Dự báo tỉ giá ngoại tệ với mô hình học cộng đồng kết hợp giải thuật tiến hóa đa mục tiêu

Forecasting of Currency Exchange Rates with Multi-Objective Evolutionary Ensemble Learning

Đinh Thị Thu Hương, Đỗ Thị Diệu My, Vũ Văn Trường, Bùi Thu Lâm

Abstract: *Time series forecasting is paid a considerable attention of the researchers. At present, in the field of machine learning, there are a lot of studies using artificial neural networks to construct the model of time series forecast in general, and foreign currency exchange rates forecast, in particular. However, determining the number of members of an ensemble is still debatable. This paper proposes the way of constructing a model and designing a multi-objective evolutionary algorithm in training neural networks ensembles in order to increase the diversity of the population. Two objectives of the selected model include: Mean Sum of Squared Errors - MSE and Diversity. We experimented the model on four data sets and compared three methods (single-objective, multi-objective and ensembles). The experimental results showed that the proposed model produced better in investigated cases.*

Keywords: *Ensemble learning, multi-objective evolutionary algorithm, diversity.*

I. GIỚI THIỆU

Các nhà khoa học đã minh chứng, trong lĩnh vực dự báo thì mô hình học cộng đồng đem lại hiệu suất tốt hơn mô hình đơn [21]. Vì vậy, xây dựng mô hình học cộng đồng là cách làm phổ biến để cải thiện hiệu suất của loại bài toán phân lớp và hồi quy. Thông thường một mô hình cộng đồng dựa trên những mô hình đơn, chẳng hạn: mạng nơron [3,15,16,21], máy vector hỗ trợ [9] hoặc cây hồi qui [14].

Những giải thuật tiến hóa đa mục tiêu đã được các nhà khoa học ứng dụng trong dự báo chuỗi thời gian. Có ba lí do khiến tối ưu tiến hóa (Evolutionary Optimiztion – EO) được quan tâm nhiều: (i) EOs không yêu cầu bất kì thông tin đặc trưng; (ii) EOs thực hiện tương đối đơn giản; (iii) EOs là cách tiếp cận đa điểm (làm việc trên quần thể) thể hiện rõ tính linh hoạt và miền giới hạn rộng để áp dụng [4],[12].

Mô hình học cộng đồng (ensemble learning) kết hợp giải thuật tiến hóa đa mục tiêu dùng mạng nơron nhân tạo (Artificial Neural Network - ANN) để huấn luyện được triển khai nhiều trong lĩnh vực dự báo chuỗi thời gian. Chẳng hạn: [13] thực hiện dự báo chuỗi thời gian với mô hình cộng đồng trên cơ sở các mô hình đơn để xây dựng dự báo lặp nhằm tìm ra được số các thành viên cộng đồng góp phần nâng cao hiệu suất dự báo; Trong [9] đề xuất dự báo chuỗi thời gian bằng giải thuật lai tiến hóa đa mục tiêu để tối ưu cấu trúc của mạng RNNs dựa trên hai mục tiêu: mục tiêu thứ nhất là các cá thể dưới một ngưỡng trên biên Pareto và mục tiêu thứ hai là dựa trên lỗi huấn luyện; [20] minh chứng việc cân bằng giữa độ đa dạng giữa các thành viên của cộng đồng và tính chính xác (đó là hai yêu cầu quan trọng để xây dựng phương pháp học cộng đồng dựa trên giải thuật tiến hóa đa mục tiêu.

Nhìn chung các nghiên cứu trước chủ yếu tập trung vào việc xem xét số lượng thành viên cộng đồng hay độ đa dạng (Diversity - DIV) của các giải pháp [18] hoặc cách chọn các hàm mục tiêu trong bài toán phân lớp mà ít đề cập tới bài toán hồi quy. Trong bài báo

này, nhóm tác giả đề xuất một mô hình học cộng đồng kết hợp giải thuật tiến hóa đa mục tiêu để tối ưu bộ trọng số của mạng nhằm nâng cao hiệu suất dự báo của bài toán hồi qui. Hai mục tiêu được lựa chọn là: MSE và DIV. Về dữ liệu, chúng tôi thử nghiệm trên 4 tập dữ liệu (HKD, JPY, EURO và USD) với khoảng thời gian 19/10/2012 đến 06/04/2015. Kết quả thử nghiệm, được so sánh với dự báo dùng mô hình ANN. Mô hình đề xuất khẳng định ưu thế của kỹ thuật học cộng đồng kết hợp với giải thuật tiến hóa.

Nội dung bài báo bao gồm: phần II trình bày lý thuyết nền tảng, phương pháp đề xuất giải quyết bài toán dự báo tỉ giá ngoại tệ được trình bày trong phần III, phần IV là kết quả thử nghiệm và cuối cùng là kết luận.

II. LÝ THUYẾT NỀN TẢNG

II.1. Dự báo chuỗi thời gian

II.1.1. Khái niệm

Một chuỗi thời gian được hiểu là một dãy rời rạc các giá trị quan sát tại các khoảng thời gian cách đều nhau $Y=\{y_1,y_2,...,y_t\}$ được xếp thứ tự diễn biến thời gian với y_1 là các giá trị quan sát tại thời điểm đầu tiên, y_2 là quan sát tại thời điểm thứ 2 và y_t là quan sát tại thời điểm thứ t [1],[2].

II.1.2. Phân tích chuỗi thời gian

Để nhận thấy sự biến động của hiện tượng qua thời gian, cần phải phân tích chuỗi thời gian. Có thể kể đến các yếu tố là nguồn gốc tạo ra đặc tính dao động, đó là: tính xu hướng, tính mùa, tính chu kì và tính ngẫu nhiên [5].

Tính xu hướng (trend): Thể hiện chiều hướng biến động, tăng hoặc giảm của hiện tượng nghiên cứu trong một thời gian dài.

Tính mùa (seasonality): Biểu hiện qua sự tăng/giảm mức độ của hiện tượng ở một số thời điểm (tháng/quí) nào đó được lặp đi lặp lại qua nhiều năm. Điều được quan tâm là kích thước của ảnh hưởng mùa: Kích thước ảnh hưởng mùa đều đặn hàng năm thì được gọi là có tính cộng (additive), còn kích thước của ảnh hưởng mùa tỉ lệ với giá trị trung bình được gọi là có tính nhân (multiplicative).

Tính chu kì (cyclical): Biến động của hiện tượng được lặp lại với thời vụ mà khoảng thời gian lớn hơn 1 năm.

Tính ngẫu nhiên (irregular): Biến động không có quy luật và hầu như không thể dự đoán được.

Những nhân tố này đã làm cho chuỗi thời gian trở nên không dừng. Bốn thành phần trên có thể kết hợp với nhau theo mô hình nhân [2]:

$$y_t = T_t.S_t.C_t.I_t \quad (1)$$

Trong đó:

y_t : giá trị quan sát ở thời điểm t.

T_t : thành phần xu hướng ở thời gian t.

S_t : thành phần thời vụ ở thời gian t.

C_t : thành phần chu kì ở thời gian t.

I_t : thành phần ngẫu nhiên ở thời gian t.

Khi xem xét biến động các thành phần của chuỗi thời gian. Trước tiên, xét đến biến động thời vụ. Với đặc tính của phương pháp số trung bình di động có tác dụng hạn chế, loại bỏ các biến động mang tính ngẫu nhiên. Vì vậy, sau khi tính số trung bình di động từ một nhóm với mức độ (chiều dài của nhóm – thường chọn giá trị mức độ chính là giá trị thời vụ) nào đó của dãy thì chuỗi số thời gian chỉ bao hàm hai thành phần xu hướng và chu kì, vì hai thành phần mùa vụ và ngẫu nhiên đã bị loại bỏ. Khi đó:

$$SI = \frac{TCSI}{TC} = \frac{y_t}{\bar{y}_t} \quad (2)$$

trong đó: $\bar{y}_t$ là số trung bình di động ứng với giá trị quan sát ở thời điểm t.

