MÔ HÌNH HAI GIAI ĐOẠN DỰ ĐOÁN GIÁ CỔ PHIẾU VỚI K-MEANS VÀ FUZZY-SVM

Article · December 2014

CITATIONS
0

READS
1,709

1 author:

Duc-Hien Nguyen
University of Da nang
13 PUBLICATIONS **22** CITATIONS

SEE PROFILE

MÔ HÌNH HAI GIAI ĐOẠN DỰ ĐOÁN GIÁ CỔ PHIẾU VỚI K-MEANS VÀ FUZZY-SVM

A TWO-STAGE MODEL FOR STOCK PRICE FORECASTING BY COMBINING K-MEANS WITH FUZZY-SVM

Nguyễn Đức Hiển
hiencit@gmail.com

Tóm tắt - Bài báo đề xuất một mô hình dự đoán giá cổ phiếu kết hợp K-Means và fuzzy – Support Vector Machines (fuzzy-SVM). Việc trích xuất tập luật mờ từ dữ liệu thô dựa vào sự kết hợp của các mô hình học thống kê chính là cơ sở của mô hình đề xuất. Kiến trúc của mô hình gồm hai giai đoạn, giai đoạn một sẽ áp dụng thuật toán K-means để phân chia không gian dữ liệu đầu vào thành nhiều cụm riêng biệt. Ở giai đoạn hai, với mỗi phân vùng của dữ liệu đầu vào, mô hình fuzzy-SVM (thuật toán f-SVM) sẽ được sử dụng để khai phá các luật mờ sử dụng cho hệ thống dự đoán. Mô hình đề xuất được áp dụng dự đoán cho một số mã cổ phiếu của thị trường chứng khoán Việt Nam. Các thông số đánh giá kết quả thực nghiệm sẽ được giới thiệu, và có sự so sánh với kết quả của một số mô hình khác.

Từ khóa – dự đoán giá cổ phiếu; mô hình mờ; mô hình mờ hướng dữ liệu; luật mờ; máy học Véc-tơ hỗ trợ; K-Means

Abstract - This paper proposed a model for stock price forecasting via a combination of K-Means and fuzzy – Support Vector machines (fuzzy-SVM). The extraction of fuzzy rules from raw data based on the combination of statistical machine learning models is the foundation of this proposed approach. The architecture of proposed model includes two stages: the first stage is using K-Means algorithm to partition the whole input space into several disjoint regions. In the second stage, the fuzzy-SVM model (f-SVM alogorithm) is used to extract fuzzy rules from each partition of input data. Then, the proposed model is applied in predicting some of stock codes of Vietnam's stock market. The experiment results are presented in comparison with the results of the other approaches.

Key words – stock price forecasting; fuzzy model; data-driven fuzzy models; fuzzy rules; support vector machine; K-Means

1. Đặt vấn đề

Vấn đề dự đoán theo chuỗi thời gian, mà đặc biệt là vấn đề dự đoán thị trường chứng khoán đã và đang thu hút được nhiều sự quan tâm nghiên cứu của các nhà khoa học. Có nhiều mô hình và giải pháp khác nhau đã được các nhà nghiên cứu đề xuất, với mục tiêu cuối cùng là nâng cao tính chính xác của kết quả dự đoán. Vấn đề dự đoán thị trường chứng khoán hiện nay chủ yếu được tiếp cận dưới hai dạng, đó là dự đoán giá cổ phiếu hoặc xu hướng của giá cổ phiếu sau *n*-ngày.

Những hướng tiếp cận phổ biến hiện nay cho vấn đề dự đoán thị trường chứng khoán là khai phá dữ liệu, ứng dụng các mô hình máy học thống kê [3]. Những nghiên cứu ở [7], [8], [14], [16], [17] đề xuất ứng dụng mạng nơ-ron nhân tạo, máy học véc-tơ hỗ trợ (SVM – Support Vector Machine), mô hình markov ẩn (HMM – Hidden Markov Model) trong dự đoán thị trường chứng khoán. Những mô hình theo hướng cải tiến và kết hợp nhiều phương thức học khác nhau để nâng cao hiệu quả dự đoán [4], [9], [11] cũng được các tác giả nghiên cứu và đề xuất.

