Một số ví dụ phân loại dùng SOM và MLP Neural Network

Cao Thăng, 2011

Tài liệu này hướng dẫn các bạn sử dụng mạng nơ ron trong một số ứng dụng thực tế. Tài liệu này cũng đã được dùng để trình bày với một số bạn sinh viên Nhật bản. Tác giả thấy nó có thể có ích cho các bạn sinh viên khác nên đã soạn lại để các bạn tham khảo. Hy vọng giúp ích gì đó cho các bạn. Do không có nhiều thời gian biên soạn nên có thể có lỗi, mong bạn đọc thông cảm và đóng góp ý kiến.

Bạn đọc có thể liên hệ với tác giả tại spiceneuro at gmail dot com hoặc http://spiceneuro.wordpress.com

Nếu quan tâm tới mạng nơ ron, các bạn có thể dùng phần mềm SpiceSOM và SpiceMLP, download tại đây

Một số dữ liệu trình bày ở đây có sẵn trong thư mục Data khi cài đặt phần mềm SpiceSOM và SpiceMLP. Tất cả các kết quả trình bày ở đây đều được sử dụng bằng SpiceSOM và SpiceMLP.

Cảm ơn các bạn.

MỤC LỤC

1. Iris Flower Data Set

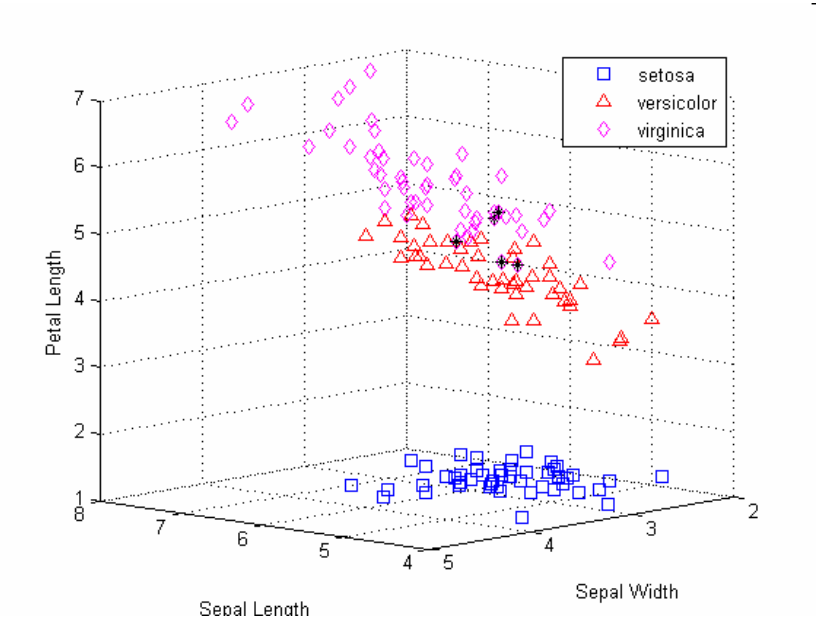
Iris dataset bao gồm dữ liệu của ba loại hoa (Iris setosa, Iris virginica và Iris versicolor), mỗi loại 50 mẫu. Các thuộc tính là độ dài và rộng của đài hoa (sepal) và cánh hoa (petal) tính theo centimeters.

Chi tiết tại http://archive.ics.uci.edu/ml/datasets/Iris

Iris setosa

Iris versicolor

Iris virginica



Hình 1. Hình minh họa hoa Iris (wikipedia)

1.1. Chuẩn bị dữ liệu

Phân loại Iris dataset với mạng nơ ron đa lớp 4 đầu vào, 1 đầu ra. Chúng ta mã hóa đầu ra của mạng như bảng 1 sau:

Table 1. Input and Output of a MLP NN

Dữ liệu	Output
Iris setosa	0.0
Iris versicolor	0.5
Iris virginica	1.0

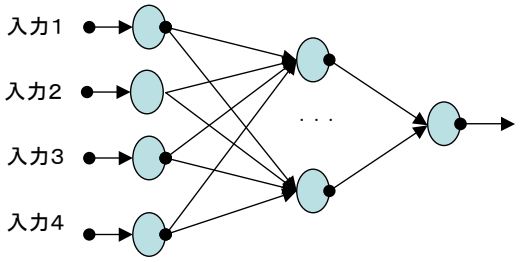
ID	Sepal Length	Sepal Width	Petal Length	Petal Width	Output	Species
1	5.1	3.5	1.4	0.2	0.0	setosa
...	...	...	...	...	...	...
51	7	3.2	4.7	1.4	0.5	versicolor
...	...	...	...	...	...	...
150	5.9	3	5.1	1.8	1.0	virginica

Phân loại Iris dataset với mạng nơ ron đa lớp 4 đầu vào, 3 đầu ra. Chúng ta mã hóa đầu ra của mạng như bảng 2 sau:

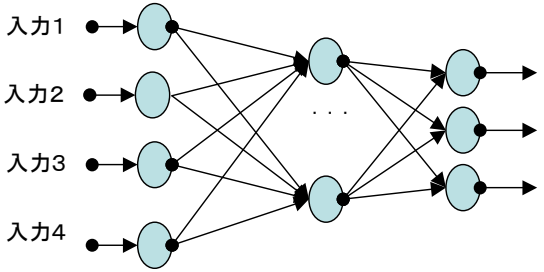
Table 2. Input and Output of a MLP NN

Dữ liệu	Output 1	Output 2	Output 3
Iris setosa	1.0	0.0	0.0
Iris versicolor	0.0	1.0	0.0
Iris virginica	0.0	0.0	1.0

ID	Sepal Length	Sepal Width	Petal Length	Petal Width	Output 1	Output 2	Output 3	Species
1	5.1	3.5	1.4	0.2	1.0	0.0	0.0	setosa
…	…	…	…	…	…	…	…	…
51	7	3.2	4.7	1.4	0.0	1.0	0.0	versicolor
…	…	…	…	…	…	…	…	…
150	5.9	3	5.1	1.8	0.0	0.0	1.0	virginica

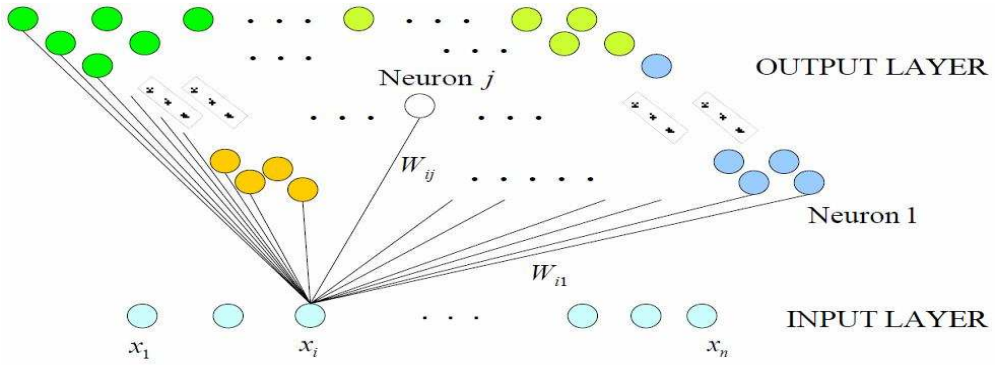


1つ出力
0.0 ＝ setosa
0.5 ＝ versicolor
1.0 ＝ virginica



3つ出力

Phân loại Iris dataset với mạng nơ ron (Self-Organizing Map). Dữ liệu cho SOM không cần đầu ra như dữ liệu cho MLP NN. Do vậy chúng ra chuẩn bị dữ liệu như hình 2 sau:



ID	Sepal Length	Sepal Width	Petal Length	Petal Width	Species
1	5.1	3.5	1.4	0.2	setosa
…	…	…	…	…	…
51	7	3.2	4.7	1.4	versicolor
…	…	…	…	…	…
150	5.9	3	5.1	1.8	virginica

Đầu vào của SOM

Fig.2. Data for SOM

1.2. Phân loại dữ liệu với SpiceSOM

Phân loại Iris flower với SOM size8x10 neurons, ta có output map và output table tương ứng như hình 3 sau. Ta thấy tại nơ ron (0, 1) và nơ ron (0, 3) đều ứng với nhãn của hai loại dữ liệu versicolor và virginica. Còn lại các nơ ron khác đều ứng với một nhãn. Như vậy mạng SOM phân loại nhầm dữ liệu versicolor và virginica tại nơ ron (0, 1) và nơ ron (0, 3) và phân loại đúng tại các nơ ron còn lại.

Trên bản đồ ta cũng thấy dữ liệu với nhãn setosa được phân biệt hẳn về một phía, còn dữ liệu với nhãn versicolor và virginica nằm gần nhau hơn. Một cách trực quan ta có thể thấy hoa loài setosa có kích thước và hình dạng khác hẳn hai loài versicolor và virginica. Và hai loài versicolor và virginica có kích thước gần như nhau và đôi khi ta không phân biệt được hai loài hoa này nếu chỉ dựa vào kích thước đài hoa và cánh hoa.

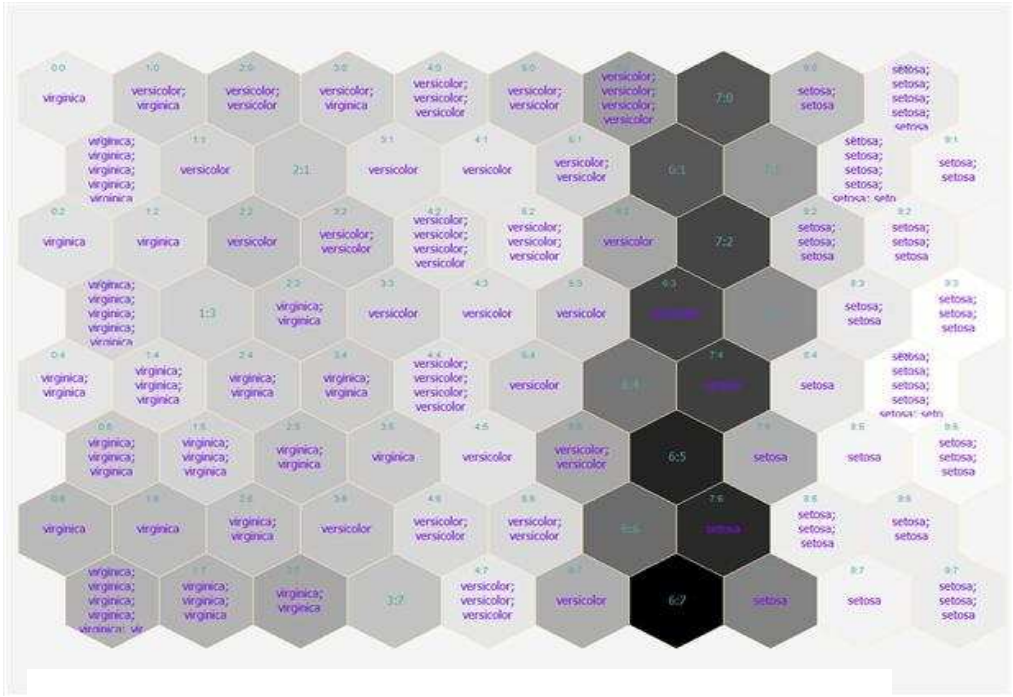


Fig.3. Output Map of SOM, trained by Spice-SOM with Iris Data

Label: virg = virginica vers = versicolor seto = setosa

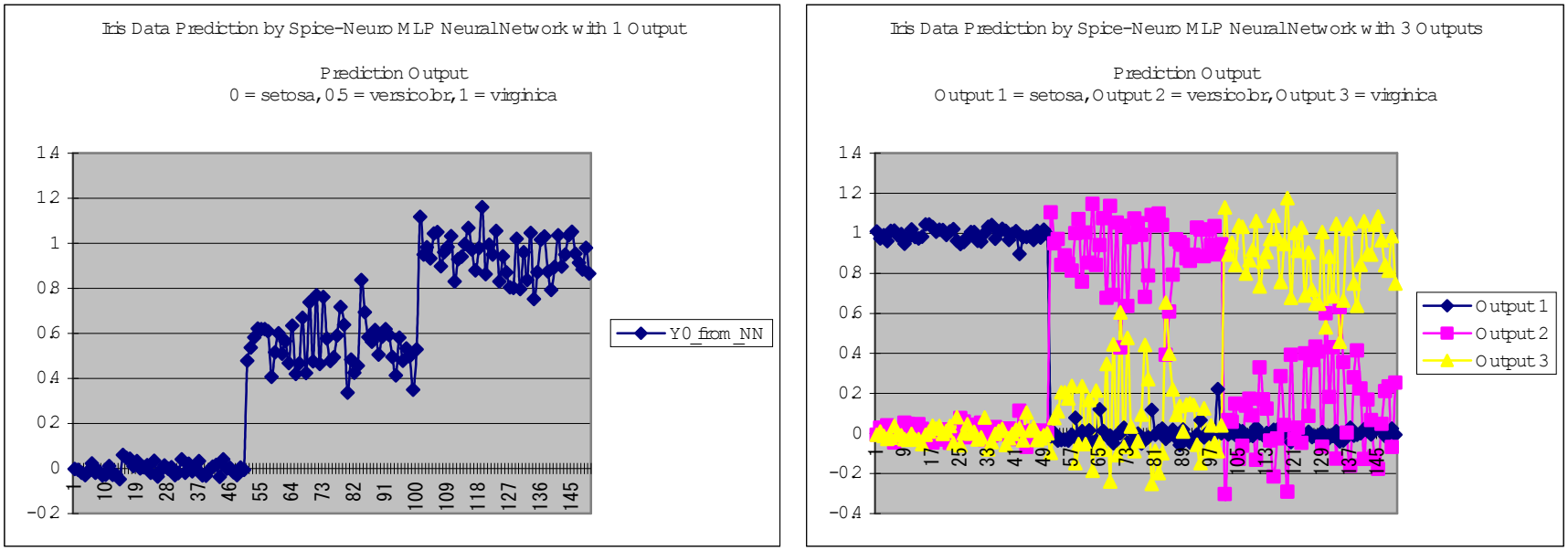
No.	X0	X1	X2	X3	X4	X5	X6	X7	X8	X9
Y0	virg	vers; virg	vers	vers; virg	vers	vers	vers		seto	seto
Y1	virg	vers		vers	vers	vers			seto	seto
Y2	virg	virg	vers	vers	vers	vers	vers		seto	seto
Y3	virg		virg	vers	vers	vers	vers		seto	seto
Y4	virg	virg	virg	virg	vers	vers		seto	seto	seto
Y5	virg	virg	virg	virg	vers	vers		seto	seto	seto
Y6	virg	virg	virg	vers	vers	vers		seto	seto	seto
Y7	virg	virg	virg		vers	vers		seto	seto	seto

1.3. Phân loại dữ liệu với SpiceNeuro

Phân loại Iris flower với mạng nơ ron MLP NN, ta có đồ thị ra của mạng như hình 4 sau.