Tiếp theo, xu hướng của chuỗi thời gian có tính chất mùa vụ thì cần loại bỏ yếu tố mùa vụ ra khỏi dãy số. Tức là:

$$CTI = \frac{TCSI}{S} = \frac{y_t}{I_s} \quad (3)$$

trong đó: I_s là chỉ số thời vụ.

Sau khi có được các yếu tố thành phần T, S, C ta có thể xác định biến động của yếu tố ngẫu nhiên bằng cách tính:

$$I_t = \frac{y_t}{TI_sI_c} \quad (4)$$

trong đó: I_c là chỉ số chu kì.

Ngoài những đặc tính trên, chuỗi thời gian còn có các đặc tính khác như: tuyến tính (linearity) và tính dừng (stationary). Tuy nhiên, rất khó gặp được một chuỗi thời gian "dừng" theo đúng nghĩa trong thực tế. Thông thường, người ta thường dùng các phép sai phân để chuyển chuỗi thời gian không dừng về chuỗi dừng [17]. Khi phân tích chuỗi thời gian, một tham số thống kê được quan tâm đó là hệ số tương quan. Với một chuỗi thời gian *x*, giả sử giá trị trung bình không thay đổi, ta có thể xấp xỉ hệ số tương quan giữa x_t và x_{t+k} như sau [1]:

$$r_k = \frac{\sum_{t=1}^{N-k}(x_t - \bar{x})(x_{t+k} - \bar{x})}{\sum_{t=1}^{N}(x_t - \bar{x})^2} \quad (5)$$

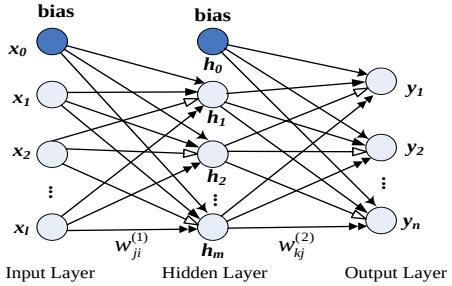
trong đó: r_k là hệ số tương quan tại độ trễ k và $\bar{x}$ giá trị trung bình của x. Mục tiêu chính của phân tích chuỗi thời gian là nhận ra và tách riêng các yếu tố ảnh hưởng để thực hiện bước dự đoán.

II.2. Chỉ số đánh giá lỗi

Về cơ bản, một mô hình dự báo chuỗi thời gian là dựa vào dữ liệu có sẵn trong quá khứ để dự đoán dữ liệu trong tương lai bằng cách khái quát hóa từ dữ liệu quá khứ để xây dựng mô hình. Tuy nhiên, không có một mô hình nào là tuyệt đối, sẽ luôn luôn có một sai số giữa giá trị thực và giá trị dự đoán. Biểu diễn dạng công thức toán học sai số đó như sau: $e_t = x_t - \hat{x}_t$, trong đó: $\hat{x}_t$ là giá trị dự đoán ở thời điểm thứ t. Có một số chỉ số đo độ lỗi thường hay dùng trong dự báo chuỗi thời gian như: SSE (Sum Squared Error), RMSE (the Root Mean Squared Error), MSE (Mean Sum of squared Errors) và MAPE (Mean Absolute Percentage Error).

II.3. Mạng nơron

Theo tài liệu [7] mạng nơron là một công cụ tính toán phổ biến trong lĩnh vực trí tuệ nhân tạo. Cấu trúc mạng bao gồm một tập các đơn vị tính toán (hay các nút) và được chia thành nhiều lớp (mỗi lớp có nhiều nút), (xem hình 1). Mức độ liên kết giữa các nút được xác định bởi một tập giá trị trọng số. Tham số bias được sử dụng để tăng độ thích nghi của mạng với bài toán đặt ra. Số lớp và các nút trong mỗi lớp phụ thuộc vào từng bài toán và được xác định bằng thử nghiệm. Số lượng nút của lớp ra bằng số biến của vector lời giải.



Hình 1. *Cấu trúc mạng nơron nhiều lớp.*

Theo [7] quá trình huấn luyện mạng nơron cần xác định hai thành phần là kiến trúc mạng và bộ trọng số liên kết. Việc xác định kiến trúc mạng thường dùng trí tuệ chuyên gia (xác định trước). Trong khi đó, xác định giá trị bộ trọng số người ta dùng là giải thuật lan truyền ngược (Back Propagation-BP). Bản chất của BP là giải thuật tìm kiếm có sử dụng kĩ thuật tìm kiếm ngược hướng gradient. Mặc dù dễ thực thi nhưng giải thuật này bộc lộ một số nhược điểm như: (1) Tại các vùng hoặc một số hướng mà bề mặt sai số bằng phẳng, các giá trị gradient nhỏ dẫn đến tốc độ hội tụ của giải thuật chậm; (2) Bề mặt sai số thường không lồi mà có nhiều vùng lõm khác nhau, nó phụ thuộc vào quá trình khởi tạo các trọng số ban đầu của mạng, điều đó dẫn đến giải thuật có thể bị tắc tại các cực trị địa phương (tắc tại các vùng lõm). Tuy nhiên, BP lại có khả năng hội tụ nên trong nghiên cứu này nhóm tác giả sử dụng BP để tinh chỉnh bộ trọng số của các mạng nơron tốt

nhất sau khi được huấn luyện qua thuật toán NSGA-II.

II.4. Bài toán tối ưu đa mục tiêu

Bài toán tối ưu đa mục tiêu có k hàm mục tiêu $\left(f(\vec{x}) = f_1(\vec{x}), f_2(\vec{x}),..., f_k(\vec{x})^T\right)$ [10] được định nghĩa như sau:

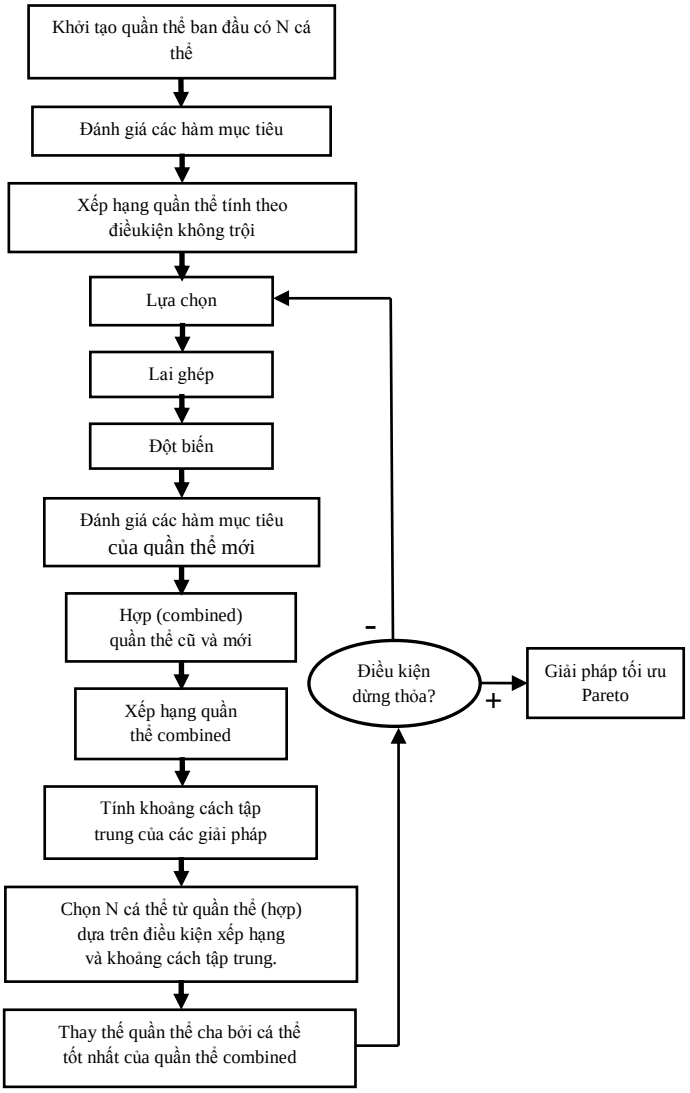
$$\begin{cases} \text{minimize } \left(f(\vec{x}) = f_1(\vec{x}), f_2(\vec{x}),..., f_k(\vec{x})^T\right) \\ g_j(\vec{x}) \geq 0 \quad (j = 1,...,m) \\ \vec{x} \in X \subset R^n \end{cases} \tag{6}$$

