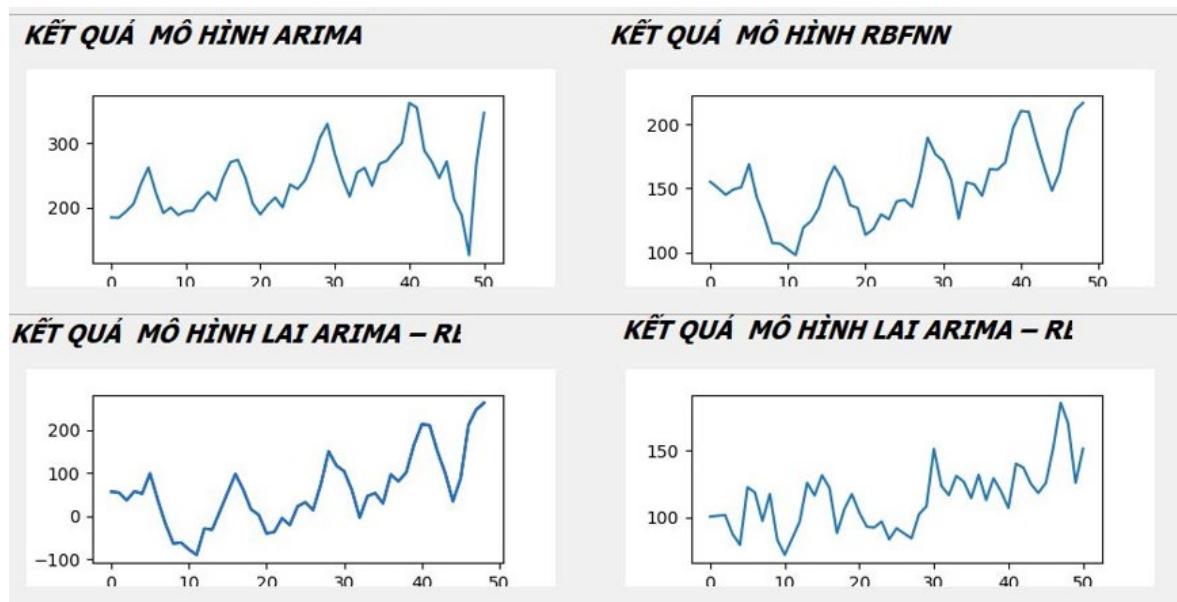
*Thực nghiệm trên tập dữ liệu Sunspots.csv khi sử dụng 64 nút đầu vào trong mô hình RBFNN.*

Kết quả thực nghiệm trên tập dữ liệu *Sunspots* cho trong bảng 4.13 thể hiện thời gian và các chỉ số đánh giá mô hình như sau:

Bảng 4.13: Kết quả thực nghiệm trên tập dữ liệu *Sunspots* với 64 nút đầu vào

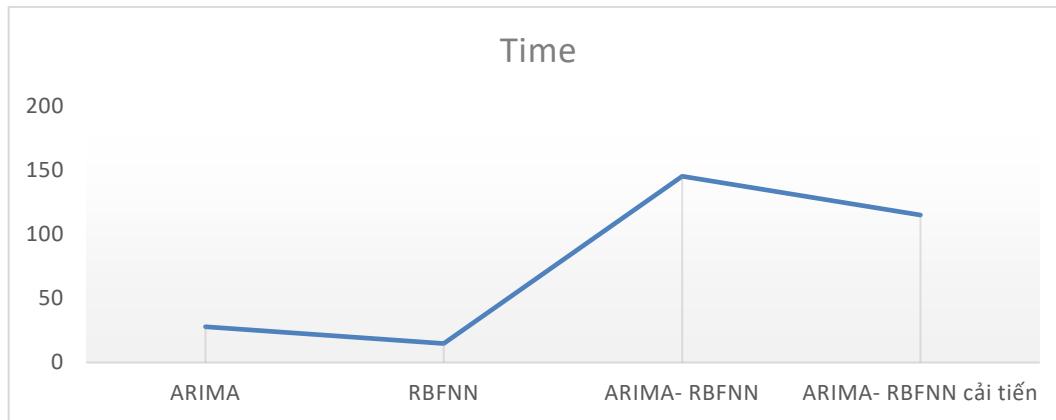
Mô hình	Sunspots 64		
	Time (s)	RMSE	MAE
ARIMA	27.9	39.04	30.45
RBFNN	14.78	118.63	106.94
ARIMA- RBFNN	145.11	38.91	28.83
ARIMA- RBFNN cải tiến	114.92	38.975	29.64

Hình 4.27 là kết quả dự báo của bốn mô hình dự báo khi sử dụng 64 nút đầu vào được thể hiện chi tiết qua các biểu đồ sau:



Hình 4.27: Biểu đồ thể hiện kết quả dự báo của các mô hình khi dùng 64 nút đầu vào trên tập dữ liệu *Sunspots*

Biểu đồ so sánh thời gian thực thi của các mô hình khi sử dụng 64 nút đầu vào được thể hiện trong hình 4.28



Hình 4.28: Thời gian thực thi của các mô hình dự báo khi dùng 64 nút đầu vào trên tập dữ liệu *Sunspots*

**Nhận xét:** Khi tăng lên 64 nút đầu vào, thời gian thực thi của các mô hình tăng so với 32 nút đầu vào và kết quả cũng giảm so với trường hợp 32 nút đầu vào. Như vậy, đối với tập dữ liệu này thì sử dụng 32 nút đầu vào tốt hơn 64 nút.

#### 4.3.2.3. Thực nghiệm trên tập dữ liệu Dentists.csv

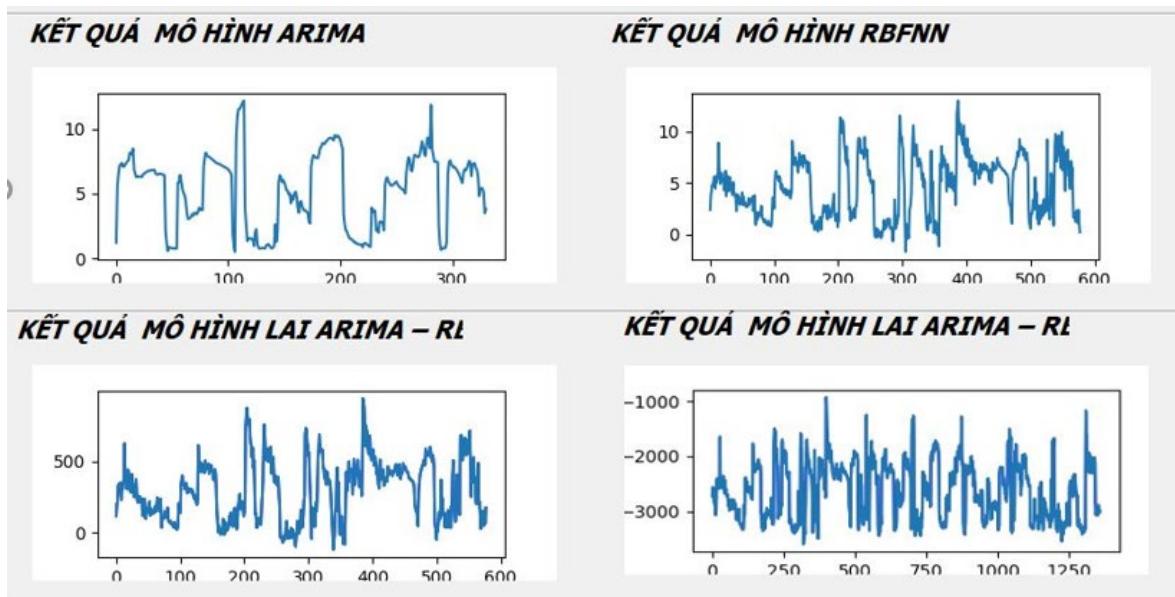
**Thực nghiệm trên tập dữ liệu Dentists.csv khi sử dụng 32 nút đầu vào trong mô hình RBFNN.**

Kết quả thực nghiệm trên tập dữ liệu *Dentists* cho trong bảng 4.14 thể hiện thời gian và các chỉ số đánh giá mô hình như sau:

