Tạp chí Khoa học và Công nghệ biển T10 (2010). Số 1. Tr 17 - 25

PHỤC HỒI DỮ LIỆU SÓNG BIỂN BẰNG MẠNG NEURON NHÂN TẠO

ĐẶNG VĂN TỎ

Trường Đại học Khoa học Tự nhiên, Thành phố Hồ Chí Minh

Tóm tắt: *Nghiên cứu này xây dựng mạng neuron nhân tạo OceanANN trên nền MATLAB để phục hồi dữ liệu sóng biển. Chương trình OceanANN được thiết kế thân thiện với người sử dụng nhờ các giao diện tiện ích. Dựa trên thuật toán Levenberg-Marquardt, OceanANN được thiết kế với 1 lớp nhập, 2 lớp ẩn và 1 lớp xuất. Tổng cộng 30 neuron cho từng lớp ẩn được sử dụng để học bản chất của chuỗi dữ liệu. Để áp dụng OceanANN, tập số liệu sóng biển thực đo ngoài khơi ở Tweed Heads (Australia) vào tháng 2 năm 1996 đã được sử dụng. Tập số liệu này được chia thành 3 phần: 70% số liệu dùng để học, 15% số liệu dùng để kiểm định và 15% số liệu còn lại được cố ý làm thất thoát để phục hồi. Mạng neuron nhân tạo rất thích hợp để sử dụng cho các số liệu sóng biển với độ phi tuyến cao và nhiễu động lớn. Các hệ số tương quan giữa số liệu tính toán và số liệu thực đo có thời khoảng quan trắc 1 giờ cho các trường hợp huấn luyện mạng, kiểm định mạng và mô phỏng mạng đều có kết quả trên 98%.*

I. GIỚI THIỆU

Kết quả tính toán của nhiều mô hình số không phải lúc nào cũng đáng tin cậy vì các số liệu thực đo không đầy đủ và có độ bất định. Vì thế, các ứng dụng thực tiễn của các mô hình số gặp nhiều hạn chế. Tuy nhiên, việc có được các số liệu thực đo, đầy đủ và đáng tin cậy không phải lúc nào cũng dễ dàng thực hiện được. Các số liệu sóng hay dòng chảy đo đạc ngoài khơi thường hay bị thất thoát, thiếu hụt vì nhiều lý do, trong đó điều kiện tự nhiên thay đổi đột ngột (như sóng to, gió lớn…) hoặc các nguyên nhân từ con người (như cắt phao, lấy cắp thiết bị…) thường hay xảy ra. Tất cả các bất định trên thường nằm ngoài dự kiến của người thực hiện nghiên cứu. Vì vậy, một chuỗi số liệu đo đạc không phải không có những chỗ gián đoạn thay vì liên tục như mong đợi. Trong khi đó, các mô hình số thường yêu cầu các số liệu đầu vào liên tục nhằm thoả mãn các điều kiện biên và điều kiện ban đầu của mô hình. Điều này cũng quan trọng trong việc sử dụng số liệu liên tục thực đo để cân chỉnh và kiểm định kết quả mô hình tính toán. Vì thế, việc phục hồi dữ liệu thất thoát trong chuỗi số liệu đo đạc gián đoạn có ý nghĩa thực tiễn quan trọng.

Bài báo này có mục đích xây dựng mạng neuron nhân tạo OceanANN trên nền MATLAB để phục hồi dữ liệu sóng biển. Cơ sở lý thuyết của mạng neuron nhân tạo, khu

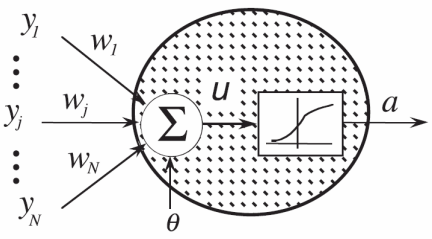
vực nghiên cứu và số liệu chuẩn bị nhằm áp dụng cụ thể thuật toán của bài toán lần lượt được trình bày kế tiếp theo sau.