trong đó: $\vec{x}$ là một vector quyết định *n* biến ($\vec{x} = (x_1, x_2,..., x_2)^T$); $\left(f(\vec{x})\right)$ biểu diễn mối quan hệ giữa tiêu chuẩn chất lượng của quá trình khảo sát và các biến độc lập $\vec{x}$ và gọi là hàm mục tiêu; $\left(g(\vec{x})\right)$ là điều kiện ràng buộc của bài toán. Trong không gian các biến, hàm số này tạo ra miền giới hạn D các khả năng cho phép của hàm $\left(f(\vec{x})\right)$.

II.5. Mô hình học cộng đồng kết hợp giải thuật tiến hóa đa mục tiêu

II.5.1. Thuật toán NSGA-II

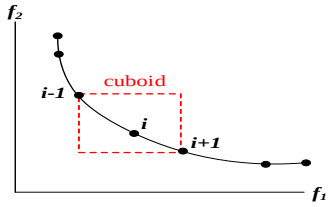
NSGA-II là thuật toán di truyền sắp xếp các lời giải không trội. Đó là một phiên bản cải tiến từ NSGA được đề xuất bởi Deb và Agarwal [10], [11]. Bản chất của thuật toán này là xem mỗi lời giải phải xác định có bao nhiêu lời giải trội hơn và tập các lời giải trội hơn đó. NSGA-II sẽ ước tính mật độ của các lời giải xung quanh một lời giải cụ thể trong quần thể bằng cách tính khoảng cách trung bình giữa hai điểm theo mỗi mục tiêu của bài toán. Giá trị này gọi là khoảng cách tập trung (crowding distance). Trong suốt quá trình lựa chọn, nó xem xét cả hai yếu tố: xếp hạng độ trội và tính khoảng cách tập trung. Những lời giải thỏa mãn sẽ thuộc biên lớp thứ nhất. Nói cách khác, đó là một thuật toán lặp nhằm giải quyết các bài toán tìm kiếm. Đặc biệt tìm kiếm ở đây được tiến hành song song trên một quần thể chứ không tìm kiếm trên một điểm. Mặt khác nhờ áp dụng các toán tử di truyền sẽ có trao đổi thông tin giữa các điểm, như vậy sẽ giảm bớt khả năng kết thúc tại một cực trị địa phương mà tìm thấy cực tiểu toàn cục. NSGA-II làm việc trên quần thể (population) gồm các phép toán lai ghép (crossover), đột biến (mutation) và lựa chọn (selection). Trong đó, việc xếp hạng quần thể con và chọn cá thể từ quần thể quyết định sự hình thành tập quần thể mới. Các bước của thuật toán được biểu diễn qua sơ đồ khối (xem Hình 2).



Hình 2. *Sơ đồ khối của thuật toán NSGA-II.*

II.5.2. Khái niệm khoảng cách tập trung

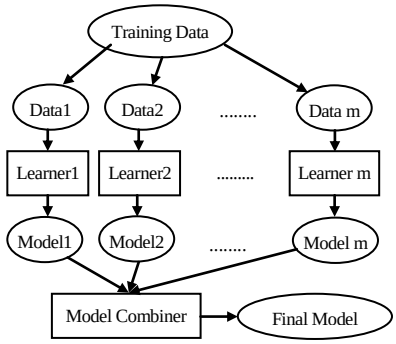
Khoảng cách tập trung của một giải pháp nằm trên một biên (front) là độ dài trung bình các cạnh của một hình hộp (cuboid). Để có thể ước tính mật độ các giải pháp xung quanh một giải pháp *i* trong quần thể, ta tính khoảng cách trung bình của hai giải pháp *i-1* và *i+1* theo từng mục tiêu (xem Hình 3).



Hình 3. *Minh họa khoảng cách tập trung quanh giải pháp i.*

II.5.3. Khái niệm học cộng đồng

Học cộng đồng là một mô hình máy học mà nhiều học viên được huấn luyện để giải quyết cùng một bài toán (phân loại hoặc hồi quy), ngược lại phương pháp tiếp cận học máy thông thường - cố gắng tìm hiểu một giả thuyết từ dữ liệu huấn luyện - phương pháp học cộng đồng cố gắng xây dựng một tập hợp các giả thuyết và kết hợp chúng để sử dụng [22] (Hình 4)



Hình 4. *Mô hình học cộng đồng.*

II.5.4. Động lực thúc đẩy sử dụng học cộng đồng

Krogh và Vedelsby [20] đã chứng minh rằng, tại một điểm dữ liệu duy nhất, bình phương lỗi của dự báo cộng đồng luôn nhỏ hơn hoặc bằng trung bình bình phương lỗi của dự báo thành phần.

$$(f_{ens}-d)^2=\sum_i w_i(f_i-d)^2-\sum_i w_i(f_i-f_{ens})^2 \qquad (7)$$

Trong đó: $f_{ens}=\sum_i w_i f_i$ mà f_{ens} là đầu ra của dự báo cộng đồng, f_i là đầu ra dự báo thứ i, w_i là trọng số của dự báo thứ i mà $\sum_i w_i=1$, còn d là giá trị thực.

Mặt khác, cũng theo [20] có thể biểu diễn như dưới đây:

$$\begin{aligned}\sum_i w_i(f_i-d)^2&=\sum_i w_i(f_i-f_{ens}+f_{ens}-d)^2\\&=\sum_i w_i[(f_i-f_{ens})^2+(f_{ens}-d)^2+2(f_i-f_{ens})(f_{ens}-d)]\\&=\sum_i w_i(f_i-f_{ens})^2+\sum_i w_i(f_{ens}-d)^2+2\sum_i w_i(f_i-f_{ens})(f_{ens}-d)\\&=\sum_i w_i(f_i-f_{ens})^2+(f_{ens}-d)^2+2(f_{ens}-d)\sum_i w_i(f_i-f_{ens})\end{aligned} \qquad (8)$$

Mà giá trị của biểu thức

$2(f_{ens}-d)\sum_i w_i(f_i-f_{ens})\to 0$ khi $\sum_i w_i=1$

Khi đó, công thức (8) chính là công thức (7). Điều này cho thấy đây chính là động lực khiến các nhà nghiên cứu đóng góp những đề xuất để tạo ra những mô hình cộng đồng nhằm tăng sự đa dạng của quần thể. Đó là vấn đề mà bài báo hướng tới.

