Tạp chí Khoa học Trường Đại học Cần Thơ
website: sj.ctu.edu.vn

MÔ HÌNH MỜ TSK DỰ ĐOÁN GIÁ CỔ PHIẾU DỰA TRÊN MÁY HỌC VÉC-TƠ HỖ TRỢ HỒI QUY

Nguyễn Đức Hiển[1] và Lê Mạnh Thạnh[2]

[1] *Trường Cao đẳng Công nghệ Thông tin, Đại học Đà Nẵng*
[2] *Đại học Huế*

Thông tin chung:
Ngày nhận: 19/09/2015
Ngày chấp nhận: 10/10/2015

Title:
The TSK fuzzy model extracted from Support-vector-machine-for-regression for stock price forecasting

Từ khóa:
Mô hình mờ, mô hình mờ TSK, máy học véc-tơ hỗ trợ, máy học véc-tơ hỗ trợ hồi qui, dự đoán giá cổ phiếu

Keywords:
Fuzzy model, TSK fuzzy model, support vector machine, support vector machine for regeression, stock price forecasting

ABSTRACT

This paper proposes a TSK fuzzy model for stock price forecasting based on Support vector machine for regression. By uniformly satisfying these conditions between TSK fuzzy models and Support vector machines for regression, we can construct an algorithm to extract TSK fuzzy model from Support vector machines. This research does not give the algorithm that allows extracting TSK fuzzy model from support vector machine, but rather proposes a solution that allows optimization of extracted fuzzy model through the adjustment of ε parameter. The proposed model is combination of the SOM clustering algorithm and fm-SVM, the algorithm to extract TSK fuzzy model from Support vector machines. The effectiveness of the proposed solutions is evaluated by the experimental results and a comparison with the results of some other models.

TÓM TẮT

Bài báo này đề xuất một mô hình mờ TSK cho bài toán dự đoán giá cổ phiếu dựa trên mô hình máy học véc-tơ hỗ trợ hồi qui. Trên cở sở thỏa mãn các điều kiện nhằm đồng nhất giữa hàm đầu ra của mô hình mờ TSK và hàm quyết định của máy học véc-tơ hỗ trợ hồi qui, chúng ta có thể xây dựng một thuật toán cho phép trích xuất mô hình mờ TSK từ máy học véc-tơ hỗ trợ. Bên cạnh đó trong nghiên cứu này chúng tôi còn đề xuất một giải pháp cho phép tối ưu hóa mô hình mờ TSK trích xuất được thông qua việc điều chỉnh tham số ε. Mô hình đề xuất là sự kết hợp của thuật toán phân cụm SOM và thuật toán trích xuất mô hình mờ TSK từ máy học Véc-tơ hỗ trợ hồi quy. Hiệu quả của giải pháp đề xuất được đánh giá thông qua các kết quả thực nghiệm và có sự so sánh với kết quả của một số mô hình khác.

1 GIỚI THIỆU

Bài toán dự đoán giá cổ phiếu 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. Bài toán dự đoán giá cổ phiếu 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 [6][15].

Một trong những hướng tiếp cận phổ biến hiện nay để giải quyết bài toán dự đoán giá cổ phiếu là trích xuất mô hình mờ dự đoán giá cổ phiếu từ dữ liệu giao dịch lịch sử, gọi là mô hình mờ hướng dữ liệu (data-driven model). Một trong những kỹ thuật trích xuất luật mờ tự động từ dữ liệu khá hiệu quả

đó là dựa vào máy học véc-tơ hỗ trợ (Support vector machines - SVM) được nhóm tác giả J.-H Chiang và P.-Y Hao nghiên cứu và công bố lần đầu tiên trong [8]. Theo hướng tiếp cận này, nhiều tác giả đã nghiên cứu đề xuất và ứng dụng các kỹ thuật rút trích các luật mờ từ SVM cho việc phát triển các mô hình mờ hướng dữ liệu cho các bài toán phân lớp [4][9], dự báo hồi quy [12][14].