Bảng 4.14: Kết quả thực nghiệm trên tập dữ liệu *Dentists* với 32 nút đầu vào

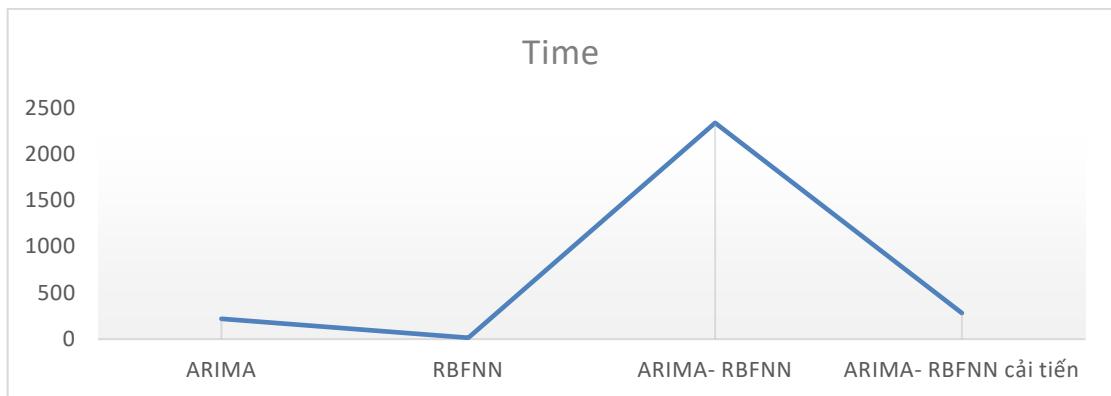
Mô hình	Dentists 32		
	Time (s)	RMSE	MAE
ARIMA	221.2	1.65	0.89
RBFNN	14.56	5.25	4.39
ARIMA- RBFNN	2337.2	1.32	0.62
ARIMA- RBFNN cải tiến	282.912	1.485	0.755

Hình 4.29 là kết quả dự báo của bốn mô hình dự báo khi sử dụng 32 nút đầu vào được thể hiện chi tiết qua các biểu đồ sau:



Hình 4.29: Biểu đồ thể hiện kết quả dự báo của các mô hình khi dùng 32 nút đầu vào trên tập dữ liệu *Dentists*

Biểu đồ so sánh thời gian thực thi của các mô hình khi sử dụng 32 nút đầu vào được thể hiện trong hình 4.30



Hình 4.30: Thời gian thực thi của các mô hình dự báo khi dùng 32 nút đầu vào trên tập dữ liệu *Dentists*

**Nhận xét:** Khi dự báo trên tập dữ liệu tương đối lớn thì kết quả dự báo có thay đổi rõ ràng theo hướng tích cực hơn, thời gian thực thi của các mô hình phụ thuộc vào độ lớn của tập dữ liệu, nhưng tập dữ liệu càng lớn thì kết quả dự báo tốt hơn nhiều so với các tập dữ liệu nhỏ hơn.

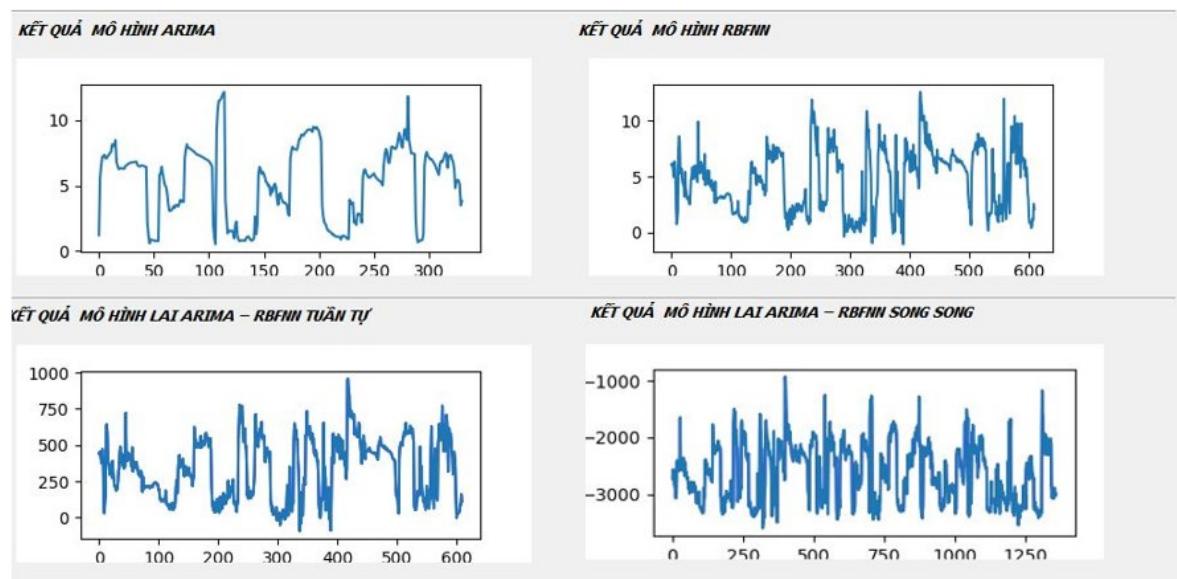
**Thực nghiệm trên tập dữ liệu *Dentists.csv* khi sử dụng 64 nút đầu vào trong mô hình RBFNN.**

Kết quả thực nghiệm trên tập dữ liệu *Dentists* cho trong bảng 4.15 thể hiện thời gian và các chỉ số đánh giá mô hình như sau:

Bảng 4.15: Kết quả thực nghiệm trên tập dữ liệu *Dentists* với 64 nút đầu vào

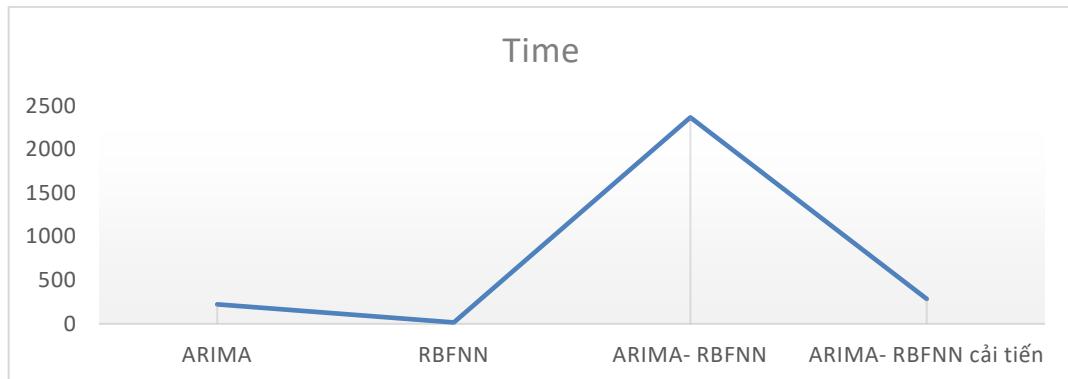
Mô hình	<i>Dentists</i> 64		
	Time (s)	RMSE	MAE
ARIMA	223.78	1.65	0.89
RBFNN	15.54	5.55	4.79
ARIMA- RBFNN	2367.21	1.32	0.63
ARIMA- RBFNN cải tiến	509.72	1.485	0.76

Hình 4.31 là kết quả dự báo của bốn mô hình dự báo khi sử dụng 64 nút đầu vào được thể hiện chi tiết qua các biểu đồ sau:



Hình 4.31: Biểu đồ thể hiện kết quả dự báo của các mô hình khi dùng 64 nút đầu vào trên tập dữ liệu *Dentists*

Biểu đồ so sánh thời gian thực thi của các mô hình khi sử dụng 64 nút đầu vào được thể hiện trong hình 4.32



Hình 4.32: Thời gian thực thi của các mô hình dự báo khi dùng 64 nút đầu vào trên tập dữ liệu *Dentists*

**Nhận xét:** Khi tăng số nút đầu vào từ 32 lên 64, thì thời gian thực thi của các mô hình đề tăng, kết quả dự báo của các mô hình gần như không thay đổi. Như vậy số nút đầu vào làm ảnh hưởng nhiều đến các mô hình theo chiều hướng giảm.

#### 4.3.2.4. Thực nghiệm trên tập dữ liệu *City\_temperature.csv*

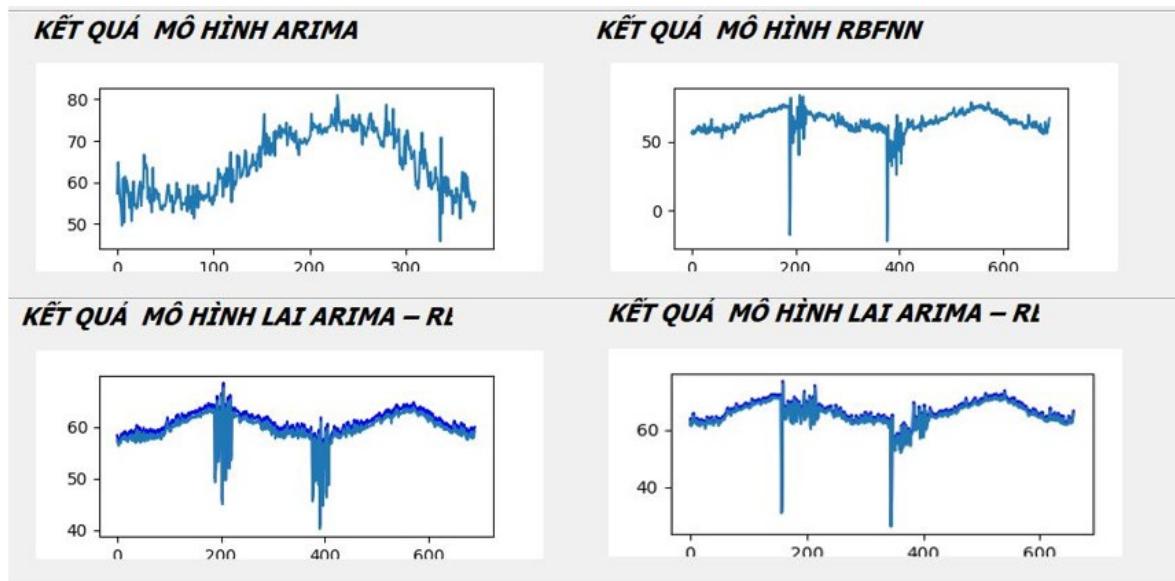
**Thực nghiệm trên tập dữ liệu *City\_temperature* khi sử dụng 32 nút đầu vào trong mô hình RBFNN.**