III. XÂY DỰNG MÔ HÌNH DỰ BÁO TỈ GIÁ

Một trong những thách thức của bài toán dự báo là dữ liệu không đầy đủ và chứa nhiễu. Kỹ thuật dùng ANN trong dự báo có nhiều ưu điểm nhưng tồn tại lớn nhất khi dùng giải thuật BP là việc khởi tạo bộ trọng số ngẫu nhiên của mạng dễ xảy ra hiện tượng cực trị địa phương.

Như chúng ta đã biết, giải thuật tiến hóa thực hiện tìm kiếm song song nên giảm được hiện tượng cực trị địa phương hướng tới cực trị toàn cục. Xuất phát từ những lí do trên, ý tưởng của tác giả sẽ tiến hóa mạng nơron và xem xét thêm mục tiêu (gọi là phương pháp tiếp cận đa mục tiêu) để kiểm soát mức độ khái quát, cân bằng tính chính xác và khái quát của mạng nơron. Khi đó, ta có được bài toán tối ưu hai mục tiêu. Trước tiên, cần phải xác định các mục tiêu của bài toán.

III.1. Xác định hàm mục tiêu

Với bài toán dự báo thì hiệu suất của dự báo được đánh giá thông qua giá trị lỗi. Mặt khác, khi dùng giải thuật NSGA-II thì một yếu tố rất quan trọng thường được xem xét tới, đó là độ đa dạng[1] của cá thể. Trong nghiên cứu này, hai hàm mục tiêu được chọn là: Trung bình tổng bình phương lỗi và độ đa dạng (Diversity-DIV).

Mục tiêu 1- MSE: MSE là sai số bình phương trung bình được tính theo công thức (9):

[1] Nếu hai mạng nơron tạo ra các lỗi khác nhau trên cùng tập dữ liệu đầu vào (input) thì được gọi là sự đa dạng [8]

$$MSE = \frac{1}{n}\Sigma_{i=1}^{n}\left(x_{thuc} - x_{dubao}\right)^2 \quad (9)$$

Trong đó: n là số giá trị dữ liệu; x_{thuc} : dữ liệu thực; x_{dubao} : dữ liệu dự báo.

x_{dubao} là biến chứa dữ liệu dự báo được tính trung bình từ M bộ dữ liệu dự báo (M là số cá thể được lựa chọn từ thuật toán NSGA-II để tham gia quá trình tinh chỉnh).

Mục tiêu 2-DIV: DIV là độ đa dạng của các cá thể trong cộng đồng. Để tránh được hiện tượng cực trị địa phương, các cá thể nên được phân bố đều trên bề mặt các lớp biên. Trong nghiên cứu này, với mỗi cá thể *i* thì độ đa dạng DIV được tính theo ba cách đo khác nhau:

Khoảng cách tập trung

$$DIV = \frac{\left(MSE_{i+1} - MSE_{i-1}\right)}{\left(MSE_{\max} - MSE_{\min}\right)} \quad (10)$$

Trong đó: MSE_{i+1} và MSE_{i-1} là giá trị của hàm mục tiêu thứ nhất của hai cá thể thứ *(i+1)* và *(i-1)*; $MSE_{\max}$ và $MSE_{\min}$ là giá trị lớn nhất và nhỏ nhất của hàm mục tiêu thứ nhất của tập các cá thể trên cùng một lớp biên với cá thể *i*.

Khoảng cách tới quần thể

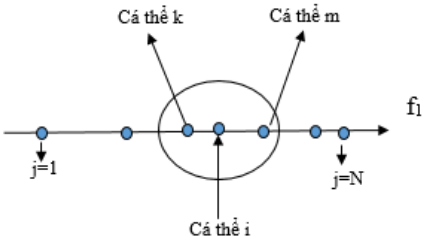
$$DIV = \frac{1}{(N-1)}\sum_{j=1, j\neq i}^{N} d(i,j) \quad (11)$$

Trong đó: $d(i,j) > 0$ là khoảng cách giữa hai cá thể *i* và *j* *và* tính theo công thức: $d(i,j) = \left|MSE_i - MSE_j\right|$

Như vậy DIV của cá thể *i* được tính bằng khoảng cách trung bình của *i* với tất cả các cá thể trong quần thể.

Khoảng cách đường tròn

Khi đó: DIV của cá thể *i* được tính bằng khoảng cách trung bình giữa các cá thể nằm trong đường tròn (trên Hình 6 là hai cá thể *k, m* nằm trong hình tròn). Đường tròn được xác định với tâm là cá thể *i*, bán kính đường tròn (R) là khoảng cách lớn nhất giữa 2 giá trị *MSE* kề nhau (để đảm bảo luôn có ít nhất hai cá thể rơi vào đường tròn này).



Hình 6. *Minh họa cách tính khoảng cách đường tròn* DIV_i **.**

III.2. Mô hình toán học của bài toán dự báo

Bài toán dự báo tỉ giá được biểu diễn như sau:

Input: Chuỗi tỉ giá $(X_t : t \in T)$ trong khoảng thời gian t (đơn vị tính là ngày)

Output: Xây dựng mô hình dự báo thỏa mãn:

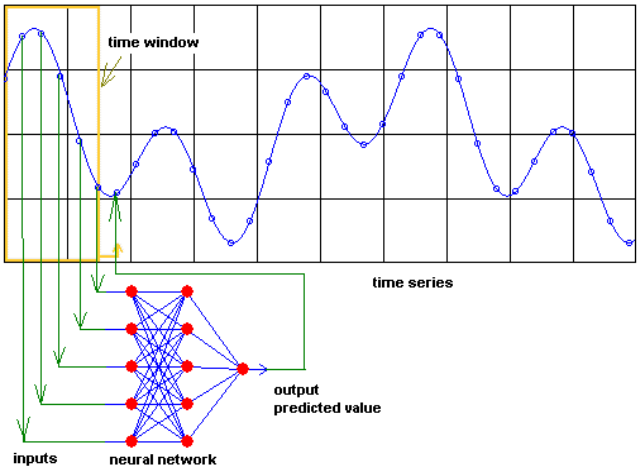
$$\begin{cases} f_1 = MSE \\ f_2 = DIV \end{cases}$$
$$\min\{f_1\}; \max\{f_2\}$$

Trong đó: *MSE* tính theo công thức (9); DIV tính theo công thức (10).

III.3. Ý tưởng

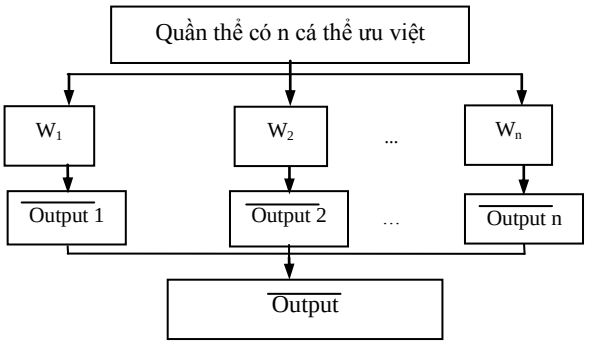
Trong nghiên cứu này, tác giả đã sử dụng một cửa sổ trượt ANN có 5 giá trị thực làm đầu vào, giải thuật BP huấn luyện mạng sẽ cho ra bộ trọng số và tính được giá trị đầu ra (chính là giá trị dự đoán của điểm dữ liệu tiếp theo) thể hiện trong Hình 7.

Tuy nhiên, khi thực hiện giải thuật BP thì bước khởi tạo bộ trọng số ngẫu nhiên ban đầu làm giá trị sai số tìm được dễ rơi vào giá trị cực trị địa phương. Vì vậy, nghiên cứu này sẽ thiết kế mô hình cộng đồng với giải thuật tiến hóa NSGA-II để tối ưu bộ trọng số của mạng. Cụ thể, thiết lập mỗi cá thể trong quần thể là một bộ gồm K bộ trọng số của K mạng nơron (hay có thể nói quần thể chứa các mạng nơron có cấu trúc nơron đầu vào, nơron ẩn và nơron đầu ra giống nhau, nhưng khác nhau bộ trọng số của mạng).