Với mạng MLP 1 output, ta thấy với ngưỡng 0.2 thì mạng phân loại đúng 100% cho dữ liệu nhãn setosa. với ngưỡng 0.8 thì có 1 dataset nhãn versicolor và 1 dataset nhãn virginica bị nhầm.

Với mạng MLP 3 outputs, ta thấy, với ngưỡng 0.5 thì mạng phân loại đúng 100% cho dữ liệu nhãn setosa, 4 dataset nhãn versicolor và 3 dataset nhãn vỉginica bị nhầm. Trong trường hợp này, mạng MLP 3 outputs phân loại kém chính xác hơn mạng MLP 1 output.



Hình 4. Output Graph of MLP with Iris Data

2. Students's Score Data

Giả sử chúng ta có bảng điểm của học sinh trong một lớp. Bằng mạng nơ ron tự tổ chức Spice-SOM, chúng ta muốn phân loại các học sinh lại dựa vào bảng điểm.

2.1. Chuẩn bị dữ liệu

Để Spice-SOM có thể đọc được dữ liệu, chúng ta chuẩn bị dữ liệu như bảng 3 sau. Lưu ý phần tên có điểm trung bình của học sinh được để trong ngoặc (), để bạn đọc dễ hiểu hơn trong output map. Bạn đọc có thể tham khảo dữ liệu này trong thư mục Data của chương trình SpiceSOM.

Bảng 3. Dữ liệu điểm học sinh cho SOM

No	English	Algebra	Geometry	…	Analysis	Power System	Management Methodology	Geological System	Name
1	5	7	7	…	7	5	5	5	Pham Kieu Anh (6.0)
2	5	9	8	…	6	8	6	5	Cung Hong Hien (7.0)
…	…	…	…	…	…	…	…	…	…
99	7	8	9	…	4	7	9	7	Vo Mai Manh (6.4)
100	7	7	7	…	5	6	7	7	Pham La Trinh (6.0)

Cho mạng SOM size 6x10 học. Ta có bản đồ ra (output map và output table) như hình 5 và 6. Trên ouput map, các bạn dễ dàng nhận thấy các học sinh có cùng thành tích được xếp gần nhau và học sinh có thành tích tốt được xếp xa học sinh có thành tích kém.

Fig 5. Output Map of SOM, trained by Spice-SOM with Students' score Data

No.	X0	X1	X2	X3	X4	X5	X6	X7	X8	X9
Y0	Ich_(8.7);Thieu_(8.8);Cong_(8.6);Dien_(8.8);Nhung_(8.5)	Ha_(8.4);Mai_(8.5);Thanh_(8.7)	Nhan_(8.5);Hien_(8.5);Minh_(8.7);Vinh_(8.5)	Linh_(8.2)	Anh_(7.3)	Nghia_(7.0);Han_(7.2);Xuan_(7.2)	Truong_(6.5)	Yen_(6.7);Cong_(6.8);Xuan_(6.9)	Bay_(6.9);Men_(6.7)	Phuong_(6.6);Quy_(6.8);Thanh_(6.5);Tuong_(6.8)
Y1	Hoa_(8.5);Linh_(8.6);Si_(8.7);Tung_(8.5)	Giang_(8.3)		Hoa_(7.5)		Hien_(7.0)				Cong_(6.3);Xuan_(6.3);Trinh_(6.0)
Y2	Hung_(8.4);Hung_(8.5);Ngan_(8.4)	Duong_(8.3)	Minh_(7.9)		Kim_(7.0)	Han_(6.7);Hoang_(6.9)			Dat_(6.1);Hoang_(6.2)	Phuong_(6.3)
Y3	Quang_(8.3);Thao_(8.3);Trang_(8.3)			Nhi_(7.5)	Tranh_(6.9)	Loan_(6.6)	Thieu_(6.1)	Thanh_(6.3); Nhan_(6.2)		Thu_(6.0);Hien_(6.0);Nghia_(6.0);Lieu_(6.0);Luan_(6.0)
Y4	Hieu_(7.4)		Lan_(7.2)	Dung_(6.9)		Linh_(6.3);Hung_(6.5)			Nga_(5.7);La_(6.0)	My_(6.0);Nghia_(5.8);Hong_(5.7)
Y5	Mai_(7.1);May_(7.1); Thu_(7.3)	Trang_(7.1); Anh_(7.1);Hue_(7.1)	Huong_(7.2)	May_(7.0);Ma_(6.9);Chi_(6.7);Han_(6.6)		Thao_(6.6);Hieu_(6.5);Lan_(6.5)	Manh_(6.4)	Anh_(6.0);Han_(5.9);La_(6.0);Thieu_(6.1);Thao_(6.0); Minh_(5.8)	Minh_(6.0);Thanh_(6.0);Yen_(5.9)	Xuan_(6.0);Sang_(5.8);Nga_(6.1);Ly_(5.6)

Fig. 6. Output Table of SOM, trained by Spice-SOM with Students' score Data

3. Face/Nonface Classification

Các bạn học nhận dạng và xử lý ảnh đã biết để phát hiện khuôn mặt, phương pháp thường dùng hiện nay là dùng Haar-like feature + Adaboost Algorithm. Dùng SOM và MLP NN cũng có độ chính xác cao nhưng tốc độ nhận dạng chậm. Ví dụ ở đây dùng MLP NN và SOM để phân loại các frame có chứa khuôn mặt, với mục đích minh họa cách sử dụng MLP NN, SOM và Output Map của SOM.

3.1. Chuẩn bị dữ liệu

Độ dài của các vector dữ liệu phụ thuộc vào feature mà các bạn sử dụng. Giả sử ta có một tập ảnh mẫu kích thước *m x n*. Nếu dùng pixel value làm feature (vector biểu diễn ảnh), ta sẽ có một vector dài *m x n* cho mỗi ảnh. Nếu dùng pixel value histogram làm feature, ta sẽ có một vector dài 256 cho mỗi ảnh. Nếu dùng các fearture khác chẳng hạn như Histogram of Oriented Gradient, vector biểu diễn sẽ phụ thuộc vào tham số mà các bạn chọn khi tạo feature. Tài liệu này không đề cập đến các feature trong nhận dạng ảnh.

Để phân loại ảnh bằng MLP, ta cần mã hóa đầu ra yêu cầu (desired output), chẳng hạn ta sử dụng 1 đầu ra với giá trị 1.0 là face, 0.0 là non face. Để phân loại ảnh bằng SOM, ta chỉ cần vector biểu diễn ảnh và nhãn, ví dụ như sau:

ID	feature vector	Output	label
1		1.0	face
2		1.0	face
…	…	…	…
n		0.0	nonfacce
n+1		0.0	nonfacce
…	…	…	…

Fig.7. Data for MLP NN

ID	feature vector	label
1		face
2		face
…	…	…
n		nonfacce
n+1		nonfacce
…	…	…

Fig.8. Data for SOM

3.2. Phân loại ảnh face-nonface bằng Spice-SOM

Các hình 9 và 10 sau minh họa Output Maps của SOM qua một số lần học với các kích thước SOM Size khác nhau, training với 400 ảnh mẫu khuôn mặt download từ http://www.cl.cam.ac.uk/research/dtg/attarchive/facedatabase.html và 700 ảnh mẫu không phải khuôn mặt được lấy ngẫu nhiên từ internet, 324 histogram of gradient inputs. Trong hình minh họa, mỗi neuron chỉ hiển thị một ảnh đầu tiên ứng với nó (trong thực tế có thể có nhiều ảnh ứng với một neuron và có neuron không có ảnh nào tương ứng).

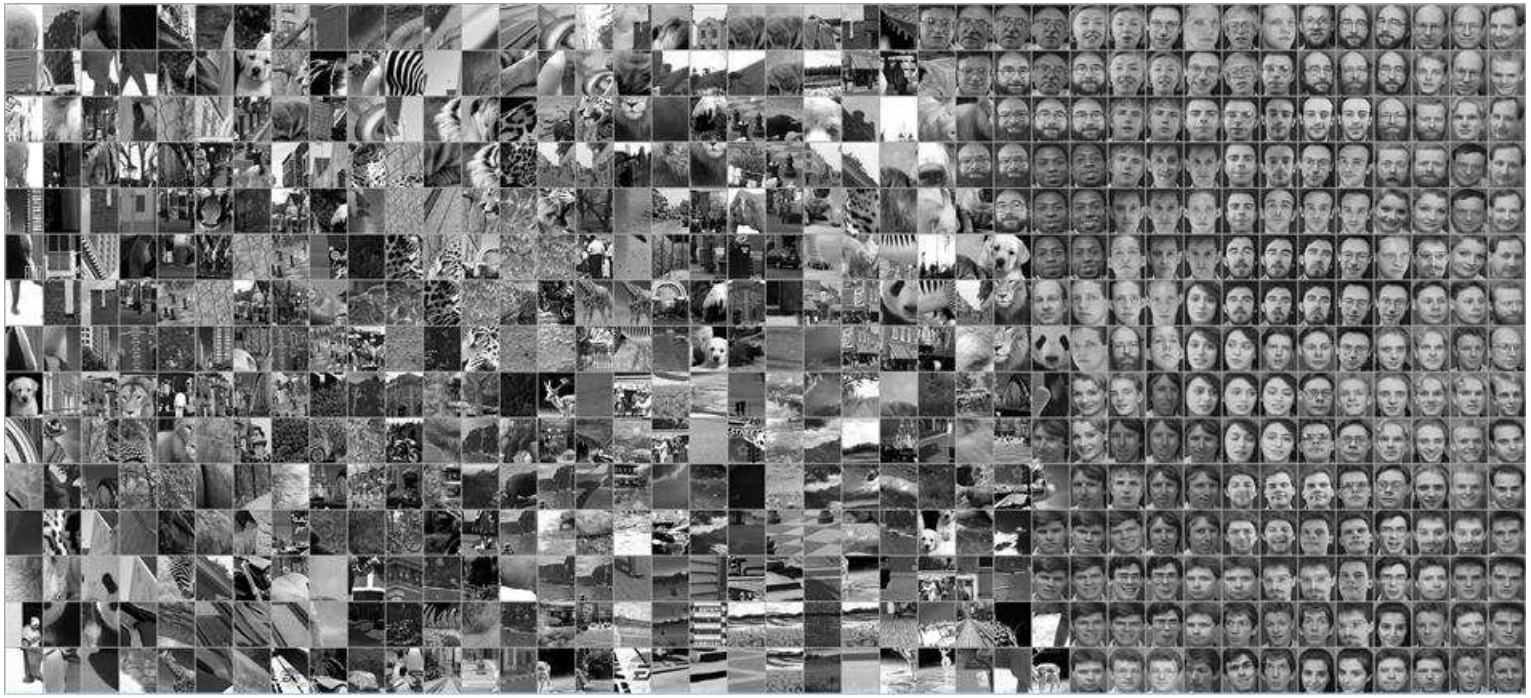


Fig. 9. Output Map of a SOM

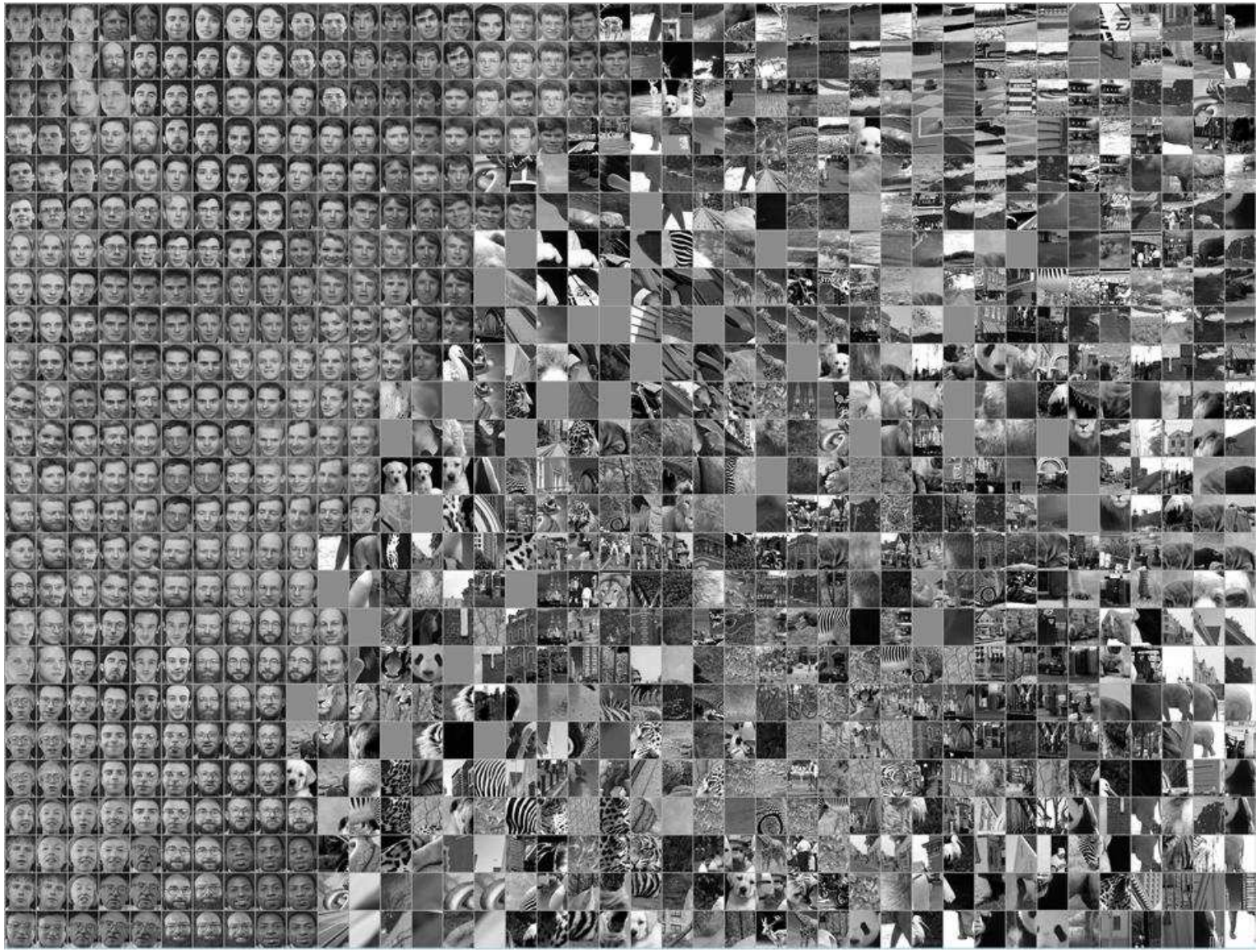


Fig.10. Output Map of a SOM

3.3. Phân loại ảnh face-nonface bằng Spice-MLP

Hình 11 minh họa lỗi trong quá trình học, giá trị đầu ra của mạng nơ ron (Actual Output from NN) và giá trị ra yêu cầu (Desired Output), với 1100 dataset (400 faces + 700 nonfaces), 324 histogram of gradient inputs, 20 hidden and 1 output neurons, Hyperbolic Tangent Activated Funtion, 605/1100 datasets for training (55%) và 495/1100 data set for testing (45%).