Mô hình dự đoán dựa trên tập mờ trích xuất được từ máy học Véc-tơ hỗ trợ được giới thiệu như là một trong những hướng nghiên cứu mới của mô hình mờ - mô hình mờ hướng dữ liệu (data-driven fuzzy models) [5], [6], [10]. Một trong những hạn chế của mô hình mờ hướng dữ liệu là vấn đề học từ động từ dữ liệu huấn luyện với kích thước lớn và thiếu tính đặc trưng. Với mục tiêu hướng đến là giải quyết vấn đề kích thước dữ liệu lớn, đồng thời tạo điều kiện thuận lợi cho chuyên gia con người có thể hiểu và phân tích được tập luật mờ học được từ dữ liệu, qua đó có thể điều chỉnh, tiến tập luật, nâng cao hiệu quả dự đoán, trong nghiên cứu này, chúng tôi đề xuất một mô hình ai giai đoạn dự đoán giá cổ phiếu dựa trên sự kết hợp K-Means và f-SVM.

Các phần tiếp theo của bài báo bao gồm: phần 2 trình bày sơ lược về mô hình trích xuất luật mờ từ SVM - thuật toán f-SVM. Trong phần 3, chúng tôi đề xuất một mô hình hai giai đoạn dự đoán giá cổ phiếu kết hợp K-Means và f-SVM. Phần 4 trình bày những kết quả thực nghiệm của mô hình đề xuất, trong đó có kết hợp so sánh với một số kết quả của các mô hình khác. Cuối cùng, trong phần 5 chúng tôi nêu lên một số kết luận và định hướng nghiên cứu tiếp theo.

2. Mô hình trích xuất luật mờ từ SVM

Máy học véc-tơ hỗ trợ SVM được Vapnik giới thiệu năm 1995, đây là mô hình học dựa trên lý thuyết học thống kê (Statistical Learning Theory) [1] và là một kỹ thuật được đề nghị để giải quyết cho các bài toán phân lớp. Một số nghiên cứu gần đây [2], [4], [6] đã đề xuất sử dụng SVM giải quyết bài toán tối ưu hóa hồi quy; đồng thời SVM cũng được sử dụng để khai phá luật mờ từ dữ liệu số [2], [3], [5]. Với vai trò giải quyết vấn đề tối ưu hóa hồi quy, lý thuyết cơ bản của SVM có thể được vắn tắt như sau [22]:

Cho một tập dữ liệu huấn luyện $\{(x_1, y_1), \dots, (x_l, y_l)\} \subset \mathcal{X} \times \mathbb{R}$, trong đó $\mathcal{X}$ xác định miền dữ liệu đầu vào. Mục tiêu của ε-SV hồi quy (ε-Support Vector Regression) là tìm siêu phẳng đi qua tất cả các phần tử dữ liệu huấn luyện, đồng thời độ sai lệch trên các y_i của cả tập dữ liệu huấn luyện là không lớn hơn ε. Trong trường hợp hồi quy phi tuyến, hàm quyết định f(x) có thể xác định như sau:

$$f(x) = \sum_{i=1}^{l} (\alpha_i - \alpha_i^*)\, K(x_i, x) + b \quad (1)$$

Sao cho:

$$\sum_{i-1}^{l} (\alpha_i - \alpha_i^*) = 0, \text{and } C \geq \alpha_i, \alpha_i^* \geq 0, \forall i, \quad (2)$$

Trong đó, C là hằng số chuẩn tắc, α_i, α_i^* là những nhân tử Lagrange; và $K(x_i, x)$ là hàm Kernel được định nghĩa như sau:

$$K(x_i, x_j) = \langle \Phi(x_j), \Phi(x_j) \rangle \quad (3)$$

với Φ là ánh xạ thuộc tính cho kernel K.

Những điểm đầu vào x_i với $(\alpha_i - \alpha_i^*) \neq 0$ gọi là những véc-tơ hỗ trợ (SVs).

Bên cạnh đó, ta biết rằng các luật mờ (fuzzy rules) được biểu diễn ở dạng IF – THEN, là cơ sở của phép suy luận mờ [2]. Giả sử có *m* luật mờ được biểu diễn như sau:

$$R_j: \text{IF } x_1 \text{ is } A_1^j \text{ and } x_2 \text{ is } A_2^j \text{ and } \dots \text{ and } x_n \text{ is } A_n^j$$
$$\text{THEN y is } B^j, \text{for j} = 1, 2, \dots, m \quad (4)$$

Trong đó $x_i (i = 1,2, \dots n)$ là các biến điều kiện; y là các biến quyết định của hệ thống mờ; A_i^j và B^j là những thuật ngữ ngữ nghĩa xác định bởi các hàm thành viên (membership functions) tương ứng $\mu_{A_i^j}(x_i)$ và $\mu_{B^j}(y)$.

Kết quả đầu ra của suy luận được xác định bằng công thức sau [22]:

$$f(x) = \frac{\sum_{j=1}^{M} \bar{z}^j \left(\prod_{i=1}^{n} \mu_{A_i^j}(x_i) \right)}{\sum_{j=1}^{M} \prod_{i=1}^{n} \mu_{A_i^j}(x_i)} \quad (5)$$

Trong đó, $\bar{z}^j$ là giá trị đầu ra khi hàm thành viên $\mu_{B^j}(y)$ đạt giá trị cực đại.

