ISSN 2354-0575

PHƯƠNG PHÁP BÌNH PHƯƠNG TỐI THIỂU TRONG DỰ BÁO

Nguyễn Kiều Hiên[1], Nguyễn Hữu Mộng[2], Hà Gia Sơn[3], Trần Ngọc Tuấn[2]
1 Trường Đại học Sao Đỏ
2 Trường Đại học Sư phạm Kỹ thuật Hưng Yên
3 Trường ĐHCN Việt-Hung

Ngày nhận: 10/01/2017
Ngày sửa chữa: 20/02/2017
Ngày xét duyệt: 20/03/2017

Tóm tắt:
Bài viết trình bày phương pháp bình phương tối thiểu cho một chuỗi thời gian cũng như trong việc kết hợp các mô hình dự báo khác nhau. Tính hiệu quả của phương pháp được minh hoạ thông qua ví dụ thực tế. Kết quả cho thấy, việc kết hợp nhiều mô hình dự báo bằng phương pháp bình phương tối thiểu cho sai số trung bình bé hơn so với các mô hình dự báo thông thường.
Từ khóa: *Mô hình dự báo, Mô hình ARIMA, chuỗi thời gian, bình phương tối thiểu.*

1. ĐẶT VẤN ĐỀ

Dự báo là sự tiên đoán có căn cứ khoa học, mang tính chất xác suất về mức độ, nội dung, các mối quan hệ, trạng thái, xu hướng phát triển của đối tượng nghiên cứu hoặc về cách thức và thời hạn đạt được các mục tiêu nhất định đã đề ra trong tương lai. Trong thời đại công nghệ thông tin và toàn cầu hóa, dự báo lại đóng vai trò quan trọng hơn khi nhu cầu về thông tin thị trường, tình hình phát triển trong tương lai càng cao. Ở nước ngoài, đã có nhiều công trình nghiên cứu về vấn đề này, đã có một hệ thống lý thuyết gồm nhiều phương pháp, qui trình cũng như nhiều mô hình để dự báo tương lai [15]. Tài liệu [12] đã phân tích và thăm dò các yếu tố của chuỗi thời gian, các mô hình của chuỗi thời gian, quy trình Box-Jenkins dành để dự báo. Tài liệu [13] nêu tổng quan về các phương pháp dự báo trong kinh doanh, phân tích kỹ các phương pháp dự báo ngắn hạn, dự báo dài hạn. Các công trình [7-11], [14] đã đưa ra ý tưởng, giải thuật và những ứng dụng để kết hợp mạng nơ ron với các mô hình khác để tăng cường hiệu quả dự báo. Trong thời gian gần đây, ở trong nước, chúng ta đã quan tâm nhiều hơn tới lĩnh vực dự báo, đã có nhiều đề tài các cấp, với những mục đích và cách tiếp cận khác nhau về dự báo, điển hình là các công trình [1-6].

Tổng hợp các công trình nghiên cứu cho thấy, ngày càng xuất hiện những mô hình có hiệu quả cao, tuy nhiên, khi dự báo, có nhiều mô hình được thiết lập và người ta thường chọn mô hình có hiệu quả cao nhất và bỏ qua các mô hình khác, điều này gây ra một sự lãng phí, chính vì vậy, tác giả bài viết này sẽ đưa ra giải pháp phối hợp nhiều mô hình dự báo để tổng hợp thành một mô hình có hiệu quả cao hơn bằng phương pháp bình phương tối thiểu. Bài viết còn trình bày một ví dụ thực tế. Kết quả của ví dụ thực tế cho thấy, giải pháp phối hợp các mô hình dự báo mà tác giả đưa ra có hiệu quả cao hơn các mô hình dự báo phổ biến khác.

2. KẾT HỢP CÁC MÔ HÌNH DỰ BÁO VÀ PHƯƠNG PHÁP BÌNH PHƯƠNG TỐI THIỂU

2.1. Một số mô hình phổ biến

Trong dự báo, số liệu trong quá khứ và hiện tại quyết định xu hướng vận động của các hiện tượng trong tương lai.

Theo các tài liệu đã công bố (xem [15]), có hai hướng dự báo là dự báo định tính và dự báo định lượng.