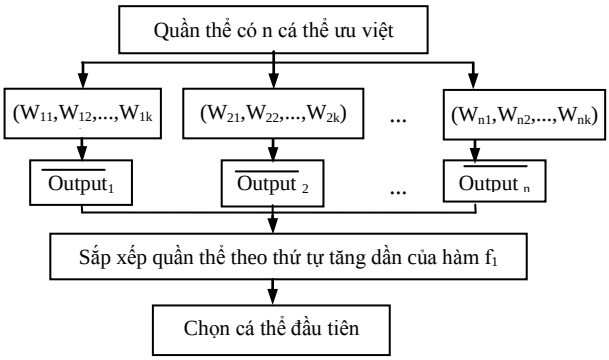
Hình 7. *Minh họa cửa sổ trượt ANN trên tập dữ liệu tỉ giá.*

Việc thiết lập này, có 2 mô hình triển khai biểu diễn trong Hình 8 và Hình 9.



$$\overline{Output} = \frac{1}{n}\sum_{i=1}^{n} Ouput_i$$

Hình 8. *Mô hình 1: mỗi cá thể là một bộ trọng số của mạng ANN (K=1).*



Hình 9. *Mô hình 2: mỗi cá thể là một bộ gồm K bộ trọng số của mạng ANN.*

Mô hình 1 là mô hình đơn giản đầu tiên khi mỗi cá thể là một mạng ANN. Ban đầu các cá thể này sẽ được gán giá trị ngẫu nhiên, sau đó qua quá trình tiến hóa, chọn lọc tự nhiên những cá thể nào đáp ứng được cả hai hàm mục tiêu thì sẽ được giữ lại, đến thế hệ cuối cùng ta giữ lại n cá thể tốt nhất nằm trên lớp biên đầu tiên của quần thể (hay n mạng nơron có sai số nhỏ nhất để ra quyết định). Lúc này, kết quả đầu ra của giải thuật NSGA-II được làm đầu vào của giải thuật BP (tức là các giá trị trọng số ban đầu không cần phải khởi tạo ngẫu nhiên). Từ bộ trọng số này, sẽ có một hệ ra quyết định với n máy ra quyết định. Tiếp tục, lấy giá trị trung bình đầu ra của n máy sẽ có được giá trị đầu ra (giá trị dự đoán của điểm tiếp theo)

Mô hình 2 là mô hình học cộng đồng với mỗi cá thể là tập K mạng nơron trong mô hình 1. Quá trình tính toán cũng giống với mô hình 1, chỉ khác biệt ở chỗ: việc đánh giá các cá thể trong quần thể dựa trên tiêu chí "làm việc theo nhóm" tức là tại mỗi thế hệ tiến hóa, từng nhóm K cá thể con sẽ được lựa chọn, sinh sản, đột biến và những nhóm cá thể nào đáp ứng tốt tất cả các hàm mục tiêu thì sẽ được ưu tiên giữ lại để tham gia vào quá trình chọn lọc tiếp theo. Cứ như vậy đến thế hệ cuối cùng, nhóm gồm K cá thể con trên lớp biên đầu tiên mà giá trị trung bình của MSE nhỏ nhất sẽ được lựa chọn.

III.4. Mô hình học cộng đồng kết hợp giải thuật NSGA-II

Giải thuật được thiết kế thể hiện trong sơ đồ khối (Hình 10). Mô tả chi tiết các bước của giải thuật gồm:

Bước 1: *Khởi tạo quần thể (kích thước quần thể là N, kích thước nhóm cá thể là K)*

- Duyệt qua tất cả các cá thể trong quần thể
  + Tính hàm mục tiêu f_1
- Duyệt qua tất cả các cá thể trong quần thể
  + Tính hàm mục tiêu f_2

Bước 2: *Vòng lặp quá trình tiến hóa qua các thế hệ*

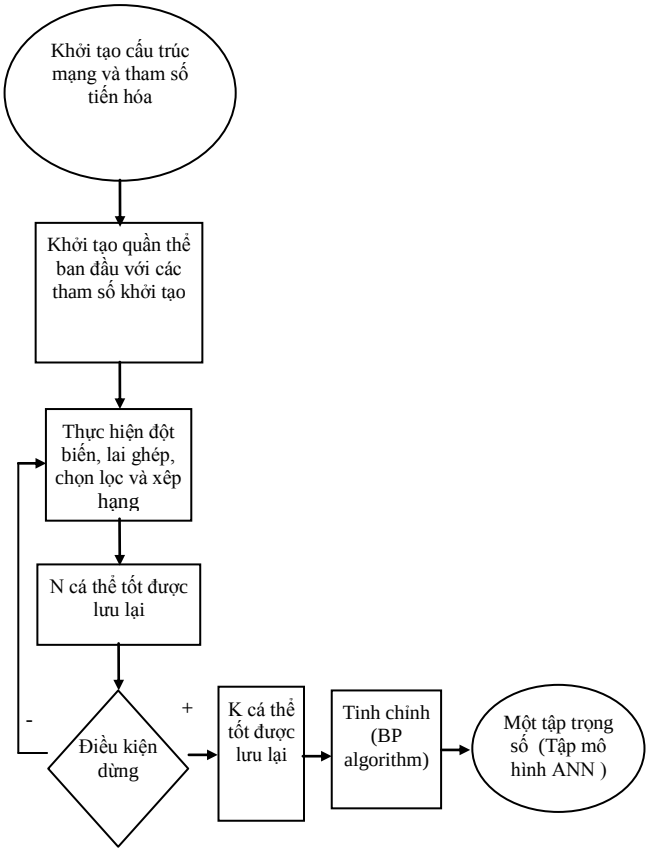
Với mỗi thế hệ thực hiện:

- Thực hiện lai ghép, đột biến tạo ra các cá thể con.
- Tính toán giá trị hàm mục tiêu f_1 cho các nhóm cá thể con.

- Tính toán giá trị hàm mục tiêu f_2 cho quần thể mới (kích thước 2*N, bao gồm cả cá thể cha mẹ và cá thể con mới được tạo ra).

- Thực hiện xếp hạng (ranking) các cá thể dựa trên tiêu chí hàm mục tiêu f_1 và f_2.

- Chọn lọc N cá thể tốt nhất (N cá thể này có thể nằm trên nhiều lớp biên khác nhau).



Hình 10. *Minh họa sơ đồ khối giải thuật tiến hóa đa mục tiêu.*

Bước 3: *Tinh chỉnh*

Kết thúc quá trình tiến hóa, tiến hành sắp xếp các nhóm cá thể trên lớp biên đầu tiên theo chiều tăng dần của hàm f_1; Sau đó lựa chọn nhóm cá thể đầu tiên (có MSE_{min}) trên lớp biên đầu tiên này.

Với mỗi cá thể trong nhóm cá thể được lựa chọn tiến hành tinh chỉnh bằng cách đưa qua ANN huấn luyện.