Để (1) và (5) bằng nhau, trước tiên chúng ta phải đồng nhất giữa hàm kernel trong (1) và hàm thành viên trong (5). Ở đây, để thỏa mãn điều kiện Mercer [15] hàm thành viên Gaussian được chọn làm hàm kernel; đồng thời giá trị của *b* trong (1) phải bằng 0.

Khi hàm Gaussian được chọn làm hàm thành viên trong (1) và hàm kernel trong (5), đồng thời số luật mờ *m* bằng với số Support vectors *l* thì (1) và (5) được biến đổithành:

$$f(x) = \sum_{i=1}^{l} (\alpha_i - \alpha_i^*) \exp\left(-\frac{1}{2} \left(\frac{x_i - x}{\sigma_i} \right)^2 \right) \quad (6)$$

và

$$f(x) = \frac{\sum_{j=1}^{l} \bar{z}^j \exp\left(-\frac{1}{2} \left(\frac{x_j - x}{\sigma_j} \right)^2 \right)}{\sum_{j=1}^{l} \exp\left(-\frac{1}{2} \left(\frac{x_j - x}{\sigma_j} \right)^2 \right)} \quad (7)$$

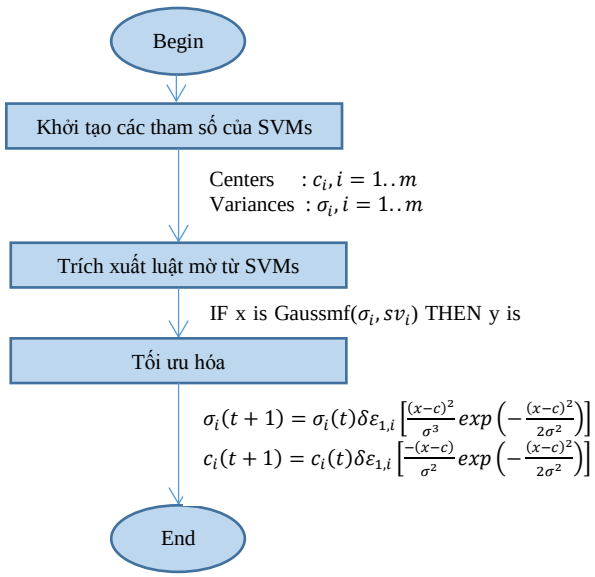
Như cách biến đổi trong [2], hàm suy luận mờ có thể viết lại như sau:

$$f(x) = \sum_{j=1}^{l} \bar{z}^j \exp\left(-\frac{1}{2} \left(\frac{x_j - x}{\sigma_j} \right)^2 \right) \quad (8)$$

và trung tâm của hàm thành viên Gaussian được chọn là

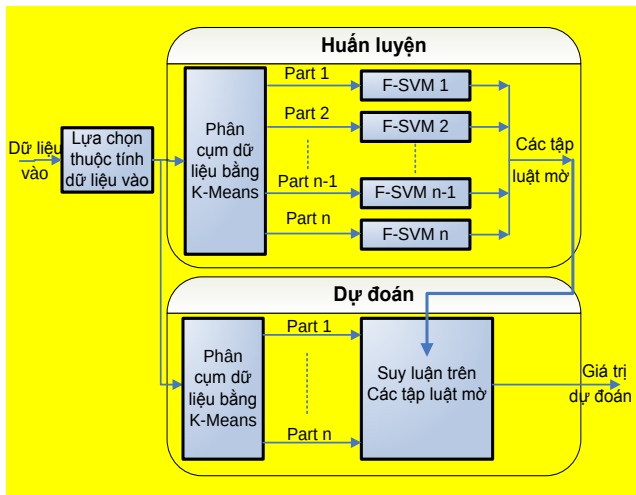
$$\bar{z}^j = (\alpha_i - \alpha_i^*) \quad (9)$$

Trên cơ sở đồng nhất hàm đầu ra của SVMs và hệ thống mờ, chúng tôi đã xây dựng được thuật toán cho phép trích xuất tập luật mờ từ máy học SVM - thuật toán f-SVM [22] (Hình 1).