Một đặc điểm đáng lưu ý của máy học véc-tơ hỗ trợ là đối với một tập dữ liệu học nhất định, nếu điều chỉnh các tham số để tăng tính chính xác của mô hình dự đoán thì số lượng véc-tơ hỗ trợ (Support Vector - SVs) cũng tăng lên [4][5][12] [17]. Nói cách khác là khi tăng hiệu suất của mô hình thì đồng nghĩa với việc làm giảm tính "có thể diễn dịch được" (intepretability) của mô hình. Như vậy, vấn đề đặt ra là làm thế nào có thể trích xuất được hệ thống mờ đảm bảo tính chính xác trong dự đoán, đồng thời đảm bảo được đặc tính "có thể diễn dịch được". Trong bài báo này, chúng tôi đề xuất giải pháp điều chỉnh giá trị tham số ε trong mô hình máy học SVM hồi qui (ε-Support Vector Regression) để sao cho có thể đảm bảo tính chính xác của mô hình dự báo đồng thời tăng "tính có thể diễn dịch được" của mô hình mờ trích xuất được.

Các phần tiếp theo của bài báo bao gồm: phần 2 trình bày sơ lược về mô hình mờ TSK, máy học véc-tơ tựa (SVM – Support Vector Machine) và điểm tương đồng của hai mô hình này; qua đó đề xuất thuật toán fm-SVM cho phép trích xuất các luật mờ từ SVMs trong đó có tích hợp các giải pháp tối ưu hóa mô hình thông qua các tham số. Trong phần 3, chúng tôi đề xuất một mô hình mờ TSK dự đoán giá cổ phiếu dựa trên sự kết hợp giữa thuật toán phân cụm SOM (Self-Organizing Map) và thuật toán trích xuất mô hình mờ fm-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 TRÍCH XUẤT MÔ HÌNH MỜ TSK TỪ MÁY HỌC VÉC-TƠ HỖ TRỢ HỒI QUY

2.1 Mô hình mờ TSK

Mô hình mờ dạng TSK [7][9][14] còn được gọi là mô hình Takagi-Sugeno, được đề xuất bởi Takagi, Sugeno, và Kang nhằm phát triển cách tiếp cận mang tính hệ thống đối với quá trình sinh luật mờ từ tập dữ liệu vào-ra cho trước. Mô hình mờ TSK được cấu thành từ một tập các luật mờ "IF – THEN", với phần kết luận của mỗi luật này là một hàm (không mờ) ánh xạ từ các tham số đầu vào tới tham số đầu ra của mô hình.

Giả sử có một hệ thống mờ TSK với *m* luật mờ được biểu diễn như sau:

$$R_j: \text{IF } x_1 \text{ is } A_1^j \text{ and } x_2 \text{ is } A_2^j \text{ and } \dots \text{ and } x_n \text{ is } A_n^j$$

$$\text{THEN } z = g_j(x_1, x_2, \dots, x_n), \text{với } j = 1, 2, \dots, m$$

Trong đó $x_i (i = 1, 2, \dots . n)$ là các biến điều kiện; z là các biến quyết định của hệ thống mờ được xác định bởi hàm không mờ $g_j(.)$; A_i^j là những thuật ngữ ngôn ngữ xác định bởi hàm thành viên tương ứng $\mu_{A_i^j}(x_i)$. Lưu ý, $\mu_{A_i^j}(x_i)$ được định nghĩa như sau:

$$\mu_{A^j}(x_i) = \prod_{i=1}^{n} \mu_{A_i^j}(x_i) \quad (1)$$

Quá trình suy luận được thực hiện như sau:

1) Kích hoạt các giá trị thành viên.

$$\prod_{i=1}^{n} \mu_{A_i^j}(x_i) \quad (2)$$

2) Kết quả đầu ra của suy luận được tính như sau:

$$f(x) = \frac{\sum_{j=1}^{m} \bar{z}^j \left(\prod_{i=1}^{n} \mu_{A_i^j}(x_i) \right)}{\sum_{j=1}^{m} \prod_{i=1}^{n} \mu_{A_i^j}(x_i)} \quad (3)$$

Trong đó, $\bar{z}^j$ là giá trị đầu ra của hàm $g_j(.)$.

