TÌM SỐ NÚT ẨN TỐI ƯU CỦA MẠNG THẦN KINH NHÂN TẠO (ANN) BẰNG LÝ THUYẾT CỰC TIỂU HOÁ MẠO HIỂM THEO CẤU TRÚC (SRM)

ThS. Lê Xuân Cầu
Trung tâm tư liệu Khí tượng Thủy văn

Tóm tắt: Vấn đề tìm nút ẩn tối ưu của ANN rất quan trọng và khó khăn trong nghiên cứu ANN và trong việc áp dụng ANN để giải các bài toán kỹ thuật. Cho đến nay, việc tìm số nút ẩn tối ưu thường được thực hiện bằng thực nghiệm. Do đó, nó càng khó khăn đối với các ANN lớn với hàng trăm, hàng nghìn nút ẩn. Bài báo này sẽ trình bày kết quả số trị của việc áp dụng lý thuyết "*Cực tiểu hoá mạo hiểm theo cấu trúc* (SRM)" để giải quyết vấn đề trên. SRM có thể dùng để phát triển một loạt các mô hình phi tuyến với độ phức tạp và dung lượng tăng dần. SRM dùng để đánh giá thể hiện tổng quan của mô hình đối với dữ liệu chưa biết.

I. Thể hiện tổng quan của ANN đối với dữ liệu chưa biết

Một trong các câu hỏi của người sử dụng các mô hình là mô hình nhận được sẽ thể hiện như thế nào trên tập các dữ liệu chưa biết và mô hình có được ứng dụng thành công trong thực tế hay không phụ thuộc rất lớn vào sai số dự báo có được của mô hình.

Khi một mô hình đã được xây dựng và một tiêu chuẩn để đánh giá chất lượng của mô hình được đưa ra (thường là sai số bình phương trung bình MSE), các tham số của mô hình thường được tìm sao cho sai số giữa kết quả tính của mô hình và các dữ liệu quan trắc được là nhỏ nhất. Vấn đề quan trọng cần phải làm sáng tỏ là mô hình với các tham số nhận được sẽ thể hiện như thế nào đối với các dữ liệu chưa biết? (hay nói một cách khác là sai số dự báo của mô hình sẽ như thế nào ?).

Một vấn đề khó khăn khi sử dụng ANN là tìm cấu hình tối ưu của ANN để thể hiện tổng quan của ANN trên dữ liệu chưa biết là tốt nhất. Hay nói một cách khác là *với một số các đầu vào và đầu ra cố định cho trước, ta phải tìm số các nút ẩn tối ưu của ANN sao cho sự thể hiện của ANN trên dữ liệu chưa biết là tốt hơn cả.*

Tác giả đã dùng SRM để thiết lập một số mô hình thống kê (hồi quy phi tuyến B-spline) và đã trình bày các kết quả đạt được khi áp dụng chúng vào nghiên cứu thuỷ văn [2]. Trên cơ sở đó tác giả áp dụng SRM để tìm cấu hình tối ưu của ANN dựa trên lý thuyết toán học "Sự chính xác gần đúng theo xác suất" (Probably Approximately Correct PAC).

Hiện nay có một số cách thường dùng để đánh giá sự thể hiện của mô hình đối với các tập dữ liệu chưa biết, chẳng hạn phương pháp kiểm tra chéo (cross-validation) và các phương pháp dựa trên cơ sở của một số lý thuyết.

a. Phương pháp kiểm tra chéo

Nếu chuỗi số liệu quan trắc đủ dài, khi đó tập số liệu có được sẽ được chia làm hai phần. Phần thứ nhất dùng để dạy ANN và phần còn lại dùng để kiểm tra độc lập chất lượng của ANN. Sự phụ thuộc của cách tiếp cận này vào các mẫu kiểm tra là sự không mong đợi bởi ba nguyên nhân sau:

1. Sự phân chia số liệu thành hai phần như trên sẽ giảm số các ví dụ để dạy ANN, do đó độ chính xác của ANN sẽ giảm. Điều đó càng không mong đợi khi chuỗi số liệu có được ngắn và chi phí cho mỗi dữ liệu quan trắc lớn. Trong thực tế nghiên cứu thuỷ văn chuỗi số liệu quan trắc được thường ngắn.