Hình 1. *Sơ đồ khối thuật toán f-SVM.*

3. Mô hình hai giai đoạn dự đoán giá cổ phiếu



Hình 2. *Mô hình hai giai đoạn*

Dự đoán giá cổ phiếu dựa vào dữ liệu quá khứ là một bài toán dự đoán chuỗi thời gian không ổn định (non-stationary), nghĩa là sự phân bố thống kê của dữ liệu không ổn định theo thời gian. Để nâng cao hiệu quả dự đoán, dữ liệu đầu vào phải được thu thập trong thời gian dài, các thuộc tính của dữ liệu phải bao phủ được càng nhiều càng tốt các trường hợp của bài toán; và như thế các thuật oán học và suy luận sẽ phải thực hiện trên một tập dữ liệu lớn. Để khắc phục trở ngại này, chúng tôi đề xuất mô hình hai giai đoạn bằng cách kết hợp kỹ thuật phân cụm K-Means và mô hình trích xuất luật mờ f-SVM. Sơ

đồ khối của mô hình được thể hiện trong Hình. 2.

Với mô hình này, trước tiên dữ liệu đầu vào sẽ được phân cụm bằng K-Means để gom những mẫu dữ liệu gần giống nhau (có phân bố thống kê gần giống nhau) vào cùng một cụm (cluster); sau đó cứ mỗi cụm dữ liệu sẽ được dùng để huấn luyện cho một mô hình f-SVM để trích xuất ra một tập luật tương ứng.

3.1. Lựa chọn thuộc tính đầu vào

Theo những kết quả nghiên của các tác giả khác về việc dự đoán thị trường chứng khoán, có nhiều cách khác nhau để lựa chọn thuộc tính đầu vào, ví dụ như: sử dụng các chỉ số kinh tế vi mô [1], sử dụng các chỉ số giá cổ phiếu hàng ngày <opening, high, low, closing price> [8], [17],… Ở mô hình này chúng tôi lựa chọn chỉ số giá cổ phiếu hàng ngày làm dữ liệu vào. Tuy nhiên, tập dữ liệu vào sẽ được tiền xử lý trước khi đưa vào huấn luyện cho mô hình.

Theo sự phân tích và đánh giá của L.J. Cao và Francis E.H. Tay trong [8], việc chuyển đổi chỉ số giá ngày thành tỷ lệ khác biệt trung bình 5 ngày (5-day relative difference in percentage of price – RDP) sẽ mang lại một số hiệu quả nhất định, đặc biệt là cải thiện được hiệu quả dự đoán. Trong mô mình này, chúng tôi lựa chọn các biến đầu vào dựa theo đề xuất và tính toán của L.J. Cao và Francis E.H. Bảng 1 thể hiện các thuộc tính lựa chọn và công thức tính của chúng.

Bảng 1: *Công thức tính toán các thuộc tính lựa chọn.*

Ký hiệu	Thuộc tính	Công thức tính
x_1	EMA100	$P_i - \overline{EMA_{100}(i)}$
x_2	RDP-5	$\big(P(i) - P(i-5)\big)/P(i-5) * 100$
x_3	RDP-10	$\big(P(i) - P(i-10)\big)/P(i-10) * 100$
x_4	RDP-15	$\big(P(i) - P(i-15)\big)/P(i-15) * 100$
x_5	RDP-20	$\big(P(i) - P(i-20)\big)/P(i-20) * 100$
y	RDP+5	$(\overline{P(i+5)} - \overline{P(i)})/\overline{P(i)} * 100$ $\overline{P(i)} = \overline{EMA_3(i)}$

Trong đó, P(i) là chỉ số giá đóng phiên của ngày thứ i, và $EMA_m(i)$ là *m-day exponential moving average* của giá đóng phiên ngày thứ i.

3.2. Phân cụm dữ liệu đầu vào bằng K-Means

Dữ liệu đầu vào của các ứng dụng khai phá dữ liệu thường là rất lớn, trong khi đó có nhiều thuật toán học là không hiệu quả với kích thước dữ liệu lớn. Một trong những hướng tiếp cận quyết vấn đề này là phân dữ liệu đầu vào thành các cụm nhỏ, áp dụng các thuật toán học trên từng cụm dữ liệu và sau đó tổng hợp các kết quả học lại [13]. Ngoài ra, một đặc điểm đáng lưu ý của giá cổ phiếu là tính không ổn định theo thời gian. Phân bố thống kê của giá cổ phiếu theo thời gian phụ thuộc vào nhiều yếu tố khác nhau như sự tăng trưởng hay suy thoái của kinh tế, tình hình chính trị, môi trường, thiên tai,… Điều đó gây nên nhiều hạn chế cho việc tìm ra những qui tắc dự đoán giá cổ phiếu dựa trên dữ liệu quá khứ. Như vậy, chính giải pháp phân cụm dữ liệu theo phân bố thống kê của chúng sẽ làm giảm tính bất ổn định khi xét trong từng cụm dữ liệu riêng biệt.