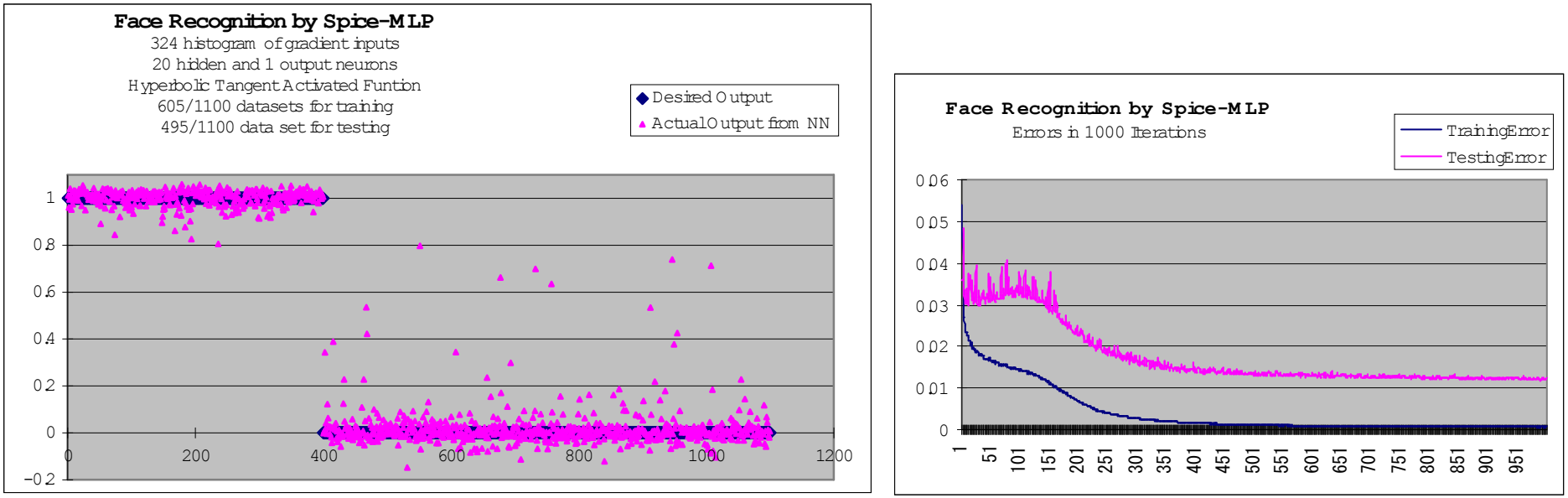
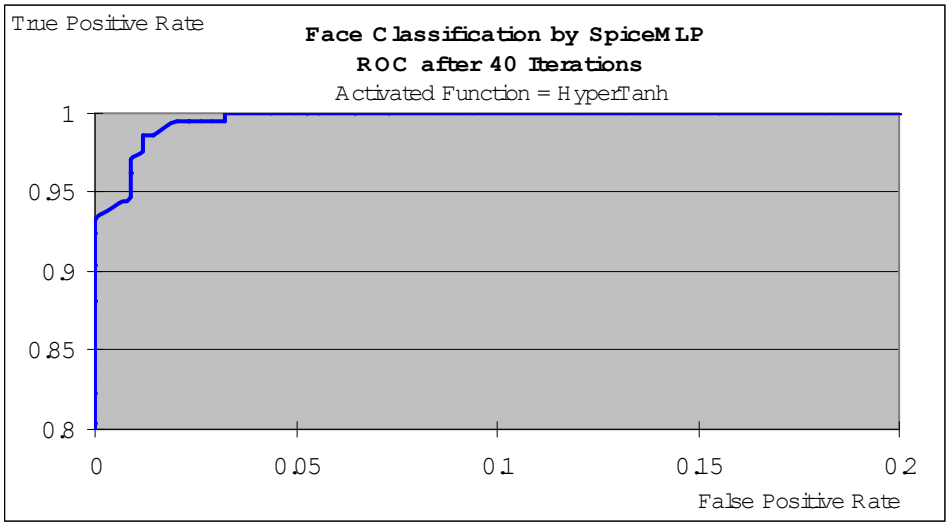


Fig.11. Outputs and Training Errors of a MLP with face/non-face data

Hình 12 sau minh họa đường đặc tính ROC (Receiver Operating Characteristic – ROC) và độ chính xác với các ngưỡng khác nhau trên tập dữ liệu kiểm tra.

Ta thấy với ngưỡng để phân biệt face/nonface là 0.52, thì mạng MLP phân loại đúng 98.7% với tập dữ liệu kiểm tra này.



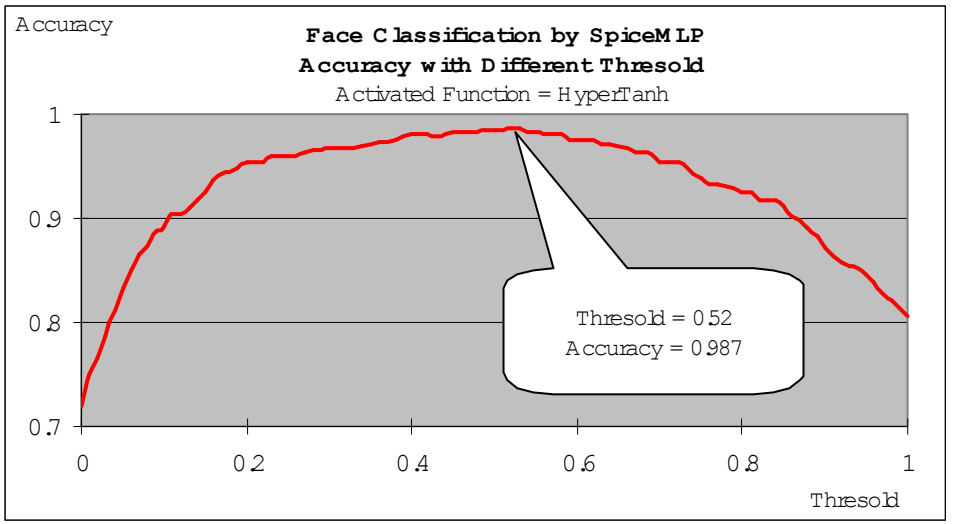


Fig.12. ROC and Accuracy Lines by a MLP with face/nonface data

4. Phân loại ảnh người đi bộ

4.1. Chuẩn bị dữ liệu

Dữ liệu minh họa ở đây được lấy từ 924 ảnh người đi bộ download từ http://www.cl.cam.ac.uk/research/dtg/attarchive/facedatabase.html và 1100 ảnh không phải người đi bộ được lấy ngẫu nhiên từ internet.

4.2. Phân loại ảnh người đi bộ bằng SpiceSom

Các hình 13 và 14 trang sau minh họa Output Maps của SOM qua một số lần học với các kích thước SOM Size khác nhau, training với 924 ảnh người đi bộ và 1100 ảnh không phải người đi bộ, 810 histogram of gradient inputs.

Fig.13. Output Map of a SOM with Pedestrian Data

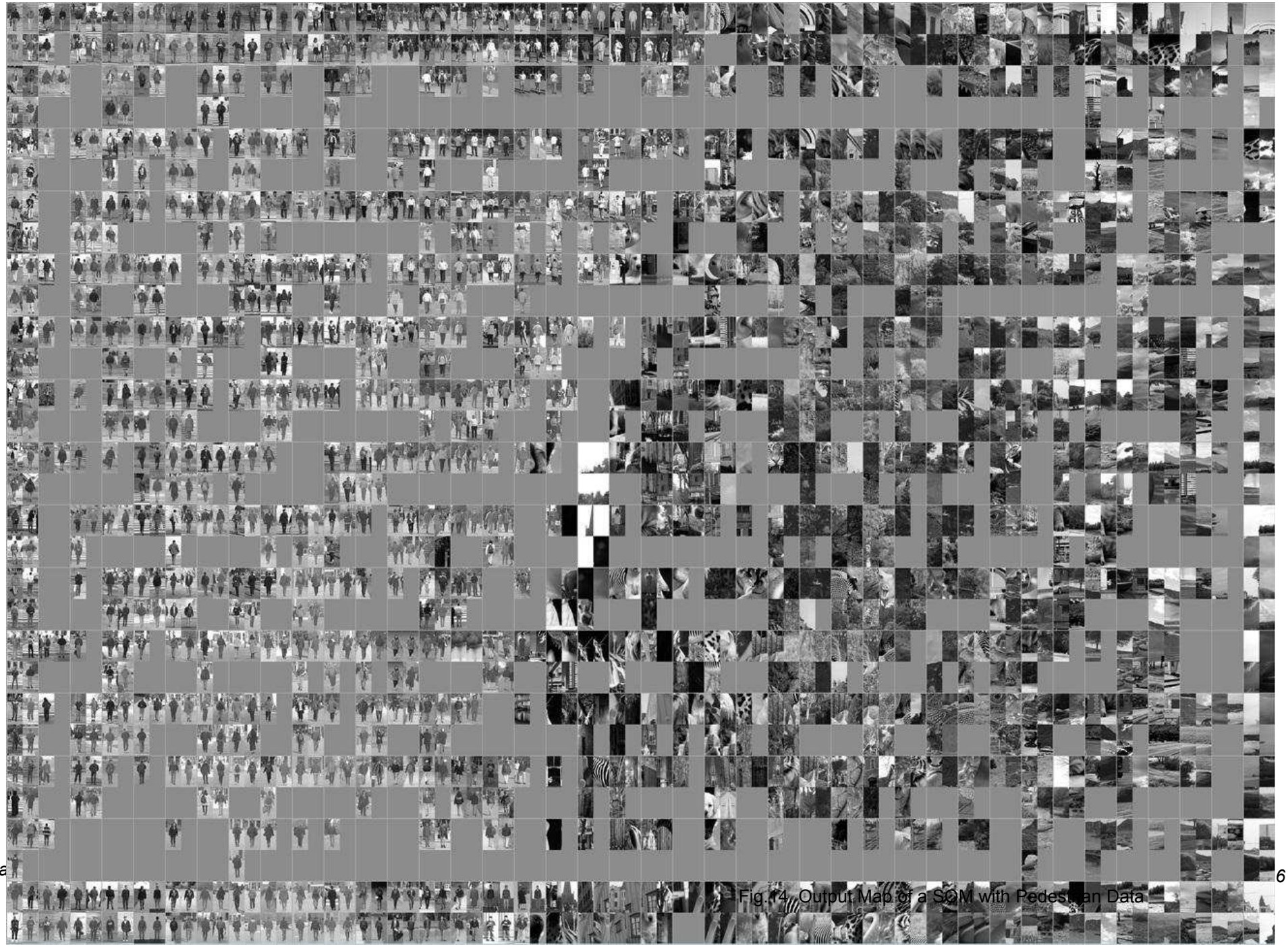


Fig.14. Output Map of a SOM with Pedestrian Data

4.3. Phân loại ảnh người đi bộ bằng SpiceMLP

Hình 15 minh họa lỗi trong quá trình học, giá trị đầu ra của mạng nơ ron (Actual Output from NN) và giá trị ra yêu cầu (Desired Output), với 2024 dataset (924 pedestrial + 1100 non-pedestrial), 810 histogram of gradient inputs, 5 hidden and 1 output neurons, Hyperbolic Tangent Activated Funtion, 1113/2024 datasets for training (55%) và 911/2024 data set for testing (45%).

