**DỰ BÁO CHUỖI DỮ LIỆU THỜI GIAN  
DÙNG MẠNG NƠ-RON**

**Lương Trần Ngọc Khiết,  
Lưu Cẩm Bó**(Sinh viên năm 3, Khoa Công Nghệ Thông Tin)  
***GVHD: ThS. Trần Sơn Hải***

**Tóm tắt**: Neural network là một phương pháp máy học phổ biến dựa trên mô phỏng hệ thần kinh con người trong việc mô phỏng và truyền tải thông tin, đặc biệt được áp dụng rộng rãi trong bài toán về dự báo, phân loại và điều khiển. Mục đích của đề tài này là đề xuất ra một mô hình máy học phù hợp cho bài toán dự báo giá vàng – giá cổ phiếu trong ngày tiếp theo dựa vào dữ liệu đã có. Đồng thời tìm hiểu về nguyên tắc hoạt động, giải thuật huấn luyện mạng neuron bằng phương pháp học dựa trên dữ liệu (Learning From Data) có tên gọi lan truyền ngược (backpropagation-BP) và áp dụng cho bài toán dự báo giá vàng – giá chứng khoán Yahoo với bộ dữ liệu mẫu thực tế để đánh giá độ chính xác dự báo và tính hiệu quả.

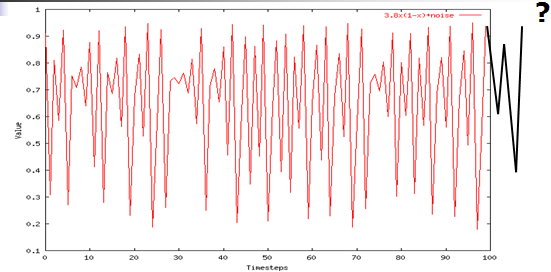
**Từ khóa**: Artificial neural networks, Time Series ForeCasting, Neural ForeCasting.

1. **Mở đầu**
   1. **Giới thiệu chung**

Hiện nay dự báo đóng vai trò rất quan trọng trong bài toán ra quyết định, lập kế hoạch của các cá nhân, tổ chức, doanh nghiệp và các quốc gia. Hầu như, trong các tất cả các lĩnh vực đều có những nghiên cứu và áp dụng dựa trên dữ liệu chuỗi thời gian. Mô hình mạng nơ-ron nhân tạo(Artificial neural networks -ANN) là một mô hình tiêu biểu trong lĩnh vực này.

Việc đề xuất một mô hình dự báo cáo có độ chính xác cao là nhu cầu tất yếu trong thực tế và hiện đang được đông đảo các nhà nghiên cứu quan tâm. Đặc biệt các mô hình dự báo thị trường kinh tế ảnh hưởng trực tiếp việc đầu tư có lợi.

* 1. **Bài toán dự báo**
* Định nghĩa bài toán dự báo
* Dự báo 1 bước: Cho tập x1, x2, x3, …, xt dự đoán giá trị xt+1.
* Dự báo n bước: TS= {X1, X2, ..., XN}. Với , trong đó là giá trị của chuỗi thời gian tại thời điểm và là độ dài của dãy . Hệ thống dự báo sẽ được cung cấp tương ứng với tập dãy kết quả truy vấn và ta sẽ cần tìm các giá trị



Hình 1 Mô phỏng bài toán

* Từ đó đưa ra những phán đoán về giá cả trong những khoảng thời gian tiếp theo các khuynh hướng tăng theo chu kỳ hay đột biến giảm xâu để lựa chọn giải pháp cho đầu tư phát triển.
  1. **Lý do chọn đề tài**
* Mạng neural đã được sử dụng nhiều trong các bài toán dự báo. Tuy nhiên để áp dụng một bài toán cụ thể ta cần xác định cụ thể cấu trúc mạng thích hợp với số lượng tầng và số nốt của mỗi tầng là một bài toán thực tế.
* Đề xuất mô hình mạng neural theo giải thuật tối ưu để dự báo các giá trị tiếp theo và đưa ra hướng quyết định mang tính lợi ích cho việc đầu tư phát triển. Đặc biệt là trong việc dự báo giá các mặt hàng được nhiều người quan tâm có sức ảnh hưởng đến thị trường kinh tế.