Thuật toán phân cụm K-Means (K-Means clustering) do MacQueen giới thiệu đầu tiên năm 1967 [12]. K-Means clustering là một thuật toán dùng trong các bài toán phân loại / nhóm *n* đối tượng thành *k* nhóm dựa trên đặc tính / thuộc tính của đối tượng (*k*, *n* nguyên, dương). Trong bài báo này chúng tôi không có mong muốn phân tích và đánh giá thuật toán K-Means, lý thuyết chi tiết về K-Means có thể tham khảo trong tài liệu [12]. Phương pháp sử dụng K-Means để phân cụm dữ liệu giá cổ phiếu cũng đã được nhiều tác giả giới thiệu sử dụng [19], [20], [21]. Trong khuôn khổ nghiên cứu này, chúng tôi đề xuất sử dụng thuật toán K-Means để phân chia các mẫu dự liệu đầu vào thánh *k* cụm riêng biệt, trong đó *k* là một hằng số được xác định trước.

Trong mô hình thực nghiệm, chúng tôi sử dụng thuật toán phân cụm K-Means trong thư viện Statistics Toolbok của bộ công cụ MATLAB 2012b. Giá trị hằng số *k* được chọn thủ công bằng cách thử nhiều lần để đạt được kết quả dự đoán tốt nhất.

3.3. Trích xuất luật mờ bằng thuật toán f-SVM

Mỗi cụm dữ liệu vào đã được phân tách bằng K-Means sẽ được đưa vào huấn luyện cho từng máy f-SVM tương ứng để trích xuất các luật mờ. Trong mô hình thực nghiệm, chúng tôi sử dụng thuật toán học SVM của thư viện LIBSVM, được phát triển bởi nhóm của Chih-Chung Chang [18], để sản sinh ra các SVs; trên cơ sở đó chúng tôi xây dựng thuật toán f-SVM để trích xuất tập luật mờ.

Các luật mờ trích xuất được sẽ có dạng như trong bảng 2.

Bảng 2: *Dạng tập luật mờ trích xuất được.*

Luật	Chi tiết
R_1	IF x_1=Gaussmf(sv_{11},σ_{11}) and … x_i = Gaussmf (sv_{1i}, σ_{1i}) and … THEN y=B_1
R_2	IF x_1=Gaussmf(sv_{21},σ_{21}) and … x_i = Gaussmf (sv_{2i}, σ_{2i}) and … THEN y=B_2
…	…
R_m	IF x_1=Gaussmf(sv_{m1},σ_{m1}) and … x_i = Gaussmf (sv_{mi}, σ_{mi}) and … THEN y=B_m

3.4. Dự đoán giá cổ phiếu dựa trên các tập luật đã trích xuất được

Các tập luật mờ trích xuất được từ các máy f-SVM tương ứng với các cụm dữ liệu huấn luyện có thể được sử dụng để suy luận dự đoán giá cổ phiếu. Với những tập luật mờ được phân thành nhiều cụm với kích thước nhỏ sẽ làm giảm độ phức tạp của các thuật toán suy luận mờ.

Những luật mờ khai phá được từ dữ liệu biểu diễn ở dạng trên vẫn có một khoảng cách nhất định đối với sự hiểu biết của chuyên gia con người; tuy nhiên việc phân cụm các luật mờ khai phá được cũng là một điều kiện để chuyên gia con người có thể ngữ nghĩa hóa và từ đó có thể hiểu và đánh giá được các luật này.

4. Kết quả thực nghiệm

Nguồn dữ liệu thực nghiệm được chọn ngẫu nhiên từ những mã cổ phiếu có lịch sử giao dịch tương đối dài bao gồm: TTC (Công ty cổ phần Gạch men Thanh Thanh), SGH (Công ty Cổ phần Khách sạn Sài Gòn), DXP (Công ty cổ phần Cảng Đoạn xá); và chỉ số của hai sàn giao dịch chứng khoán Việt Nam VNINDEX và HASTC (bảng 3). Các dữ liệu trên được lấy từ nguồn dữ liệu lịch sử của 2 sàn chứng khoán Việt Nam, thông qua website http://www.cophieu68.vn/.

Bảng 3: *Nguồn dữ liệu thực nghiệm.*

Tên cổ phiếu	Thời gian	Dữ liệu training	Dữ liệu testing
Công ty cổ phần Gạch men Thanh Thanh (TTC)	08/08/2006 - 16/04/2014	1520	200
Công ty Cổ phần Khách sạn Sài Gòn (SGH),	16/07/2001 - 08/04/2014	1780	200
Công ty cổ phần Cảng Đoạn xá (DXP)	16/12/2005 - 16/04/2014	1610	200
VNINDEX	28/07/2000 - 16/04/2014	2800	200
HASTC	01/01/2006 – 16/04/2014	1700	200

Các tập dữ liệu training sẽ được dùng để trích xuất các tập luật mờ. Bảng 4 thể hiện một nhóm luật mờ trích xuất được từ dữ liệu training của mã cổ phiếu TTC.