II. PHƯƠNG PHÁP VÀ TÀI LIỆU

1. Cơ sở lý thuyết mạng neuron nhân tạo (ANN)

Mạng neuron nhân tạo ANN là một mô hình thống kê toán học dựa trên sự mô phỏng các hoạt động của hệ thần kinh sinh học. ANN không cố gắng để mô phỏng các hoạt động tinh tế của bộ não, nhưng chúng cố gắng tái tạo các hoạt động logic của bộ não bằng cách tập hợp nhiều dữ liệu đầu vào có dạng neuron thần kinh để thực hiện các quá trình tính toán hay nhận thức.

Hầu hết các mạng neuron đều có sơ đồ chung là neuron và cấu trúc liên kết mạng (hình 1). Mỗi một neuron bao gồm hai phần: hàm số mạng (net function) và hàm số kích hoạt (activation function). Hàm số mạng xác định phương thức liên kết của dữ liệu nhập $\{y_j ; 1 \le j \le N\}$ với nhau trong neuron. Trong mô hình neuron này, mối liên kết tuyến tính có trọng số $u = \sum_{j}^{N} w_j y_j + \theta$ được áp dụng, với w_j là trong số $\{w_j ; 1 \le j \le N\}$ và θ là độ lệch dùng để mô phỏng ngưỡng của neuron. Dữ liệu xuất của neuron được ký hiệu là a_i, nó liên kết với dữ liệu nhập u_i của mạng neuron nhờ hàm kích hoạt (hay phép biến đổi tuyến tính hoặc phi tuyến f): $a = f(u)$. Nhiều hàm mạng và hàm kích hoạt khác nhau được sử dụng để thiết lập cấu trúc mạng neuron khác nhau. Chi tiết có thể tham khảo trong các tài liệu mạng neuron [1].

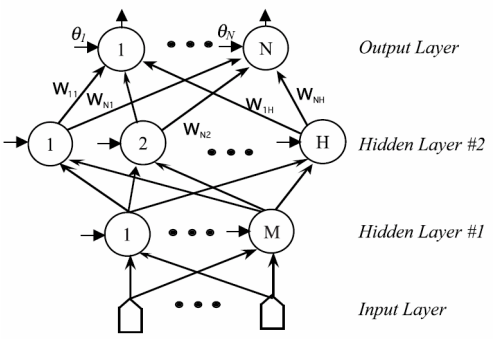


Hình 1: Mô hình neuron

1.1. Mạng nhận thức nhiều lớp

Mạng nhận thức nhiều lớp (MLP) là một mạng lan truyền tiến (feed-forward) được xếp thành nhiều lớp. Mỗi một neuron trong mạng MLP có một hàm kích hoạt phi tuyến, thường là hàm sigmoid hay hàm tanhyerbolic. Cấu hình tiêu biểu của mạng MLP được thể hiện trong hình 2. Trong hình này, các hình tròn mô tả các neuron và chúng được sắp xếp

theo từng lớp. Mạng neuron này có 3 lớp: lớp nhập (hình ngũ giác), lớp ẩn (hình tròn M và H) và lớp xuất (hình tròn N). Kết quả của các lớp ẩn thường không nhìn thấy được.



Hình 2: Cấu hình mạng neuron có 3 lớp

Một trong những điều kiện áp dụng thành công mô hình MLP là chọn đúng các ma trận trọng số. Phương pháp phổ biến để chọn đúng các ma trận trọng số là sử dụng phương pháp tối ưu hoá hay phương pháp huấn luyện lan truyền ngược sai số (error back-propagation training method). Phương pháp này được biết như sau. Nếu gọi bình phương sai số của mạng:

$$E = \sum_{k=1}^{K} \left[e(k)\right]^2 = \sum_{k=1}^{K} \left[d(k) - z(k)\right]^2 = \sum_{k=1}^{K} \left[d(k) - f(\mathbf{W} \cdot \mathbf{x}(k))\right]^2 \quad (1)$$

Trong đó W là một ma trận trọng số, x là vector dữ liệu nhập, $d(k)$ là dữ liệu huấn luyện $\{d(k);\ 1 \le k \le K\}$, $z(k)$ là dữ liệu xuất $\{z(k);\ 1 \le k \le K\}$ và $e(k) = d(k) - z(k)$.