1. **Mục đích nghiên cứu, đối tượng nghiên cứu, phạm vi nghiên cứu, phương pháp nghiên cứu**
   1. **Mục đích nghiên cứu**

Nghiên cứu xây dựng và phát triển mô hình máy học dựa trên mạng nơ-ron và một chương trình máy tính mô phỏng thực nghiệm trên dữ liệu với nhiều cấu trúc mạng mạng khác nhau với việc dự báo giá cho một số mặt hàng trên thị trường kinh tế như vàng bạc, chứng khoán, ngoại tệ…. Từ kết quả thực nghiệm đưa kết luận mô hình phù hợp nhất.

* 1. **Đối tượng**

Đề tài lựa chọn đối tượng đầu vào để nghiên cứu gồm

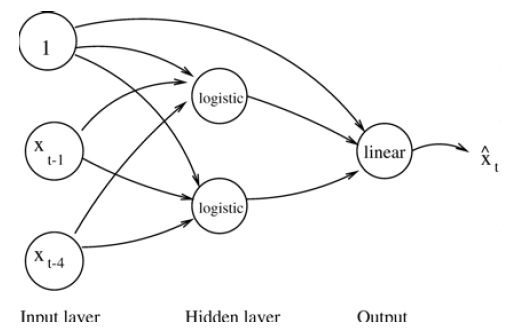
* Bài toán dự báo.
* Dữ liệu chuỗi thời gian.
* Công cụ máy học ANN.
  1. **Phạm vi nghiên cứu**

Đề xuất một phương pháp cho dự báo dữ liệu biến động giá cho thị trường vàng và chứng khoán (chủ yếu là giá cổ phiếu của yahoo) trong thời gian tới bằng mạng nơ ron nhân tạo (ANN) bằng thuật toán máy học lan truyền ngược **Backpropagation**.

* 1. **Phương pháp nghiên cứu**

Các phương pháp nghiên cứu được áp dụng nhằm xác định kết quả gồm

* Mô hình hóa đề xuất mô hình máy học dựa trên ANN cho bài toán dự báo giá vàng và giá cổ phiếu Yahoo.
* Cài đặt thực nghiệm trên máy tính
* Phân tích số liệu, đánh giá và đưa kết luận.

1. **Kết quả nghiên cứu**
   1. **Mạng nơ-ron nhân tạo**

Mạng nơ-ron nhân tạo (ANN) là một lĩnh vực nghiên cứu rất lớn trong lĩnh vực trí tuệ nhân tạo, ANN được xem như một hệ thống kết nối tập hợp các ngõ vào (inputs) đến tập hợp các ngõ ra (outputs) qua một hay nhiều lớp nơ-ron, các lớp này được gọi là các lớp ẩn.

Hình 2 Kiến trúc ANN

Đối với các ANN một lớp ẩn có H nơ-ron, phương trình tổng quát để tính giá trị dự báo (*ngõ ra*) sử dụng đến các mẫu quan sát quá khứ ,, …, làm *ngõ vào* được viết dưới dạng sau:

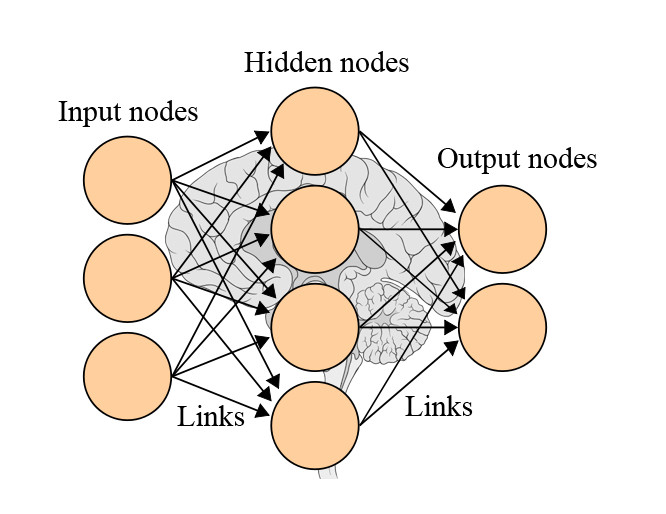
(2.15)

Trong đó:

* biểu thị các trọng số cho kết nối giữa hằng số ngõ vào và các nơ-ron lớp ẩn
* là trọng số kết nối trực tiếp giữa ngõ vào hằng số và ngõ ra
* và là các trọng số của các kết nối khác giữa các ngõ vào và các nơ-ron lớp ẩn giữa các nơ-ron lớp ẩn với ngõ ra.
* và là hai hàm kích hoạt lần lượt được sử dụng tại ngõ ra và tại các nơ-ron lớp ẩn.