Bảng 4: *Một nhóm luật mờ trích xuất được ứng với mã cổ phiếu TTC.*

Luật	Chi tiết
R1	IF x1=Gaussmf(0.09,-0.11) and x2 = Gaussmf (0.09,-0.12) and x3=Gaussmf(0.09,-0.04) and x4=Gaussmf(0.09,-0.10) and x5=Gaussmf(0.09,-0.09) THEN y=0.10
R2	IF x1=Gaussmf(0.10,-0.01) and x2 = Gaussmf (0.09,-0.06) and x3=Gaussmf(0.10,0.04) and x4=Gaussmf(0.10,-0.10) and x5=Gaussmf(0.10,-0.12) THEN y=0.57
R3	IF x1=Gaussmf(0.09,0.02) and x2 = Gaussmf (0.10,0.02) and x3=Gaussmf(0.09,0.08) and x4=Gaussmf(0.10,-0.08) and x5=Gaussmf(0.10,-0.13) THEN y=-0.02
R4	IF x1=Gaussmf(0.10,-0.04) and x2 = Gaussmf (0.10,-0.08) and x3=Gaussmf(0.10,0.02) and x4=Gaussmf(0.09,-0.08) and x5=Gaussmf(0.09,-0.11) THEN y=-0.29
R5	IF x1=Gaussmf(0.10,-0.03) and x2 = Gaussmf (0.09,-0.06) and x3=Gaussmf(0.10,0.03) and x4=Gaussmf(0.09,-0.10) and x5=Gaussmf(0.09,-0.13) THEN y=-0.38

Bằng cách sử dụng hàm AVALFIS trong thư viện công cụ Matlab Fuzzy Logic, chúng tôi đã thử nghiệm suy luận dựa trên các tập luật sản xuất được đối với các tập dữ liệu testing. Bên cạnh đó chúng tôi cũng thử nghiệm dự đoán trên cùng bộ dữ liệu đó với các mô hình được đề xuất bởi các tác giả khác, bao gồm RBN, SVM và mô hình kết hợp K-Means+SVM. Mô hình RBN được xây dựng dựa trên mạng neural hồi qui Generalized là một kiểu của Radial Basis Network (RBN). Mạng neural hồi qui Generalized được đề xuất giải quyết bài toán dự đoán trong [7], [14], [16]. Mô hình K-Means+SVM là mô hình dựa trên sự kết hợp của K-Means và SVM, được đề xuất để dự đoán xu hướng cổ phiếu trong [19]. Hiệu quả của các mô hình được so sánh và đánh giá dựa trên ba thông số, gồm NMSE (Nomalized Mean Squared Error), MAE (Mean Absolute Error), và DS (Directional Symmetry). Trong đó NMSE và MAE đo lường độ lệch giữa giá trị thực tế và giá trị dự đoán, DS đo lường tỷ lệ dự đoán đúng xu hướng của giá trị RDP+5. Giá trị tương ứng của NMSE và MAE là nhỏ và của DS là lớn chứng tỏ rằng mô hình dự đoán tốt.

Bảng 5a: *Kết quả dự đoán theo các mô hình RBN, SVM.*

Mã cổ phiếu	RBN			SVM		
	NMSE	MAE	DS	NMSE	MAE	DS
HASTC	0.9039	0.0184	39.30	0.9278	0.0191	38.31
VN INDEX	1.0910	0.0115	34.31	1.0725	0.0110	34.33
TTC	1.2211	0.0391	39.80	1.2687	0.0394	38.90
SGH	1.1120	0.0604	38.46	1.1015	0.0576	38.31
DXP	1.2197	0.0244	39.80	1.2073	0.0242	39.83

Bảng 5b: *Kết quả dự đoán theo các mô hình K-Means+SVM, K-Means+F-SVM.*

Mã cổ phiếu	Số cụm	K-Means+SVM			K-Means+F-SVM		
		NMSE	MAE	DS	NMSE	MAE	DS
HASTC	6	0.9057	0.0188	41.71	0.7601	0.0164	44.72
VN INDEX	6	1.1726	0.0109	42.68	1.1408	0.0108	42.21
TTC	6	1.1358	0.0392	42.71	1.1390	0.0391	42.81
SGH	6	1.0792	0.0573	41.71	1.0909	0.0646	42.71
DXP	6	1.1138	0.0258	45.72	1.1281	0.0254	45.22