Bước 4: *Tính toán giá trị MSE cộng đồng*

Phân tích:

Trong giải thuật mà nhóm tác giả đề xuất:

Với giải thuật NSGA-II nguyên bản thì sau bước 2 chúng ta có thể lựa chọn được những cá thể trên lớp biên đầu tiên để thực hiện tính toán giá trị MSE trung bình như trong bước 4. Còn trong giải thuật này do xuất phát từ ý tưởng lựa chọn hàm mục tiêu f_2 (DIV) của nhóm tác giả là để việc lựa chọn các cá thể diễn ra trên toàn không gian tìm kiếm, tránh việc hội tụ sớm (nếu chỉ đơn thuần sử dụng 1 hàm mục tiêu MSE) vào các cực trị địa phương. Ở đây, hàm DIV được tính toán dựa trên khoảng cách giữa các giá trị MSE của từng cá thể. Điều này dẫn đến phải thực hiện tách rời 2 lần duyệt cho mỗi lần tính hàm mục tiêu MSE và DIV.

Cụ thể, trong thuật toán đề xuất : Bước 1, 2 được thực hiện theo đúng ý tưởng của giải thuật đa mục tiêu nhưng có 2 điểm khác biệt: Thứ nhất, ở bước 1 hàm f_2 phải được thực hiện sau khi đã tính toán được hàm f_1 cho tất cả các cá thể. Thứ hai, ở bước 2 sau khi tạo ra tập các cá thể con, các cá thể này sẽ được tính toán giá trị hàm mục tiêu f_1. Lúc này mới tiến hành gộp cả cá thể cha mẹ và cá thể con thành một quần thể có kích thước lớn gấp hai lần, tiến hành sắp xếp lại các cá thể theo giá trị hàm f_1 sau đó tính toán lại giá trị hàm f_2 trên quần thể mới này chứ không tính toán riêng trên các cá thể con mới tạo ra bởi độ đa dạng ở đây được tính theo khoảng cách tập trung giữa cá thể trước và sau cá thể được xét. Sau khi tính toán xong f_1, f_2 mới tiến hành xếp hạng quần thể để tạo thành các lớp biên khác nhau và chọn lựa N cá thể tốt nhất vào quần thể mới.

Bước 3 (đưa thêm bước 3): Tại mỗi thế hệ sẽ có thao tác xếp hạng (ranking) dựa trên tiêu chí là các hàm mục tiêu. Cá thể nào có hạng càng nhỏ thì các giá trị hàm mục tiêu của nó càng tốt. Những cá thể có cùng hạng sẽ cùng nằm trên một lớp biên (front) do đó các cá thể nằm trên lớp biên đầu tiên sẽ là những cá thể tốt nhất. Như vậy, mục đích là sử dụng các cá thể như những bộ trọng số khởi tạo ban đầu được đưa qua ANN để hiệu chỉnh cho kết quả tốt hơn nữa. Sau khi

đưa qua ANN huấn luyện giá trị hàm f_1 của các cá thể đã thay đổi, lúc này tiến hành tính toán lại hàm f_2 và thực hiện xếp hạng lại tập cá thể này thành các lớp biên khác nhau. Kết quả lựa chọn các cá thể trên lớp biên đầu tiên đảm bảo đưa ra những kết quả tối ưu nhất trên cả hai hàm mục tiêu.

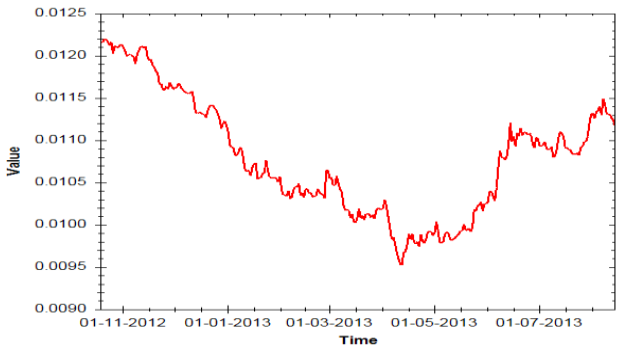
Bước 4, sau khi đã lựa chọn được cộng đồng các cá thể tốt nhất từ bước 3, trong đó mỗi cá thể *i* sẽ cho ta một ANN với bộ trọng số W_i lấy từ cá thể này. Lúc này tiến hành đưa tập dữ liệu thực qua từng ANN này để tính ra từng tập dữ liệu dự báo. Giá trị MSE sẽ được tính dựa trên tập dữ liệu thực và trung bình của tập dữ liệu dự báo được sinh ra bởi cộng đồng các cá thể này.

(*) Độ phức tạp thuật toán đề xuất: Với N là số cá thể trong quần thể; M là số hàm mục tiêu; C là số cá thể con trong mỗi cá thể, T là số thế hệ, B là số vòng lặp trong quá trình tinh chỉnh. (với M=2; C=1 hoặc 3; N<B<=T). Độ phức tạp của việc tính toán hàm mục tiêu f_1 là O(NC); của hàm f_2 là $O(N^2)$; bước thực hiện xếp hạng (Ranking) là $O(MN^2)$. Do đó độ phức tạp tính chung cho cả bước 2 (vòng lặp tiền hóa) là $O(TMN^2)$. Bước tinh chỉnh có độ phức tạp O(NB). Như vậy thuật toán độ phức tạp là $O(TMN^2)$.

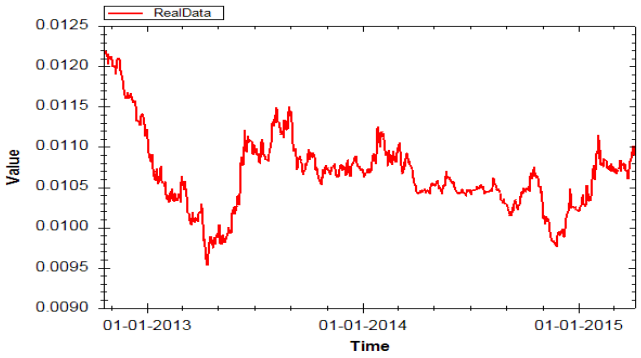
IV. KẾT QUẢ THỬ NGHIỆM VÀ PHÂN TÍCH

IV.1. Mô tả dữ liệu

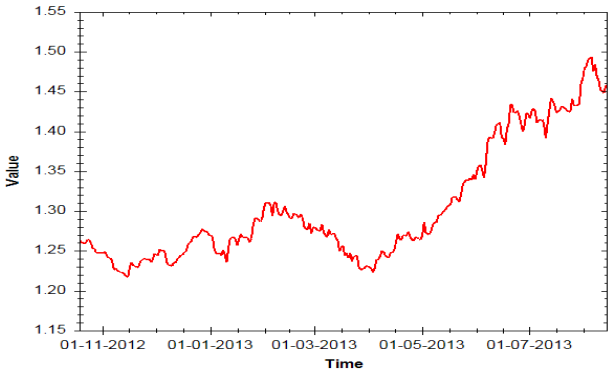
Dữ liệu được sử dụng là 4 loại tiền tệ (đô la Hồng Kông (HKD), YEN Nhật (JPY), tiền tệ của một nhóm các quốc gia thuộc liên minh Châu Âu (EURO), đô la Mỹ (USD)) so với đô la Úc (AUD). Để đánh giá hiệu suất của mô hình đề xuất, dữ liệu được chia thành 2 tập: tập dữ liệu huấn luyện (train) và tập dữ liệu kiểm tra (test). Trong thử nghiệm này, tác giả chọn tỉ lệ dữ liệu huấn luyện và kiểm tra là: 70%-30% (Hình 11–Hình 14).