Mạng neuron nhân tạo truyền thẳng với ít nhất một lớp ẩn và đủ số đơn vị cho lớp ẩn có thể xấp xỉ bất kỳ *hàm khả đánh giá* (measurable function) tuyến tính hay phi tuyến nào.

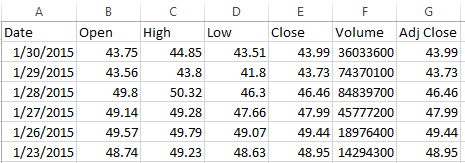
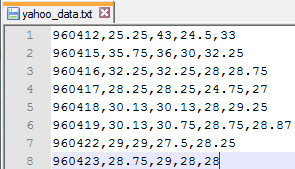
Mạng neuron học cấu hình mạng từ dữ liệu chuỗi thời gian bằng cách ánh xạ từ một vectơ dữ liệu đầu vào sang dữ liệu đầu ra. Một số lượng dữ liệu liên tiếp của dữ liệu chuỗi thời gian (cửa sổ đầu vào Xt-s, Xt-s+1, …, Xt) được ánh xạ sang khoảng thích hợp và được sử dụng như dữ liệu đầu vào của tầng nhập. Trong giai đoạn truyền tiến, những giá trị đó được truyền qua tầng ẩn rồi đến các đơn vị đầu ra. Khi truyền tới đơn vị đầu ra, giá trị lỗi được tính toán dựa vào sự khác biệt giữa giá trị đầu ra với giá trị của dữ liệu chuỗi thời gian tại thời điểm t+1. Sau đó, giá trị lỗi này được truyền ngược lại tới các kết nối giữa tầng ẩn và tầng đầu ra, kết nối giữa tầng đầu vào và tầng ẩn để cập nhập lại trọng số của các kết nối này.



Hình 3 Mô hình lan truyền ANN

Hình 4 Mô hình học với chuỗi thời gian

* 1. **Các bước thực hiện**
* Quy hoạch dữ liệu đầu vào:
* Lựa chọn thành phần chỉ số quan trọng có ảnh hưởng đến biến đang xem xét.
* Thu thập dữ liệu.
* Tiền xử lý dữ liệu

Hình 5 Dữ liệu sau xử lý (dạng file .txt) Hình 6 Dữ liệu trước xử lý (dạng file .csv)

* Phân chia tập dữ liệu thành tập huấn luyện, kiểm tra và kiểm định.
* Xây dựng cấu trúc mạng neuron :
* Xác định cấu trúc bao gồm số lớp, số neuron trong từng lớp. Thực tế mạng neuron với ***một tầng đầu vào, một tầng ẩn, một tầng đầu ra*** cùng với sự thay đổi số đơn vị tại mỗi tầng là đủ để xấp xỉ bất kì một hàm liên tục nào [7].
* Thông thường để tìm giá trị tối ưu cho số đơn vị ở lớp ẩn căn cứ vào sai số nhỏ nhất khi kiểm tra với tập dữ liệu chuẩn sau khi đã huấn luyện. Dựa vào tiêu chí trên đã xác định được số lớp của mạng neuron theo hình sau*:*

Low

High

Open

Data

Date

Input

**Hidden**

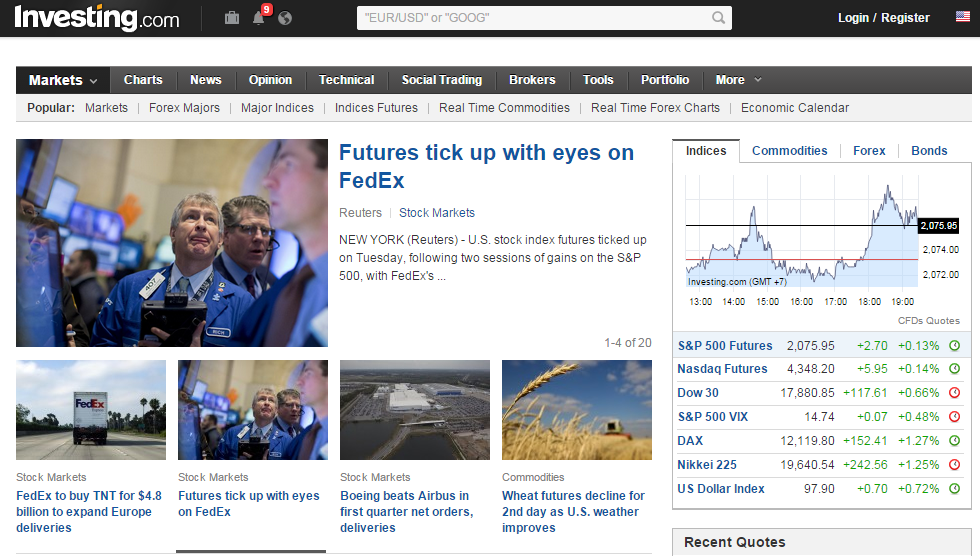
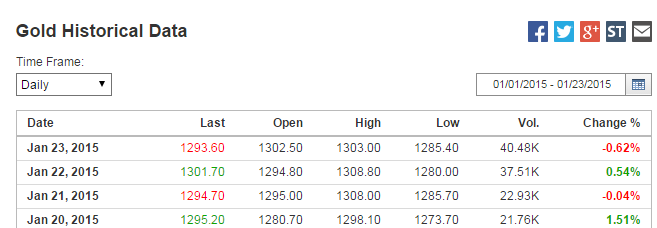
Output