2.2 Máy học véc-tơ hỗ trợ hồi quy

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][3] và là một kỹ thuật được đề nghị để giải quyết cho các bài toán phân lớp. Từ đó, nhiều nghiên cứu đã đề xuất sử dụng SVM giải quyết bài toán tối ưu hóa hồi quy [6][11] [15][16]. 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 [1][3]:

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. Với ε-Support Vector Regression, bài toán tối ưu hóa ràng buộc cần giải quyết là:

$$\min_{w,b,\xi,\xi^*} \frac{1}{2} w^T w + C \sum_{i=1}^{l} (\xi_i + \xi_i^*) \quad (4)$$

Sao cho: $(w^T.\Phi(x_i)+b)-y_i \le \varepsilon+\xi_i$,

$$y_i-(w^T.\Phi(x_i)+b) \le \varepsilon+\xi_i^*,$$

$$\xi_i,\xi_i^* \ge 0, \text{và } i=1,2,\dots,l$$

Và đưa đến bài toán Quadratic Programming:

$$\max_{\alpha,\alpha^*} -\frac{1}{2}\sum_{i,j}(\alpha_i-\alpha_i^*)(\alpha_j-\alpha_j^*)(\Phi(x_i)^T.\Phi(x_j)) - \varepsilon\sum_{i=1}^{l}(\alpha_i+\alpha_i^*)-\sum_{i=1}^{l}y_i(\alpha_i+\alpha_i^*) \quad (5)$$

Sao cho:

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

Trong đó, C là tham số chuẩn tắc, ε là sai số cho phép, ξ_i, ξ_i^* là những biến lỏng, và α_i, α_i^* là những nhân tử Lagrange.

Véc-tơ w có dạng:

$$w=\sum_{i=1}^{l}(\alpha_i-\alpha_i^*).x_i \quad (6)$$

Và hàm quyết định là:

$$f(x)=\sum_{i=1}^{l}(\alpha_i-\alpha_i^*)\,(\Phi(x_i)^T.\Phi(x_j))+b \quad (7)$$

Gọi $K(x_i,x_j)=\Phi(x_i)^T.\Phi(x_j)$ là hàm nhân của không gian đầu vào; và hàm quyết định (7) được viết lại như sau:

$$f(x)=\sum_{i=1}^{l}(\alpha_i-\alpha_i^*)K(x_i,x)+b \quad (8)$$

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

2.3 Trích xuất mô hình mờ TSK

Xét hàm đầu ra của mô hình mờ TSK (3) và hàm quyết định của mô hình máy học Véc-tơ hồi quy (8). Để (3) và (8) đồng nhất với nhau, trước tiên chúng ta phải đồng nhất giữa hàm nhân trong (8) và hàm thành viên trong (3). Ở đây, để thỏa mãn điều kiện Mercer [13] hàm thành viên Gauss được chọn làm hàm nhân; đồng thời giá trị của b trong (8) phải bằng 0.

Khi hàm Gauss được chọn làm hàm thành viên và hàm nhân, đồng thời số luật mờ bằng với số véc-tơ hỗ trợ (m = l) thì (3) và (8) trở thành:

$$f(x)=\sum_{i=1}^{l}(\alpha_i-\alpha_i^*)\exp\left(-\frac{1}{2}\left(\frac{x_i-x}{\sigma_i}\right)^2\right) \quad (9)$$

và

$$f(x)=\frac{\sum_{j=1}^{l}\bar{z}^j\exp\left(-\frac{1}{2}\left(\frac{x_j-x}{\sigma_j}\right)^2\right)}{\sum_{j=1}^{l}\exp\left(-\frac{1}{2}\left(\frac{x_j-x}{\sigma_j}\right)^2\right)} \quad (10)$$

Như cách biến đổi trong [8], hàm suy luận mờ (10) có thể viết lại như sau:

$$f(x)=\sum_{j=1}^{l}\bar{z}^j\exp\left(-\frac{1}{2}\left(\frac{x_j-x}{\sigma_j}\right)^2\right) \quad (11)$$

Và chúng ta chọn: $\bar{z}^j=(\alpha_i-\alpha_i^*)$ (12)

Như vậy, trên cơ sở thỏa mãn các điều kiện để đồng nhất hàm đầu ra của SVMs và hệ thống mờ TSK, chúng ta có thể trích xuất được mô hình mờ TSK từ máy học Véc-tơ hỗ trợ.