2. Kiểm tra chéo không phải là một phương pháp có sự bảo đảm và có nguyên lý. Cấu hình của ANN phụ thuộc vào các mẫu kiểm tra được lấy ra. Mức độ biến đổi giữa các mẫu có thể là rất đáng kể. Các mẫu có thể có độ lệch lớn (highly skewed) đến nỗi mô hình nhận được bị sai lệch.
3. Do số liệu quan trắc được chỉ là một tập số liệu có được trong tập số liệu chung của một quá trình cho nên các sai số bình phương trung bình MSE cho tập số liệu để dạy và kiểm tra ANN mang tính ngẫu nhiên. Việc lấy các giá trị ngẫu nhiên MSE làm chỉ tiêu đánh giá chất lượng của ANN là một điều khó khăn và rất khó có thể dùng nó để tìm cấu hình tối ưu của ANN.
4. Sai số dự báo của một mô hình không những phụ thuộc vào sai số bình phương trung bình mà còn phụ thuộc vào dung lượng và độ phức tạp của mô hình (cấu trúc của mô hình), độ phức tạp của dữ liệu. Sai số kiểm tra của phương pháp kiểm tra chéo không tính đến các vấn đề này.

Một phương pháp tốt hơn là phương pháp kiểm tra chéo trượt (leave-one-out cross- validation). Một tập gồm L mẫu sẽ được chia ra làm hai tập. Một tập có (L-1) mẫu dùng để dạy và một tập kiểm tra gồm có 1 mẫu. Giá trị trung bình của sai số bình phương cho mẫu để ra ngoài (leave out) trên L khả năng phân chia như vậy sẽ được tính. ANN với giá trị trung bình sai số bình phương nhận được này nhỏ nhất sẽ là ANN tốt hơn cả.

b, Dự báo sự thể hiện của ANN trên dữ liệu chưa biết dựa trên một số lý thuyết

Hiện nay có 3 ứng dụng cơ bản để đánh giá sự thể hiện của ANN.

John Moody [7] đưa ra độ đo mà nó được gọi là *sai số dự báo tổng quan* (Generalized Prediction Error GPE). GPE dựa trên quan hệ giữa số các tham số hiệu quả của mô hình, sự biến đổi trong dữ liệu và độ dài của số liệu có được.

David Mackay [6] đề nghị sử dụng cách tiếp cận Bayesian (dựa vào xác suất mô hình là đúng khi cho trước số liệu quan trắc).

V. N. Vapnik [7, 8, 9, 10] đề nghị sử dụng phương pháp cực tiểu mạo hiểm theo cấu trúc (Structural Risk Minimisation SRM). ***Phương pháp này được dùng để dự báo sự thể hiện tổng quan của ANN trên dữ liệu chưa biết mà không cần dựa vào sự dự báo trên tập dữ liệu kiểm tra. Nó rất có ích bởi vì trong thế giới thực số liệu quan trắc được bị giới hạn trong đa số các bài toán. SRM cho phép phát triển một chuỗi các ANN với độ phức tạp và dung lượng tăng dần mà mỗi ANN được dạy hội tụ trên dữ liệu quan trắc được. Chất lượng của những ANN được đánh giá dựa trên sai số của ANN và khoảng tin cậy của các sai số đó. Khoảng tin cậy của sai số đồng biến với kích thước VC*** (Vapnik-Chervonenkis dimension) ***của ANN.***

Nói một cách dễ hiểu về lý thuyết SRM như sau:

Cấu trúc của mô hình là một cách sắp xếp mô hình theo một trật tự nào đó thành các lớp sao cho độ phức tạp và dung lượng của mô hình càng tăng, chẳng hạn mô hình được sắp xếp theo số các tham số mô hình tăng dần hay theo độ lớn của chuẩn vectơ các tham số hoặc số các biến đầu vào v.v.. Như vậy, chúng ta có thể đo dung lượng của mô hình bằng kích thước VC.

Sai số thực nghiệm của mô hình (sai số bình phương trung bình) mang tính ngẫu nhiên. Nó phụ thuộc vào tính ngẫu nhiên của tập số liệu được lấy ra, độ dài của số liệu quan trắc được và các tham số của mô hình.

Theo SRM thì *sai số dự báo* của mô hình được tính như là sai số bảo đảm của mô hình(bằng tổng của sai số thực nghiệm và khoảng tin cậy của sai số thực nghiệm). Nó phụ thuộc vào các yếu tố sau:

- Sai số thực nghiệm,
- Cấu trúc của mô hình (dung lượng của mô hình),

- Độ phức tạp của dữ liệu,
- Mức bảo đảm sai số của mô hình mà người làm mô hình cho trước.