pr

Hình 7 Mô hình mạng nơ-ron thực nghiệm

*Theo đề xuất* (<http://www.heatonresearch.com> ) *việc kiểm tra các cấu trúc mạng có định dạng số nút lớp ẩn giao động trong phạm vi từ 1 đến một bội số số nút đầu vào. Do thời gian hạn hẹp nên trong đề tài này với 4 nút đầu vào nhóm sẽ tiến hành thực nghiệm đến 2 lần số đầu vào (trên các sơ đồ mạng từ 4-1-1 , 4-2-1, ... 4-7-1, 4-8-1.)*

* Đánh giá khả năng xấp xỉ dùng các hàm như trung bình lỗi (MAE), tổng bình phương lỗi (sum of squared errors).
* Huấn luyện mạng.
  1. **Mô hình ANN cho bài toán dự báo**
     1. **Mô hình dự báo giá vàng**
* Quy hoạch dữ liệu đầu vào



Hình 8 Giao diện trang web Investing và phần dữ liệu hiện thị

* Xây dựng cấu trúc mạng

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Hiden Node | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 |
| Error (MAE) | 52.79 | 39.09 | 35.90 | 41.81 | 43.51 | 54.09 | 55.42 | 61.19 |
| Training Time (miliseconds) | 296.7 | 599.7 | 933.6 | 1250.3 | 1358.9 | 1606.7 | 1726.5 | 1831.9 |

Hình 9 Bảng tổng hợp trung bình lỗi mô hình mạng 4-1-1 tới 4-8-1

Hình 10 Biểu đồ tương quan MAE- TraningTime số lượng nút ẩn của bài toán giá Vàng

Hình 11 Biểu đồ tương quan MAE- LearningRate

H3

H3Pr

H2Pr

HH1

H1Pr

LH3

LH2

LH1

HH3

HH2

OH3

OH2

OH1

DH3

DH2

DH1

L

H

O

Data

D

Input

**Hidden**

Output

Pr

H1

H2

Hình 12 Mô hình mạng nơ-ron bài toán dự báo giá Vàng

*Trong đó :*

* *D – O – H – L lần lượt là giá trị ngày (Date) – giá mở cửa (Open)- giá cao nhất (High)- giá thấp nhất (Low) của data là (đã được xử lý đúng định dạng) đưa vào mạng để huấn luyện.*
* *H1- H2- H3 lần lượt là các nút mạng ở tầng ẩn.*
* *Pr là nút đầu ra, giá trị cần tìm.*
* *Các trọng số liên quan gồm*
  + *DH1, DH2, DH3 là các trọng số nối từ nút D tới các nút H1 – H2 – H3 tầng ẩn*
  + *Tương tự { OH1,OH2,OH3 }, { HH1,HH2,HH3 }, { LH1,LH2,LH3 } là các trọng số nối từ nút O tới các nút H1 – H2 – H3, H tới các nút H1 – H2 – H3, L tới các nút H1 – H2 – H3 ở tầng ẩn.*

*H1Pr, H2Pr, H3Pr là trọng số từ các nút H1- H2 – H3 tới nút Pr tầng đầu ra*

Tóm tắt việc huấn luyện mạng

Cho trước K mẫu dữ liệu

Bước 1: chọn trước giá trị

Bước 2: khởi tạo mẫu ngẫu nhiên w, bắt đầu với mẫu thứ nhất k=1 và gán sai lệch E=0

Bước 3: bắt đầu quá trình học, gán. Đầu ra của mạng neuron tính theo: Trong đó là giá trị ra của lớp ẩn, được tính bởi:

Bước 4: cập nhật trọng số neuron đầu ra: với

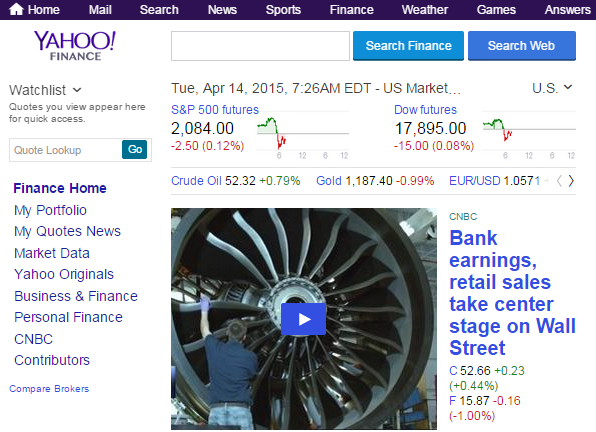
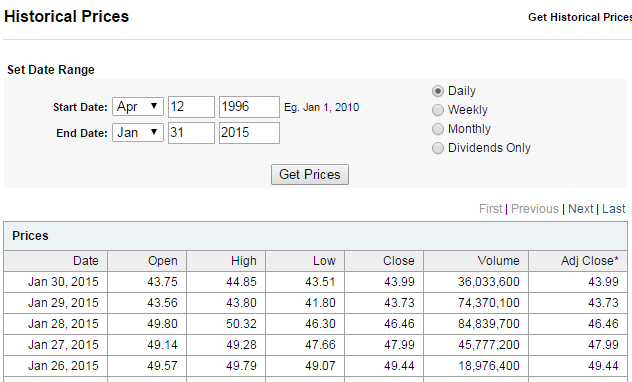
Bước 5: cập nhật trọng số neuron ẩn

Bước 6: tính sai lệch bằng cách cộng thêm sai lệch hiện tại

Bước 7: nếu k<K thì k=k+1 và trở lại bước 3. Nếu không thì qua bước 8

Bước 8: nếu thì kết thúc quá trình học. Còn nếu gán E=0 và bắt đầu một chu kỳ học mới bằng cách trở lại bước 3

* + 1. **Mô hình dự báo giá cổ phiếu Yahoo**
* Quy hoạch dữ liệu

Hình 13 Giao diện trang web http://finance.yahoo.com/ cung cấp giá cổ phiếu Yahoo

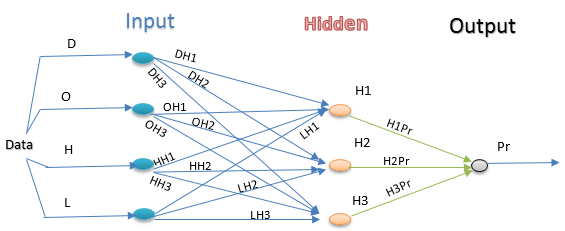
* Xây dựng cấu trúc mạng

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Hiden Node | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 |
| Error (MAE) | 13.53 | 9.11 | 4.26 | 5.19 | 7.44 | 5.39 | 5.27 | 5.98 |
| Training Time(miliseconds) | 913.70 | 3350.80 | 3587.50 | 6103.70 | 7218.60 | 8819.90 | 10374.40 | 8552.70 |

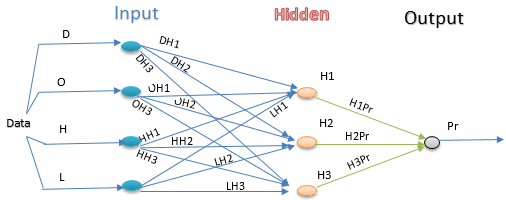
Hình 14 Bảng tổng hợp MAE-TrainingTime với số lượng nút ẩn bài toán dự báo giá cổ phiếu Yahoo

Hình 15 Biểu đồ tương quan MAE-TrainingTime với số lượng nút ẩn bài toán dự báo giá cổ phiếu Yahoo

So với biểu đồ tương quan MAE- Training Time giá Vàng thì ở giá cổ phiếu Yahoo cũng có phần giá trị lỗi giảm dần từ ở mạng 1 nút ẩn tới 3 nút ẩn và sau đó thay đổi lên xuống ở các mạng có câu trúc tếp theo từ 4 trở đi.