2.4 Tối ưu hóa tham số của các hàm thành viên

Những tham số của hàm thành viên có thể được tối ưu hóa dùng những thuật toán gradient descent hoặc thuật toán di truyền (GAs) [8][9]. Trong trường hợp này, để nhận được tập mờ tối ưu, chúng tôi cập nhật gía trị các tham số của hàm thành viên theo các hàm thích nghi sau đây:

$$\sigma_i(t+1)=\sigma_i(t)\delta\varepsilon_{1,i}\left[\frac{(x-c)^2}{\sigma^3}\exp\left(-\frac{(x-c)^2}{2\sigma^2}\right)\right] \quad (13)$$

$$c_i(t+1)=c_i(t)\delta\varepsilon_{1,i}\left[\frac{-(x-c)}{\sigma^2}\exp\left(-\frac{(x-c)^2}{2\sigma^2}\right)\right] \quad (14)$$

2.5 Tối ưu hóa mô hình bằng tham số ε

Một trong những đặc điểm của mô hình mờ là "tính có thể diễn dịch được" [7]. Tuy nhiên, đối với mô hình máy học véc-tơ hỗ trợ nếu 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 diễn dịch được" của hệ thống mờ giảm đi.

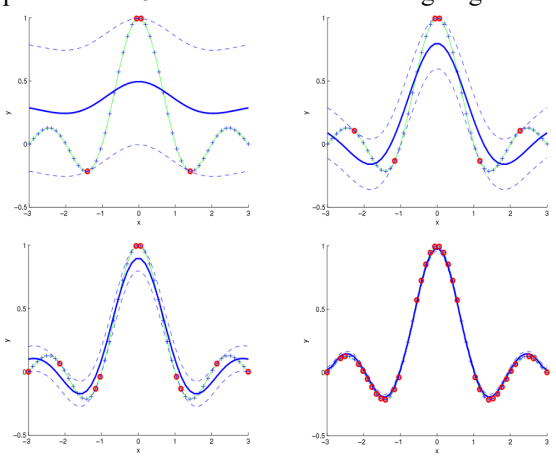
Xét kết quả thực nghiệm mô hình máy học véc-tơ hồi quy trên hàm hồi qui phi tuyến Sinc(x) được cho bởi công thức sau:

$$Sinc(x) = \begin{cases} \frac{\sin(x)}{x} & if\ x \neq 0 \\ 1 & if\ x = 0 \end{cases} \quad (15)$$

Tập dữ liệu huấn luyện được xác định trong phạm vi từ -3π đến +3π.

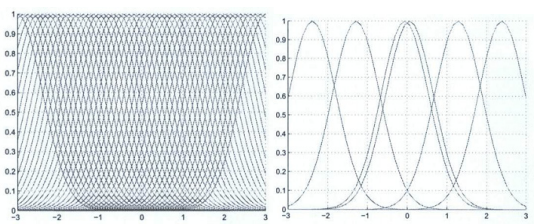
Trong quá trình xác định cấu trúc SVM, chúng tôi sử dụng tham số ε để điều chỉnh số lượng SVs. Theo kết quả thể hiện ở Hình 1 chúng ta thấy: khi giá trị của tham số ε giảm đi thì số lượng SVs cũng tăng lên, đồng thời độ chính xác của kết quả dự đoán cũng tăng lên (đường đậm nét là đường dự đoán hồi quy, đường đánh dấu + là đường biểu diễn giá trị dữ liệu đúng).