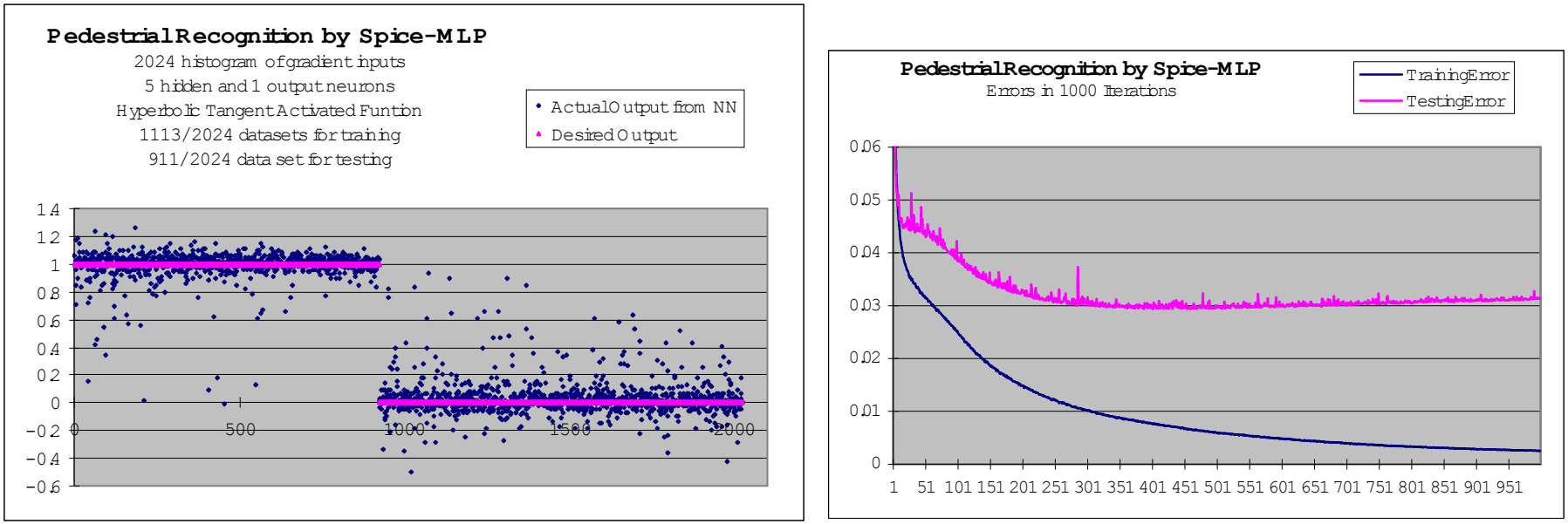


Fig.15. Outputs and Training Errors of a MLP with Pedestrial Data

Các hình nhỏ trong hình 16 sau minh họa đường đặc tính ROC trên tập dữ liệu kiểm tra khi mạng MLP vừa khởi tạo (chưa train hay 0 iteration) và sau khi đào tạo qua 1, 2, 3, 10, 20 iterations. Ta thấy khi mạng MLP vừa khởi tạo, đường ROC giống như đường ROC của phép chọn ngẫu nhiên. Mạng hội tụ khá nhanh sau một số ít lần lặp (iterations).

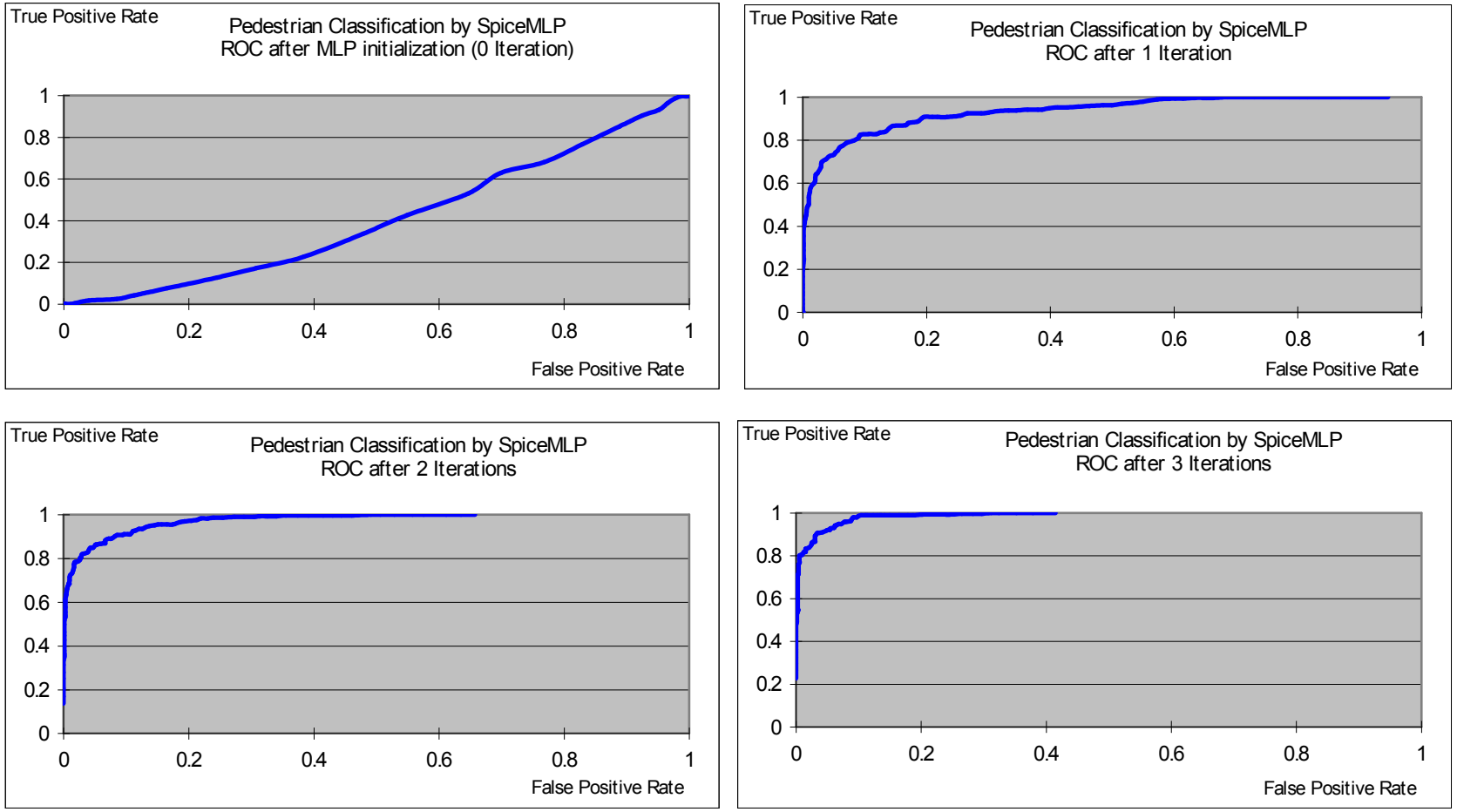


Fig.16. ROC after several training iteration (continued in next page)

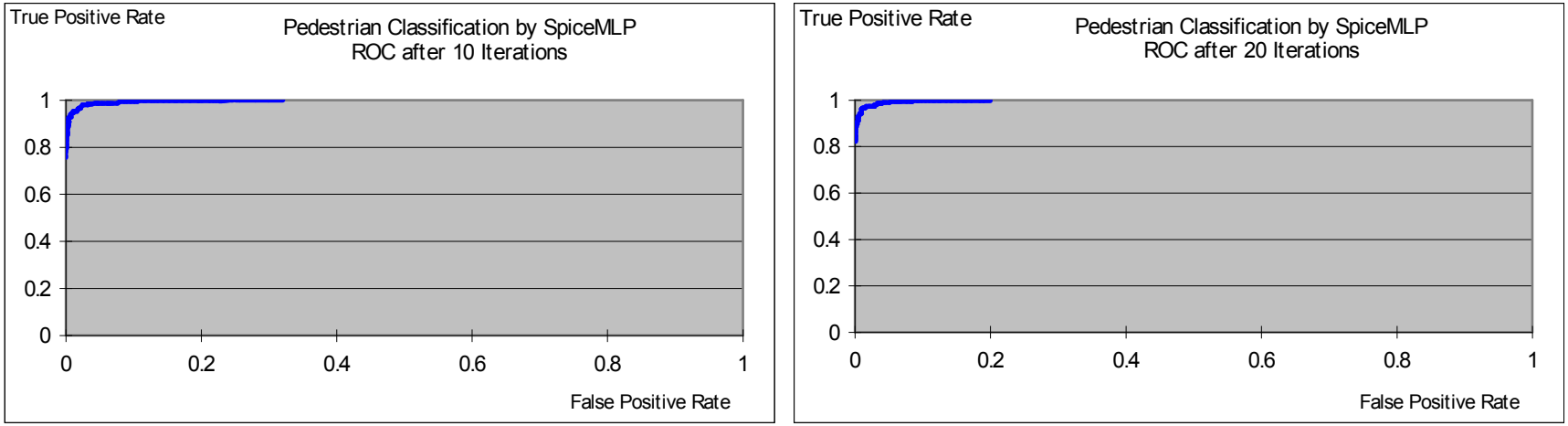


Fig.16. ROC after several training iteration (continued)

Hình 17 sau minh họa đường đặc tính ROC và độ chính xác với các ngưỡng khác nhau trên tập dữ liệu kiểm tra sau 30 iterations. Với ngưỡng để phân biệt pedestrial/non-pedestrial là 0.75, thì mạng MLP phân loại đúng 98.2% với tập dữ liệu kiểm tra này.

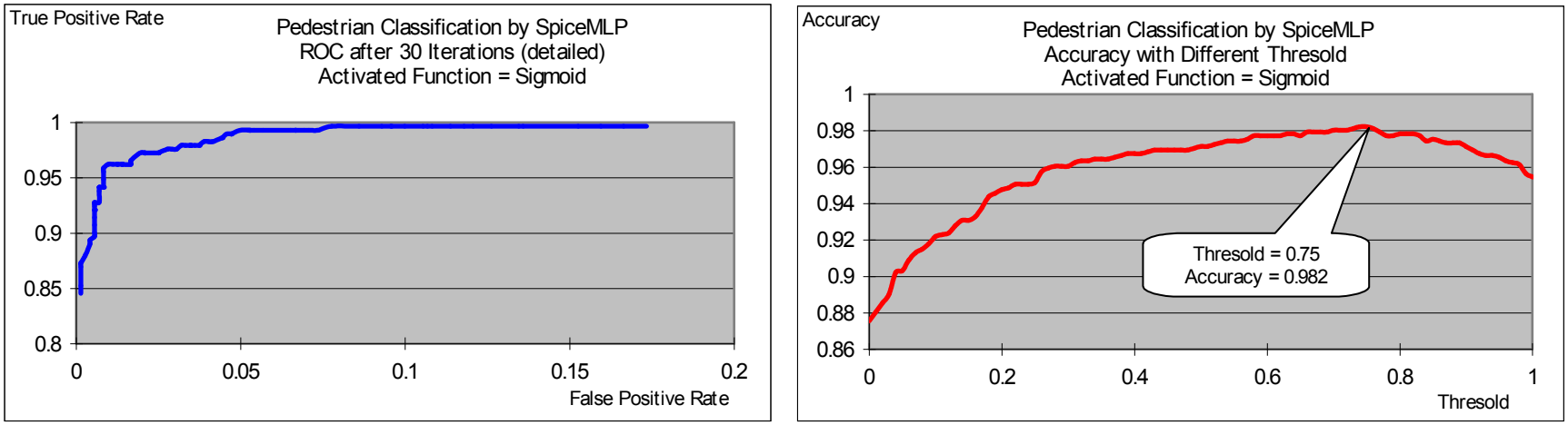


Fig.17. ROC and Accuracy Lines by a MLP with pedestrial data

Fig.18.a. Một số ảnh phân loại nhầm non-pedestrial -> pedestrial với ngưỡng phân biệt pedestrial/non-pedestrial là 0.5

Fig. 18.b. Một số ảnh phân loại nhầm pedestrian -> non-pedestrian với ngưỡng phân biệt pedestrian/non-pedestrian là 0.5

5. Phân loại ảnh xe hơi

Tương tự như phân loại ảnh khuôn mặt và người đi bộ, các hình sau minh họa Output Maps của SOM qua một số lần học với các kích thước SOM Size khác nhau, training với các ảnh mẫu xe hơi download từ http://cogcomp.cs.illinois.edu/Data/Car/