Kết quả thực nghiệm trên tập dữ liệu *City\_temperature* cho trong bảng 4.16 thể hiện thời gian và các chỉ số đánh giá mô hình như sau:

Bảng 4.16: Kết quả thực nghiệm trên tập dữ liệu *City\_temperature* với 32 nút đầu vào

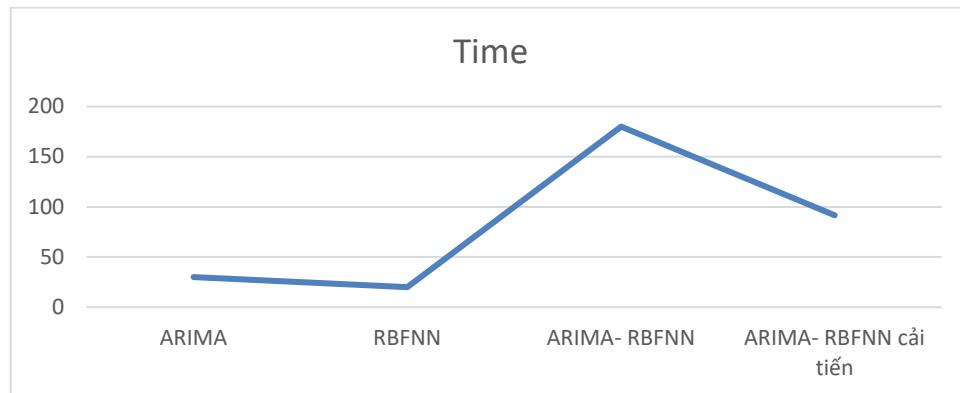
Mô hình	<i>City_temperature</i> 32		
	Time (s)	RMSE	MAE
ARIMA	332.34	4.19	3.42
RBFNN	19.45	44.84	33.63
ARIMA- RBFNN	4374.45	3.46	2.68
ARIMA- RBFNN cải tiến	742.48	3.825	3.05

Hình 4.33 là kết quả dự báo của bốn mô hình dự báo khi sử dụng 32 nút đầu vào được thể hiện chi tiết qua các biểu đồ sau:



Hình 4.33: Biểu đồ thể hiện kết quả dự báo của các mô hình khi dùng 32 nút đầu vào trên tập dữ liệu *City\_temperature*

Biểu đồ so sánh thời gian thực thi của các mô hình khi sử dụng 32 nút đầu vào được thể hiện trong hình 4.34



Hình 4.34: Thời gian thực thi của các mô hình dự báo khi dùng 32 nút đầu vào trên tập dữ liệu *City\_temperature*

**Nhận xét:** Kết quả dự báo trên tập dữ liệu lớn luôn cho kết quả tốt, nhưng về mặt thời gian thì cũng mất rất nhiều, trong trường hợp này thì mô hình ARIMA-RBFNN cài tiến thấy rõ thời gian thực thi hiệu quả hơn nhiều so với mô hình ARIMA-RBFNN

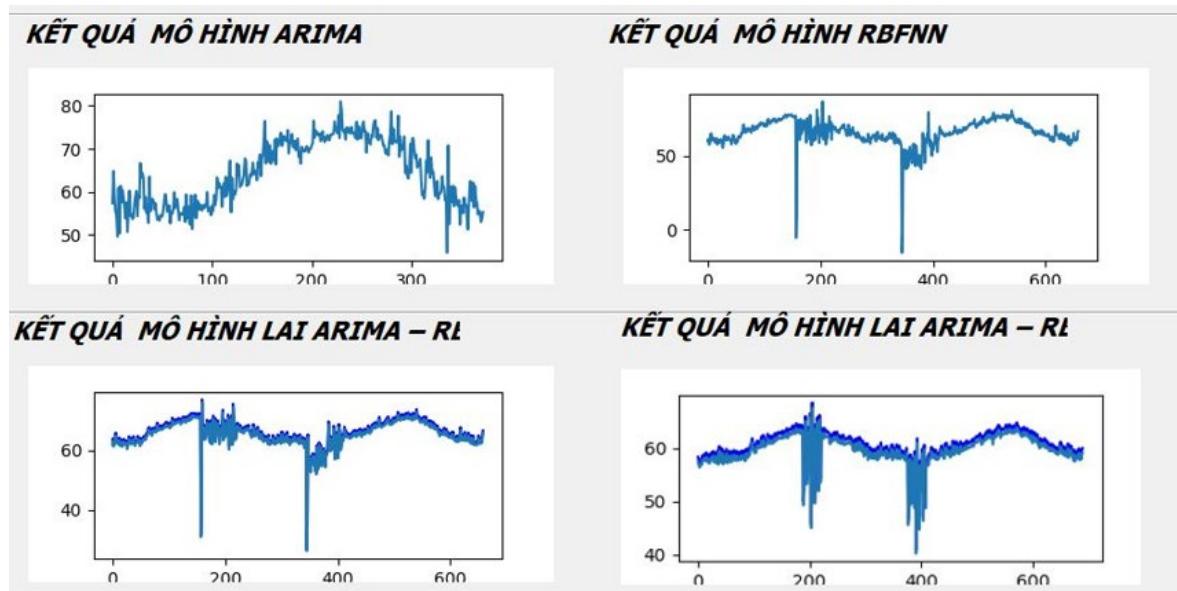
**Thực nghiệm trên tập dữ liệu *City\_temperature* khi sử dụng 64 nút đầu vào trong mô hình RBFNN.**

Kết quả thực nghiệm trên tập dữ liệu *City\_temperature* cho trong bảng 4.17 thể hiện thời gian và các chỉ số đánh giá mô hình như sau:

Bảng 4.17: Kết quả thực nghiệm trên tập dữ liệu *City\_temperature* với 64 nút đầu vào

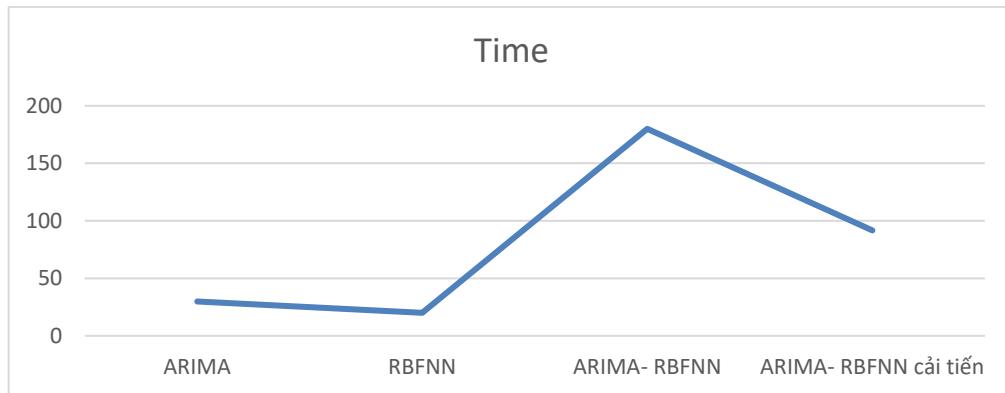
Mô hình	<i>City_temperature</i> 64		
	Time (s)	RMSE	MAE
ARIMA	228.77	4.19	3.42
RBFNN	15.25	45.32	34.62
ARIMA- RBFNN	2543	3.46	2.68
ARIMA- RBFNN cải tiến	518.54	3.825	3.05

Hình 4.35 là kết quả dự báo của bốn mô hình dự báo khi sử dụng 64 nút đầu vào được thể hiện chi tiết qua các biểu đồ sau:



Hình 4.35: Biểu đồ thể hiện kết quả dự báo của các mô hình khi dùng 64 nút đầu vào trên tập dữ liệu *City\_temperature*

Biểu đồ so sánh thời gian thực thi của các mô hình khi sử dụng 64 nút đầu vào được thể hiện trong hình 4.36



Hình 4.36: Thời gian thực thi của các mô hình dự báo khi dùng 64 nút đầu vào trên tập dữ liệu *City\_temperature*

**Nhận xét:** Khi tăng số nút đầu vào trên tập dữ liệu lớn thì thời gian thực thi của các mô hình giảm, nhưng kết quả dự báo lại không tốt so với số nút đầu vào nhỏ hơn

#### 4.4. Nhận xét kết quả thực nghiệm ở các tập dữ liệu.

Sau khi thực nghiệm trên bốn tập dữ liệu, mỗi tập dữ liệu thực nghiệm hai trường hợp và cho kết quả như sau:

- Về thời gian thực thi và tài nguyên sử dụng

Các trường hợp thực nghiệm cho thấy, thời gian thực thi của các mô hình phụ thuộc nhiều vào tập dữ liệu, khi tập dữ liệu nhỏ thì thời gian thực thi của các mô hình không thay đổi nhiều, mô hình ARIMA-RBFNN cải tiến so với các mô hình khác thì không hiệu quả, tuy nhiên khi dự báo trên tập dữ liệu càng lớn thì sự chênh lệch về mặt thời gian giữa mô hình ARIMA-RBFNN cải tiến với mô hình ARIMA-RBFNN càng lớn. Điều này cho thấy mô hình ARIMA-RBFNN cải tiến có thể ứng dụng trong nhiều lĩnh vực, vì hiện nay trong các lĩnh vực thì lượng dữ liệu ngày càng nhiều.