Bằng cách giữ cố định giá trị tham số C = 10. Khi giá trị ε = 0.0, sẽ có 50 SVs nhận được từ mô hình, đồng nghĩa với việc chúng ta nhận được 50 luật mờ (chú ý rằng, trong trường hợp này tất cả các mẫu dữ liệu huấn luyện được chọn làm SVs đầu ra). Hình 2a thể hiện phân bố của 50 hàm thành viên mờ tương ứng trong trường hợp này. Khi tăng giá trị tham số ε = 0.1, thì có 6 SVs nhận được tương ứng với 6 luật mờ. Hình 3b thể hiện phân bố của 6 hàm thành viên mờ tương ứng.



Hình 1: Mối quan hệ giữa số lượng SVs và tham số ε

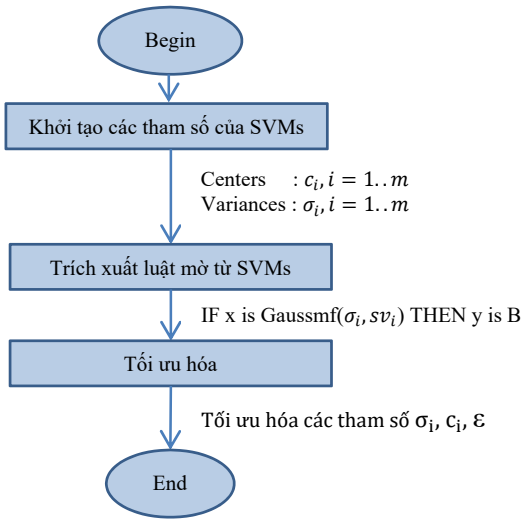
Giá trị của ε tương ứng theo thứ tự các hình vẽ là 0.5, 0.2, 0.1 và 0.01



(a) (b)

Hình 2: Phân bố của 50 và 6 hàm thành viên mờ

Từ những phân tích trên, chúng tôi đã đề xuất thuật toán fm-SVM cho phép trích xuất mô hình mờ TSK từ máy học véc-tơ hỗ trợ như thể hiện ở Hình 3.

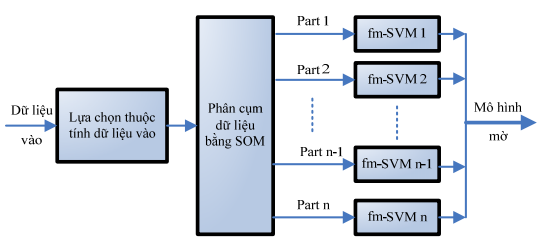


Hình 3: Sơ đồ khối thuật toán fm-SVM

Trong thuật toán này, ngoài việc tối ưu hóa các tham số của hàm thành viên, chúng ta có thể điều chỉnh giá trị tham số ε để nhận được tập luật mờ tối ưu.

3 MÔ HÌNH MỜ TSK DỰ ĐOÁN GIÁ CỔ PHIẾU

Trong phần này, chúng tôi đề xuất một mô hình trích xuất luật mờ dự đoán giá cổ phiếu bằng cách sử dụng thuật toán fm-SVM đã đề xuất. Sơ đồ khối của mô hình được thể hiện trong Hình 4.



Hình 4: Mô hình mờ dự đoán giá cổ phiếu

3.1 Lựa chọn thuộc tính đầu vào

Theo những kết quả nghiên của các tác giả khác về việc dự đoán giá cổ phiếu có nhiều cách khác nhau để lựa chọn thuộc tính đầu vào, ví dụ như: sử dụng các chỉ số kinh tế vi mô, hoặc sử dụng các chỉ số giá cổ phiếu hàng ngày <opening, high, low, closing price> [5][6][11] [15],… 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 [6][11], 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. Tay. Bảng 1 thể hiện các thuộc tính lựa chọn và công thức tính của chúng.