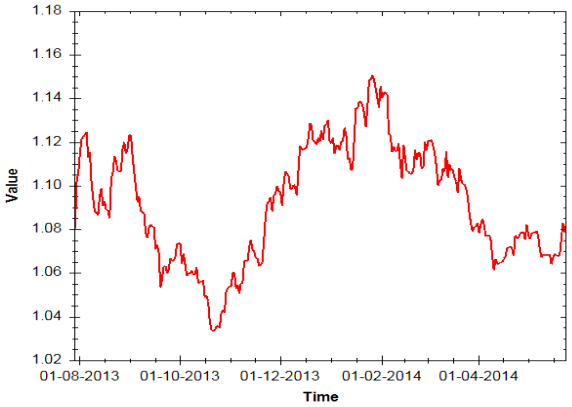
Hình 11. *Dữ liệu tỉ giá JPY/AUD.*



Hình 12. *Dữ liệu tỉ giá HKD/AUD.*



Hình 13. *Dữ liệu tỉ giá EURO/AUD.*



Hình 14. *Dữ liệu tỉ giá USD/AUD.*

IV.2. Tham số cài đặt thử nghiệm

Trên cơ sở mô hình đề xuất và bốn tập dữ liệu đã trình bày ở trên, tác giả tiến hành cài đặt chương trình dự báo tỉ giá. Chương trình có các chức năng như: Huấn luyện dữ liệu, phân tích dữ liệu và dự báo dữ liệu. Để đánh giá hiệu suất của mô hình, chúng tôi cài đặt thêm phương pháp dự báo mô hình mạng ANN để đối sánh. Bảng 1, Bảng 2 mô tả tham số cài đặt. Trong đó, điều kiện dừng (trong hình 10) chính là số thế hệ tiến hóa (1000).

Bảng 1. *Các tham số cài đặt mô hình ANN*

Mô hình ANN	
Hệ số học	0.03
Số nút của lớp vào	5
Số nút của lớp ẩn	10
Số nút ra	1
Số lần lặp của BP	100000
Tỉ lệ dữ liệu kiểm tra (%)	30

Bảng 2. *Các tham số cài đặt mô hình đề xuất*

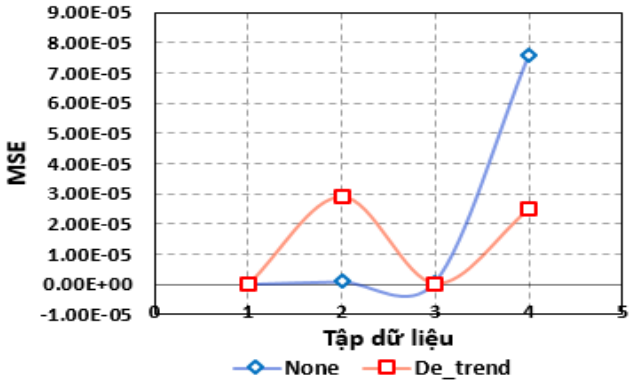
Mô hình dựa trên học cộng đồng kết hợp với giải thuật tiến hóa	
Kích thước quần thể	100
Hệ số đột biến	0.9
Số thế hệ	1000
Số nút của lớp vào	5
Số nút của lớp ẩn	10
Số nút ra	1
Số lần lặp của BP	1000
Tỉ lệ dữ liệu kiểm tra (%)	30

IV.3. Kết quả thử nghiệm

Trước hết, để khảo sát sự ảnh hưởng của đặc tính chuỗi thời gian thì chúng tôi thử nghiệm trên 4 tập dữ liệu với lựa chọn không khử (none) và khử các đặc tính: log, tính mùa và xu hướng (de_trend). Trong ba đặc tính khử thì khử xu hướng thể hiện vượt trội, kết quả được thể hiện trong Bảng 4. Kết quả được trực quan hóa qua Hình 15.

Bảng 4*. MSE khử xu hướng và không khử xu hướng của mô hình đề xuất.*

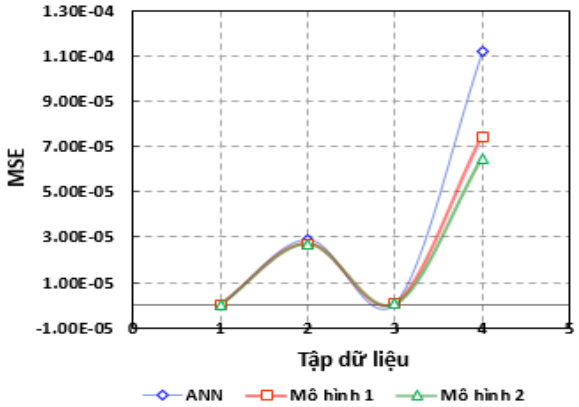
Đặc tính	MSE $_{JPY}$	MSE$_{USD}$	MSE$_{HKD}$	MSE$_{EURO}$
None	6.58E-9	8.21E-07	8.13E-7	7.61E-5
De_trend	3.87E-9	2.90E-05	2.64E-7	2.48E-5



Hình 15. *Biểu đồ so sánh MSE khử xu hướng và không khử xu hướng của mô hình đề xuất.*

Bảng 5*. So sánh giá trị MSE của mô hình đề xuất với một số mô hình khác.*

Mô hình	MSE $_{JPY}$	MSE$_{USD}$	MSE$_{HKD}$	MSE$_{EURO}$
ANN	6.72E-9	2.89E-5	7.30E-7	1.12E-4
Mô hình 1 (NSGA-II, k=1)	5.87E-9	2.72E-5	5.77E-7	7.44E-5
Mô hình 2 (NSGA-II, k=3)	5.63E-9	2.70E-5	5.59E-7	6.48E-5



Hình 16. *Biểu đồ so sánh MSE của 3 phương pháp.*

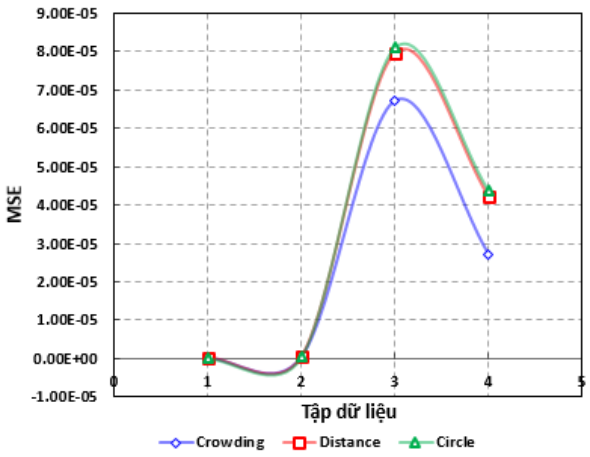
Tiếp theo, chúng tôi tiến hành chạy thực nghiệm 30 lần ngẫu nghiên và thu được kết quả ở Bảng 5. Kết quả này được trực quan hóa qua Hình 16.

Ngoài ra, trong nghiên cứu này còn tiến hành chạy thực nghiệm 30 lần ngẫu nghiên để so sánh 3 cách tính khoảng cách khác nhau với các tham số trong Bảng 3 để đánh giá hiệu suất của mô hình đề xuất. Kết quả thu được trình bày trong Bảng 6.

***Bảng 6**. So sánh giá trị MSE của 3 phương pháp tính độ đa dạng*

Phương pháp	MSE $_{JPY}$	MSE_{USD}	MSE_{HKD}	MSE_{EURO}
Crowding	5.69E-09	5.22E-07	6.74E-05	2.71E-05
Distance	7.85E-09	5.26E-07	7.97E-05	4.21E-05
Circle	8.09E-09	5.42E-07	8.11E-05	4.37E-05

Kết quả được trực quan hóa qua Hình 17.



***Hình 17.** Biểu đồ so sánh MSE_{Train} của các độ đa dạng.*

Từ kết quả thực nghiệm cho thấy một số nhận xét: đặc tính xu hướng có ảnh hưởng đến chuỗi thời gian tỉ giá; Trong ba phương pháp chọn hàm mục tiêu thứ hai khác nhau thì phương pháp tính khoảng cách tập trung cho giá trị lỗi nhỏ; Trong ba mô hình so sánh ở bảng 5 thì mô hình đề xuất có giá trị lỗi nhỏ nhất.