Mô hình ARIMA-RBFNN cải tiến nhanh hơn mô hình ARIMA-RBFNN trong các trường hợp thực nghiệm. Tuy nhiên về mức độ sử dụng tài nguyên thì mô hình ARIMA-RBFNN cải tiến sử dụng nhiều hơn. Qua bảng 4.10 thống kê mức độ sử dụng CPU của các mô hình, thì mô hình ARIMA-RBFNN cải tiến tốt hơn về mặt thời gian nhưng chưa hiệu quả về sử dụng tài nguyên.

Bảng 4.18: Thống kê sử dụng tài nguyên của các mô hình dự báo

THỐNG KÊ SỬ DỤNG TÀI NGUYÊN CPU (%)					
Dataset	Độ dài	ARIMA	RBFNN	ARIMA-RBFNN	ARIMA-RBFNN cải tiến
Sunspots	150	48	45	47	83
AirPassengers	145	54	48	51	81
Dentists	2135	55	48	48	87
City temperature	5328	52	46	58	86

#### - Về độ chính xác

Trong thực nghiệm này sử dụng hai giá trị RMSE và MAE để đánh giá độ chính xác của các mô hình dự báo. Qua các thực nghiệm cho thấy mô hình ARIMA-RBFNN luôn cho kết quả thấp hơn mô hình ARIMA-RBFNN cải tiến, và các mô hình dự báo này cho kết quả tốt khi tập dữ liệu lớn. Nguyên nhân do tập dữ liệu đầu vào chưa được xử lý. Đối với mô hình ARIMA-RBFNN đã tận dụng được ưu điểm của từng mô hình, ARIMA xử lý tốt các thành phần tuyến tính trong tập dữ liệu và RBFNN xử lý tốt các thành phần phi tuyến còn lại (sau khi ARIMA đã thực thi) của tập dữ liệu, còn mô hình ARIMA-RBFNN cải tiến thực hiện song song trên hai tập dữ liệu riêng biệt do đó kết quả chưa như mong đợi. Điều này cho thấy nếu tập dữ liệu được xử lý trước khi thực hiện mô hình ARIMA-RBFNN cải tiến bằng cách phân tách thành hai tập dữ liệu riêng biệt, một tập chứa các thành phần tuyến tính của dữ liệu làm đầu vào cho mô hình ARIMA, tập còn lại chứa các thành phần phi tuyến của dữ liệu làm đầu vào cho mô hình RBFNN. Khi đó kết quả dự báo sẽ thay đổi theo hướng tích cực hơn.

## Chương 5

# KẾT LUẬN VÀ HƯỚNG PHÁT TRIỂN

*Chương này tóm tắt lại các kết quả đã đạt được của luận văn, những đóng góp, hạn chế và hướng phát triển trong tương lai.*

### 5.1. Kết quả đạt được

Qua nghiên cứu và thực nghiệm, đề tài cơ bản đã đạt được các mục tiêu cũng như các nhiệm vụ đã đề ra.

- Tìm hiểu các mô hình dự báo trên chuỗi thời gian: ARIMA, RBFNN, mô hình lai ARIMA-RBFNN, ...

- Xây dựng mô hình lai ARIMA-RBFNN cải tiến, bằng cách thực hiện song song hai mô hình.

- Cài đặt các mô hình dự báo chuỗi thời gian ARIMA, RBFNN, mô hình lai ARIMA-RBFNN và mô hình lai ARIMA-RBFNN cải tiến lên hệ thống.

- Thực nghiệm sử dụng các mô hình dự báo đã cài đặt trên bốn tập dữ liệu

### 5.2 Các mặt hạn chế

Khi cài đặt và thực nghiệm các mô hình dự báo trên các tập dữ liệu, từ kết quả cho thấy còn một số hạn chế

- Chưa thực nghiệm được trên nhiều tập dữ liệu lớn khác nhau.

- Kết quả chưa tốt với các tập dữ liệu nhỏ khi sử dụng mô hình ARIMA-RBFNN cải tiến.

- Đề tài mới thực nghiệm song song trên nhiều chương trình (multi processing), chưa thực hiện song song trên nhiều CPU (multi processor)

### 5.3. Hướng phát triển

Đề tài đã cải tiến mô hình lai ARIMA-RBFNN, mặc dù vẫn còn nhiều hạn chế, nhưng những hạn chế này có thể khắc phục

- Tiền xử lý dữ liệu, bằng cách phân tách tập dữ liệu thành hai thành phần riêng biệt tuyến tính và phi tuyến, sau đó thực hiện song song hai mô hình ARIMA và RBFNN. Khi đó mô hình ARIMA-RBFNN cải tiến sẽ cho kết quả tốt hơn hiện tại.

- Thực nghiệm trên nhiều tập dữ liệu lớn

- Thực hiện song song trên nhiều CPU (multi processor)
- Cải tiến cách xác định các tham số cho mô hình ARIMA và RBFNN.

## TÀI LIỆU THAM KHẢO

- [1]. H.S. Behera, Sibarama Panigrahi. “A hybrid ETS–ANN model for time series forecasting”. Engineering Applications of Artificial Intelligence, Volume 66, Pages 49-59 (2017).
- [2]. [http://en.wikipedia.org/wiki/Time\\_series](http://en.wikipedia.org/wiki/Time_series). Time series.
- [3]. Habiluddina, Ahmad Jawahirb. “Comparing of ARIMA and RBFNN for short-term forecasting”. International Journal of Advances in Intelligent Informatics ISSN: 2442-6571 Vol. 1, No 1, pp. 15-22 (2015).
- [4]. N. Vijay and G.C. Mishra. “Time Series Forecasting Using ARIMA and ANN Models for Production of Pearl Millet (BAJRA) Crop of Karnataka, India”. International Journal of Current Microbiology and Applied Sciences ISSN: 2319-7706 Volume 7 Number 12 (2018).
- [5]. Jonathan D. Cryer, Kung-Sik Chan. Time Series Analysis, Springer Texts in Statistics (2008).
- [6]. L. Zhang, G. X. Zhang, and R. R. Li. “Water Quality Analysis and Prediction Using Hybrid Time Series and Neural Network Models”. JAST\_Volume 18\_Issue 4\_Pages 975-983 (2018).
- [7]. Rakhlin; Kalinin; Shvets; Iglovikov. “Automatic Instrument Segmentation in Robot-Assisted Surgery using Deep Learning”. IEEE International Conference on Machine Learning and Applications (ICMLA), (2018).
- [8]. Li Wang Haofei Zou Jia Su Ling Li Sohail Chaudhry. “An ARIMA-ANN Hybrid Model for Time Series Forecasting”. Systems Research and Behavioral Science Syst. Res. 30, 244–259, (2013).
- [9]. Parviz, L., Kholghi, M. and Hoofifar. “A Comparison of the Efficiency of Parameter Estimation Methods in the Context of Stream Flow Forecasting”. J. Agr. Sci. Tech., 12(1): 47-60, (2010).
- [10]. Ross Ihaka. *Time Series Analysis*, Lecture Notes for 475.726, Statistics Department, University of Auckland, (2005).