- Dự báo định tính là dự báo dựa trên phán đoán chủ quan, trực giác của người ra quyết định. Phương pháp phổ biến là lấy phiếu thăm dò và thu thập ý kiến như lấy ý kiến các nhà phân phối, người tiêu dùng, chuyên gia…. Nhược điểm chung của phương pháp này là mang tính chủ quan, kinh nghiệm và cảm tính.

- Dự báo định lượng là phương pháp dự báo dựa vào các mô hình dự báo. Các mô hình dự báo được xây dựng bằng các công cụ toán học và dựa vào các qui luật tự nhiên nên khắc phục được phần nào tính chủ quan và cảm tính của người làm dự báo. Ta xét một số mô hình phổ biến sau.

Mô hình chuỗi thời gian tự hồi quy hoàn toàn có dạng sau:

$$Y_t = \alpha_1 Y_{t-1} + \alpha_2 Y_{t-2} + ... + \alpha_p Y_{t-p} + u_t, \quad (2.1)$$

trong đó, Y_t là quan sát thứ t đối với biến phụ thuộc và u_t là thành phần sai số, Y_{t-i} là quan sát thứ i trước Y_t, $\alpha_1, \alpha_2, ..., \alpha_p$ là các hệ số trong các lần quan sát trước thời điểm t.

Mô hình trung bình trượt (MA – Moving Average) có dạng

$$Y_t = v_t - \beta_1 v_{t-1} - \beta_2 v_{t-2} - ... - \beta_q v_{t-q}, \quad (2.2)$$

trong đó, v_t là nhiễu trắng tại thời điểm t, v_{t-i} là nhiễu trắng tại thời điểm $t-i$, $\beta_1, \beta_2, ..., \beta_q$ là các hệ số

nhiễu trắng. Do đó, Y_t là tổ hợp tuyến tính của các biến ngẫu nhiên nhiễu trắng.

Mô hình ARMA (Auto Regressive Moving Average) là mô hình kết hợp hai mô hình tự hồi quy và mô hình trung bình trượt. Do đó, mô hình ARMA (p, q) có dạng tổng quát là

$$Y_t = \alpha_1 Y_{t-1} + \alpha_2 Y_{t-2} + ... + \alpha_p Y_{t-p} + u_t + \\ + v_t - \beta_1 v_{t-1} - \beta_2 v_{t-2} - ... - \beta_q v_{t-q}, \quad (2.3)$$

Xét một chuỗi thời gian không có tính dừng. Khi đó, ta có thể lập được một chuỗi thời gian dừng từ các sai phân cấp d (sau d lần lấy sai phân) rồi áp dụng mô hình ARMA (p, q) ta sẽ được mô hình ARIMA(p,q,d). Đây là mô hình dự báo phổ biến, được sử dụng nhiều.

2.1. Phương pháp bình phương tối thiểu

Trong lý thuyết xử lý các dữ liệu thống kê người ta đã đưa ra phương pháp bình phương tối thiểu xây dựng một hàm (thường là đa thức bậc k) để xấp xỉ quá trình đang xét được cho bởi một bảng các giá trị như sau (chuỗi thời gian) như sau:

x	x_1	x_2	...	x_n
y	y_1	y_2	...	y_n

Với bảng cho trước này, tồn tại duy nhất một đa thức bậc m

$$\varphi_m(x) = a_m x^m + a_{m-1} x^{m-1} + ... + a_1 x + a_0 \quad (2.4)$$

thoả mãn điều kiện

$$\sum_{i=1}^{n} \left(\varphi_m(x_i) - y_i\right)^2 \to \min. \quad (2.5)$$

Đa thức xấp xỉ $\varphi_m(x)$ được xác định một cách duy nhất bởi các hệ số a_0, a_1,..., a_m được xác định duy nhất từ hệ phương trình đại số tuyến tính sau

$$\sum_{i=1}^{n} \left(\varphi_m(x_i) - y_i\right) x_i^{m-k} = 0, \quad k = 0, 1, .. m. \quad (2.6)$$

Do đó, để xác định đa thức $\varphi_m(x)$ ta phải giải hệ (2.6) với m+1 phương trình với m+1 ẩn a_0, a_1, ..., a_m.