II. Ứng dụng của SRM tìm số nút ẩn tối ưu của ANN

Một công việc chính nhưng khó khăn hiện nay khi áp dụng ANN vào giải các bài toán thực tế là tìm số nút ẩn tối ưu (optimal number of hidden nodes) của ANN. Đa số người sử dụng ANN tìm số nút ẩn tối ưu bằng thực nghiệm. Người ta thử ANN với số các nút ẩn khác nhau và dùng phương pháp kiểm tra chéo để kết thúc quá trình tìm tham số. Trong các ANN lớn với hàng chục , hàng trăm, hàng nghìn nút ẩn thì công việc đó không dễ dàng. Ngoài các mặt hạn chế của phương pháp kiểm tra chéo đã nêu ở trên, kết quả cuối cùng phụ thuộc nhiều vào kinh nghiệm của người sử dụng ANN và cách làm thủ công như vậy tốn nhiều thời gian và không thể tự động được.

Bài toán dạy ANN được đề ra như là bài toán gần đúng hàm. Quá trình học của ANN gồm 3 thành phần:

1. Vectơ X được lấy ra ngẫu nhiên từ một phân bố cố định nhưng chưa biết.
2. Các vectơ đầu ra Y tương ứng với các vectơ đầu vào X tuân theo hàm phân bố cố định nhưng chưa biết.
3. Khả năng của ANN có thể thực hiện như là một tập các hàm f(x, w) , $w \subset W$ (w là vectơ tham số trong không gian tham số W).

Chúng ta đưa ra một khái niệm về cấu trúc lồng nhau của các tập con các hàm gần đúng:

$$S_b = \{f(x, w), w \subset W_b\}$$

sao cho

$$S_1 \subset S_2 \subset .. .. \subset S_n$$

các kích thước VC của các tập con các hàm thoả mãn điều kiện:

$$h_1 < h_2 < .. .. < h_n$$

như vậy kích thước VC của ANN có thể điều khiển được.

Có một vài cách định nghĩa cấu trúc của ANN:

1. Cấu trúc của ANN được xác định như là cấu hình của ANN (Số các nút ẩn tăng dần).
2. Cấu trúc của ANN được xác định bởi quy trình học của ANN.
3. Cấu trúc của ANN được xác định bởi quá trình tiền xử lý dữ liệu.

Cấu trúc của ANN ở đây được xác định với sự tăng dần các nút ẩn của ANN. Ký hiệu hidsANN1, hidsANN2,.. là số các nút ẩn của các mạng thần kinh ANN1, ANN2,.. Khi đó cấu trúc của ANN có thể được xác định như sau:

- **if** hidsANN1< hidsANN2 **then** $ANN1 \subset ANN2$
- Các ANN sẽ được sắp xếp sao cho:

$$ANN1 \subset ANN2 \subset .. .. \subset ANN5..$$

Quá trình thực hiện SRM bao gồm 2 bước:

a, Cực tiểu hoá mạo hiểm thực nghiệm,

b, Cực tiểu hoá mạo hiểm theo cấu trúc.

Quá trình cực tiểu hoá mạo hiểm theo cấu trúc được tiến hành như sau:

Mạo hiểm thực nghiệm của ANN (sai số bình phương trung bình) được cực tiểu hoá đối với mỗi phần tử S_i của cấu trúc. Phần tử của cấu trúc S^* mà nó cực tiểu mạo hiểm bảo đảm (sai số bảo đảm) sẽ là phần tử tối ưu. Trong đó mạo hiểm bảo đảm được xác định như là tổng của mạo hiểm thực nghiệm và khoảng tin cậy.

Phương pháp SRM cho ANN có thể dùng để giải các bài toán phân lớp và các bài toán thiết lập mối quan hệ tương quan [1, 4, 5, 7, 8, 9, 10, 11].

2.1. ANN với SRM cho bài toán phân lớp

Khi dùng ANN để giải các bài toán phân lớp, bất đẳng thức [9] sau đây là đúng:

$$\Pr ob\left\{ Sup_{w \subset W}\left(\frac{p(w) - \nu(w)}{p(w)}\right) > \varepsilon \right\} < \left(\frac{2.L.e}{h}\right)^{h}.\exp\left(-\frac{\varepsilon^2.L}{4}\right) \qquad (2.1)$$

ở đây:
ε là một số dương nhỏ tuỳ ý,
h là VC-dimension,
p(w) là hàm mạo hiểm lý thuyết,
ν(w) là hàm mạo hiểm thực nghiệm,
L là độ dài của số liệu quan trắc được,
e là cơ số lôgarit tự nhiên,
Prob được ký hiệu là xác suất.