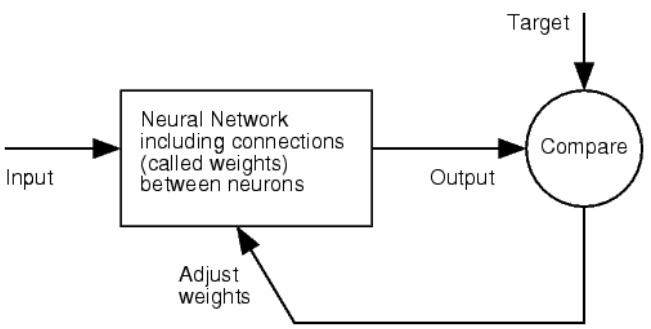
Như vậy, mục tiêu tìm ma trận trọng số tối ưu W sẽ tương ứng với việc cực tiểu hoá bình phương sai số E. Điều này dẫn đến bài toán tối ưu hóa bình phương tối thiểu phi tuyến. Hiện có khá nhiều phương pháp để giải bài toán tối ưu loại này, về cơ bản chúng có thể được thực hiện bởi công thức lặp nhờ các thuật toán tối ưu sau: $\mathbf{W}(t+1) = \mathbf{W}(t) + \Delta\mathbf{W}(t)$. Ở đây $\Delta W(t)$ là độ hiệu chỉnh của các trọng số hiện thời $W(t)$. Các thuật toán khác nhau chủ yếu sẽ khác nhau về dạng của $\Delta W(t)$. Nhiều thuật toán như phương pháp gradient liên hiệp, phương pháp Newton… thường hay được sử dụng [1].

1.2. Mạng lan truyền ngược (Back Propagation - BP)

Mạng lan truyền ngược tiêu biểu thường sử dụng thuật toán gradient hướng xuống (gradient descent) giống như phép học Widrow-Hoff. Trong mạng này, các trọng số được thay đổi hay di chuyển dọc theo giá trị âm của gradient của hàm thực hiện. Thuật ngữ lan truyền ngược được sử dụng vì nó liên quan đến phương cách tính toán gradient của các

mạng neuron nhiều lớp phi tuyến. Hiện nay có khá nhiều biến thể của thuật toán cơ bản lan truyền ngược được xây dựng, chúng hầu hết dựa trên các kỹ thuật cơ bản của tối ưu hóa như phương pháp gradient liên hợp hoặc phương pháp Newton.

Trong thực tế để tiến hành thiết kế hoặc sử dụng các mạng neuron lan truyền ngược để học hay huấn luyện các mạng truyền thẳng nhằm giải một bài toán cụ thể nào đó, các bước cơ bản sau đây thường được tiến hành: a) Tập hợp các dữ liệu được học hoặc huấn luyện; b) Xây dựng mạng neuron; c) Huấn luyện mạng; d) Ứng dụng mạng neuron để mô phỏng các dữ liệu mới. Sơ đồ khối của mạng lan truyền ngược được biết như sau (hình 3).