Như vậy, đa thức $\varphi_m(x)$ cũng có thể được coi là một mô hình dự báo. Đa thức $\varphi_m(x)$ được xác định từ hệ (2.6) sẽ cho ta tổng các bình phương của các sai số tại các điểm quan sát là nhỏ nhất trong vô số các đa thức bậc m.

Ý tưởng của phương pháp bình phương tối thiểu trên đây cũng có thể áp dụng để xây dựng một mô hình dự báo kết hợp nhiều mô hình dự báo với nhau.

Giả sử ta có m mô hình dự báo khác nhau. Đại lượng dự báo tính theo các mô hình kí hiệu là $X_1(t), X_2(t), ..., X_m(t)$. Khi đó, ta có thể xét một mô hình dự báo mới như sau

$$F(t) = \beta_0 + \beta_1 X_1(t) + ... + \beta_m X_m(t), \quad (2.7)$$

trong đó $\beta_0, \beta_1, ..., \beta_m$ là các trọng số cần xác định. $F(t)$ được coi là mô hình kết hợp các mô hình dự báo đã có $X_1(t), X_2(t), ..., X_m(t)$. Do các trọng số là tuỳ ý nên ta có vô số mô hình kết hợp. Vì vậy, ta sẽ tìm một mô hình dự báo $F(t)$ sao cho

$$\sum_{i=1}^{n} \left(Y_i - F(t_i)\right)^2 \to \min. \quad (2.8)$$

Để xác định $F(t)$ thoả mãn điều kiện (2.8) ta coi biểu thức $\sum_{i=1}^{n} \left(Y_i - F(t_i)\right)^2$ là hàm m + 1 biến và kí hiệu là $E(\beta_0, \beta_1, ..., \beta_m)$. Hàm này xác định dương và chắc chắn có cực tiểu. Điểm cực tiểu $\left(\widehat{\beta_0}, \widehat{\beta_1}, ..., \widehat{\beta_m}\right)$ thoả mãn điều kiện sau đây:

$$\begin{cases} \frac{\partial E}{\partial \beta_0} = -2 \sum_{i=1}^{n} \left(\beta_0 + \beta_1 X_1(t_i) + ... + \beta_m X_m(t_i) - Y_i\right) = 0, \\ \frac{\partial E}{\partial \beta_1} = -2 \sum_{i=1}^{n} \left(\beta_0 + \beta_1 X_1(t_i) + ... + \beta_m X_m(t_i) - Y_i\right) X_{1i} = 0, \\ ..... \\ \frac{\partial E}{\partial \beta_m} = -2 \sum_{i=1}^{n} \left(\beta_0 + \beta_1 X_1(t_i) + ... + \beta_m X_m(t_i) - Y_i\right) X_{mi} = 0. \end{cases} \quad (2.9)$$

Đây là hệ m + 1 phương trình đại số tuyến tính với m + 1 ẩn $\beta_0, \beta_1, ..., \beta_m$ có thể giải được bằng nhiều phương pháp đúng khác nhau, tuy nhiên, khi m lớn thì hệ này chỉ có thể giải được bằng các phương pháp gần đúng. Hệ (2.9) có thể viết gọn lại dưới dạng tổng quát của hệ phương trình đại số tuyến tính.

3. VÍ DỤ ÁP DỤNG

Xét ví dụ gồm 18 quan sát như trong bài báo [4]. Trong bài báo này tác giả xây dựng 2 mô hình dự báo là SARIMA và SCARIMA. Đối với mỗi mô hình tác giả đã tính giá trị dự báo tại các mốc thời gian và ME của các mô hình. Ở đây, chúng tôi sử dụng chuỗi thời gian này và xét trực tiếp chuỗi thời gian đã cho bằng phương pháp bình phương tối thiểu, mô hình kết hợp ARMA và mô hình kết hợp SARIMA với SCARIMA theo phương pháp bình phương tối thiểu.