- [11]. Roy Batchelor. *Box-Jenkins Analysis*. Lecture Notes, ESCP-EAP, Paris, 2004
- [12]. Wu, Zonghan, et al. "Connecting the Dots: Multivariate Time Series Forecasting with Graph Neural Networks." arXiv preprint arXiv:2005.11650 (2020).
- [13]. Lâm Hoàng Vũ. "Dự báo chuỗi thời gian sử dụng mô hình ARIMA và giải thuật di truyền". Luận văn thạc sĩ, Trường Đại học Bách Khoa, TPHCM, (2012).
- [14]. Nguyễn Chí Thành - Hà Gia Sơn. "Kết hợp mạng Nơron FIR và mô hình ARIMA theo hình thức động để nâng cao hiệu quả dự báo chuỗi thời gian". Tạp chí Nghiên cứu KH&CN quân sự, Số Đặc san CNTT, 12 – 2017.
- [15]. <https://www.kaggle.com>: Nguồn dữ liệu trong thực nghiệm.
- [16]. S. K. Lahiri, K.C. Ghanta. "Artificial neural network model with the parameter tuning assisted by a differential evolution technique: the study of the hold up of the slurry flow in a pipeline". Chemical Industry & Chemical Engineering Quarterly 15 (2) 103–117, (2009).
- [17]. Miroslav R. Radovanović. Optimal Selection of ANN Training and Architectural Parameters Using Taguchi Method: A Case Study. Full Professor University of Niš Faculty of Mechanical Engineering.

## **PHỤ LỤC**

Hướng dẫn sử dụng chương trình Demo

### 1. Môi trường thực nghiệm

Hệ thống thực nghiệm trên máy tính Dell Inspiron 15, Inter® core™ i5-5300U CPU @ 2.3 GHz, 16 GB RAM trở lên, hệ điều hành Windows 10.

Sử dụng tốt trên phần mềm Python 3.77, sử dụng các thư viện Keras, Tensorflow, PyQt5, matplotlib, ...

### 2. Các bước thực hiện Demo

- Từ cửa sổ Command Prompt thực hiện chương trình bằng cách gọi ứng dụng Demo: ArimaRbfnn.py

- Chọn tập dữ liệu dự báo
- Chọn số nút ẩn để Demo
- Chọn mô hình cần thực nghiệm (Check vào )
- Bấm vào nút “Thực hiện”
- Quy trình tương tự khi thực hiện các mô hình khác
- Bấm “Kết thúc” để hoàn tất thực nghiệm

# DỰ BÁO TRÊN CHUỖI THỜI GIAN SỬ DỤNG MÔ HÌNH LAI ARIMA VÀ RBF NEURAL NETWORK

## A HYBRID ARIMA AND RBF NEURAL NETWORK MODEL FOR TIMES SERIES FORCASTING

Phạm Chí Công, Nguyễn Thành Sơn  
Trường đại học Sư phạm Kỹ thuật TP.HCM

### TÓM TẮT

Hiện nay trong khai phá dữ liệu lớn được quan tâm nhiều nhất trong lĩnh vực khoa học dữ liệu, một trong những vấn đề được đặt lên hàng đầu là những bài toán về dự báo, hầu hết trong các lĩnh vực hoạt động xã hội hiện nay thì vấn đề dự báo đóng góp một phần không nhỏ trong sự tồn tại và phát triển. Người ta đưa ra rất nhiều các kỹ thuật trong khai phá dữ liệu để dự báo, nhưng Bài toán dự báo sử dụng chuỗi số thời gian luôn là một đề tài “nóng” luôn được quan tâm.

Các nhà nghiên cứu đã đưa ra rất nhiều các phương pháp nhằm sử dụng nguồn dữ liệu lớn hiện nay để phục vụ cho các vấn đề về dự báo. Trong luận văn này, chúng tôi cũng đi vào nghiên cứu các phương pháp dự báo sao cho cải thiện được kết quả so với các mô hình dự báo khác. Dựa trên mô hình dự báo ARIMA và RBFNN, L.Zhang và cộng sự [6] đã đưa ra mô hình dự báo lai ARIMA-RBFNN và đã cho kết quả dự báo tốt hơn khi thực hiện từng mô hình. Tuy nhiên, lượng dữ liệu ngày càng lớn nên thời gian thực thi của mô hình sẽ lâu hơn. Do đó, việc cải tiến mô hình ARIMA-RBFNN để thời gian thực thi nhanh hơn là một vấn đề cần quan tâm.

Trong đề tài này, chúng tôi cải tiến mô hình ARIMA-RBFNN đã được L.Zhang và cộng sự giới thiệu, nhằm mục đích cải thiện thời gian thực thi và kết quả của dự báo tốt hơn.

**Từ khóa:** RBFNN; ARIMA; Mô hình cải tiến ARIMA-RBFNN; Chuỗi thời gian.

### ABSTRACT

Today, big data mining is most interested in data science. One of the top issues is the problem of prediction. Mostly, in the social activities today, the problem of forecasting plays a significant part in the existence and development. Many techniques are given in data mining for prediction, but the forecasting problem using time series is a "hot" topic that is always interested.

Researchers have come up with a variety of methods to use current large data sources to serve the problems of forecasting. In this thesis, we also study prediction methods to improve the results compared to other predictive models. Based on prediction models ARIMA and RBFNN, L.Zhang et al [6] gave the ARIMA-RBFNN hybrid prediction model and gave better predictive results when performing each model. However, the amount of data is increasing, so the execution time of the model will be longer. Therefore, improving the ARIMA-RBFNN model for faster execution time is a matter of concern.

In this study, we refine the ARIMA-RBFNN model introduced by L.Zhang et al in order to improve the execution time and the results of forecasts better.

**Keywords:** Radial Basis Function Neural Networks; ARIMA; Hybrid ARIMA - RBFNN; Times Series.

### 1. GIỚI THIỆU

Sự phát triển mạnh mẽ của công nghệ và sự bùng nổ của thông tin số trong những năm gần đây, nó đã góp phần không nhỏ vào sự phát triển của xã hội. Với sự đa dạng và lượng dữ liệu không lồ là nguồn tài nguyên vô giá nếu chúng ta biết khai thác và sử dụng những

thông tin có ích trong đó. Vấn đề đặt ra là khai thác và lưu trữ dữ liệu hiện nay như thế nào, Các phương pháp khai thác dữ liệu truyền thống ngày càng không phù hợp và không đáp ứng được nhu cầu thực tế. Do đó, các công nghệ khai phá dữ liệu mới ra đời đã cho phép chúng ta khai thác được những tri thức hữu

dụng bằng cách trích xuất những thông tin có mối quan hệ hoặc có mối tương quan nhất định từ một kho dữ liệu lớn (Big Data), mà bình thường chúng ta không nhận diện và sử dụng được, từ đó chúng ta giải quyết được các bài toán tìm kiếm, dự báo các xu thế, các hành vi trong tương lai, và nhiều tính năng thông minh khác.

Một trong những vấn đề quan trọng nhất hiện nay trong khai phá dữ liệu lớn là những bài toán về dự báo, hầu hết trong các lĩnh vực hoạt động xã hội hiện nay thì vấn đề dự báo đóng góp một phần không nhỏ trong sự tồn tại và phát triển. Hiện nay người ta đưa ra rất nhiều các kỹ thuật trong khai phá dữ liệu để dự báo, nhưng Bài toán dự báo sử dụng chuỗi số thời gian luôn là một đề tài “nóng” luôn được quan tâm.

## 2. CÁC CÔNG TRÌNH ĐÃ NGHIÊN CỨU LIÊN QUAN

2.1. Nguyễn Chí Thành và Hà Gia Sơn (2017): Nghiên cứu “Kết hợp mạng Neron FIR và mô hình ARIMA theo hình thức động để nâng cao hiệu quả dự báo chuỗi thời gian”

Tác giả dùng dữ liệu mẫu để ước lượng các mô hình, sau đó dự báo các giá trị của biến phụ thuộc, dùng các giá trị này để xây dựng tập các trọng số, tạo các giá trị dự báo ngoài mẫu từ các mô hình riêng biệt và sử dụng các trọng số đã tìm được.

2.2. L. Zhang, G. X. Zhang, and R. R. Li (2018): “Phân tích và dự đoán chất lượng nước bằng cách sử dụng mô hình lai ARIMA và mạng Neron RBF”

Tác giả xem xét một chuỗi thời gian ( $y_t$ ) được cấu thành từ cấu trúc tự tương quan tuyến tính ( $L_t$ ) và thành phần phi tuyến ( $N_t$ ). Đó là:  $y_t = L_t + N_t$

+ Sử dụng mô hình ARIMA để dự đoán  $y_t$ . Kết quả dự đoán là  $\hat{L}_t$ , và et là phần dư giữa các chuỗi của mô hình ARIMA.

$$et = y_t - \hat{L}_t$$

+ et được dùng làm đầu vào của mô hình RBFNN

2.3. Li Wang, Haofei Zou, Jia Su, Ling Li and Sohail Chaudhry: “Mô hình lai ARIMA-ANN

trong dự báo chuỗi thời gian”. Tác giả đề xuất một mô hình lai đặc biệt trong việc tích hợp các lợi thế của ARIMA và ANN trong việc mô hình hóa các hành vi tuyến tính và phi tuyến trong tập dữ liệu.