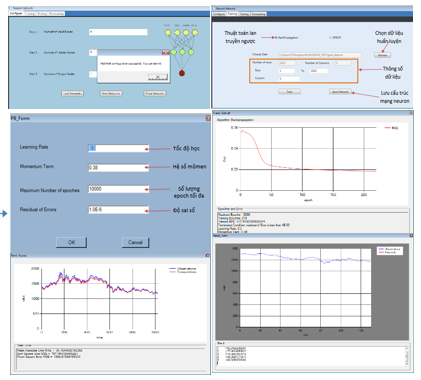
Vậy mô hình cho bài toán dự báo giá cổ phiếu Yahoo theo đề tài là 4-3-1.

1. **Kết luận và hướng phát triển**
   1. **Kết luận**

Về mặt lý thuyết: đề xuất mô hình mạng neuron theo thuật toán lan truyền ngược có cấu trúc gồm lớp Input gồm 4 nút D- O – H- L, lớp Hidden gồm 3 nút H1 –H2- H3 và một nút Pr đầu ra tầng Output.

Về mặt ứng dụng, đề tài đã xây dựng hiện thực một chương trình dự báo chuỗi thời gian bằng mạng neuron nhân tạo viết bằng ứng dụng WPF C# sử dụng Visual Studio 2013, thử nghiệm trên máy Laptop Dell Inspiron 3420 Core i5 3120M 2.50GHZ, RAM 4GB trên các bộ dữ liệu mẫu về giá vàng thế giới, giá cổ phiếu Yahoo trên sàn chứng khoán.

Hình 16 Cấu trúc mạng nơ-ron 4-3-1 cho các bài toán



Hình 17 Các màn hình chính trong chương trình

Đề tài vẫn còn hạn chế là chưa kết nối trực tiếp với các dữ liệu của các sàn chứng khoán và sàn giao dịch Vàng ở Việt Nam do hoạt động chưa công khai rõ ràng nên không có dữ liệu đầy đủ cho mô hình mạng của đề tài.

* 1. **Hướng phát triển**

Cài đặt và hoàn thiện thêm giải thuật RPROP để so sánh đối chiếu kết quả.

Đề xuất một mô hình lai (hỗn hợp) giữa các mô hình đang có (giữa RPROP và BR) để tận dụng ưu điểm, loại bỏ khuyết điểm nhằm mang lại hiệu quả cao nhất trong dự báo.

Nâng cấp tối ưu hệ số trong thuật toán

**TÀI LIỆU THAM KHẢO**

* **Nguồn thông tin tham khảo**

1. Nguyễn Đặng Kim Khánh – Dương Tuấn Anh, “Time series discord discovery using WAT algorithm and iSAX representation”.
2. Le Hoang Thai,Tran Son Hai,Nguyen Thanh Thuy, “*Image Classification using Support Vector Machine and Artificial Neural Network*”, 2012.
3. Lê Minh Trung – Nguyễn Khắc văn , “*Fuzzy Support Vector Machines*”, 2014.
4. Bishop,C. (1995) "*Neural Networks for Pattern Recognition*",   
   Oxford University Press, Oxford, UK, pp.116-149.(Thuật lan truyền ngược <http://homepages.gold.ac.uk/nikolaev/311practbp.htm> )
5. Box, G.E.P., and G.M. Jenkins (1994), *Time Series Analysis: Forecasting and Control*, Revised Edition, Holden Day, San Francisco.
6. D. Randall Wilson -In *Proceedings of the 2001 International Joint Conference on Neural Networks (IJCNN’01)*, 115-119.

(<http://axon.cs.byu.edu/papers/wilson.ijcnn2001.pdf> xác định learingrate)

1. E.A.Wan (1993),*Time Series Prediction by Using a Connectionist Network with Internal Delay Line*, *pages 195-217. Addison Wesley*.

* **Hỗ trợ vẽ đồ họa**:

1. <http://dynamicdatadisplay.codeplex.com/>
2. <http://mathnetnumerics.codeplex.com/>.

**Giáo viên hướng dẫn:** ThS. Trần Sơn Hải  
(Số điện thoại: 01263891175 - Email: [haits@hcmup.edu.vn](mailto:haits@hcmup.edu.vn) )

**Nhóm sinh viên thực hiện:**

* Lương Trần Ngọc Khiết

(Số điện thoại: 01687469465 - Email: [khietltn@gmail.com](mailto:khietltn@gmail.com) )

* Lưu Cẩm Bó

(Số điện thoại: 090.8299693 - Email: [luucambo@yahoo.com.vn](mailto:luucambo@yahoo.com.vn))