Bảng 1: Các thuộc tính lựa chọn

Ký hiệu	Thuộc tính	Công thức tính
x_1	EMA100	$P_i - \overline{EMA_{100}(\iota)}$
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	$\overline{(P(\iota+5) - P(\iota))}/\overline{P(\iota)} * 100$ $\overline{P(\iota)} = \overline{EMA_3(\iota)}$

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 SOM

Gần đây, nhiều nghiên cứu của các tác giả khác đã đề xuất sử dụng SOM như là một giải pháp khá hiệu quả để phân cụm dữ liệu, đặc biệt là đối với dữ liệu thị trường chứng khoán [6][15]. Trong nghiên cứu này, chúng tôi sử dụng SOM để phân dữ liệu đầu vào thành các cụm theo sự tương đương phân bố thống kê của các điểm dữ liệu. Kết quả phân cụm bởi SOM sẽ giúp giải quyết được hai vấn đề [6]:

1) Kích thước dữ liệu trong từng cụm sẽ nhỏ hơn làm tăng tốc độ học của mô hình.

2) Dữ liệu trong các cụm có sự tương đương trong phân bố thống kê, như vậy sẽ hạn chế được các trường hợp nhiễu.

3.3 Trích xuất mô hình mờ bằng fm-SVM

Mỗi cụm dữ liệu vào đã được phân tách bằng SOM sẽ được đưa vào huấn luyện cho từng máy fm-SVM tương ứng để trích xuất các luật mờ. Các tập luật mờ trích xuất được từ các máy fm-SVM tương ứng với các cụm dữ liệu huấn luyện có thể được sử dụng để suy luận dự đoán giá cổ phiếu. Những luật mờ khai phá được từ dữ liệu đã được phân thành các cụm riêng biệt và được cải thiện tính "có thể diễn dịch được", như vậy các chuyên gia con người có thể diễn dịch thành luật ngôn ngữ và từ đó có thể hiểu và đánh giá được các luật này.

4 KẾT QUẢ THỰC NGHIỆM

Để đánh giá mô hình đề xuất, chúng tôi xây dựng một hệ thống thử nghiệm dựa trên bộ công cụ Matlab. Thuật toán học SVM của thư viện LIBSVM được phát triển bởi nhóm của Chih-Wei Hsu [2], được sử dụng để sản sinh ra các SVs từ dữ liệu huấn luyện, làm cơ sở để xây dựng thuật toán trích xuất các luật mờ fm-SVM. Việc phân cụm dữ liệu đầu vào được thực hiện dựa trên bộ công cụ SOM được phát triển bởi Juha Vesanto và các đồng sự [10]. Sau cùng, chúng tôi sử dụng hàm AVALFIS trong thư viện công cụ Matlab Fuzzy Logic để suy luận dự báo giá cổ phiếu dựa vào các luật mờ sản xuất được.

Bảng 2: 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

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 2). 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/.

Các tập dữ liệu training sẽ được dùng để trích xuất các tập luật mờ. Bảng 3 thể hiện một nhóm luật mờ trích xuất được từ dữ liệu training của mã cổ phiếu TTC.

Bảng 3: 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

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 SVM, mô hình kết hợp SOM+SVM và SOM+f-SVM. Mô hình SOM+SVM là mô hình dựa trên sự kết hợp của SOM và SVM, được đề xuất để dự đoán xu hướng cổ phiếu trong [6][15]. Mô hình SOM+f-SVM là mô hình kết hợp SOM với f-SVM thuần túy (chưa điều chỉnh tham số ε). 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.

Kết quả thực nghiệm dự đoán trên 200 mẫu dữ liệu testing được thể hiện trong Bảng 4.

So sánh giá trị các thông số MNSE và MAE trong Bảng 4 ta thấy, trên cả 5 mã cổ phiếu, giá trị các thông số MNSE và MAE của mô hình SOM+fm-SVM đề xuất là nhỏ hơn so với mô hình 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 mô hình SVM, điều này chứng tỏ tỷ lệ dự đoán đúng xu hướng của mô hình đề xuất cao hơn.