Hình 3: Sơ đồ khối mạng lan truyền ngược

1.3. Thuật toán lan truyền ngược Levenberg-Marquardt

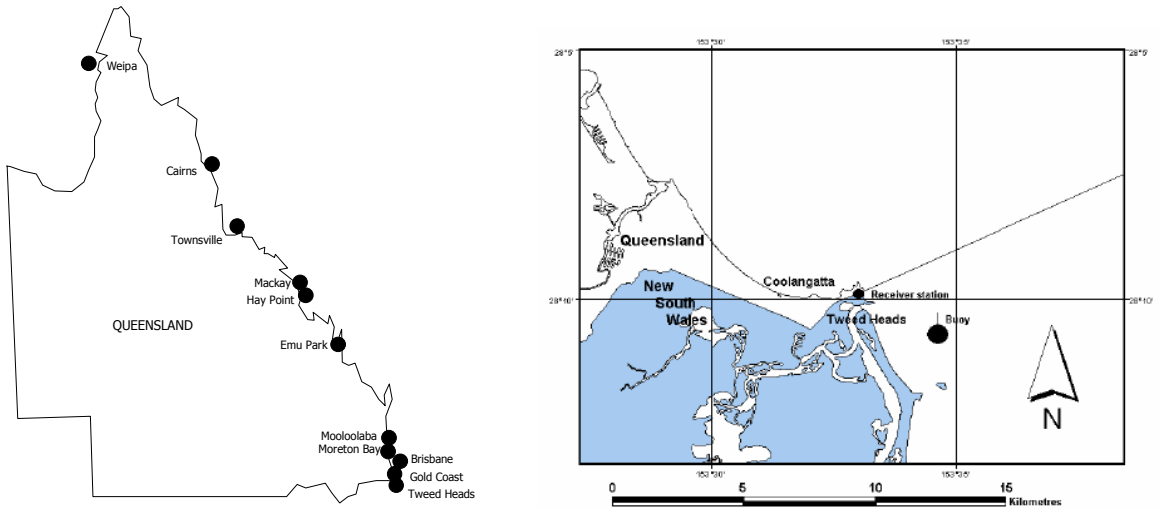
Một số thuật toán huấn luyện lan truyền ngược như gradient hướng dốc xuống có tốc độ hội tụ chậm. Vì vậy, một trong những thuật toán cải thiện tốc độ hội tụ hay tốc độ học của mạng neuron từ 10 cho đến 100 lần là mạng huấn luyện lan truyền ngược theo thuật toán Levenberg-Marquardt [2, 3]. Thuật toán Levenberg-Marquardt được xây dựng có tốc độ huấn luyện nhanh cấp 2 mà không cần tính đến ma trận Hessian giống như phương pháp Newton. Nếu hàm thực thi có dạng tổng các bình phương, lúc đó ma trận Hessian có thể được xấp xỉ như sau: $\mathbf{H} = \mathbf{J}^T \cdot \mathbf{J}$ và $\mathbf{G} = \mathbf{J}^T \cdot \mathbf{e}$, trong đó J là ma trận Jacobian chứa các đạo hàm bậc nhất của các sai số mạng đối với trọng số W và độ lệch b và e vector sai số của mạng. Ma trận Jacobian có thể được tính thông qua kỹ thuật lan truyền ngược chuẩn khi ấy việc tính toán đơn giản hơn việc tính toán ma trận Hessian [2].

2. Nguồn số liệu

2.1. Khu vực nghiên cứu

Australia có một mạng lưới phao đo sóng hiện đại được thiết lập dọc theo bờ biển Queensland. Trong đó, phao đo sóng ở Tweed Heads thuộc dự án TRESBP do hai bang New South Wales (NSW) và Queensland (QLD) cùng lắp đặt để sử dụng. Ở Tweed Heads