Trong Bảng 1 chuỗi thời gian là các giá trị ở cột thứ 2. Các giá trị này ta kí hiệu là y_1, y_2,..., y_n (n = 18). Cột thời gian là các ngày tháng trong năm. Ta đặt

$$t_1 = 1, t_2 = 2, ..., t_n = 18$$

Trong trường hợp này ta xấp xi hàm dự báo bằng một hàm tuyến tính $y = ax + b$, trong đó, các hệ số a, b được xác định từ hệ phương trình sau đây:

$$\begin{cases} \left(\sum_{i=1}^{n} t_i^2\right) a + \left(\sum_{i=1}^{n} t_i\right) b = \sum_{i=1}^{n} t_i y_i, \\ \left(\sum_{i=1}^{n} t_i\right) a + nb = \sum_{i=1}^{n} y_i. \end{cases}$$

hay $\begin{cases} 2109a + 171b = 0{,}487; \\ 171a + 18b = 0{,}0628. \end{cases}$

Giải hệ này ta được a = -0,0002778; b = 0,00612799.

Kết quả tính các giá trị dự báo tương ứng tại các mốc thời gian đã cho cũng như sai lệch tuyệt đối trung bình được cho trong Bảng 1.

Bảng 1.

T	y	t2	ty	Ydb	E
1	0,0125	1	0,0125	0,00585	0,00665
2	0,0132	4	0,0264	0,005572	0,007628
3	-0,0019	9	-0,0057	0,005295	0,007195
4	0,0002	16	0,0008	0,005017	0,004817
5	-0,0006	25	-0,003	0,004739	0,005339
6	0,0005	36	0,003	0,004461	0,003961
7	0,0003	49	0,0021	0,004183	0,003883
8	0,0008	64	0,0064	0,003906	0,003106
9	0,0106	81	0,0954	0,003628	0,006972
10	0,0049	100	0,049	0,00335	0,00155
11	0,0034	121	0,0374	0,003072	0,000328
12	0,0051	144	0,0612	0,002794	0,002306
13	0,0069	169	0,0897	0,002517	0,004383
14	0,0055	196	0,077	0,002239	0,003261
15	-0,0044	225	-0,066	0,001961	0,006361
16	0,0008	256	0,0128	0,001683	0,000883
17	0,0020	289	0,034	0,001405	0,000595
18	0,0030	324	0,054	0,001128	0,001872
171	0,0628	2109	0,487	-0,04138	0,104176
				ME	0,009737

Kết quả tính theo mô hình kết hợp ARMA được cho trong cột Y_{ARMA} ở Bảng 2. Đối với mô hình kết hợp SARIMA với ARIMA bằng phương pháp bình phương tối thiểu ta có $m = 2$, $n = 18$ và hệ phương trình xác định các trọng số $\beta_0, \beta_1, \beta_2$ là

$$\begin{cases} n\beta_0 + A_{12}\beta_1 + A_{13}\beta_2 = B_1, \\ A_{21}\beta_0 + A_{22}\beta_1 + A_{23}\beta_2 = B_2, \\ A_{31}\beta_0 + A_{32}\beta_1 + A_{33}\beta_2 = B_3. \end{cases} \quad (3.1)$$

trong đó,

$$A_{12} = A_{21} = \sum_{i=1}^{n} X_{1i},\ A_{13} = A_{31} = \sum_{i=1}^{n} X_{2i};$$

$$A_{22} = \sum_{i=1}^{n} X_{1i}^2,\ A_{33} = \sum_{i=1}^{n} X_{2i}^2;$$

$$A_{23} = A_{32} = \sum_{i=1}^{n} X_{1i}X_{2i};$$

$$B_1 = \sum_{i=1}^{n} Y_i,\ B_2 = \sum_{i=1}^{n} X_{1i}Y_i,\ B_3 = \sum_{i=1}^{n} X_{2i}Y_i.$$

Để tính các hệ số của hệ phương trình (3.1) ta lập bảng sau

T_i	X_{1i}	X_{2i}	X_{1i}^2	X_{2i}^2	$X_{1i}X_{2i}$	Y_i	$X_{1i}Y_i$	$X_{2i}Y_i$
T_1	X_{11}	X_{21}	X_{11}^2	X_{21}^2	$X_{11}X_{21}$	Y_1	$X_{11}Y_1$	$X_{21}Y_1$
….	....	....	....	....	....	....	....	....
T_n	X_{1n}	X_{2n}	X_{1n}^2	X_{2n}^2	$X_{11}X_{2n}$	Y_n	$X_{1n}Y_n$	$X_{2n}Y_n$
$\sum_1^n$	A_{12}	A_{13}	A_{22}	A_{33}	A_{23}	B_1	B_2	B_3

Bảng này được tính dễ dàng trong Excell, trong đó dòng cuối cùng là tổng của các cột tương ứng cho ta các giá trị cần tính trên.