Bảng 4: Kết quả dự đoán trên 200 mẫu dữ liệu thử nghiệm

Mã cổ phiếu	SVM			SOM+SVM				SOM+f-SVM				SOM+fm-SVM		
	NMSE	MAE	DS	NMSE	MAE	DS	Số luật	NMSE	MAE	DS	Số luật	NMSE	MAE	DS
HASTC	0.9278	0.0191	38.31	0.9057	0.0188	41.71	561	0.7601	0.0164	44.72	6*25	0.7615	0.0181	44.02
VN INDEX	1.0725	0.0110	34.33	1.1726	0.0109	42.68	816	1.1408	0.0108	42.21	6*31	1.1401	0.0115	42.31
TTC	1.2687	0.0394	38.90	1.1358	0.0392	42.71	476	1.1390	0.0391	42.81	6*22	1.1452	0.0411	42.75
SGH	1.1015	0.0576	38.31	1.0792	0.0573	41.71	691	1.0909	0.0646	42.71	6*27	1.0851	0.0602	41.85
DXP	1.2073	0.0242	39.83	1.1138	0.0258	45.72	652	1.1281	0.0254	45.22	6*27	1.1390	0.0301	45.43

So sánh kết quả của mô hình SOM+fm-SVM đề xuất với mô hình SOM+SVM và SOM+f-SVM trong Bảng 4, 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 các thuật toán f-SVM và fm-SVM 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 SOM với f-SVM và fm-SVM sẽ thừa hưởng hiệu quả của mô hình SOM+SVM là điều tất yếu. Tuy nhiên, so với mô hình dự đoán SOM+SVM thì các mô hình mờ TSK có những ưu điểm sau:

1) Mô hình dự đoán SOM+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 trình suy luận sẽ được thực hiện trên tập luật này. Đối với người dùng cuối thì mô hình suy luận dựa trên một tập luật mờ sẽ dễ hiểu và sáng tỏ hơn.

2) Ngoài ra, việc áp dụng SOM 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 dễ dàng đọc hiểu và phân tích các luật mờ học được.

Điểm cải thiện của mô hình dựa trên fm-SVM so với mô hình dựa trên f-SVM chính là số luật mờ trích xuất được trong từng mô hình dự đoán. Ví dụ, đối với mã cổ phiếu HATC, tổng số luật mờ theo mô hình SOM+f-SVM là 561, trong theo mô hình SOM+fm-SVM chỉ là 6*25. Như vậy, số luật mờ của mô hình đề xuất đã giảm đi rất nhiều so với mô hình SOM+f-SVM, trong khi tính chính xác của kết quả dự đoán vẫn được đảm bảo.

5 KẾT LUẬN

Trong nghiên cứu này đề xuất một mô hình dự đoán giá cổ phiếu dựa trên sự kết hợp của SOM và fm-SVM. Kết quả thực nghiệm trên dữ liệu thử nghiệm cho thấy mô hình đề xuất thật sự mang lại hiệu quả thể hiện ở chỗ: độ chính xác của kết quả dự đoán cao hơn hoặc tương tương đương so với các mô hình khác, thể hiện qua các giá trị của các thông số NMSE, MAE và DS, trong khi đó thì số lượng luật mờ của các mô hình được rút gọn đáng kể. Như đã trình bày ở phần 4 của bài báo, một trong những hiệu quả mang lại của việc rút gọn và gom cụm các luật mờ trích xuất được là sẽ giảm độ phức tạp trong quá trình suy luận, đồng thời giúp cho việc diễn dịch và phân tích các luật này dễ dàng hơn.