Từ (2.1) suy ra rằng cho trước một số dương η thì đối với mọi w ⊂ W bất đẳng thức sau đây là đúng với một xác suất bằng (1-η):

$$p(w) < \nu(w) + C(L/h, \nu(w), \eta) \qquad (2.2)$$

ở đây C(L/h, ν(w), η) là khoảng tin cậy và nó được xác định theo công thức sau:

$$C(L/h, \nu(w), \eta) = 2.\left(\frac{h(\ln(2L/h)+1) - \ln(\eta)}{L}\right)\left(1 + \sqrt{1 + \frac{\nu(w).L}{h.(\ln(2L/h)+1) - \ln(\eta)}}\right)$$

Như vậy số nút ẩn tối ưu của ANN là số nút ẩn sao cho sai số bảo đảm :
$Err_{guaranteed}$ = ν(w) +C(L/h, ν(w), η) đạt giá trị cực tiểu. Khi đó đối với sai số dự báo của mô hình Err, bất đẳng thức sau là đúng với xác suất (1-η):

$$Err < Err_{guaranteed}$$

2.2. ANN với SRM cho bài toán thiết lập mối quan hệ tương quan

Đối với bài toán thiết lập mối quan hệ tương quan sai số $Err_{guaranteed}$ với xác suất bảo đảm (1-η) được tính theo công thức:

$$Err_{guaranteed}\left(N_{hid}, w, L\right) = \left[\frac{Err_{empirical}\left(N_{hid}, w, L\right)}{1 - \sqrt{\frac{N_{hid}\left(\ln\left(\frac{L}{N_{hid}}\right)+1\right) - \ln(\eta)}{L}}}\right] \qquad (2.3)$$

ở đây:
N_{hid} là số các nút ẩn của ANN,
L là độ dài của chuỗi số liệu,
w là vectơ tham số,
$Err_{guaranteed}$ là sai số bảo đảm,

$Err_{empirical}$ là sai số thực nghiệm (sai số bình phương quân bình MSE).

Điều đó có nghĩa là đối với sai số dự báo của mô hình Err, bất đẳng thức (Err < $Err_{guaranteed}$) là đúng với xác suất (1-η) hay nói một cách khác:

$$Prob\left(Err < Err_{guaranteed}\right) = (1-\eta)$$

Số nút ẩn tối ưu của ANN là số nút ẩn sao cho sai số bảo đảm $Err_{guaranteed}$ đạt giá trị cực tiểu.

III. Trường hợp nghiên cứu tìm số nút ẩn tối ưu của ANN khi xác định quan hệ tương quan Q = f(H, dH/dt)

Khả năng tìm một ANN có cấu hình tối ưu bằng SRM được thể hiện trong bài toán thiết lập mối quan hệ tương quan Q = f(H, dH/dt). Có tất cả 93 mẫu số liệu thực đo: lưu lượng nước Q (m^3/s), mực nước H(cm) và cường suất mực nước dH/dt (cm/ h) trong năm 1971 tại trạm thuỷ văn Sơn Tây, sông Hồng.
Cấu hình của ANN sẽ được tìm bằng 2 phương pháp:

1) Phương pháp kiểm tra chéo,
2) Phương pháp cực tiểu hoá mạo hiểm theo cấu trúc SRM.

Tập số liệu quan trắc được chia ra thành 2 tập con một cách ngẫu nhiên. Tập thứ nhất gồm 72 mẫu (ví dụ) và tập thứ 2 gồm 21 mẫu. Một chương trình máy tính ANN&SRM được xây dựng. Các ANN với các nút ẩn biến đổi từ 3 đến 17 sẽ được đưa vào để học. Cho trước xác suất bảo đảm của sai số là (1-η) = 95% và số lần lặp lớn nhất để dạy ANN là 20000. Kết quả tính sai số và một số chỉ tiêu phụ khác được tính cho mỗi ANN. Sau đây là một số chỉ tiêu phụ sẽ được tính toán cho mỗi lần lặp:

- Sai số bình phương trung bình MSE và sai số tương đối sigma, khoảng tin cậy của sai số, sai số bảo đảm (2.3) cho tập số liệu để dạy ANN,
- Sai số bình phương trung bình MSE và sai số tương đối sigma cho tập số liệu kiểm tra ANN,
- Tổng sai số âm dương,
- Số các điểm lệch âm và dương,
- Chuẩn vectơ tham số,
- Giá trị tuyệt đối lớn nhất của tham số.