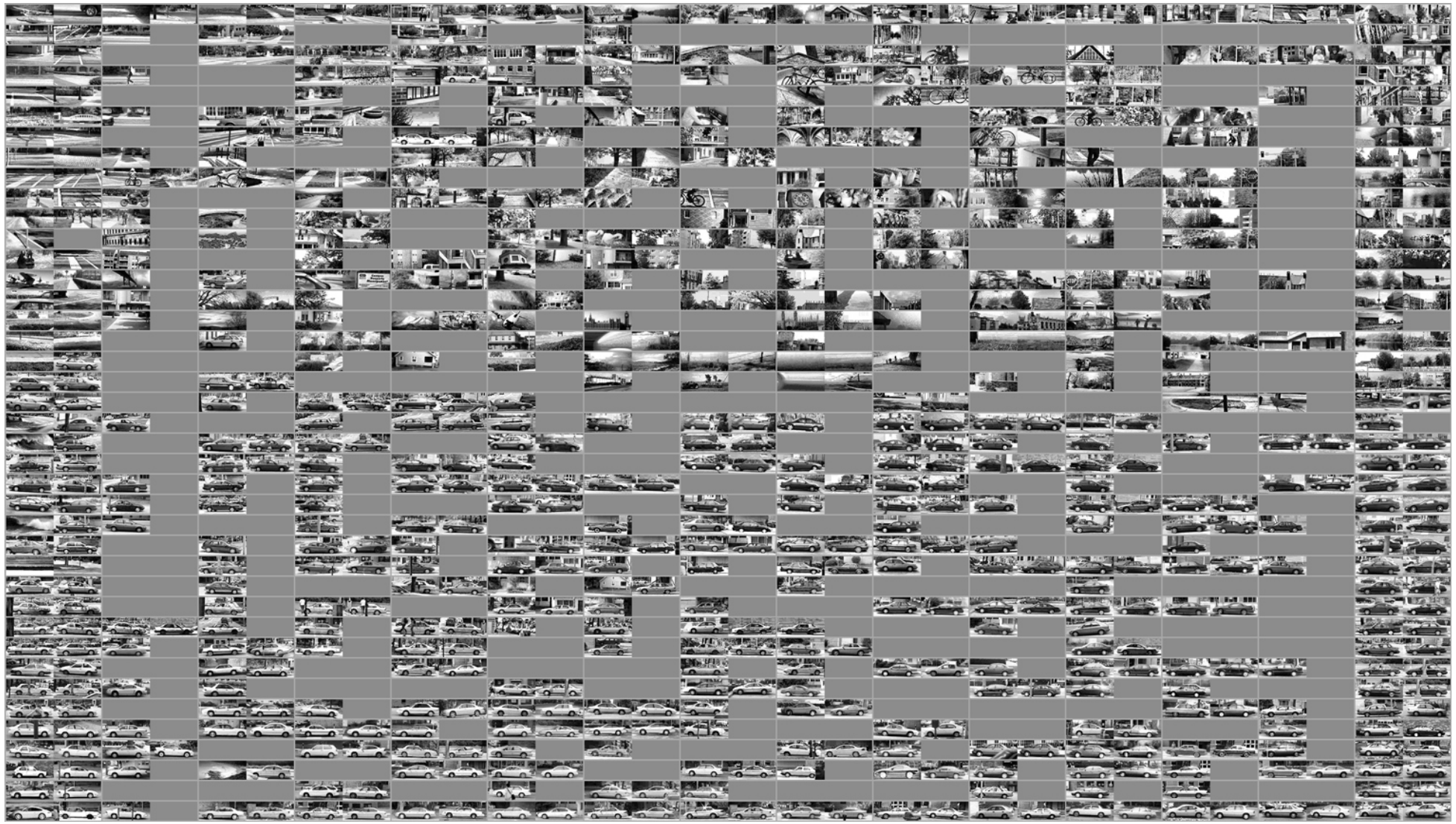


Fig.19.a. Output Map of a SOM with Car Data

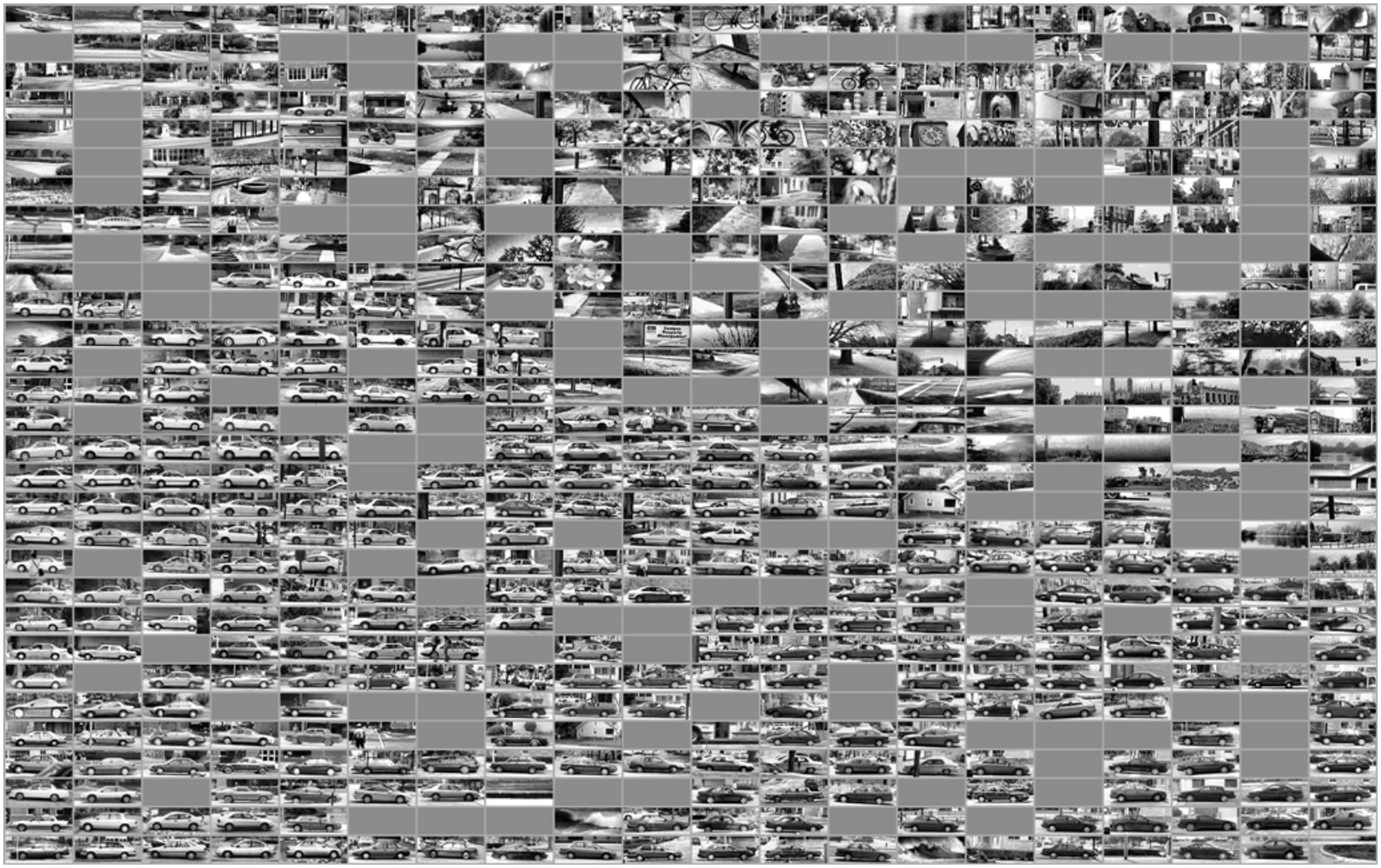


Fig.19.b. Output Map of a SOM with Car Data

6. Dự báo chứng khoán

Một trong những ứng dụng thú vị của mạng nơ ron là dự báo chứng khoán. Dựa vào các số liệu thống kê có sẵn của thị trường, mạng nơ ron có thể dự báo khá chính xác về giá chứng khoán trong những ngày tiếp theo.

Để dự báo chứng khoán bằng mạng nơ ron được chính xác, việc quan trọng nhất là tìm các dữ liệu thích hợp của thị trường bao gồm giá chứng khoán cho tới các thông tin kinh tế vĩ mô, vi mô, mã hóa các thông tin đó một cách hợp lý để mạng nơ ron có thể học và tổng quát hóa được.

Có nhiều phương pháp dự báo chứng khoán bằng mạng nơ ron. Tài liệu này trình bày với các bạn phương pháp đơn giản nhất là dựa vào giá trong thời gian đã qua để dự báo giá trong thời gian tới.

6.1. Chuẩn bị dữ liệu

Bạn có thể download được giá của NASDAQ Stock Data tại http://www.dailyfinance.com/historical-stock-prices/ Ví dụ từ ngày 3_12_1990 tới ngày 1_12_2010, dữ liệu có dạng như bảng 4 sau:

Table 4. NASDAQ Stock Price

High	Low	Open	Close	Day_Month_Year
362.2798	359.0498	361.3198	361.3198	3_12_1990
364.2898	360.4199	364.1099	364.1099	4_12_1990
370.9700	364.0198	370.8699	370.8699	5_12_1990
378.0298	370.9199	372.2898	372.2898	6_12_1990
372.3499	369.0198	371.5398	371.5398	7_12_1990
…	…	…	…	…
2545.4100	2515.4800	2519.8900	2543.1200	24_11_2010
2541.4900	2522.4000	2525.9100	2534.5600	26_11_2010
2531.0300	2496.8300	2522.2400	2525.2200	29_11_2010
2510.7100	2488.6100	2497.1200	2498.2300	30_11_2010
2558.2900	2534.8100	2535.1900	2549.4300	1_12_2010

Từ dữ liệu này, bạn muốn dự báo giá của thị trường trong ngày tiếp theo. Chẳng hạn như bạn muốn dự đoán giá trung bình (Open + Close)/2.0 của ngày tới dựa vào giá trung bình của những ngày đã qua.

Dữ liệu bạn có hiện tại là dữ liệu dạng time series, để mạng nơ ron MLP có thể học được, bạn cần chuẩn bị lại như sau.

Từ các cột dữ liệu giá Open và Close mà bạn đang có, bạn tạo cột mới có giá trị trung bình của 2 cột dữ liệu trên.

Bạn muốn sử dụng giá trung bình của 15 ngày đã qua để dự đoán giá của 1 ngày tới, thị trường NASDAQ. Bạn tạo một hàng (row) gồm 16 dữ liệu, 15 dữ liệu đầu của hàng này là giá của 15 ngày liên tiếp, dữ liệu thứ 16 (cuối cùng) là giá của ngày tiếp theo ngày thứ 15 nói trên. Như vậy bạn sẽ sử dụng mạng nơ ron có 15 đầu vào và 1 đầu ra. Dữ liệu của bạn có dạng như bảng 5 sau:

Table 5. Giá của NASDAQ Stock, được chuẩn bị lại để mạng MLP NN có thể học được

Today - 14	Today - 13	Today - 12	…	Today - 2	Today - 1	Today	Tomorrow	Label
361.3198	364.1099	370.8699	…	371.22	372.2998	373.5999	372.4099	21_12_1990
364.1099	370.8699	372.2898	…	372.2998	373.5999	372.4099	372.3999	24_12_1990
370.8699	372.2898	371.5398	…	373.5999	372.4099	372.3999	371.0498	26_12_1990
372.2898	371.5398	371.47	…	372.4099	372.3999	371.0498	371.2	27_12_1990
…	…	…	…	…	…	…	…	…
2573.305	2578.305	2575.455	…	2520.705	2499.58	2531.505	2530.235	24_11_2010
2578.305	2575.455	2575.03	…	2499.58	2531.505	2530.235	2523.73	26_11_2010
2575.455	2575.03	2571.545	…	2531.505	2530.235	2523.73	2497.675	29_11_2010
2575.03	2571.545	2544.88	…	2530.235	2523.73	2497.675	2542.31	30_11_2010

Tương tự, nếu bạn muốn sử dụng giá của 20 ngày đã qua để dự đoán giá của 3 ngày tới, bạn tạo dữ liệu như trên và sẽ sử dụng mạng nơ ron 20 đầu vào và 3 đầu ra, ...

Ví dụ sau sử dụng giá trung bình của 15 ngày đã qua để dự đoán giá trung bình của của 1 ngày tới, tức là 15 đầu vào, 1 đầu ra, từ ngày 21_12_1990 tới ngày 30_11_2010. Sử dụng chương trình SpiceMLP để dự đoán. Số Datasets là 5026. Ở đây 80% dữ liệu (4021) datasets được dùng để học, 20% dữ liệu (1005) datasets được dùng để kiểm tra. Số lần lặp khi học là 1000.

```
Number of trained data: 4021.       Number of tested data: 1005
Taken iterations: 1000              Number of Inputs: 15
Number of Outputs: 1
```

Dữ liệu này có trong thư mục Data khi các bạn cài đặt chương trình SpiceMLP (hay còn gọi là Spice Neuro).

6.2. Tìm các thông số thích hợp cho mạng nơ ron

Để mạng nơ ron dự báo tốt, cần chọn các thông số thích hợp cho mạng. Thông số thích hợp thường phụ thuộc nhiều vào dữ liệu của bạn, một thông số có thể tốt cho dữ liệu này nhưng lại kém khi sử dụng ở dữ liệu khác. Ở đây giới thiệu với các bạn phương pháp đơn giản nhất: với cùng dữ liệu học và kiểm tra, thay đổi một thông số để tìm giá trị tối ưu tương đối. Lưu ý, trước khi đào tạo mạng, bạn cần chuẩn hóa dữ liệu vào và ra. Ví dụ ở đây dùng hàm Linear để chuẩn hóa. Trước hết, ta tìm số nơ ron lớp ẩn sao cho (có vẻ) hợp lý nhất.

Table 6. Lỗi khi sử dụng cùng hàm biến đổi HyperTanh cho lớp ẩn và lớp ra, thay đổi số nơ ron ẩn.

Number of Hidden Neurons	Errors	
	Training	Testing
15	9.01E-05	9.35E-05
12	9.22E-05	1.12E-04
10	8.62E-05	1.04E-04
8	9.01E-05	9.36E-05
6	6.54E-05	6.94E-05
4	6.49E-05	6.65E-05
3	4.43E-05	4.25E-05
2	4.28E-05	4.01E-05

Các bạn thấy, số nơ ron lớp ẩn là 2 có vẻ tốt hơn. Tiếp theo, ta tìm hàm biến đổi của lớp ra.

Table 7. Lỗi khi sử dụng cùng hàm biến đổi HyperTanh cho lớp ẩn, số nơ ron ẩn là 2, thay đổi hàm biến đổi của lớp ra.

Activated Function		Errors	
Hidden Layer	Output Layer	Training	Testing
HyperTanh	Identity	3.94E-05	3.38E-05
HyperTanh	Sigmoid	7.35E-05	1.08E-04
HyperTanh	ArcTan	6.97E-05	6.41E-05
HyperTanh	ArcSinh	6.24E-05	6.47E-05
HyperTanh	Sin	5.05E-05	4.83E-05
HyperTanh	Gaussian	5.48E-05	5.77E-05
HyperTanh	XSinX	5.54E-05	5.53E-05

Các bạn thấy, hàm biến đổi lớp ra là Identity có vẻ tốt hơn. Và sau cùng, ta tìm hàm biến đổi cho lớp ẩn.

Bảng 8. Lỗi khi sử dụng cùng hàm biến đổi Identity cho lớp ra, số nơ ron ẩn là 2, thay đổi hàm biến đổi của lớp ẩn.

Activated Function		Errors	
Hidden Layer	Output Layer	Training	Testing
Sigmoid	Identity	4.33E-05	3.93E-05
HyperTanh	Identity	3.94E-05	3.38E-05
ArcTan	Identity	4.03E-05	4.37E-05
ArcSinh	Identity	4.05E-05	3.99E-05
Sin	Identity	4.07E-05	4.18E-05
Gaussian	Identity	4.43E-05	4.25E-05
XsinX	Identity	4.72E-05	4.98E-05

Các bạn thấy, hàm biến đổi lớp ẩn là HyperTanh có vẻ tốt hơn.