Với các kết quả thực nghiệm dự đoán trên 200 mẫu dữ liệu testing thể hiện trong Bảng 5 ta thấy, trên cả 5 mã cổ phiếu, giá trị các thông số MNSE và MAE của mô hình K-Means+f-SVM (bảng 5a) đề xuất là nhỏ hơn so với các mô hình RBN và SVM, điều này chứng tỏ độ sai lệch giữa giá trị dự đoán và giá trị thực tế của mô hình đề xuất là ít hơn so với hai mô hình kia. Bên cạnh đó, ta cũng thấy giá trị thông số DS của mô hình đề xuất lớn hơn so với các mô hình RBN và SVM, điều này chứng tỏ tỷ lệ dự đoán đúng xu hướng của mô hình đề xuất cao hơn so với hai mô hình kia.

So sánh kết quả của mô hình K-Means+f-SVM đề

xuất với mô hình K-Means+SVM (bảng 5b), ta thấy giá trị của những thông số của cả hai mô hình là tương đương. Điều này cũng dễ dàng lý giải được, bởi vì thuật toán f-SVM đề xuất đã rút trích ra tập luật mờ dùng cho mô hình dự đoán từ các máy SVMs, và như vậy mô hình dự đoán đề xuất kết hợp K-Means và f-SVM sẽ thừa hưởng hiệu quả của mô hình K-Means+SVM là điều tất yếu. Tuy nhiên, so với mô hình dự đoán K-Means+SVM mô hình dự đoán đề xuất có những ưu điểm sau: 1) Mô hình dự đoán K-Means+SVM là một mô hình "hộp đen" đối với người dùng cuối, trong khi mô hình đề xuất cho phép trích xuất ra một tập luật mờ và quá trì suy luận sẽ được thực hiện trên tập luật này. Đối với người dung cuối thì mô hình suy luận dựa trên một tập luật mờ sẽ dễ hiểu và sáng tỏ hơn. 2) Trên cở sở tập luật mờ trích xuất được, những chuyên gia con người có thể đọc hiểu và điều chỉnh, bổ sung tập luật này bằng các luật chuyên gia để nâng cao hiệu quả suy luận dựa trên tập luật. 3) Ngoài ra việc áp dụng K-Means để phân cụm dữ liệu đầu vào thành từng tập nhỏ riêng biệt, bên cạnh hiệu quả mang lại là giảm kích thước dữ liệu vào và từ đó làm giảm độ phức tạp của thuật toán, tập luật sinh ra cũng sẽ được phân thành các cụm riêng biệt tương ứng, điều này cũng sẽ góp phần giúp cho chuyên gia con người đọc hiểu và phân tích các luật mờ học được dễ dàng hơn.

5. Kết luận

Trong nghiên cứu này đề xuất một mô hình dự đoán giá cổ phiếu dựa trên sự kết hợp của K-Means và f-SVM. Kết quả thực nghiệm trên dữ liệu thử nghiệm cho thấy mô hình đề xuất thật sự mạng lại hiệu quả dự đoán cao hơn so với các mô hình đơn như RBN, SVM trước đó của các tác giả khác, thể hiện qua các giá trị tốt hơn của các thông số NMSE, MAE và DS. Đồng thời, với giải pháp kết hợp phân cụm bằng K-Means trong mô hình đã giúp cải thiện đáng kể thời gian thực hiện các thuật toán trong mô hình. Mặt khác, như đã trình bày ở phần 4.2 của bài báo, một trong những hiệu quả mang lại của mô hình đề xuất là việc gom cụm các luật mờ trích xuất được, là một hình thức chia nhỏ tập luật, sẽ giúp cho việc phân tích các luật này dễ dàng hơn.

Bên cạnh những ưu điểm nêu trên, mô hình đề xuất cũng còn những tồn tại nhất định, một trong những vấn đề tồn tại đó chính là ở thuật toán trích xuất luật mờ từ máy học SVM. Cụ thể là đối với máy học SVM, nếu chúng ta tăng tính chính xác của mô hình thì số lượng SVs cũng tăng lên, đồng nghĩa với số lượng luật mờ cũng tăng lên. Điều này làm cho tính phức tạp của hệ thống tăng lên và đặc biệt là tính "có thể hiểu được" của tập luật mở giảm đi, gây nên sự khó khăn cho chuyên gia con người để có thể hiểu và phân tích các luật này. Việc nghiên cứu tìm tìm giải pháp cải thiện tính "có thể hiểu được" của tập luật mờ trích xuất được từ SVMs cũng chính là một trong những định hướng nghiên cứu tiếp theo của chúng tôi.