Các tiêu chuẩn tổng sai số âm dương, số các điểm lệch âm và dương, chuẩn vectơ tham số, giá trị tuyệt đối lớn nhất của tham số cho phép ta xét mặt hồi quy có đi qua trung tâm các điểm thực nghiệm hay không.
Mạng thần kinh nhân tạo với *n1* đầu vào, *n2* nút ẩn, *n3* đầu ra sẽ được ký hiệu là ANN(n1:n2:n3).
Sau một số bước lặp cho trước chương trình ANN&SRM sẽ in ra các kết quả tính dưới dạng sau:

----------Đối với Số liệu để dạy----------
Số các nút ẩn= 4
Số lần lặp m=20000
Sai số quân phương trung bình Sicma = 6.2843 %
Tổng độ lệch dư = 5.22485689
Số điểm lệch dương N+= 34
Số điểm lệch âm N-=38
Chuẩn của vectơ tham số = 20.612353
Tham số lớn nhất (về trị tuyệt đối)= 12.477661

Tổng số các tham số = 17
Số các tham số khác không = 17(100.000 %)
Biên bảo đảm C = 219728.117619
*Sai số bình phương trung bình (MSE)= 213080.03552318
Sai số quân phương trung bình (RMSE)= 461.60593099
*Sai số bảo đảm(GuaranteedError)= 432808.153142
---Đối với số liệu kiểm tra---
Số các ví dụ=21
MSE= 298978,13681221
RMSE= 546.78893260
Sigma= 9.2820 %

Do đó ta có thể so sánh các kết quả nhận được bằng phương pháp kiểm tra chéo và phương pháp SRM.

Kết quả tính các sai số cho ANN được ghi trong bảng 3.1.

Bảng 3.1.
(Số các bước lặp N =20000)

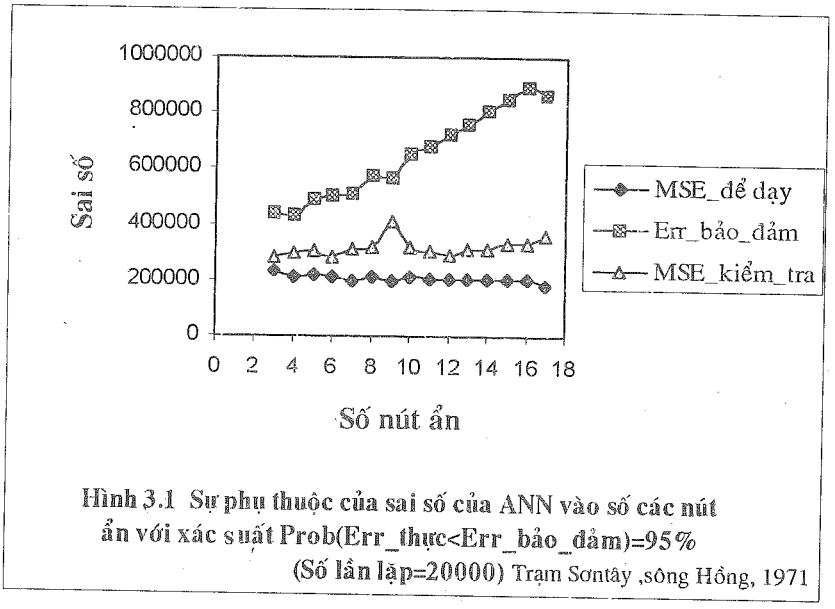
	Sai số bảo đảm của ANN với xác suất Prob(Err_thực<Err_bảo_đảm) = 95%					
	(72 ví dụ để dạy)				(21 ví dụ kiểm tra)	
Số nút ẩn	MSE	C1	Err_bảo đảm	Sigma %	MSE	Sigma %
3	235227	203980	439208	6,13	280811	9,06
4	213080	219728	432808	6,28	298978	9,28
5	222098	265294	487393	6,71	305427	9,2
6	212060	288328	500389	6,14	282208	9,36
7	201028	307272	508300	6,29	311805	9.6
8	212820	362314	575134	6,39	320579	9.37
9	196022	369004	565066	5,84	408959	9,39
10	212528	439991	652520	6,3	320341	9.52
11	208111	471645	679757	5,94	302427	8.8
12	208335	514974	723310	5,87	290403	9,03
13	205882	553411	759294	5,94	314362	8.91
14	205822	600159	805982	5,91	313306	8.96
15	205415	648458	853873	5,87	336813	9.34
16	203126	693083	896210	5,79	331409	9.38
17	184067	677950	862017	5,73	360946	9.36