Việc phân tích ngữ nghĩa tập luật mờ trích xuất từ dữ liệu, còn gọi là luật mờ hướng dữ liệu, sẽ giúp cho các chuyên gia con người đánh giá được tập luật; qua đó có thể lựa chọn một số ít luật chuyên gia để bổ sung vào tập luật mờ hướng dữ liệu. Vấn đề khó khăn gặp phải chính là việc đồng bộ giữa phân hoạch mờ hướng dữ liệu và phân hoạch mờ theo chuyên gia; đây chính là cơ sở để có thể tích hợp luật chuyên gia với luật mờ hướng dữ liệu. Trong những nghiên cứu tiếp theo, chúng tôi sẽ nghiên cứu các giải pháp làm sáng tỏ phân hoạch mờ của tập luật mờ hướng dữ liệu, đồng bộ với phân hoạch mờ theo chuyên gia, từ đó có thể tích hợp luật chuyên gia với tập luật mờ hướng dữ liệu nhằm nâng cao hiệu quả dự đoán.

TÀI LIỆU THAM KHẢO

1. Alex J. Smola, Bernhard Scholkopf, 2004. A Tutorial on Support Vector Regression, Statistics and Computing 14: 199–222 .
2. Chih-Wei Hsu, Chih-Chung Chang, Chih-Jen lin, 2010. A practical Guide to Support Vector Classification, http://www.csie.ntu.edu.tw/~ cjlin/libsvm/
3. Corinna Cortes and Vladimir Vapnik, 1995. Support-Vector Networks. Machine Learning, 20: 273-297.
4. David Martens, Johan Huysmans, Rudy Setiono, Jan Vanthienen, Bart Baesens, 2008. Rule Extraction from Support Vector Machines - An Overview of issues and application in credit scoring, Studies in Computational Intelligence (SCI) 80: 33–63.
5. Duc-Hien Nguyen, Manh-Thanh Le, 2013. A two-stage architecture for stock price forecasting by combining SOM and fuzzy-SVM, International Journal of Computer Science and Information Security (IJCSIS), USA, ISSN: 1947-5500, Vol. 12 No. 8: 20-25.
6. Francis Eng Hock Tay and Li Yuan Cao, 2001. Improved financial time series forecasting by combining Support Vector Machines with self-organizing feature map, Intelligent Data Analysis 5, IOS press: 339-354.
7. John Yen, Reza Langari,1999. Fuzzy logic: Intelligence, Control, and Information, Prentice hall, Uper dadle river, New Jersey.
8. J.-H Chiang and P.-Y Hao, 2004. Support vector learning mechanism for fuzzy rule-based modeling: a new approach, IEEE Trans. On Fuzzy Systems, vol. 12: 1-12.
9. J.L. Castro, L.D. Flores-Hidalgo, C.J. Mantas and J.M. Puche, 2007. Extraction of fuzzy rules from support vector machines, Elsevier. Fuzzy Sets and Systems, 158: 2057 – 2077.
10. Juha Vesanto, Johan Himberg, Esa Alhoniemi, Jaha Parhankangas, 2000. SOM Toolbox for Matlab 5, http://www.cis.hut.fi/projects/som-toolbox/.
11. L.J.Cao and Francis E.H.Tay, 2003. Support vector machine with adaptive parameters in Financial time series forecasting, IEEE trans. on neural network,vol. 14, no. 6.
12. Nahla Barakat, Andrew P. Bradley, 2010. Rule extraction from support vector machines: A review, Neurocomputing – ELSEVIER, 74: 178–190.
13. R. Courant, D. Hilbert,1953. Methods of Mathematical Physics, Wiley, New York.
14. S. Chen, J. Wang and D. Wang, 2008. Extraction of fuzzy rules by using support vector machines, IEEE, Computer society: 438-441.

15. Sheng-Hsun Hsu, JJ Po-An Hsieh, Ting-Chih CHih, Kuei-Chu Hsu, 2009. A two-stage architecture for stock price forecasting by integrating self-organizing map and support vector regression, Expert system with applications 36: 7947-7951.
16. Wang-Hsin Hsu, Yi-Yuan Chiang, Wen-Yen Lin, Wei-Chen Tai, and Jung-Shyr Wu, 2009. SVM-based Fuzzy Inference System (SVM-FIS) for Frequency Calibration in Wireless Networks, CIT'09 Proceedings of the 3rd international conference on comminications and information technology: 207-213.
17. Nguyễn Đức Hiển, 2013. 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.