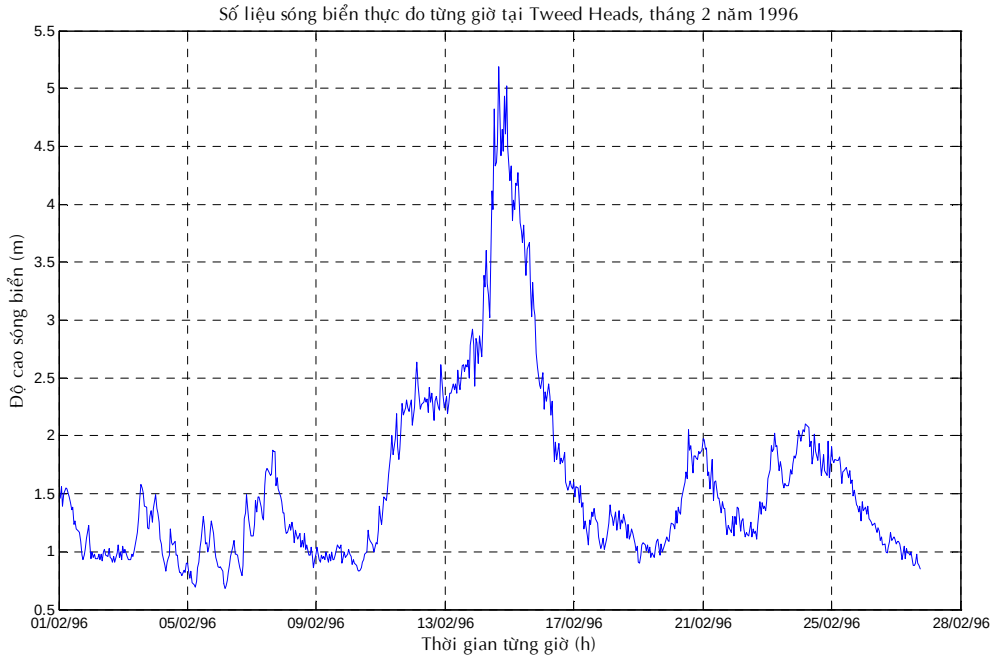
phao đo sóng có tên Waverider được đặt ở độ sâu 25 m, có kinh độ 28^{o}10.745’ và vĩ độ 153^{o}34.597’, cách bờ khoảng 2100 m (hình 4). Số liệu sóng tại Tweed Heads được lấy mẫu với thời khoảng 0.78 sec (tần số 1.28 Hz) và được ghi thành từng nhóm (burst) 2048 điểm (hay khoảng 26 phút) một lần ghi liên tục. Khoảng thời gian đo giữa hai lần đo là 1 giờ (hay obs quan trắc = 1 giờ). Vì thế số liệu sóng thu thập từ phao đo sóng Waverider khá tốt. Khi có bão, số liệu sóng đo đạc sẽ dầy đặt hơn. Mặc dù vậy, sự liên tục về số liệu theo yêu cầu đầu vào của mô hình toán, ví dụ 10 phút có một số liệu sóng chẳng hạn, là không thể có.



Hình 4: Mạng đo đạc và phao đo sóng tại Tweed Heads (NSW), Australia

2.2. Chuẩn bị số liệu

Trong bài báo này, số liệu sóng thực đo từng giờ tại Tweed Heads vào tháng 02 năm 1996 (hình 5) sẽ được sử dụng cho mô hình OceanANN. Để có thể áp dụng OceanANN phục hồi dữ liệu thất thoát và kiểm tra tính khả thi của mô hình, tổng số 670 số liệu của tháng 2 năm 1996 sẽ được chia ra thành 470 số liệu (≈ 70% tổng số số liệu) được mô hình OceanANN dùng để huấn luyện (train) hay học theo thuật toán có giám sát (supervise), 100 số liệu (≈15% tổng số liệu) sẽ được dùng để kiểm định mô hình (verification), 100 số liệu (≈15% số liệu) sẽ được chọn ngẫu nhiên và cố ý làm thất thoát trong chuỗi số liệu tổng cộng. Sau đó mô hình OceanANN sẽ được sử dụng để mô phỏng (test/model) và phục hồi lại số liệu thất thoát vừa giả định mất đi. Số liệu thất thoát được mô phỏng bằng mô hình OceanANN sẽ được so sánh với 15% số liệu thực đo để đánh giá khả năng ứng dụng cũng như độ tin cây của mô hình. Tất cả các đánh giá sẽ dựa trên hệ số tương quan.