Trong đó  $L_t$  đại diện cho thành phần tuyến tính và  $N_t$  là thành phần phi tuyến

Tác giả sử dụng mô hình cộng ( $L+N$ ) và mô hình nhân ( $L^*N$ ) để kết hợp hai mô hình trong phân tích chuỗi thời gian. Các biểu thức toán học cho hai trường hợp này được thể hiện bằng phương trình:

$$\text{Mô hình cộng: } \hat{y}_t = \hat{L}_t + \hat{N}_t$$

$$\text{Mô hình nhân: } \hat{y}_t = \hat{L}_t * \hat{N}_t$$

## 3. PHƯƠNG PHÁP NGHIÊN CỨU

Trong công trình này, bốn phương pháp dự báo bao gồm ARIMA, RBFNN, mô hình lai ARIMA-RBFNN, và mô hình lai ARIMA-RBFNN cải tiến. Các mô hình được trình bày ở các phần tiếp theo.

### 3.1 Mô hình dự báo ARIMA

Mô hình tự hồi quy tích hợp với trung bình di động(ARIMA) là một mô hình tuyến tính có khả năng biểu diễn cả chuỗi thời gian tĩnh lẫn không tĩnh. Mô hình ARIMA dựa vào các mẫu tự tương quan trong bản thân của chuỗi thời gian để sinh ra dự đoán. Hệ thống các phương pháp dùng để xác định, kiểm tra và cải tiến mô hình ARIMA có sự đóng góp rất lớn của hai nhà thống kê, G.E.P.Box và G.M.Jenkins. Do đó việc mô hình và dự đoán dựa trên mô hình ARIMA còn được gọi là phương pháp luận Box-Jenkins [1]

Các mô hình chỉ mô tả chuỗi dùng hoặc những chuỗi đã sai phân hóa, nên mô hình ARIMA( $p,d,q$ ) thể hiện những chuỗi dữ liệu không dùng, đã được sai phân (ở đây,  $d$  chỉ mức độ sai phân).

Khi chuỗi thời gian dùng được lựa chọn (hàm tự tương quan ACF giảm đột ngột hoặc giảm đều nhanh), chúng ta có thể chỉ ra một mô hình dự định bằng cách nghiên cứu xu hướng của hàm tự tương quan ACF và hàm tự tương quan từng phần PACF.

Theo lý thuyết, nếu hàm tự tương quan

ACF giảm đột biến và hàm tự tương quan từng phần PACF giảm mạnh thì chúng ta có mô hình tự tương quan. Nếu hàm tự tương quan ACF và hàm tự tương quan từng phần PACF đều giảm đột ngọt thì chúng ta có mô hình hỗn hợp.

Về mặt lý thuyết, không có trường hợp hàm tự tương quan ACF và hàm tự tương quan từng phần cùng giảm đột ngọt. Trong thực tế, hàm tự tương quan ACF và hàm tự tương quan từng phần PACF giảm đột biến khá nhanh. Trong trường hợp này, chúng ta nên phân biệt hàm nào giảm đột biến nhanh hơn, hàm còn lại được xem là giảm đều. Do đó lúc sẽ có trường hợp giảm đột biến đồng thời khi quan sát biểu đồ hàm tự tương quan ACF và hàm tự tương quan từng phần PACF, biện pháp khắc phục là tìm vài dạng hàm dự định khác nhau cho chuỗi thời gian dừng. Sau đó, kiểm tra độ chính xác mô hình tốt nhất. Hình 3.2: Sơ đồ mô phỏng mô hình ARIMA

Mô hình ARIMA (1, 1, 1) :  $y(t) - y(t-1) = a_0 + a_1(y(t-1) - y(t-2) + e(t) + b_1e(t-1))$

Hoặc  $z(t) = a_0 + a_1z(t-1) + e(t) + b_1e(t-1)$ ,

Với  $z(t) = y(t) - y(t-1)$  ở sai phân đầu tiên:  $d = 1$ .

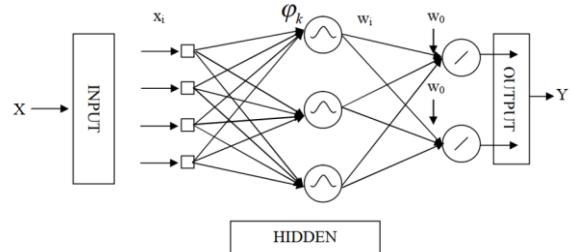
Tương tự ARIMA(1,2,1):  $h(t) = a_0 + a_1z(t-1) + e(t) + b_1e(t-1)$ ,

Với  $h(t) = z(t) - z(t-1)$  ở sai phân thứ hai:  $d = 2$ .

Tuy nhiên, trong thực hành  $d$  lớn hơn 2 rất ít được sử dụng.

### 3.2 Mô hình dự báo RBFNN (Radial Basis Function Neural Networks)

RBFNN là một loại mạng Neural nhân tạo truyền thống gồm có ba lớp. Nó bao gồm  $n$  nút của lớp đầu vào cho vector đầu vào  $x \in \mathbb{R}^n$ ,  $N$  neuron ẩn (giá trị của neuron ẩn thứ  $k$  chính là giá trị trả về của hàm cơ sở bán kính  $\varphi_k$ ) và  $m$  neuron đầu ra.



Hình 3.1: Mô hình RBFNN

Mô hình RBFNN có thể biểu diễn bằng công thức toán học sau:

$$\varphi(x) = \sum_{k=1}^N (w_k \varphi_k(x) + w_{0k}) = \sum_{k=1}^N \left( w_k e^{-\frac{\|x - v^k\|^2}{\sigma^2_k}} + w_{0k} \right) = y^j$$

Với tầng ẩn thì thường dùng hàm tổng là hàm  $S = \|x - w\|^2$ , còn hàm chuyển là hàm Gauss  $\varphi(v) = e^{-v}$

Tầng ra thì dùng hàm tổng là hàm  $S = \sum_{i=1}^N w_i x_i$ , hàm chuyển là hàm tuyến tính  $\varphi(v) = av$

Có nhiều cách huấn luyện mạng RBFNN. Có thể tách riêng một pha để xác định các tham số độ rộng  $\sigma_k$  của mỗi hàm bán kính và sau đó tìm các tham số  $w_k$  (phương pháp 2 pha) hoặc huấn luyện 1 lần nhờ tìm cực tiểu sai số tổng các bình phương.

### 3.3 Mô hình lai ARIMA-RBFNN

Theo [6] tác giả L. Zhang và cộng sự, xem xét một chuỗi thời gian ( $y_t$ ) được cấu thành từ cấu trúc tự tương quan tuyến tính ( $L_t$ ) và thành phần phi tuyến ( $N_t$ ). Đó là:

$$y_t = L_t + N_t$$

Tác giả dự đoán chuỗi thời gian bằng mô hình lai ARIMA và RBFNN như sau:

+ Mô hình ARIMA (Box et al., 1994) đã được sử dụng để dự đoán  $y_t$  và để cho  $\hat{L}_t$  biểu thị kết quả dự đoán. Các  $e_t$  là phần dư giữa các chuỗi của mô hình ARIMA.

$$e_t = y_t - \hat{L}_t$$

+  $e_t$  được coi là đầu vào của mô hình RBFNN (Moody and Darken, 1989), sau đó mô hình RBFNN có thể được biểu thị như sau:

$$e_t = f(e_{t-1}, e_{t-2}, \dots, e_{t-n}) + \varepsilon_t$$

Trong đó,  $f$  là hàm phi tuyến được xác định bởi mạng nơ ron và  $\varepsilon_t$  là lỗi ngẫu nhiên.

Kết quả đầu ra của RBFNN được định

nghĩa là  $\hat{N}_t$ .

+ Hai mô hình được kết hợp để dự báo và kết quả dự đoán từ các mô hình lai ARIMA-RBFNN được biểu thị như sau:

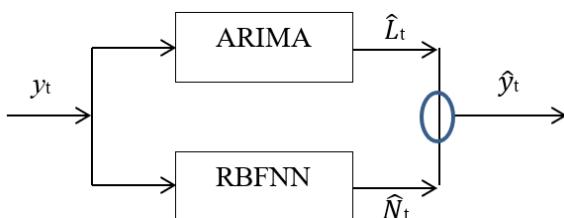
$$\hat{y}_t = \hat{L}_t + \hat{N}_t$$

Vì vậy, kết quả dự đoán được tạo ra thông qua mô hình lai ARIMA-RBFNN đã thu được thông qua sự kết hợp giữa dự đoán tuyến tính của ARIMA và kết quả dự đoán phi tuyến được dự đoán bởi mô hình RBFNN (thông qua phần lỗi của mô hình ARIMA).

### 3.4 Mô hình lai ARIMA-RBFNN cải tiến

Trong các mô hình lai ARIMA-RBFNN hiện nay, đa số các tác giả thường phân tích chuỗi thời gian thành 2 thành phần, tuyến tính và phi tuyến tính, sau đó sử dụng mô hình ARIMA để dự báo trên chuỗi thời gian, kết quả của mô hình ARIMA sẽ gồm 2 phần, phần kết quả dự báo và phần lỗi (Thành phần phi tuyến), phần lỗi này tiếp tục được sử dụng để dự báo bằng mô hình RBFNN. Kết quả cuối cùng các tác giả sử dụng phết cộng hoặc nhân hai kết quả của hai mô hình.