Trong bảng 3.1 ngoài các sai số tính cho 2 tập số liệu dạy và kiểm tra, chương trình ANN&SRM còn tính khoảng tin cậy của sai số thực nghiệm C , sai số bảo đảm Err_guaranteed của ANN .

Một số tiêu chuẩn phụ được ghi trong bảng 3.2
Bảng 3.2.

Các tham số phụ cho ANN						
Số nút ẩn	Tổng âm dương	Số lệch dương	Số lệch âm	Number In Tolerance	Chuẩn vectơ tham số	Giá trị tuyệt đối tham số lớn nhất
12	-91	32	40	63%	28,32	11.72
4	5.22	33	39	58%	20.61	12.47

Sai số của ANN phụ thuộc vào số các nút ẩn thể hiện ở hình 3.1



Hình 3.1 Sự phụ thuộc của sai số của ANN vào số các nút ẩn với xác suất Prob(Err_thực<Err_bảo_đảm)=95%
(Số lần lặp=20000) Trạm Sơntây ,sông Hồng, 1971

Sau đây là ví dụ chọn ANN nào tốt hơn cả trong hai ANN : ANN(2:4:1) và ANN(2:12:1)

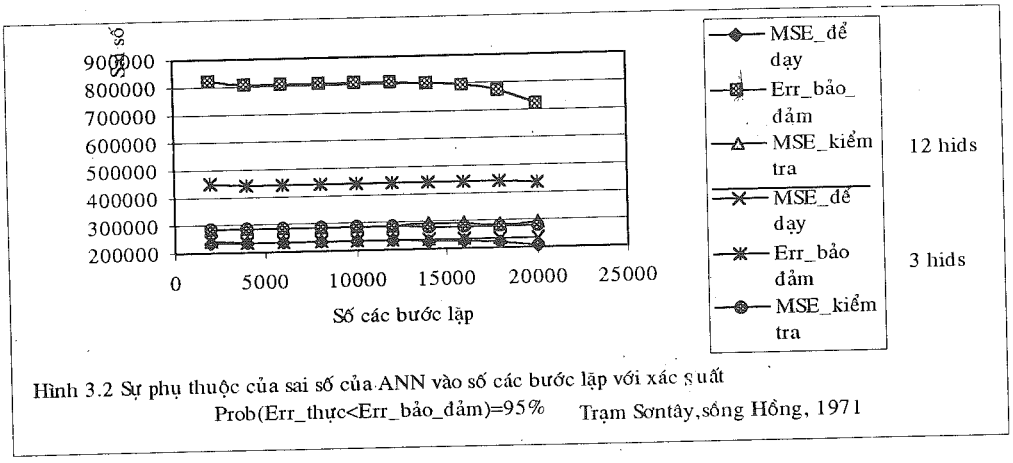
Kết quả trong bảng 3.1 và hình 3.1 chỉ ra rằng sự biến đổi của sai số kiểm tra theo số nút ẩn thể hiện tính ngẫu nhiên của chúng. Vì vậy, nếu chỉ dùng MSE của ANN cho các số liệu dạy và kiểm tra thì rất khó xác định số nút ẩn tối ưu. Rõ ràng, sai số quân phương trung bình tương đối RMSE (Sigma) của ANN(2:12:1) tính từ tập số liệu kiểm tra là 9,03 còn của ANN(2:4:1) là 9,28. Sự chênh lệch này không đáng kể về số trị. Nhưng nếu ta lấy các RMSE (hoặc MSE) làm tiêu chuẩn để chọn số nút ẩn tối ưu thì số nút tối ưu của ANN là 12.