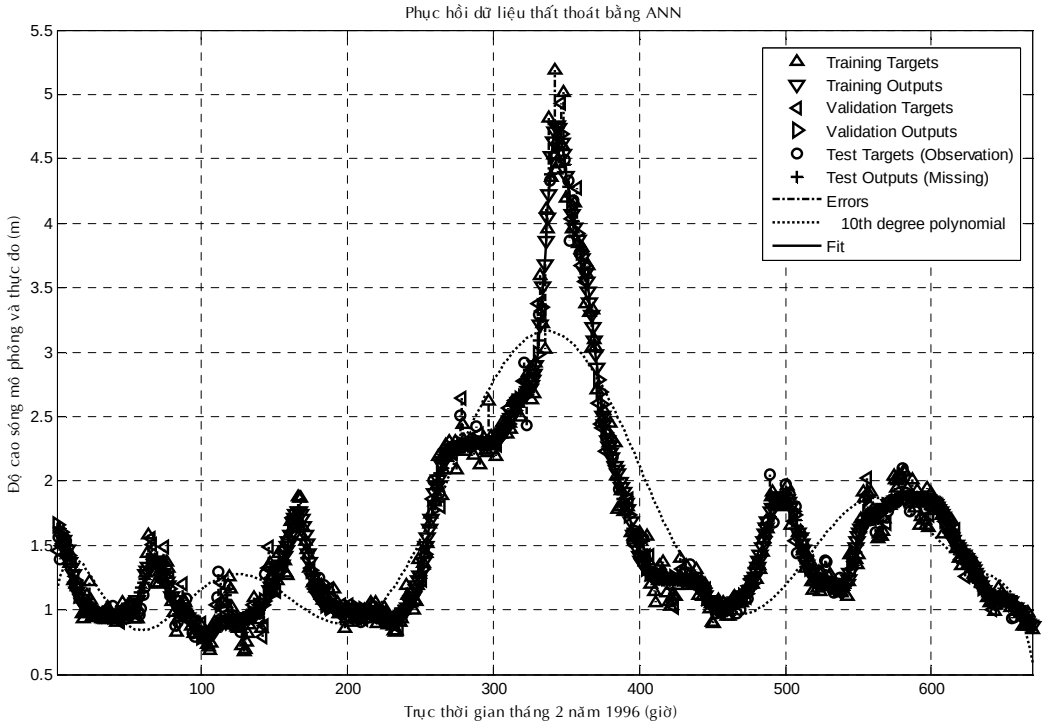


Hình 5: Số liệu sóng thực đo vào tháng 2 năm 1996 tại Tweed Heads (NSW)

III. KẾT QUẢ TÍNH TOÁN

Với các số liệu đã chuẩn bị, OceanANN sẵn sàng được sử dụng và kết quả tính toán được trình bày trong hình 6, hình 7.