Tài liệu tham khảo

[1] Christan Pierdzioch, Jorg Dopke, Daniel Hartmann, *Forecasting stock market volatility with macroeconomic variables in real time*. Journal of Economics and Business 60, 256-276 (2008)

[2] Corinna Cortes and Vladimir Vapnik, *Support-Vector Networks.* Machine Learning, 20, 273-297 (1995)

[3] Hajizadeh E., Ardakani H. D., Shahrabi J., *Application Of Data Mining Techniques In Stock Markets: A Survey*. Journal of Economics and International Finance Vol. 2(7), 109-118 (2010)

[4] Francis Eng Hock Tay and Li Yuan Cao, *Improved financial time series forecasting by combining Support Vector Machines with self-organizing feature map*. Intelligent Data Analysis 5, 339-354, IOS press (2001)

[5] J.-H Chiang and P.-Y Hao, *Support vector learning mechanism for fuzzy rule-based modeling: a new approach*. IEEE Trans. On Fuzzy Systems, vol. 12, pp. 1-12 (2004)

[6] J.L. Castro, L.D. Flores-Hidalgo, C.J. Mantas and J.M. Puche, *Extraction of fuzzy rules from support vector machines*. Elsevier. Fuzzy Sets and Systems, 158, 2057 – 2077 (2007)

[7] Kreesuradej W., Wunsch D., Lane M. *Time-delay Neural Network For Small Time Series Data Sets*. in World Cong. Neural Networks, San Diego, CA (1994)

[8] L.J.Cao and Francis E.H.Tay, *Support vector machine with adaptive parameters in Financial time series forecasting*, IEEE trans. on neural network,vol. 14, no. 6 (2003)

[9] Md. Rafiul Hassan, Baikunth Nath, Michael Kirley, *A fusion model of HMM, ANN and GA for stock market forecasting*, Expert Systems with Applications 33, 171–18 (2007)

[10] S. Chen, J. Wang and D. Wang, *Extraction of fuzzy rules by using support vector machines*. IEEE, Computer society, pp. 438-441 (2008)

[11] Sheng-Hsun Hsu, JJ Po-An Hsieh, Ting-Chih CHih, Kuei-Chu Hsu, *A two-stage architecture for stock price forecasting by integrating self-organizing map and support vector regression*, Expert system with applications 36, 7947-7951 (2009)

[12] MacQueen J. B., *Some Methods for classification and Analysis of Multivariate Observations*, Proceedings of 5-th Berkeley Symposium on Mathematical Statistics and Probability, Berkeley, University of California Press, 1:281-297 (1967).

[13] T.G. Dietterich, *Machine learning research: Four current directions*, AI Magazine, 18(4), 97-136 (1997)

[14] Younes Chtioui, Suranjan Panigrahi, Leonard Francl, *A generalized regression neural network and its application for leaf wetness prediction to forecast plant disease*, Chemometrics and Intelligent Laboratory System 48, 47-58 (1999)

[15] R. Courant, D. Hilbert, *Methods of Mathematical Physics*. Wiley, New York (1953)

[16] Iffat A. Gheyas, Leslies S. Smith, *A Neural network approach to time series forecasting*. Proceeding of the World congress on Engineering 2009 Vol II (2009)

[17] Md. Rafiul Hassan and Baikunth Nath, *Stock market forecasting using Hidden markov model: A new approach*. 5th International conference on intelligent system design and applications (ISDA'05) (2005)

[18] Chih-Wei Hsu, Chih-Chung Chang, and Chih-Jen lin, *A practical Guide to Support Vector Classification*. http://www.csie.ntu.edu.tw/~cjlin/libsvm/ (2010)

[19] Vạn Duy Thanh Long, Lê Minh Duy, Nguyễn Hoàng Tú Anh, *Phương pháp dự đoán xu hướng cổ phiếu dựa trên việc kết hợp K-mean và SVM với ước lượng xác suất lớp*, Đại học quốc gia – Tp HCM (2011)

[20] Keerthiram Murugesan Jun Zhang, *Hybrid Bisect K-Means Clustering Algorithm*, Proceeding of BCGIN '11 Proceedings of the 2011 International Conference on Business Computing and Global Informatization, 216-219 (2011)

[21] Pei-Chann Chang and Chin-Yuan Fan , *A Hybrid System Integrating a Wavelet and TSK Fuzzy Rules for Stock Price Forecasting*, IEEE Transactions on systems, MAN, And Cybernetics – Part C: Applications and reviews, vol. 38, No. 6 (2008)

[22] Nguyễn Đức Hiển, *Ứng dụng mô hình máy học Véc-tơ tựa (SVM) trong phân tích dữ liệu điểm sinh viên*. Tạp chí Khoa học và Công nghệ - Đại học Đà Nẵng. 12(73).2013, 33-37 (2013)

(BBT nhận bài: …/…/2014, phản biện xong: …/…/2014))