Như vậy ta sẽ chọn số nơ ron lớp ẩn là 2, hàm biến đổi lớp ra là Identity, hàm biến đổi lớp ẩn là HyperTanh.

6.3. Đào tạo mạng

Tiến hành đào tạo mạng vài lần và chọn lần đào tạo có lỗi training error và testing error nhỏ nhất. Thông tin về mạng học và đồ thị lỗi của bạn sẽ có dạng sau.

Thông tin của lần học cuối cùng

```
Hàm biến đổi cho lớp ẩn: HyperTanh
Hàm biến đổi cho lớp ra: Identity
Tỷ lệ học cuối cùng: 0.03308719
Giá trị MSE của Dữ liệu học: 4.138118E-05
Giá trị MSE của Dữ liệu kiểm tra: 3.423549E-05
Số lượng dữ liệu đã học: 4021
Số lượng dữ liệu đã kiểm tra: 1005
Số lần lặp: 1000
```

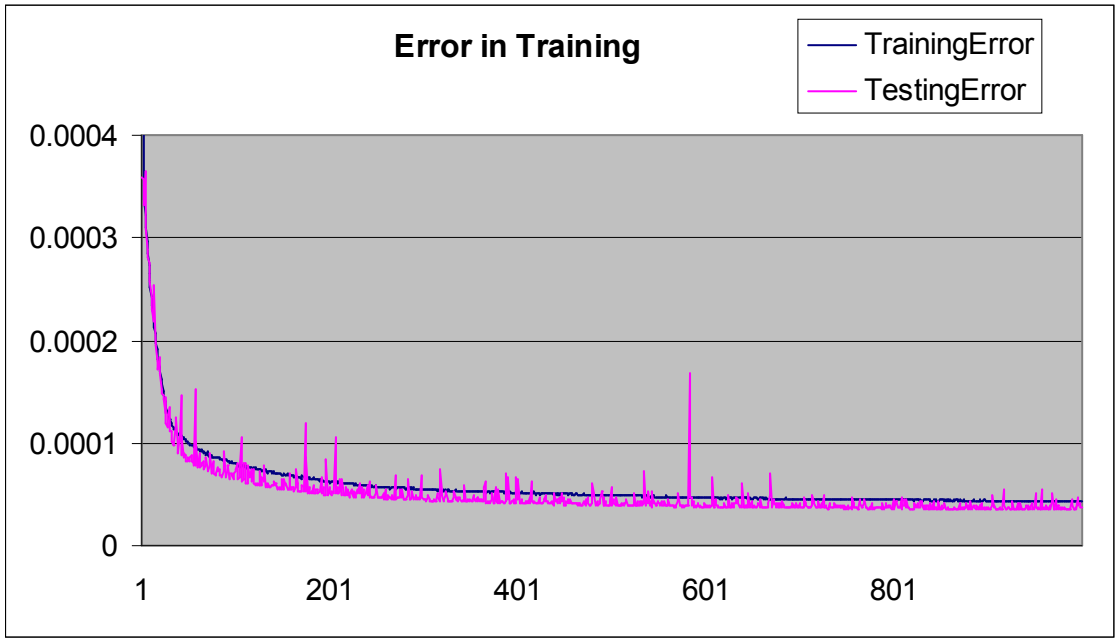


Fig. 20. Training Error when learning NASDAQ Stock prices

6.4. Kiểm tra dữ liệu được mô hình hóa (modeling)

Sau khi mạng học xong, kiểm tra dữ liệu học trong phần "Xem dữ liệu", đầu ra của dữ liệu học (training data) do mạng MLP đưa ra (NN Outputs) có dạng như hình 21 sau:

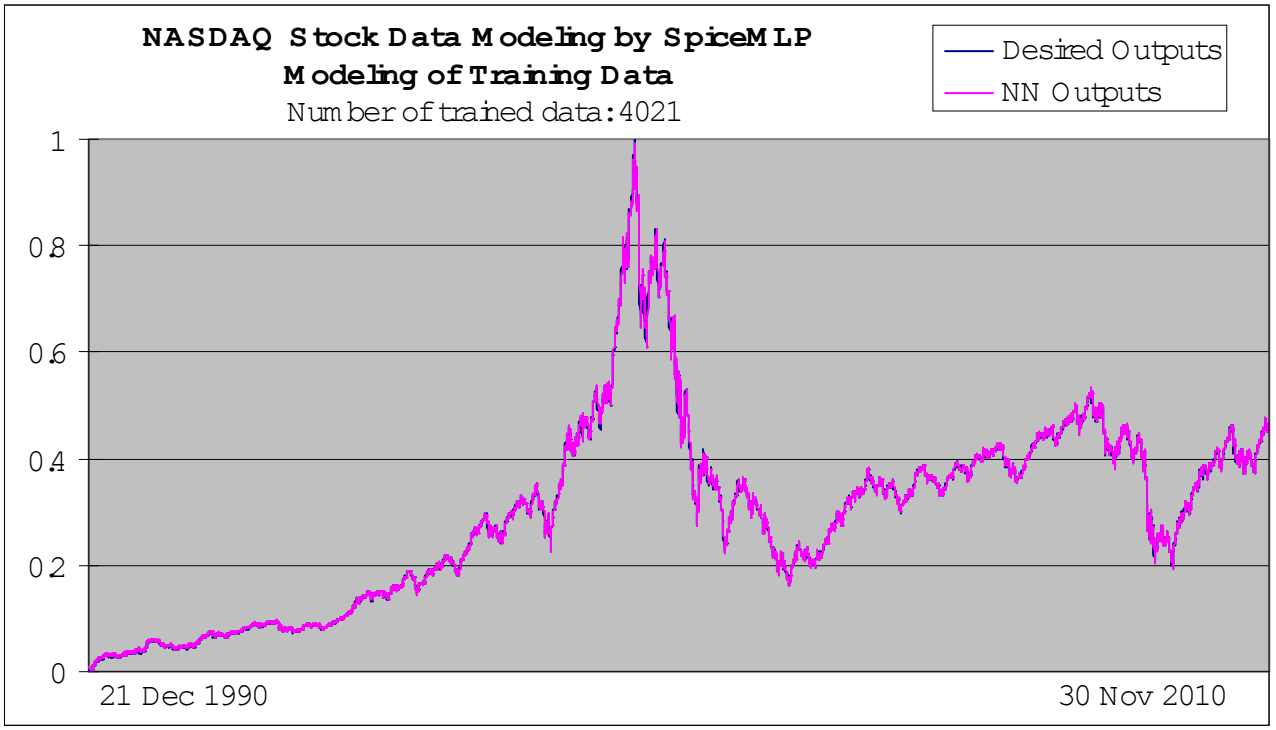


Fig. 21. Outputs of Training Data (NASDAQ Stock prices)

Ta thấy với dữ liệu học, đầu ra của mạng gần trùng khớp với đầu ra yêu cầu (tức là đầu ra thực của dữ liệu học).
Với dữ liệu kiểm tra (testing data), bạn có đồ thị dạng hình 22 sau:

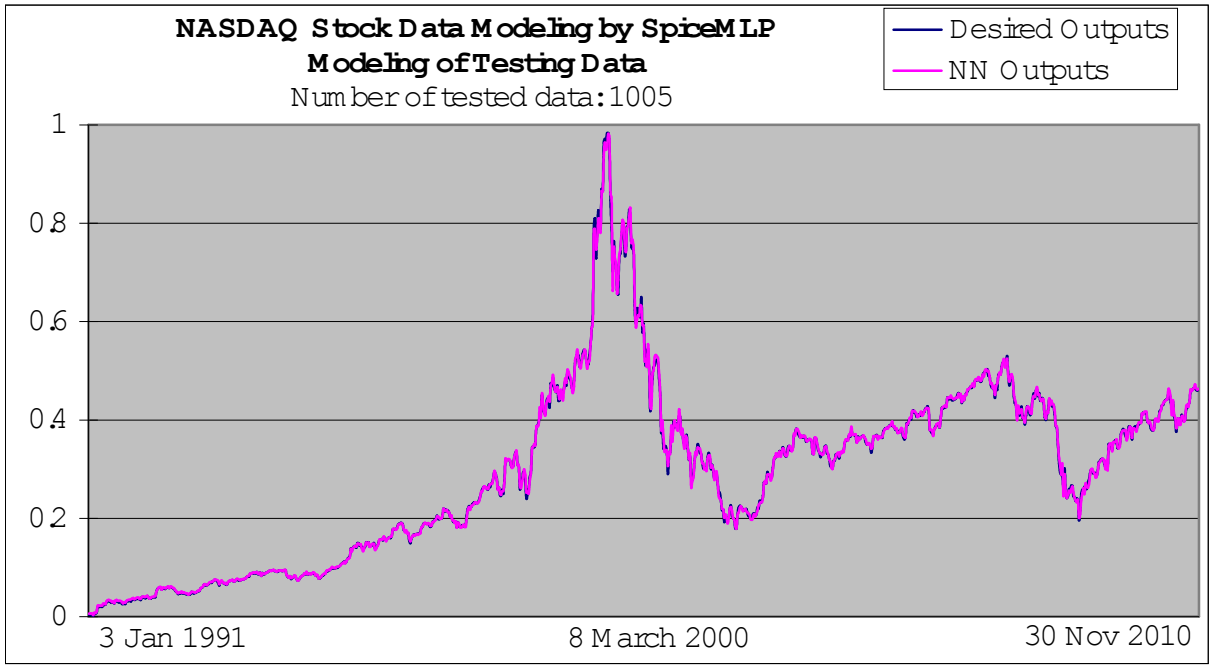


Fig. 22.a. Outputs of Testing Data (NASDAQ Stock prices)

Xem trong khoảng thời gian ngắn, bạn có đồ thị như hình ở trang sau.

Với dữ liệu kiểm tra, đầu ra của mạng cũng xấp xỉ đầu ra yêu cầu (tức là đầu ra thực của dữ liệu học). Bạn dễ nhận thấy mạng nơ ron MLP học khá tốt. Tuy nhiên tại một số điểm vẫn còn lỗi nhỏ.

Câu hỏi dành cho bạn đọc

- ✧ Làm thế nào để giảm lỗi trong dự báo bằng mạng MLP?
- ✧ Liệu còn có thể dùng những dữ liệu khác để dự báo chứng khoán?
- ✧ Liệu có thể áp dụng phương pháp trên cho các bài toán time series khác?
- ✧ Bài toán không phải time series thì chuẩn bị dữ liệu như thế nào?
- ✧ Ở ví dụ trên sử dụng giá trung bình của 15 ngày đã qua để dự đoán giá trung bình của của 1 ngày tới. Làm thể nào để sử dụng giá trung bình của

30 ngày đã qua để dự đoán 3 giá trung bình của của 3 ngày tới?

✧ Với dữ liệu cho đến hôm nay, làm thế nào để dự đoán giá của ngày mai?

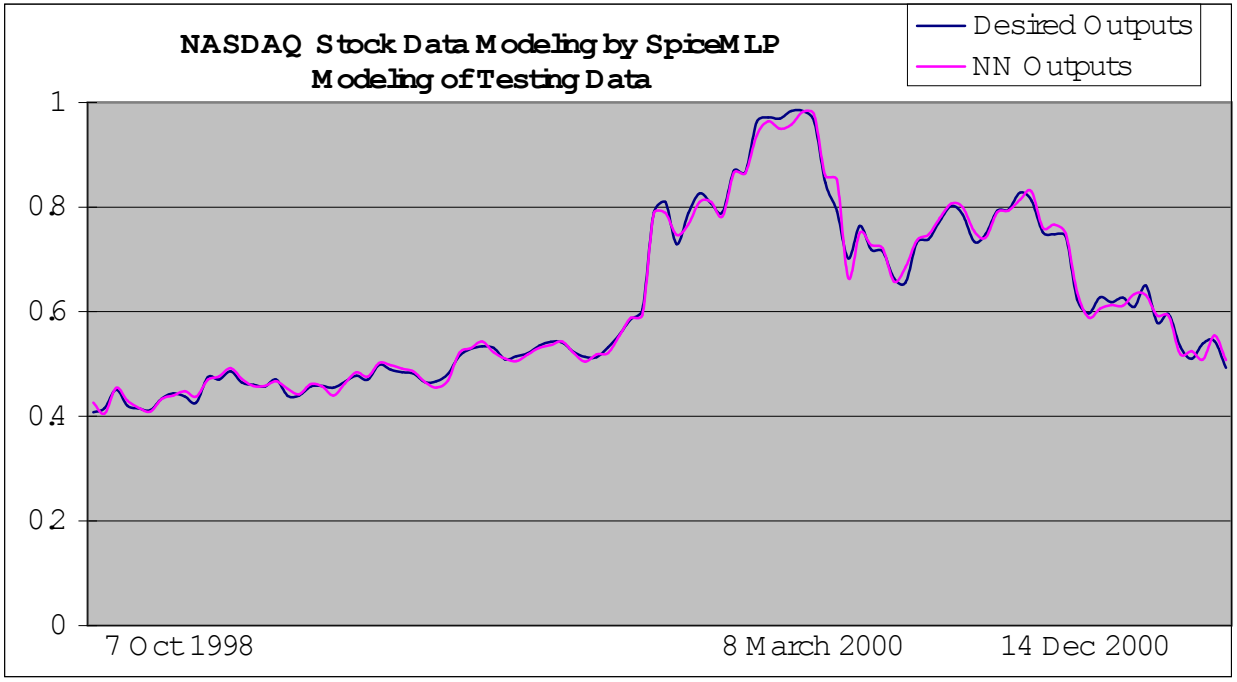


Fig. 22.b. Outputs of Testing Data (NASDAQ Stock prices)

7. Dự báo tỷ giá

Giả sử bạn muốn dự báo tỷ giá Canada Dollar/US Dollar và Canada Dollar/Japanese Yen. Bạn có thể download được dữ liệu này tại http://www.bankofcanada.ca/rates/exchange/10-year-lookup/

Tỷ giá Canada Dollar/US Dollar và Canada Dollar/Japanese Yen download được từ địa chỉ trên có dạng như sau

USD->CAD	CAD->USD	Date	JPY->CAD	CAD->JPY	Date
1.5657	0.6387	2001/10/12	0.012948	77.232005	2001/10/12
1.5579	0.6419	2001/10/15	0.012883	77.621672	2001/10/15
1.5619	0.6402	2001/10/16	0.01288	77.639752	2001/10/16
…	…	…	…	…	…
1.5981	0.6257	2001/11/8	0.01331	75.13148	2001/11/8
1.6021	0.6242	2001/11/9	0.013325	75.046904	2001/11/9
Bank holiday	Bank holiday	2001/11/12	Bank holiday	Bank holiday	2001/11/12
1.5981	0.6257	2001/11/13	0.013158	75.999392	2001/11/13
1.5916	0.6283	2001/11/14	0.013082	76.440911	2001/11/14
1.5868	0.6302	2001/11/15	0.012961	77.154541	2001/11/15
…	…	…	…	…	…

Giả sử bạn dùng hai cột dữ liệu CAD->USD và CAD->JPY, dùng dữ liệu của 30 ngày đã qua để dự báo tỷ giá của ngày thứ 5 tới. Nghĩa là ngày hiện tại của bạn là Today, bạn dùng dữ liệu Today-30, Today-29, ..., Today-1,Today để dự báo tỷ giá của Today+5. Như vậy bạn sẽ có 60 đầu vào (30 đầu vào cho CAD->USD và 30 đầu vào cho CAD->JPY), 2 đầu ra cho CAD->USD và CAD->JPY. Ví dụ sau dùng dữ liệu của 15 ngày đã qua để dự báo tỷ giá của ngày thứ 5 tới với hai tỷ giá CAD->USD và CAD->JPY. Nghĩa là bạn sẽ có 30 đầu vào, 2 đầu ra.