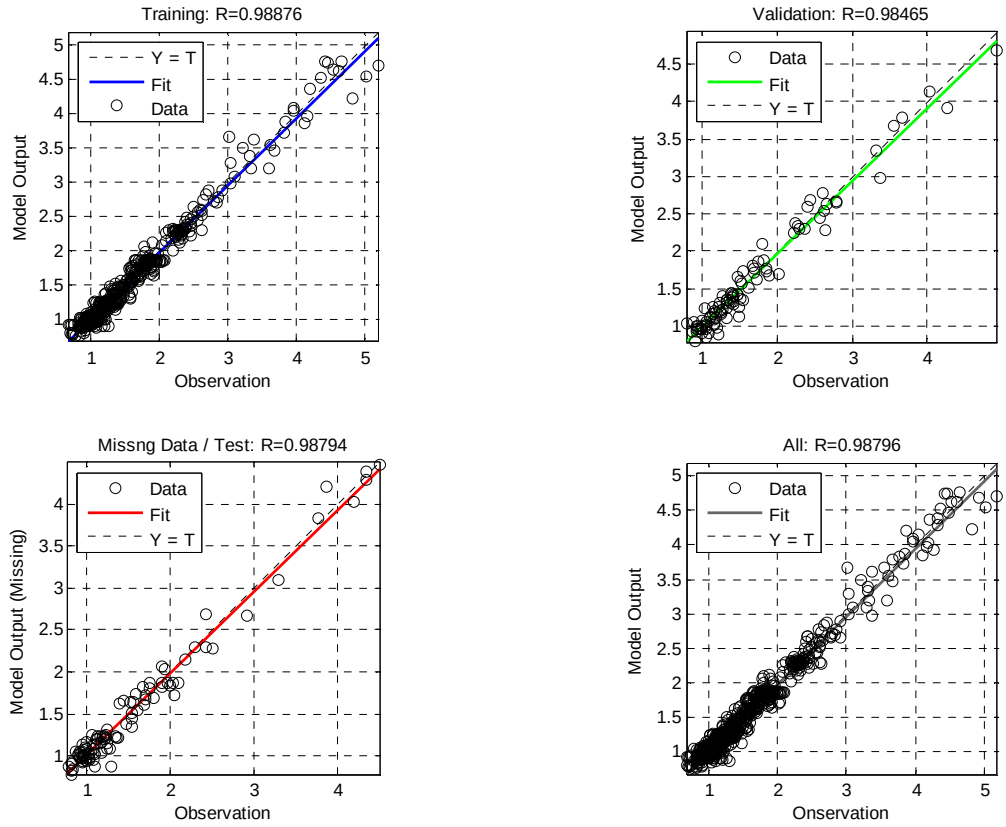
Hình 6 trình bày kết quả tính toán tổng hợp của mô hình OceanANN, trong đó các ký hiệu hình tam giác có đỉnh hướng lên trên và hình tam giác có đỉnh hướng xuống dưới chỉ số liệu nhập và xuất trong lúc học có giám sát khi mô hình đang được xây dựng. Ký hiệu hình tam giác có đỉnh hướng về phía trái và hình tam giác có đỉnh hướng về phía phải chỉ số liệu nhập và xuất trong quá trình kiểm định mô hình. Ký hiệu hình tròn và chữ thập chỉ số liệu được mô phỏng hay số liệu thất thoát được phục hồi. Chữ thập ký hiệu số liệu thất thoát được phục hồi và vòng tròn là số liệu đo đạc tương ứng với số liệu giả thuyết bị thất thoát để so sánh. Đường nét đứt gãy chỉ khoảng sai số trong các quá trình tính toán. Đường chấm chấm chỉ đường cong khớp các số liệu thực đo bằng đa thức bậc 10. Đường cong liền nét là đường cong mô phỏng từ OceanANN với 30 nút ẩn nhằm phục hồi các số liệu thất thoát. Kết quả tính toán cho thấy có sự phù hợp rất tốt của số liệu được phục hồi và số liệu thực đo. Chi tiết mô tả về trạng thái học có giám sát của mô hình OceanANN và các đánh giá sai số MSE được trình bày trong tài liệu tham khảo [4].



Hình 6: Kết quả tính tổng hợp trong việc phục hồi dữ liệu sóng biển

Hình 7 mô tả hệ số tương quan của các phép tính trong mô hình. Hình nhỏ góc trái trên cùng mô tả giá trị thực đo và giá trị tính toán của mạng OceanANN. Hình vẽ này cho thấy giá trị thực đo và tính toán từ phép học có sự phù hợp khá cao với hệ số tương quan R = 0.999. Đối với hình nhỏ bên trên góc phải, mạng oceanANN cũng cho kết quả khá tốt. Ở đây số liệu thực đo và số liệu tính toán từ mô hình để kiểm định có hệ số tương quan R = 0.985. Hình nhỏ bên dưới góc phải mô tả số liệu được mô phỏng và số liệu thực đo. Ở đây số liệu mô phỏng được giả sử như là số liệu thất thoát trong chuỗi số liệu cần được phục hồi. Kết quả cho thấy số liệu cần được phục hồi gần giống với số liệu thực đo với hệ số tương quan R = 0.998. Như vậy, mạng OceanANN có thể được sử dụng để phục hồi số liệu bị thất thoát với độ tin cậy cao. Sau cùng là hình nhỏ ở dưới góc phải chỉ số liệu tính toán và số liệu thực đo bằng mô hình OceanANN cho tất cả ba trường hợp riêng lẻ vừa diễn giải trên.