V. KẾT LUẬN VÀ HƯỚNG PHÁT TRIỂN

Trong bài báo này, hiệu quả của mô hình học cộng đồng đã được kiểm tra. Mô hình đề xuất đã tối ưu bộ trọng số của mạng nơron dùng giải thuật NSGA-II (có sử dụng học cộng đồng) nhằm tránh được giá trị lỗi tìm được rơi vào cực trị địa phương. Khi sử dụng thuật toán NSGA-II để xây dựng mô hình, chúng tôi đã sử dụng hai hàm mục tiêu: MSE và DIV (với 3 phương pháp tính độ đa dạng khác nhau). Các kết quả của mô hình được thực nghiệm trên bốn tập dữ liệu tiền tệ. Các thực nghiệm cho thấy kỹ thuật học cộng đồng giúp mạng nơron có giá trị lỗi tốt hơn và cải thiện đáng kể hiệu suất của bài toán dự báo. Từ bài báo, chúng tôi thấy có hai vấn đề phát sinh: Thứ nhất, tiếp tục tiến hành thực nghiệm trên nhiều chuỗi thời gian khác trong thực tế; Thứ hai, thay đổi các hàm mục tiêu và cải tiến các toán tử của NSGA-II nhằm bảo tồn tính đa dạng của quần thể.

TÀI LIỆU THAM KHẢO

[1] HÀ VĂN SƠN, *Giáo trình nguyên lí thống kê kinh tế*, NXB Thống kê, 2010.

[2] NGUYỄN QUANG DONG, *Kinh tế lượng*, NXB Khoa học & Kỹ thuật, 2008.

[3] A. KROGH and J. VEDELSBY, *"Neural network ensembles, cross validation, and active learning"*, in Advances in Neural Information Processing Systems,Vol. 7, The MIT Press, 1995, 231–238.

[4] A. ZHOU and ET AL, "*Multiobjective evolutionary algorithms: A survey of the state of the art*," Swarm and Evolutionary Computation, Vol. 1, No. 1, 2011, 32 – 49.

[5] C. CHATFIELD, *The Analysis of Time Series: an Introduction*, 6th edition, Chapman & Hall/CRC, 2004.

[6] C. SMITH and Y. JIN, *"Evolutionary Multi-Objective Generation of Recurrent Neural Network Ensembles for Time Series Prediction"*, Journal Neurocomputing, 2014, 302-311.

[7] E. ALPAYDIN, *Introduction to machine learning*, 2nd edition, *MIT Press*, 2010.

[8] G. BROWN and ET AL, *"Diversity creationmethods: A survey and categorisation"*, Journal of Information Fusion, Vol 6, 2004, 5-20.

[9] G. VALENTINI, T. DIETTERICH, *"Bias-variance analysis and ensembles of svm"*, 3rd International Workshop on Multiple Classifier Systems, Springer-Verlag Berlin Heidelberg, Vol. 2364, 2002, 222–231.

[10] HOZAIRI and ET AL, *"Implementation of nondomanated sorting genetic algorithm-II (NSGA-II) for Multiobjective Optimization Problem on distribution of Indonesian Navy Warship"*, Journal of Theoretical and Applied Information Technology, Vol 6, No1, 2014, 274-281.

[11] K. DEB and ET AL, *"A Fast Elitist Multi-objective Genetic Algorithm: NSGA-II"*, IEEE Transactions on Evolutionary Computation, 2002, 182 – 197.

[12] K. DEB, *"Multi-Objective Optimization Using Evolutionary Algorithms: An Introduction"*, Springer London, 2001, 3-34.

[13] K. GEBHARD and ET AL, *Introduction to Modern Time Series Analysis*, 2nd editor, Springer-verlag, 2013.

[14] L. BREIMAN, *"Bagging predictors",* Journal of Machine Learning, Vol. 24, No. 2, 1996, 123–140.

[15] L. HANSEN, P. SALAMON, "*Neural Network Ensembles",* IEEE Trans. on Pattern Analysis and Machine Intelligence, Vol. 12, No. 10, 1990, 993– 1001.

[16] M. PERRONE, L. COOPER, *"When Networks Disagree: Ensemble Methods for Hybrid Neural Networks", Artificial Neural Networks for Speech and Vision*, Chapman & Hall, 1993, 126–142.

[17] M. ŠTĚPNIČKA and ET AL, "*Forecasting seasonal time series with computationalintelligence: contribution of a combination of distinct methods"*, Proceedings of Conference on Eusflat-Lfa, Atlantis Press, 2011, 461-471.

[18] P. ADHVARYU, M. PANCHAL, *"A Review on Diverse Ensemble Methods for Classification"*, IOSR Journal of Computer Engineering, Vol 1, No. 4, 2012, 27-32.

[19] S. CHIAM and ET AL, **"***Multiobjective Evolutionary Neural Networks for Time Series Forecasting"*, Lecture Notes in Computer Science, Vol 4403, Springer-Verlag Berlin Heidelberg, 2007, 346-360.

[20] S. GU and Y. JIN, *"Generating Diverse and Accurate ClassifierEnsembles Using Multi-Objective Optimization",* Proceedings of Conference on Computational Intelligence in Multi-Criteria Decision-Making, 2014.

[21] U. NAFTALY and ET AL, *"Optimal ensemble averaging of neural networks",* Network: Computation in Neural System, Vol. 8, No 3, 1997, 283–296.

[22] Z. ZHOU, *"Ensemble learning"*, Berlin: Springer, 2009, 270-273.

[23] http://www.rba.gov.au/statistics/hist-exchange-rates/index.html

Ngày nhận bài: 22/06/2015

SƠ LƯỢC VỀ TÁC GIẢ

ĐINH THỊ THU HƯƠNG

Sinh năm 1973 tại Hà Nội.

Tốt nghiệp ĐH ngành Tin học, Trường ĐH Khoa học Tự nhiên, ĐH Quốc gia Hà Nội, năm 1999; Cao học chuyên ngành Khoa học máy tính, Học Viện Kĩ thuật Quân sự, năm 2007.

Hiện đang công tác tại Khoa CNTT, ĐH Sài Gòn và đang là NCS tại Học viện Kĩ thuật Quân sự.

Hướng nghiên cứu: Học máy và tính toán tiến hóa.

E-mail: huongdtt2011@gmail.com

ĐỖ THỊ DIỆU MY

Sinh năm 1992 tại Bắc Giang.

Đang là học viên Học viện Kĩ thuật Quân sự.

E-mail: dieumy46.mta@gmail.com

BÙI THU LÂM

Nhận bằng Tiến sĩ chuyên ngành Khoa học máy tính, ĐH New South Wales, Australia, năm 2007.

Hiện đang công tác tại Khoa CNTT, Học viện Kĩ thuật Quân sự.

Hướng nghiên cứu: Tính toán thông minh và ứng dụng trong hỗ trợ quyết định.

E-mail: lam.bui07@gmail.com

VŨ VĂN TRƯỜNG

Sinh năm 1985 tại Bắc Ninh.

Tốt nghiệp ĐH ngành CNTT, Cao học ngành Hệ thống thông tin tại Học viện Kĩ thuật Quân sự, năm 2010 và 2014.

Hiện đang công tác tại Viện Kỹ thuật Công trình Đặc biệt, Học viện Kĩ thuật Quân Sự.

Hướng nghiên cứu: GIS, Computer Vision, Simulation, Evolutionary Computations.

E-mail: truongvv@mta.edu.vn