Mặt khác, các tiêu chuẩn phụ trong bảng 3.2 chỉ ra rằng mặt hồi quy của ANN(2:4:2) đi qua trung tâm trường các điểm thực nghiệm và cong trơn hơn mặt cong ANN(2:12:2), bởi vì tổng âm dương nhỏ hơn, các điểm thực nghiệm so với mặt cong sẽ phân tán đều hơn, chuẩn vectơ tham số nhỏ hơn. Một điều được nhiều nhà nghiên cứu thừa nhận là ANN với cấu hình đơn giản (Mô hình nên có ít tham số) thường cho thể hiện tốt.

Sự biến đổi MSE của ANN cho tập số liệu kiểm tra theo số các nút ẩn không lớn (hình 3.1) nên trong thực hành nếu ta lấy ANN với MSE nhỏ để chọn số nút ẩn tối ưu là rất khó khăn và có thể chọn sai cấu hình của ANN (như ví dụ trên số nút ẩn tối ưu có thể bằng 12).

Bây giờ ta sử dụng SRM để tìm số nút tối ưu của ANN. *Theo SRM thì mạng thần kinh nhân tạo ANN với sai số bảo đảm Err_guaranteed nhỏ nhất sẽ là ANN với số nút tối ưu.* Kết quả bảng 3.1 chỉ ra rằng Err_guaranteed của ANN (2:4:1) có giá trị nhỏ nhất và nó bằng 432808 , cho nên số nút ẩn tối ưu sẽ là 4. Kết quả này phù hợp với các nhận xét rút ra từ các tiêu chuẩn phụ. Điều đó dễ dàng nhận thấy trên đồ thị hình 3.1 và kết quả trích lũ bằng ANN [3] chứng minh điều đó.

Sai số của ANN phụ thuộc vào số các bước lặp (thời gian dạy ANN) thể hiện ở hình 3.2



Hình 3.2 Sự phụ thuộc của sai số của ANN vào số các bước lặp với xác suất Prob(Err_thực<Err_bảo_đảm)=95% Trạm Sơntây,sông Hồng, 1971

Hình 3.2 chỉ ra sự sự phụ thuộc của việc tìm số nút tối ưu của ANN với các nút ẩn bằng 4 và 12 vào số các bước lặp (thời gian dạy ANN). Số nút ẩn tối ưu được tìm bằng cả hai phương pháp kiểm tra chéo và SRM. Khi dùng phương pháp kiểm tra chéo vì sai số của ANN(2:4:1) và ANN(2:12:1) cho các tập số liệu để dạy và kiểm tra khác nhau rất ít , trong khi đó nếu ta lấy sai số bảo đảm của ANN với SRM thì ANN với số nút ẩn bằng 4 tốt hơn ANN với số nút ẩn bằng 12.

Qua ví dụ tìm số nút ẩn tối ưu của ANN bằng cả 2 phương pháp như đã trình bày ở trên ta có một số nhận xét như sau:

1. Sử dụng phương pháp kiểm tra chéo (phương pháp thường hay dùng) để tìm số nút ẩn tối ưu phải dựa vào kinh nghiệm của người sử dụng. Việc đó rất khó khăn (đặc biệt trong các ANN lớn), tốn nhiều công sức và hay dẫn tới kết quả ANN với cấu hình được chọn không phải là tốt nhất. Điều đó càng không mong đợi khi chuỗi số liệu khí tượng thuỷ văn thường ngắn.
2. Khi sử dụng SRM ta có thể tìm số nút ẩn tối ưu của ANN và tính được sai số bảo đảm của ANN nhận được. Do đó, người áp dụng ANN có thể biết được sai số dự báo của ANN.

IV. Kết luận

Vấn đề tìm số nút ẩn tối ưu của ANN là bài toán rất quan trọng khi nghiên cứu áp dụng ANN để giải các bài toán kỹ thuật. Phương pháp kiểm tra chéo tìm số nút ẩn tối ưu của ANN đòi hỏi người sử dụng ANN phải có trình độ hiểu biết sâu về ANN và giảm đáng kể tiến độ ứng dụng rộng rãi ANN. Kết quả trình bày ở trên chỉ ra những tồn tại của phương pháp kiểm tra chéo trong việc tìm số nút ẩn tối ưu của ANN.