Tuy nhiên, hiện tại các tác giả đang thực hiện tuần tự từng mô hình, sau đó gộp kết quả lại. Để kiểm nghiệm về thời gian cũng như kết quả dự báo, chúng ta sử dụng mô hình lai cải tiến bằng cách thực hiện song song hai mô hình, từ kết quả đạt được, chúng ta sẽ xem xét để đề xuất mô hình tốt hơn.



Hình 3.2: mô hình ARIMA-RBFNN cải tiến

Gọi  $\hat{L}_t$  là giá trị dự báo của mô hình ARIMA,  $\hat{N}_t$  là giá trị dự báo của mô hình RBFNN, giá trị dự báo của  $y$  được tính như sau:

$$\hat{y} = \alpha \hat{L}_t + (1 - \alpha) \hat{N}_t \quad \alpha \in (0,1)$$

Để xác định tham số trọng số  $\alpha$ , chúng ta sẽ tìm giá trị của  $\alpha$  để hệ số dự báo lỗi MSE là

nhỏ nhất.

$$MSE = \sum_{i=1}^n (Y_i - Y_{hybrid,i})^2 = \sum_{i=1}^n (Y_i - [\alpha Y_{NN,i} + (1 - \alpha) Y_{DTW,i}])^2$$

Trong đó  $Y_i$  là giá trị thực tế tại thời điểm  $i$ ,  $Y_{NN,i}$  là giá trị dự báo tại thời điểm  $i$  được tạo bởi ANN và  $Y_{DTW,i}$  là giá trị dự báo tại thời điểm  $i$  được tạo bởi khớp mẫu trong DTW. Đây là một hàm bậc hai, do đó chúng ta có thể rút ra giá trị của  $\alpha$  làm cho lỗi dự báo MSE nhỏ nhất như sau:

$$\alpha = \frac{\sum_{i=1}^n (Y_{NN,i} - Y_{DTW,i})(Y_i - Y_{DTW,i})}{\sum_{i=1}^n (Y_{NN,i} - Y_{DTW,i})^2}$$

Vì  $\alpha$  nằm trong phạm vi  $[0, 1]$ , nếu giá trị tính toán của  $\alpha$  là âm, chúng ta có thể chọn giá trị của nó là 0 và nếu giá trị tính toán của lớn hơn 1, chúng ta có thể chọn giá trị của nó là 1.

## 4. KẾT QUẢ VÀ PHÂN TÍCH THỰC NGHIỆM

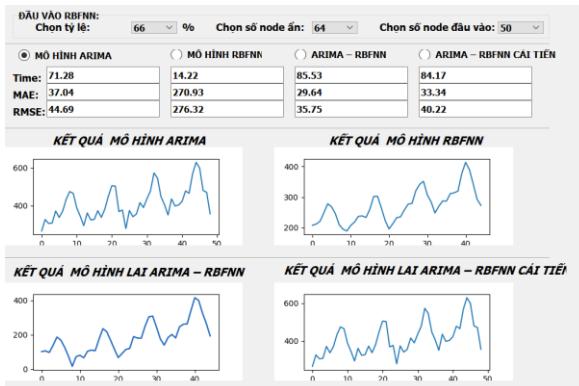
Nghiên cứu này được thực nghiệm trên bốn bộ dữ liệu thực tế gồm: Dữ liệu AirPassengers, Sunspots, Dentists, City\_temperature. Các bộ dữ liệu này được cộng đồng mạng về khai phá dữ liệu công bố [15].

Trong phần thực nghiệm này, tập dữ liệu được chia thành 2 phần, 66% dùng cho training và 34% dùng cho testing. Tuy nhiên tỷ lệ này có thể thay đổi cho các tập dữ liệu khác nhau. (trong demo cho phép tự chọn tỷ lệ cho hai phần training và testing).

### 4.1. Thực nghiệm trên tập dữ liệu AirPassengers

Bảng 4.1: Kết quả thực nghiệm trên tập dữ liệu Arpassanger

Mô hình	AirPassengers 64		
	Time (s)	RMSE	MAE
ARIMA	268.16	44.81	37.3
RBFNN	33.74	152.07	116.09
ARIMA- RBFNN	291.32	35.85	29.84
ARIMA- RBFNN cải tiến	129.62	40.33	32.54
Tỷ lệ ARIMA- RBFNN cải tiến so với ARIMA- RBFNN	44.49%	112.50%	109.05%
Chênh lệch tỷ số giữa ARIMA- RBFNN cải tiến so với ARIMA- RBFNN	55.51%	-12.50%	-9.05%



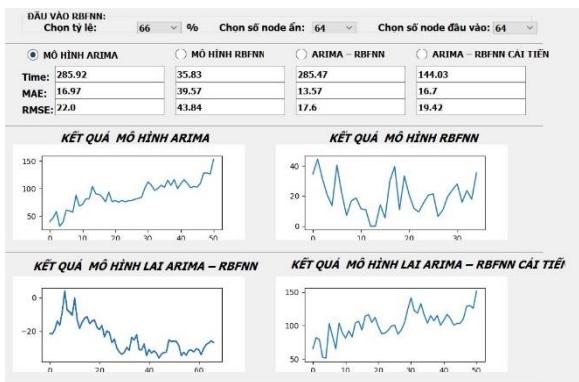
Hình 4.1: Biểu đồ thể hiện kết quả dự báo của các mô hình

Với kết quả thực nghiệm chạy trên tập dữ liệu AirPassengers cho thấy thời gian thực thi của ARIMA-RBFNN cải tiến tốt hơn mô hình ARIMA-RBFNN, tuy nhiên xét về mặt chính xác thì chưa được cải thiện cụ thể như trong bảng 4.1 cho thấy, thời gian cải thiện được 55.51%, tuy nhiên độ chính xác lại giảm 12.5%

## 4.2. Thực nghiệm trên tập dữ liệu Sunspots

Bảng 4.2: Kết quả thực nghiệm trên tập dữ liệu Sunspots

Mô hình	Sunspots 64		
	Time (s)	RMSE	MAE
ARIMA	285.92	22.0	16.97
RBFNN	35.83	43.84	39.57
ARIMA- RBFNN	285.47	17.6	13.57
ARIMA- RBFNN cải tiến	144.03	19.42	16.7



Hình 4.2: Biểu đồ thể hiện kết quả dự báo của các mô hình

Với kết quả thực nghiệm chạy trên tập dữ liệu Sunspots cho thấy mô hình ARIMA-RBFNN hiệu quả nhất, trong khi đó mô hình ARIMA-RBFNN cải tiến lại cho kết quả không tốt về thời gian thực thi vẫn tốt hơn các mô hình khác, nguyên nhân có thể do tập dữ

liệu không theo một xu hướng nhất định.

## 4.3. Thực nghiệm trên tập dữ liệu Dentists

Bảng 4.3: Kết quả thực nghiệm trên tập dữ liệu Dentists

Mô hình	Dentists 64		
	Time(s)	RMSE	MAE
ARIMA	2273.96	76.4	5.01
RBFNN	28.22	2.98	2.55
ARIMA- RBFNN	2249.31	61.12	4.01
ARIMA- RBFNN cải tiến	1489.94	72.51	4.58



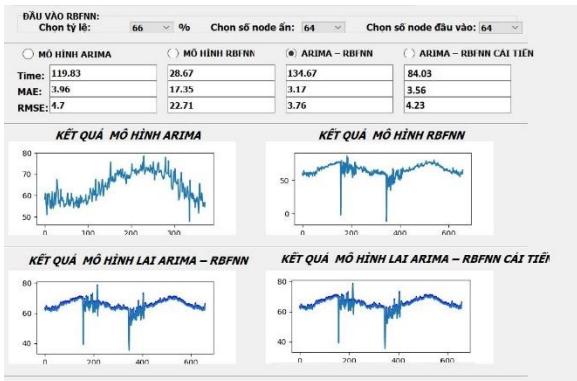
Hình 4.3: Biểu đồ thể hiện kết quả dự báo của các mô hình

Khi thực nghiệm trên tập dữ liệu được tăng lên về kích thước (độ dài chuỗi 2135), thì các mô hình dự báo cho kết quả tốt hơn rất nhiều so với các tập dữ liệu nhỏ hơn (độ dài chuỗi 150). Xét về thời gian thực thi thì mô hình ARIMA-RBFNN cải tiến nhanh hơn mô hình ARIMA-RBFNN, trong khi đó thì kết quả dự báo của mô hình ARIMA-RBFNN cải tiến lại thấp hơn ARIMA-RBFNN.