Với dữ liệu như trên, bạn bỏ các dòng "Bank Holiday" đi, sau đó chuẩn bị dữ liệu như sau

	CAD->USD					CAD->JPY					CAD->USD	CAD->JPY	LABEL
ID	Today-14	Today-13	…	Today-1	Today	Today-14	Today-13	…	Today-1	Today	Today+5	Today+5	Today Date
1	0.6387	0.6419		0.6302	0.6285	77.232005	77.621672		77.23797	76.581406	0.6257	75.13148	2001/11/1
2	…	…	…	…	…	…	…	…	…	…	…	…	…

Bạn chú ý thấy CAD->USD có giá trị trong [0.6199,1.0905] và CAD->JPY có giá trị trong [68.918, 123.885]. Vì hai khoảng giá trị khác nhau khá nhiều, nếu bạn chuẩn hóa dữ liệu một lần với tất cả dữ liệu, CAD->USD sẽ có giá trị nhỏ và CAD->JPY có giá trị lớn. Như vậy mạng MLP học sẽ không tốt.

Để mạng MLP học tốt, bạn nên chuẩn hóa riêng CAD->USD và CAD->JPY. Trong thư mục Data của chương trình Spice-MLP có chứa dữ liệu chưa chuẩn hóa (CAD_USD_JPN_2489_data_30inputs_2outputs.csv) và dữ liệu đã được chuẩn hóa theo phương pháp Linear (CAD_USD_JPN_Normalized_2489_data_30inputs_2outputs.csv), với 30 đầu vào, 2 đầu ra và số dữ liệu là 2489, từ ngày 2001/11/1 tới ngày 2011/10/3.

Chọn ngẫu nhiên 70% dữ liệu (1742 datasets) làm dữ liệu học (training data) và 30% dữ liệu (747 datasets) làm dữ liệu kiểm tra (testing data). Cho mạng MLP học, ta có thông tin như sau:

```
Activated Function for Hidden Layer: HyperTanh

Activated Function for Output Layer: Linear

Final Learning rate: 0.003811921

Final MSE of Training Set: 0.0004247656

Final MSE of Testing Set: 0.0004671731

Number of trained data: 1742

Number of tested data: 747

Taken iterations: 1000
```

Lưu dữ liệu đã học bởi mạng MLP vào file csv và vẽ đồ thị, ta có đồ thì đầu ra của mạng như hình 23 sau.

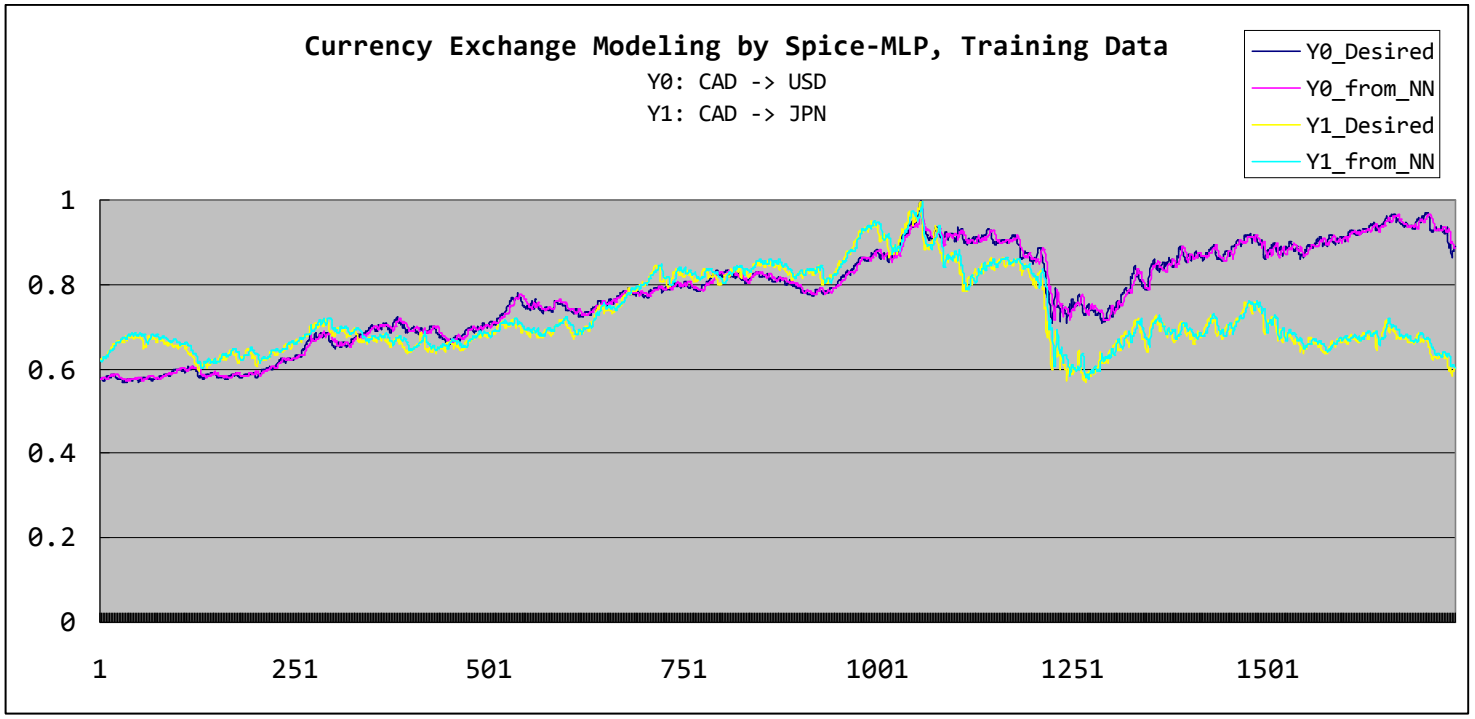


Fig. 23.a. Outputs of Training Data (CAD->USD and CAD->JPY)

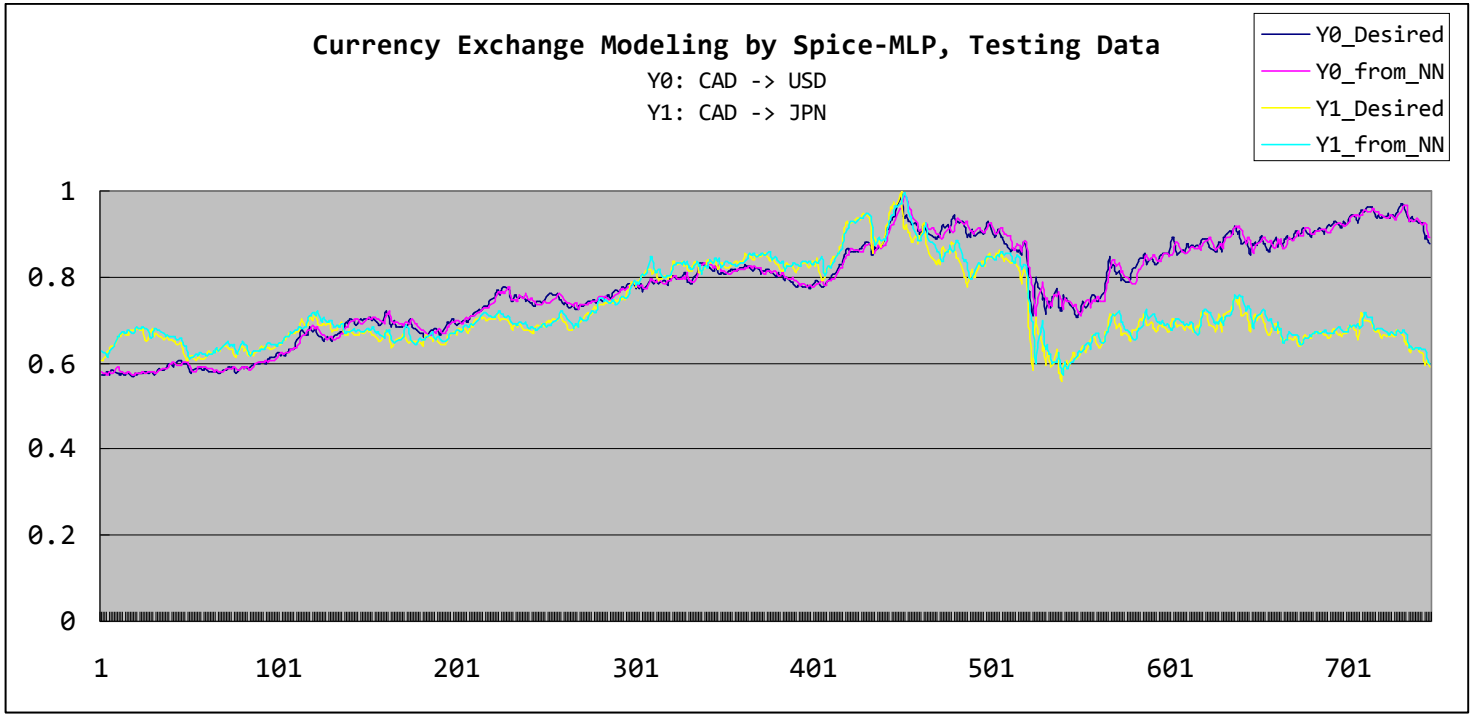


Fig. 23.b. Outputs of Testing Data (CAD->USD and CAD->JPY)

Theo đồ thị trên, ta thấy nói chung mạng MLP học tốt với cả hai dữ liệu CAD->USD và CAD->JPY, tuy nhiên tại một số điểm đầu ra của mạng hơi lệch so với đầu ra yêu cầu.

Câu hỏi dành cho bạn đọc:

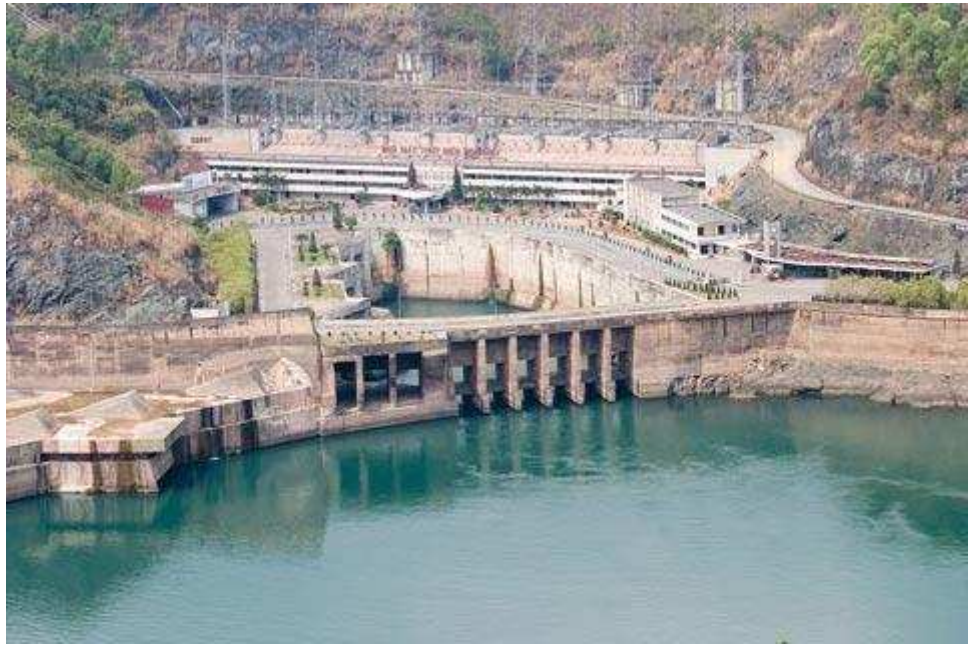
- ✧ Điều chỉnh dữ liệu và các thông số của mạng để mạng MLP học được với độ chính xác mà bạn mong muốn.
- ✧ Ví dụ trên dùng dữ liệu của 15 ngày đã qua để dự báo tỷ giá của ngày thứ 5 tới. Vậy muốn dùng dữ liệu của 60 ngày đã qua để dự báo tỷ giá của ngày thứ 10 và ngày thứ 15 tới thì làm thế nào?