SRM có thể áp dụng để **tìm số nút ẩn tối ưu của ANN và tính sai số dự báo của ANN**. SRM là phương pháp khách quan dựa trên cơ sở của lý thuyết toán học. Nó giúp ta có được phương pháp tự động tìm số nút ẩn tối ưu của ANN, vì vậy sẽ rất thuận tiện cho người áp dụng ANN để giải các bài toán nghiên cứu khí tượng thuỷ văn. SRM tránh được các tồn tại của phương pháp kiểm tra chéo.

TÀI LIỆU THAM KHẢO

1. Barlett, Vapnik-Chervonenkis. Dimention bounds for two and three-layer networks. Neural Computation 5(3): 371-373, 1993.
2. L. X. Cau. Using residual method, Cubic Splines and Structural theory for constructing Non-linear multiple regression.- Journal of Hydrology and Meteorology, 4(424), 1996.

(xem tiếp trang 60)

III- TÌNH HÌNH HẢI VĂN

1. Gió và sóng

- Vùng biển phía bắc: Hướng gió chủ yếu là nam và đông nam.Ven bờ tốc độ gió trung bình 4-6m/s (cấp 3-cấp 4). Ngoài khơi gió mạnh nhất 19-21m/s (cấp 8-cấp 9). Hướng sóng chủ yếu là nam và đông nam. Ven bờ độ cao sóng trung bình 0,25 -0,50m (cấp II). Ngoài khơi sóng cao nhất 3,0-4,0m (cấp V - cấp VI).

-Vùng biển phía nam: Hướng gió chủ yếu là tây nam và nam. Ven bờ tốc độ gió trung bình 5-7m/s (cấp 3- cấp 4). Ngoài khơi Vũng Tàu, Côn Đảo, Trường Sa gió mạnh nhất 25 - 27 m/s (cấp 10). Hướng sóng chủ yếu là tây nam và nam. Ven bờ độ cao sóng trung bình 0,75 - 1,00m (cấp III). Ngoài khơi Vũng Tàu, Côn Đảo, Trường Sa sóng cao nhất 4,0 - 5,0m (cấp VI).

2. Nhiệt độ nước biển

- Vùng biển phía bắc: Nhiệt độ nước biển tầng mặt trung bình 31-32°C, cao nhất 33-34°C, thấp nhất 29-30°C.

- Vùng biển phía nam : Nhiệt độ nước biển tầng mặt trung bình 27 - 29°C, cao nhất 30- 32 °C, thấp nhất 24 - 26 °C.

3. Độ mặn nước biển

- Vùng biển phía bắc: Độ mặn nước biển tầng mặt trung bình 22-24‰, cao nhất 25-27‰, thấp nhất 18-20‰.

- Vùng biển phía nam: Độ mặn nước biển tầng mặt trung bình 31-32‰, cao nhất 33-34‰, thấp nhất 29-30‰.

Trung tâm quốc gia dự báo KTTV và Trung tâm KTTV biển biên soạn

---

(tiếp theo trang 40)

3. L.X. Cầu. ứng dụng của mạng thần kinh nhân tạo (ANN) xử lý dữ liệu khí tượng thuỷ văn.- Tạp chí Khí tượng Thuỷ văn, 4(460), 1999.
4. Guyon, V. Vapnik, B. Boser, L. Bottou, and S. A. Solla. Structural Risk Minimization for Character Recognition.- Advances in Neural Information Processing System 4, ed. Moody, 1992.
5. Haussler. Decision theoretic generalisations of the PAC model for neural net and another learning applications.- Inform. Comput., 100(1):78-150, Sept. 1992.
6. Mackey, J. C. David. A practical Bayesian Framwork for Back-propagation.- Neural computation 4(3): 448-472, 1992.
7. J. E. Moody. Note on Generalisation, Reglarization an Architecture selection in Non-linear learning system. In First IEEE-SP workshop on Neural Networks for Signal Processing. New York:IEEE Computer Society Press, 1991.
8. V.N. Vapnik. Estimation of dependences based on empirical data.- Springer-Verlag, New York, 1982.
9. V.N. Vapnik. Principle of Risk Minimisation for learning theory.- Advances in Neural Information Processing System 4, ed. Moody, 1992.
10. V. N. Vapnik. The Nature of Statistical Learning Theory.- Springer, New York, 1995.
11. Vapnik and A. Ja. Chervonenkis. 'Necessary and sufficient conditions for consistency of the method of empirical risk minimization'[in Russian].- Yearbook of the Academy of Sciences of the USSR on Recognition, Classification, and forecasting, 2, 217-249, Nauka (Moscow), 1989.