## 4.4. Thực nghiệm trên tập dữ liệu City\_temperature

Bảng 4.4: Kết quả thực nghiệm trên tập dữ liệu City\_temperature

Mô hình	City_temperature 64		
	Time(s)	RMSE	MAE
ARIMA	119.83	4.7	3.96
RBFNN	28.67	22.71	17.35
ARIMA- RBFNN	134.67	3.76	3.17
ARIMA- RBFNN cải tiến	84.03	4.23	3.56



Hình 4.4: Biểu đồ thể hiện kết quả dự báo của các mô hình

Khi tập dữ liệu đủ lớn thì thời gian thực thi của mô hình ARIMA-RBFNN cải tiến là tốt nhất, nhanh gấp gần 3 lần so với mô hình ARIMA-RBFNN tuy nhiên về độ chính xác lại không bằng ARIMA-RBFNN điều đó cho thấy vấn đề tiền xử lý dữ liệu là điều rất quan trọng và nó ảnh hưởng rất nhiều tới kết quả dự báo của các mô hình.

## 5. KẾT LUẬN

### 5.1. Kết quả đạt được

Sau khi thực nghiệm trên bốn tập dữ liệu, mỗi tập dữ liệu thực nghiệm hai trường hợp và cho kết quả như sau:

- Về thời gian thực thi và tài nguyên sử dụng:

Các trường hợp thực nghiệm cho thấy, thời gian thực thi của các mô hình phụ thuộc nhiều vào tập dữ liệu, khi tập dữ liệu nhỏ thì thời gian thực thi của các mô hình không thay đổi nhiều, mô hình ARIMA-RBFNN cải tiến so với các mô hình khác thì không hiệu quả, tuy nhiên khi dự báo trên tập dữ liệu càng lớn thì sự chênh lệch về mặt thời gian giữa mô hình ARIMA-RBFNN cải tiến với mô hình ARIMA-RBFNN càng lớn. Điều này cho thấy mô hình ARIMA-RBFNN cải tiến có thể ứng dụng trong nhiều lĩnh vực, vì hiện nay trong các lĩnh vực thì lượng dữ liệu ngày càng nhiều.

Mô hình ARIMA-RBFNN cải tiến nhanh hơn mô hình ARIMA-RBFNN trong các trường hợp thực nghiệm. Tuy nhiên về mức độ sử dụng tài nguyên thì mô hình ARIMA-RBFNN cải tiến sử dụng nhiều hơn. Qua bảng 4.10 thống kê mức độ sử dụng CPU

của các mô hình, thì mô hình ARIMA-RBFNN cải tiến tốt hơn về mặt thời gian nhưng chưa hiệu quả về sử dụng tài nguyên.

#### - Về độ chính xác

Trong thực nghiệm này sử dụng hai giá trị RMSE và MAE để đánh giá độ chính xác của các mô hình dự báo. Qua các thực nghiệm cho thấy mô hình ARIMA-RBFNN luôn cho kết quả thấp hơn mô hình ARIMA-RBFNN cải tiến, và các mô hình dự báo này cho kết quả tốt khi tập dữ liệu lớn. Nguyên nhân do tập dữ liệu đầu vào chưa được xử lý. Đối với mô hình ARIMA-RBFNN đã tận dụng được ưu điểm của từng mô hình, ARIMA xử lý tốt các thành phần tuyến tính trong tập dữ liệu và RBFNN xử lý tốt các thành phần phi tuyến còn lại (sau khi ARIMA đã thực thi) của tập dữ liệu, còn mô hình ARIMA-RBFNN cải tiến thực hiện song song trên hai tập dữ liệu riêng biệt do đó kết quả chưa như mong đợi. Điều này cho thấy nếu tập dữ liệu được xử lý trước khi thực hiện mô hình ARIMA-RBFNN cải tiến bằng cách phân tách thành hai tập dữ liệu riêng biệt, một tập chứa các thành phần tuyến tính của dữ liệu làm đầu vào cho mô hình ARIMA, tập còn lại chứa các thành phần phi tuyến của dữ liệu làm đầu vào cho mô hình RBFNN. Khi đó kết quả dự báo sẽ thay đổi theo hướng tích cực hơn.

### 5.2. Các mặt hạn chế

Khi cài đặt và thực nghiệm các mô hình dự báo trên các tập dữ liệu, từ kết quả cho thấy còn một số hạn chế

- Chưa thực nghiệm được trên nhiều tập dữ liệu lớn khác nhau.

- Kết quả chưa tốt với các tập dữ liệu nhỏ khi sử dụng mô hình ARIMA-RBFNN cải tiến.

- Đề tài mới thực nghiệm song song trên nhiều chương trình (multi processing), chưa thực hiện song song trên nhiều CPU (multi processor)

### 5.3. Hướng phát triển

Đề tài đã cải tiến mô hình lai ARIMA-RBFNN, mặc dù vẫn còn nhiều hạn chế, nhưng những hạn chế này có thể khắc phục:

- Tiên xử lý dữ liệu, bằng cách phân tách tập dữ liệu thành hai thành phần riêng biệt tuyến tính và phi tuyến, sau đó thực hiện song song hai mô hình ARIMA và RBFNN. Khi đó mô hình ARIMA-RBFNN cải tiến sẽ cho kết quả tốt hơn hiện tại.

- Thực nghiệm trên nhiều tập dữ liệu lớn

- Thực hiện song song trên nhiều CPU (multi processor)

- Cải tiến cách xác định các tham số cho mô hình ARIMA và RBFNN.

## TÀI LIỆU THAM KHẢO

- [1]. H.S. Behera, Sibarama Panigrahi. "A hybrid ETS–ANN model for time series forecasting". Engineering Applications of Artificial Intelligence, Volume 66, Pages 49-59 (2017).
- [2]. [http://en.wikipedia.org/wiki/Time\\_series](http://en.wikipedia.org/wiki/Time_series). Time series.
- [3]. Habiluddina, Ahmad Jawahirb. "Comparing of ARIMA and RBFNN for short-term forecasting". International Journal of Advances in Intelligent Informatics ISSN: 2442-6571 Vol. 1, No 1, pp. 15-22 (2015).
- [4]. N. Vijay and G.C. Mishra. "Time Series Forecasting Using ARIMA and ANN Models for Production of Pearl Millet (BAJRA) Crop of Karnataka, India". International Journal of Current Microbiology and Applied Sciences ISSN: 2319-7706 Volume 7 Number 12 (2018).
- [5]. Jonathan D. Cryer, Kung-Sik Chan. Time Series Analysis, Springer Texts in Statistics (2008).
- [6]. L. Zhang, G. X. Zhang, and R. R. Li. "Water Quality Analysis and Prediction Using Hybrid Time Series and Neural Network Models". JAST\_Volume 18\_Issue 4\_Pages 975-983 (2018).
- [7]. Rakhlin; Kalinin; Shvets; Iglovikov. "Automatic Instrument Segmentation in Robot-Assisted Surgery using Deep Learning". IEEE International Conference on Machine Learning and Applications (ICMLA), (2018).
- [8]. Li Wang Haofei Zou Jia Su Ling Li Sohail Chaudhry. "An ARIMA - ANN Hybrid Model for Time Series Forecasting" . Systems Research and Behavioral Science Syst. Res. 30, 244-259, (2013).
- [9]. Parviz, L., Kholghi, M. and Hoorfar. "A Comparison of the Efficiency of Parameter Estimation Methods in the Context of Stream Flow Forecasting". J. Agr. Sci. Tech., 12(1): 47-60, (2010).
- [10]. Ross Ihaka. Time Series Analysis, Lecture Notes for 475.726, Statistics Department, University of Auckland, (2005).
- [11]. Roy Batchelor. Box-Jenkins Analysis. Lecture Notes, ESCP-EAP, Paris, 2004
- [12]. Wu, Zonghan, et al. "Connecting the Dots: Multivariate Time Series Forecasting with Graph Neural Networks." arXiv preprint arXiv:2005.11650 (2020).

- [13]. Lâm Hoàng Vũ. “Dự báo chuỗi thời gian sử dụng mô hình ARIMA và giải thuật di truyền”. Luận văn thạc sĩ, Trường Đại học Bách Khoa, TPHCM, (2012).
- [14]. Nguyễn Chí Thành - Hà Gia Sơn. “Kết hợp mạng Noron FIR và mô hình ARIMA theo hình thức động để nâng cao hiệu quả dự báo chuỗi thời gian”. Tạp chí Nghiên cứu KH&CN quân sự, Số Đặc san CNTT, 12 – 2017.
- [15]. <https://www.kaggle.com>: Nguồn dữ liệu trong thực nghiệm.
- [16]. S. K. Lahiri, K.C. Ghanta. “Artificial neural network model with the parameter tuning assisted by a differential evolution technique: the study of the hold up of the slurry flow in a pipeline”. Chemical Industry & Chemical Engineering Quarterly 15 (2) 103–117, (2009).
- [17]. Miroslav R. Radovanović. Optimal Selection of ANN Training and Architectural Parameters Using Taguchi Method: A Case Study. Full Professor University of Niš Faculty of Mechanical Engineering.

**Tác giả chịu trách nhiệm bài viết:**

Họ tên: Phạm Chí Công

Đơn vị: Trường Đại Học Sư Phạm Kỹ Thuật TPHCM

Điện thoại: 0938 065 567

Email: 1981301@student.hcmute.edu.vn