Từ hệ (3.1) ta được

$$\beta_0 = \frac{B_1 - A_{12}\beta_1 - A_{13}\beta_2}{n}$$

Thay biểu thức này vào 2 phương trình sau ta được hệ 2 phương trình của β_1, β_2. Giải hệ hai phương trình này ta được

$$\beta_1 = \frac{B_3A_{22} - B_2A_{32}}{A_{22}A_{33} - (A_{23})^2};\ \beta_2 = \frac{B_2A_{33} - B_3A_{23}}{A_{22}A_{33} - (A_{23})^2}.$$

Thay các giá trị số của các biến từ Bảng 2 ta được $\beta_0 = -0.00033, \beta_1 = 0.42813, \beta_2 = 0.40231$.

Các giá trị dự báo tương ứng theo mô hình (2.7) cho trường hợp này được cho trong cột Y_{KH}. Các sai số tại các điểm dự báo và sai số trung bình được cho trong các cột E.

Bảng 2

STT	T	Y	sarima	scarima	E_{sarima}	$E_{scarima}$	Y_{ARMA}	Y_{KH}	E_{ARMA}	E_{KH}
1	2013M1	0.01250	0.01000	0.01020	0.00250	0.00230	0.00810	0.00436	0.00440	0.00814
2	2013M2	0.01320	0.02120	0.01790	0.00800	0.00470	0.01588	0.00805	0.00268	0.00515
3	2013M3	-0.00190	0.00100	0.00170	0.00290	0.00360	0.00078	0.00040	0.00268	0.00230
4	2013M4	0.00020	0.00000	0.00280	0.00020	0.00260	0.00091	0.00091	0.00071	0.00071
5	2013M5	-0.00060	0.00350	0.00530	0.00410	0.00590	0.00339	0.00207	0.00399	0.00267

6	2013M6	0.00050	0.00130	0.00150	0.00080	0.00100	0.00080	0.00031	0.00030	0.00019
7	2013M7	0.00030	-0.00130	0.00110	0.00160	0.00080	-0.00036	0.00013	0.00066	0.00017
8	2013M8	0.00080	0.00240	0.00300	0.00160	0.00220	0.00191	0.00101	0.00111	0.00021
9	2013M9	0.01060	0.00570	0.00540	0.00490	0.00520	0.00426	0.00212	0.00634	0.00848
10	2013M10	0.00490	0.00460	0.00470	0.00030	0.00020	0.00352	0.00180	0.00138	0.00310
11	2013M11	0.00340	0.00640	0.00500	0.00300	0.00160	0.00434	0.00194	0.00094	0.00146
12	2013M12	0.00510	0.00830	0.00590	0.00320	0.00080	0.00547	0.00237	0.00037	0.00273
13	2014M1	0.00690	0.01020	0.01120	0.00330	0.00430	0.00864	0.00483	0.00174	0.00207
14	2014M2	0.00550	0.01590	0.01620	0.01040	0.01070	0.01310	0.00719	0.00760	0.00169
15	2014M3	-0.00440	-0.00350	-0.00200	0.00090	0.00240	-0.00263	-0.00130	0.00177	0.00310
16	2014M4	0.00080	-0.00150	0.00110	0.00230	0.00030	-0.00044	0.00013	0.00124	0.00067
17	2014M5	0.00200	0.00460	0.00560	0.00260	0.00360	0.00394	0.00221	0.00194	0.00021
18	2014M6	0.00300	0.00190	0.00270	0.00110	0.00030	0.00158	0.00087	0.00142	0.00213
ME (Sai số trung bình)					**0.00298**	**0.00292**			**0.00229**	**0.00201**

3. KẾT LUẬN

Từ các kết quả ví dụ ta chưa thể kết luận được phương pháp bình phương tối thiểu áp dụng trực tiếp cho chuỗi thời gian và áp dụng kết hợp các mô hình dự báo là tốt nhất hay hơn hẳn các mô hình dự báo khác. Tuy nhiên, chúng tôi hy vọng, đề xuất của chúng tôi sẽ cho các nhà phân tích, nghiên cứu một phương pháp đơn giản và có thể rất hiệu quả.