Tóm lại, mô hình vừa được xây dựng hoàn toàn học được bản chất của số liệu với các hệ số tương quan rất cao (R ~ 0.99). Vì thế nó có thể được dùng để phục hồi số liệu thất thoát hoặc có thể dùng để dự báo số liệu tương lai khi chưa có số liệu đo đạc dùng cho mô hình số.



Hình 7: Kết quả hệ số tương quan của huấn luyện, kiểm định, mô phỏng và tổng hợp của mô hình OceanANN

IV. KẾT LUẬN

Nghiên cứu này đã xây dựng được một mạng trí tuệ nhân tạo có tên OceanANN trên nền MATLAB dùng cho việc xử lý số liệu sóng biển ngoài khơi. Dựa trên mạng lan truyền ngược với thuật toán Levenberg-Marquardt, OceanANN được thiết kế với 1 lớp nhập, 2 lớp ẩn và 1 lớp xuất. Tổng cộng 30 neuron cho từng lớp ẩn được sử dụng để học bản chất của chuỗi dữ liệu.

Để ứng dụng OceanANN, tập số liệu sóng biển ngoài khơi ở Tweed Heads, Australia vào tháng 2 năm 1996 đã được thu thập. Tập số liệu này được chia thành 3 phần: 70% số liệu dùng để học, 15% số liệu dùng để kiểm định và 15% số liệu còn lại được cố ý làm thất thoát để phục hồi từ OceanANN.

OceanANN khớp dữ liệu sóng tốt hơn nhiều so với việc khớp dữ liệu bằng đa thức bậc 10. Các số liệu sóng có độ phi tuyến cao và nhiễu động lớn chỉ thích hợp với việc sử

dụng các mạng neuron để tính toán.

OceanANN đã được sử dụng để phục hồi dữ liệu sóng thất thoát tại Australia với kết quả rất tốt. Với thời khoảng ghi số liệu sóng liên tiếp cách nhau 1 giờ, các hệ số tương quan giữa số liệu tính toán và số liệu thực đo cho các trường hợp huấn luyện mạng, kiểm định mạng và mô phỏng mạng đều có kết quả trên 98%. Vì thế, OceanANN có thể dùng để phục hồi dữ liệu thất thoát hoặc mô phỏng số liệu trong tương lai với độ tin cậy cao.

TÀI LIỆU THAM KHẢO

1. **Hu, Y. H. and Hwang, J. N., 2002.** Handbook of neural network signal processing: CRC Press.
2. **Hagan, M. T. and Menhaj, M., 1999.** Training feed-forward networks with the Marquardt algorithm, IEEE Transactions on Neural Networks, Vol. 5, pp. 989-993.
3. **MathWorks, 2008.** Neural Network Toolbox for Matlab, *www.mathworks.com.*
4. **Đặng Văn Tỏ, 2009.** Phục hồi dữ liệu thất thoát bằng phương pháp mạng neuron, Đề tài nghiên cứu cấp trường, Đại học Khoa học Tự nhiên Tp. HCM.

RECOVERY OF WAVE DATA USING AN ARTIFICIAL NEURAL NETWORK

DANG VAN TO

Summary: *This study aims at developing an artificial neural network OceanANN in MATLAB to recover missing wave data. The program OceanANN was designed friendly to the user for their convenient and useful interfaces. Based on the Levenberg-Marquardt algorithm, OceanANN was established with 1 input - layer, 2 hidden - layers and 1 output - layer. Each hidden layer consisting of 30 neurons was used to learn the nature of data series. For the application of OceanANN, the wave data recorded offshore of the Tweed Heads (Australia) in February 1996 was employed. The data set was divided into 3 segments: 70% of the data for learning, 15% for verification and the rest 15% of the data deliberately removed for OceanANN to recover. Artificial neural networks are suitable tools for the application on wave data of high non-linearity and fluctuations. The regression coefficients showing the relationship between the computed data and recorded data with hourly observations for the learning process, verification and testing process are always higher than 98%.*

Ngày nhận bài: 21 - 10 - 2009

Người nhận xét: TS. Lê Đình Mầu