8. Dự báo lưu lượng nước hồ Hòa Bình

Dựa vào dữ liệu về lưu lượng nước về hồ thủy điện trong quá khứ, ta có thể dự báo lưu lượng nước về hồ trong tương lai gần. Trong thư mục Data của chương trình Spice-MLP có file dữ liệu Hoabinh_water_level_3_input_1_output.csv. Đây là file dữ liệu về dự báo lưu lượng nước tương lai trước 10 ngày Q(t+10) của hồ Hòa Bình dựa vào các lưu lượng nước tại thời điểm hiện tại và quá khứ. Dữ liệu có 3 đầu vào gồm lưu lượng nước hiện tại Q(t), lưu lượng nước trước đó 10 ngày Q(t-10) và lưu lượng nước trước đó 20 ngày Q(t-20). Số dữ liệu là 570, trong đó 480 mẫu học (từ line 2 tới line 481) và 90 mẫu kiểm tra (từ line 482 tới line 571). Dữ liệu này do bạn Phạm Thị Hoàng Nhung, trường ĐH Thủy lợi cung cấp, bạn đọc có thể tham khảo luận văn Master của Phạm Thị Hoàng Nhung (1997) về " khảo sát một số phương pháp học máy tiên tiến, thực hiện việc kết hợp giữa phương pháp học máy mạng neuron với thuật toán gene và ứng dụng vào bài toán dự báo lưu lượng nước đến hồ Hòa Bình". Xin cảm ơn bạn Phạm Thị Hoàng Nhung đã cho phép sử dụng dữ liệu lưu lượng nước hồ Hòa Bình để minh họa trong tài liệu này.

Thông tin vắn tắt về nhà máy thủy điện Hòa bình trên wiki như sau: Nhà máy Thủy điện Hoà Bình được xây dựng tại hồ Hòa Bình, tỉnh Hòa Bình, trên dòng sông Đà thuộc miền bắc Việt Nam. Cho đến nay đây là công trình thủy điện lớn nhất Việt Nam và Đông Nam Á. Nhà máy do Liên Xô giúp đỡ xây dựng và vận hành. Công trình khởi công xây dựng ngày 6 tháng 11 năm 1979, khánh thành ngày 20 tháng 12 năm 1994. Công suất sản sinh điện năng theo thiết kế là 1.920 megawatt, gồm 8 tổ máy, mỗi tổ máy có công suất 240.000 kilowatt. Sản lượng điện hàng năm là 8,16 tỷ kilowatt giờ (KWh).

Ảnh về hồ thủy điện trên internet như sau:



Với dữ liệu Hoabinh_water_level_3_input_1_output.csv, các bạn load dữ liệu như Fig.24, chuẩn hóa dữ liệu theo phương pháp Linear, chia dữ liệu học và kiểm tra như Fig.25.

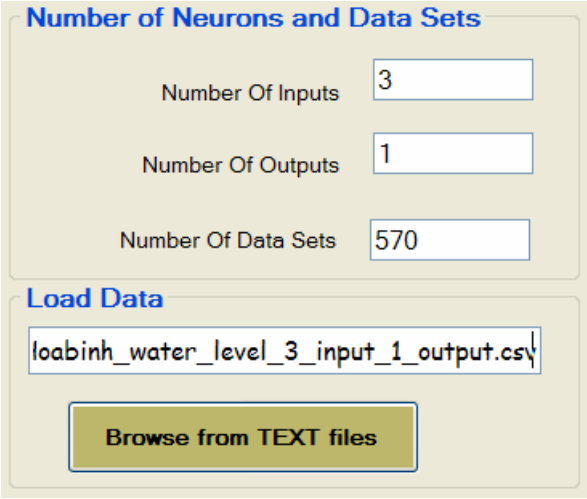


Fig. 24. Load Data with Hoabinh_water_level_3_input_1_output.csv

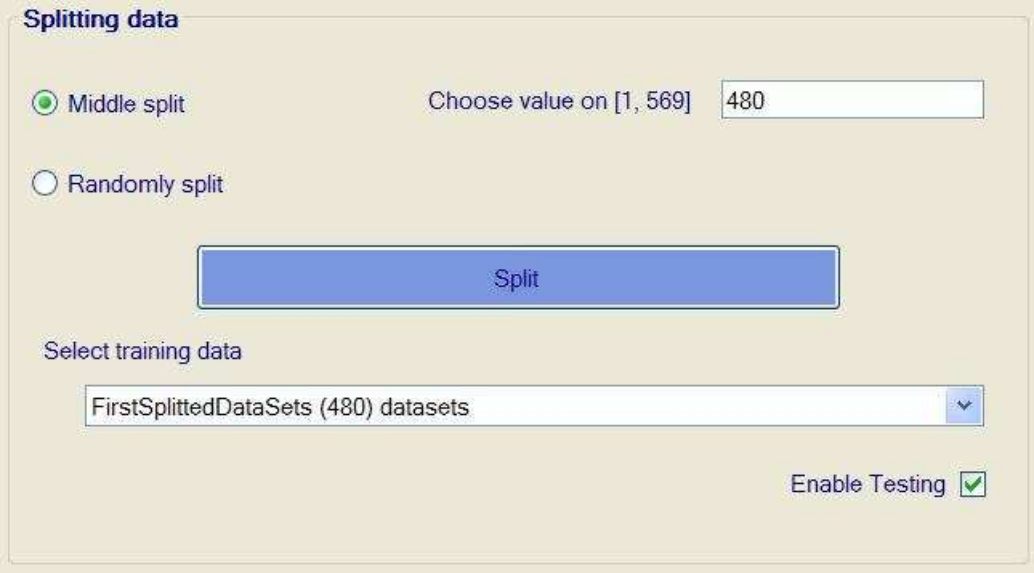


Fig. 25. Split Data for training and testing

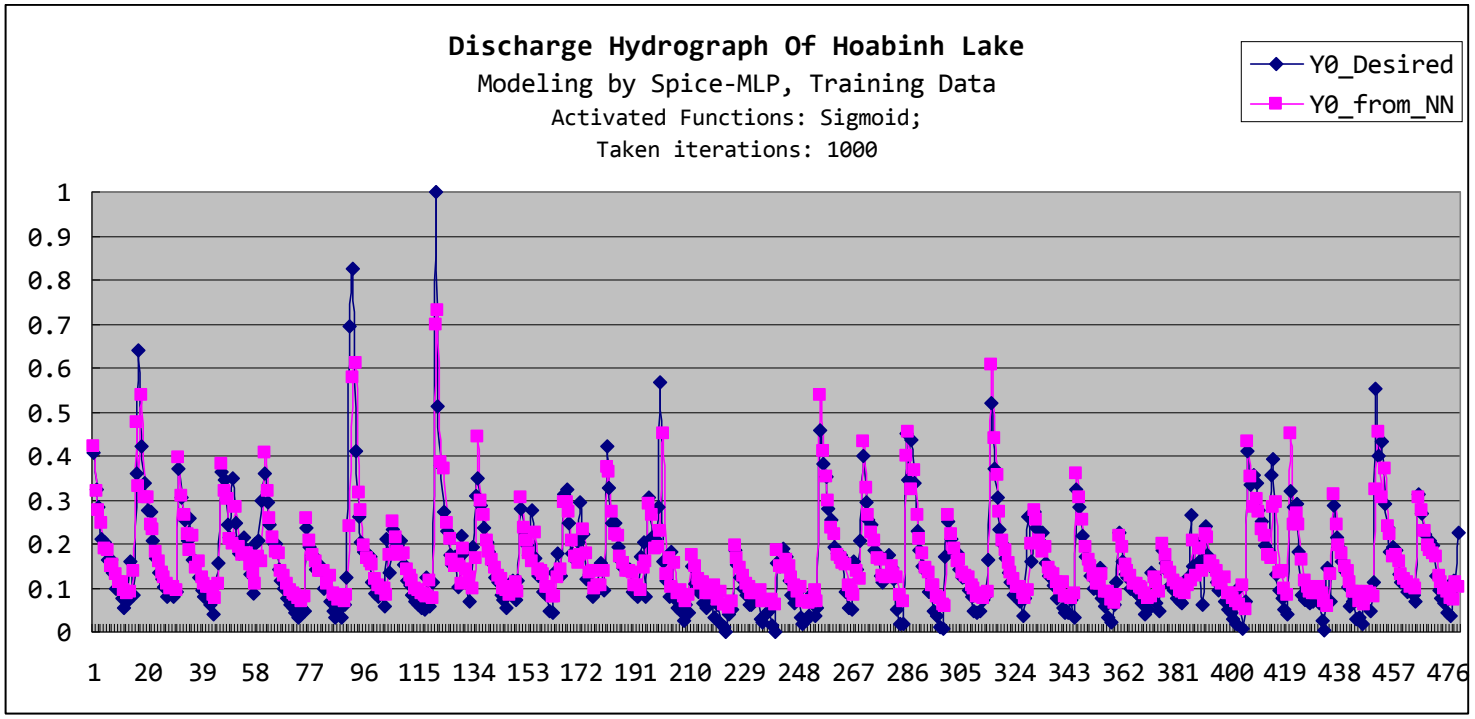


Fig. 26. Discharge Hydrograph Of Hoabinh Lake, Training Data

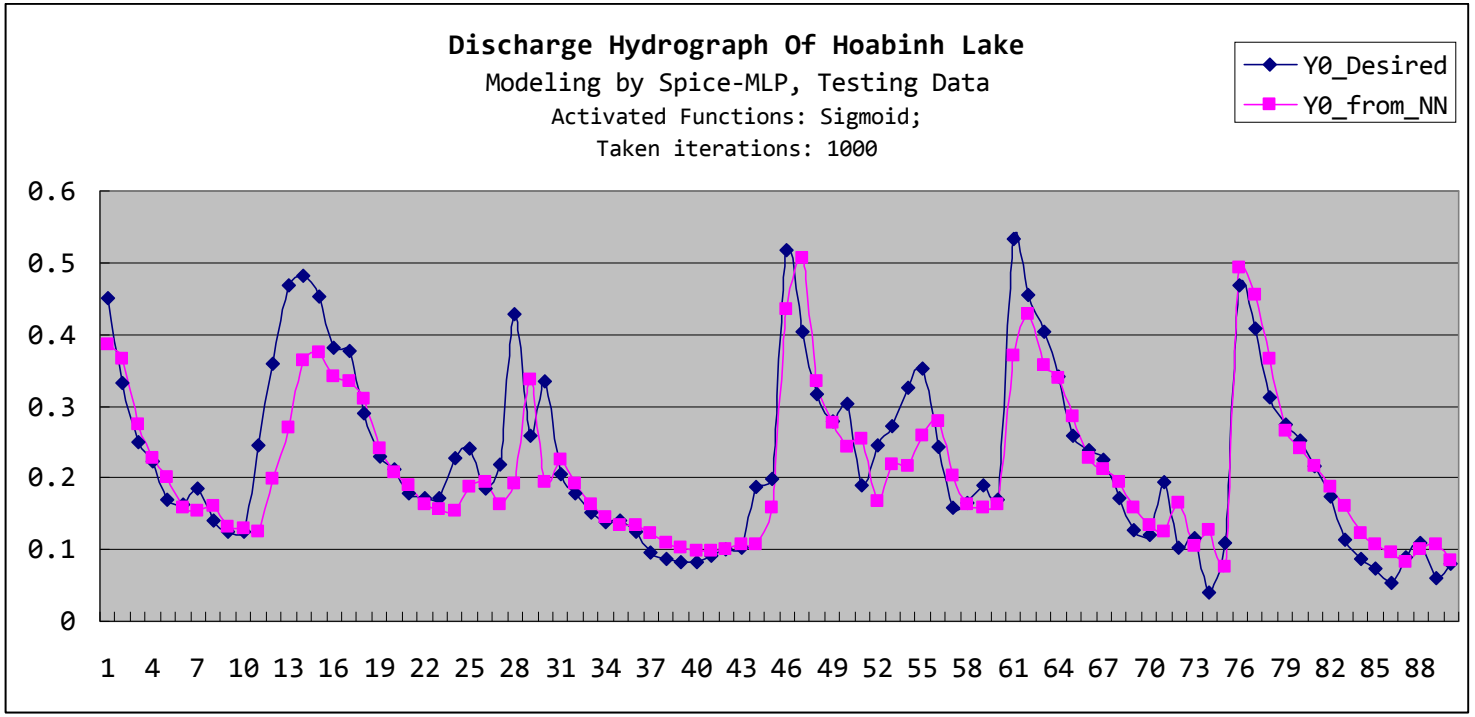


Fig. 27. Discharge Hydrograph Of Hoabinh Lake, Testing Data

Chọn hàm activated functions cho lớp ẩn và lớp ra, cho mạng học, sau một số iteration, lưu dữ liệu vào một file csv, bạn sẽ có đồ thị đầu ra của mạng với dữ liệu học như Fig.26, dữ liệu kiểm tra như Fig.27. và thông tin về mạng học như Fig.28.

Chúng ta thấy mạng học khá tốt, tuy nhiên có một số điểm đầu ra của mạng và đầu ra mong muốn có độ lệch khá lớn. Làm thế nào để làm giảm độ lệch này? xin mời bạn đọc nghiên cứu. Chúc các bạn may mắn.

```
SPICE-NEURO by Cao Thang 2004-2011
Last trained information;
 Activated Function for Hidden Layer: Sigmoid;
 Activated Function for Output Layer: Sigmoid;
 Final Learning rate: 0.03572268;
 Final MSE of Training Set: 0.003680485;
 Final MSE of Testing Set: 0.003794955;
 Number of trained data: 480;
 Number of tested data: 90;
 Taken iterations: 1000;
```

Fig. 28. Discharge Hydrograph Of Hoabinh Lake modeling by Spice-MLP, Training Information

9. Phân loại ảnh Áo dài

Hình 29 là hai output maps của 110 ảnh áo dài, bởi Spice-SOM trong hai lần học khác nhau.

Fig. 29. Vietnamese Aodai grouped by SOM in two learning

10. Kết luận

Tài liệu này hướng dẫn các bạn cách sử dụng mạng nơ ron trong các ứng dụng thực tế. Tác giả hy vọng nó có ích các bạn. Cảm ơn các bạn đã đọc. Chúc các bạn may mắn trong học tập, công việc và enjoy cuộc sống.