Tài liệu tham khảo

[1]. Đỗ Quang Giám, Vũ Thị Hân (2012), *Xây dựng mô hình Arima cho dự báo khách du lịch quốc tế đến Việt nam*, Tạp chí Khoa học và Phát triển: Tập 10, số 2: 364 - 370, Trường ĐH Nông Nghiệp Hà Nội.

[2]. Vũ Thị Gương (2012), *Kỹ thuật khai phá dữ liệu chuỗi thời gian áp dụng trong dự báo chứng khoán*, luận án Thạc sĩ khoa học CNTT, Học viện Bưu chính Viễn Thông, Hà nội.

[3]. Nguyễn Trung Hòa (2007), "*Một số thuật toán mô phỏng và phân tích chuỗi thời gian*", Luận án Tiến sỹ Toán ứng dụng, trường ĐHBK Hà Nội, Hà Nội.

[4]. Nguyễn Khắc Hiếu, 2014, "*Mô hình ARIMA và dự báo lạm phát 6 tháng cuối năm 2014*", Tạp chí Kinh Tế và Dự Báo số 16, tháng 8-2014 .

[5]. Phạm Văn Khánh (2008), "*Phân tích thống kê dự báo và mô phỏng một số chuỗi thời gian*", Luận án Tiến sỹ Toán ứng dụng, ĐH Quốc gia Hà Nội, Hà Nội.

[6]. Võ Văn Tài, "*Dự báo sản lượng lúa Việt Nam bằng các mô hình toán học*", Tạp chí Khoa học 2012:23b 125-134.

[7]. C. Lee Giles, Steve Lawrence, A. C. Tsoi (2001), "*Noisy Time Series Prediction using a Recurrent Neural Network and Grammatical Inference*" - Machine Learning, Volume 44, Number 1/2, July/August, pp. 161–183,

[8]. Eric A Wan (2003), "*Finite Impulse Response Neural Networks for Autoregressive Time Series Prediction*", Proceedings of the NATO Advanced Workshop on Time Series Prediction and Analysis, Sante Fe, NM.

[9]. Eric A Wan (2004), *Finite Impulse Response Neural Networks with Application in Time Series Prediction - A Dissertation Submitted to the Department of Electrical Engineering and the Committee on Graduate Studies of Stanford University in Partial Fulfillment of the Requirements for the Degree of Doctor of Philosophy.*

[10]. Ho Joon Kim (2005), "*Time Series Prediction using an Interval Arithmetic FIR Network*", Neural Information Processing - Letters and Reviews Vol.8, No.3, September.

[11]. Luis Aburto, Richard Weber (2012), "*Demand Forecast in a Supermarket using a Hybrid Intelligent System*", Department of Industrial Engineering, University of Chile, pp 143-151.

[12]. Michael Falk , Frank Marohn (2012), "*A First Course on Time Series Analysis - Examples with SAS*", by Chair of Statistics, University of Wurzburg.

[13]. Michael K. Evans (2002), *Practical Bususiness Forecasting*, Blackwell Publishers Ltd, a Blackwell Publishing company. Bodmin, Cornwall.
[14]. Marek Hlav'acek (2009), *Seasonal Time Series Modeling via Neural Networks with Swithching Units*, PHD Czech Technical University Prague.
[15]. N.Gujarati (2004), *Basic Econometrics*, Fourth Edition-The McGraw−Hill Companies.

A LEAST SQUARES METHOD IN FORECASTING MODELS

Abstract:
This paper presents a least squares method as well as combines different models of forecasting on time series. The result is illustrated on practical examples. The part of application is combined with the data of actual process. As a result, the combined method is better than single forecast model.
Keywords: *Model of forecasting, ARIMA method, time series, least squares method.*