Một Giải Pháp Cải Thiện Hiệu Quả Phân Cụm Bằng SOM Trong Phân Cụm Dữ Liệu Tài Chính

Nguyễn Đức Hiển

Trường Đại học Công nghệ thông tin và Truyền thông Việt – Hàn, Đại học Đà Nẵng
ndhien@vku.udn.vn

Tóm tắt. Trong bài toán dự báo dữ liệu chuỗi thời gian tài chính, kỹ thuật phân cụm SOM được sử dụng để phân cụm dữ liệu đầu vào trước khi đưa vào huấn luyện các mô hình dự báo. Những thực nghiệm trên dữ liệu thực tế cho thấy kết quả phân cụm bằng SOM chưa thật sự tốt. Bài báo này đề xuất một giải pháp điều chỉnh phân cụm sau khi huấn luyện mạng SOM nhằm cải thiện hiệu quả phân cụm dữ liệu trong bài toán phân cụm dữ liệu tài chính. Hiệu quả của thuật toán SOM* đề xuất được đánh giá trên cơ sở so sánh với thuật toán SOM nguyên thủy, qua thực nghiệm trên tập dữ liệu thật lấy từ sàn chứng khoán Yahoo Finance.

Từ khóa: SOM, dự báo giá cổ phiếu, mô hình dự báo.

Abstract. In terms of the time series forecasting problem, the SOM clustering technique is used to cluster the input data before entering and utilizing it to train the forecasting model. The results of the conducted experiments on the real data showed that the clustering outputs by the SOM are not fully optimized. This paper proposes a solution to customize the clusters which are trained by the SOM to improve the effectiveness of the data clustering in the financial data clustering problem. The effectiveness of the proposed algorithm — SOM*, is assessed based on the comparison with the original SOM algorithm, through experiments on real data sets extracting from the Yahoo Finance stock exchange.

Keywords: SOM, Stock price forecasting, Forecasting models.

1 Đặt vấn đề

Vấn đề dự báo theo chuỗi thời gian, mà đặc biệt là vấn đề dự báo giá cổ phiếu đã và đang thu hút được nhiều sự quan tâm nghiên cứu của các nhà khoa học. Những hướng tiếp cận phổ biến hiện nay cho vấn đề dự báo dữ liệu thời gian tài chính là khai phá dữ liệu, ứng dụng các mô hình máy học thống kê [3][5][6]. Những nghiên cứu gần đây chủ yếu tập trung vào hướng cải tiến và kết hợp nhiều phương thức học khác nhau để nâng cao hiệu quả dự báo, như mô hình kết hợp SVM và SOM (Self-Organizing Map) [1][4], kết hợp HNN, AMN và GA [5], kết hợp K-means và SVM [15]. Mô hình kết hợp giữa hệ thống mờ (Fuzzy modeling) và SVM là một hướng nghiên cứu mới của mô hình mờ, gọi là mô hình mờ hướng dữ liệu (data-driven models) [10], [11], [12], [13], [14], nó cho phép trích xuất các luật mờ từ SVMs để làm cơ sở cho hệ thống dự báo mờ. Một trong những thách thức của mô hình hướng dữ liệu là vấn đề học tự động từ dữ liệu huấn luyện với kích thước lớn và thiếu tính đặc trưng, và tiếp đến là sự bùng nổ tập luật mờ học được cũng là điều khó tránh khỏi.

Một trong những hướng nghiên cứu nhằm giải quyết vấn đề kích thước dữ liệu lớn trong mô hình hướng dữ liệu là kết hợp một giải thuật phân cụm dữ liệu, như k-Means, SOM (Self-Organizing Map), SVM,… để chuyển thành các bài toán với kích thước dữ liệu nhỏ hơn [1], [4], [15]. Một trong những kết quả gần đây của nhóm nghiên cứu đã đề xuất một mô hình kết hợp SOM và SVMs để trích xuất các luật mờ cho bài toán dự báo dữ liệu chuỗi thời gian tài chính. Việc sử dụng SOM để phân cụm dữ liệu trước khi thực hiện trích xuất mô hình mờ sẽ giúp giải quyết được hai vấn đề [1], [4], [10], [15]:

1) Kích thước dữ liệu trong từng phân cụm sẽ nhỏ hơn làm tăng tốc độ huấn luyện mô hình.

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

Tuy nhiên, Kỹ thuật phân cụm SOM là một kỹ thuật máy học không giám sát được ứng dụng nhiều trong các bài toán phân cụm dữ liệu [1], [7], [9]. Nhiều nghiên cứu gần đây đã khẳng định kỹ thuật phân cụm SOM mang lại hiệu quả trong các trường hợp giả quyết bài toán khai phá dữ liệu với các tập dữ liệu lớn [6], [10], [14]. Tuy nhiên thực tế kết quả huấn luyện mạng SOM phụ thuộc vào tập dữ liệu huấn luyện đôi khi thiếu tính đặc trưng, không bao phủ được không gian bài toán, dẫn đến một số trường hợp dữ liệu bị phân cụm lệch. Trong nghiên cứu này, tác giả đề xuất giải pháp điều chỉnh kết quả phân cụm của SOM để đảm bảo các phân cụm dữ liệu được phân bố tốt hơn, qua đó có thể cải thiện hiệu quả huấn luyện và ứng dụng mô hình dự báo. Phần tiếp theo của bài báo sẽ giới thiệu về mô hình dự báo dữ liệu chuỗi thời gian tài chính có ứng dụng SOM trong việc phân cụm dữ liệu đầu vào. Phần thứ 3 của bài báo trình bày giải pháp điều chỉnh kết quả phân cụm bằng SOM trong mô hình. Phần thứ 4 là một số kết quả thực nghiệm và bàn luận. Phần cuối là kết luận và một số đề xuất.

2 Mô hình dự báo sử dụng kỹ thuật phân cụm SOM

Để giải quyết bài toán dự báo dữ liệu chuỗi thời gian tài chính, nhóm nghiên cứu đã đề xuất mô hình lai ghép giữa kỹ thuật phân cụm SOM và thuật toán trích xuất mô hình mờ từ máy học véc-tơ hỗ trợ hồi quy [10 … 14], thể hiện ở hình vẽ Fig.1. Theo đó, tập dữ liệu đầu vào được phân chia thành các cụm tách rời bằng kỹ thuật phân cụm SOM trước khi ứng dụng thuật toán trích xuất mô hình mờ dựa trên máy học véc-tơ hỗ trợ để trích xuất ra các mô hình mờ.

Quá trình thực hiện dự báo giá cổ phiếu theo mô hình đề xuất được thể hiện qua hai đoạn như sau:

Giai đoạn 1: Huấn luyện mô hình bằng tập dữ liệu huấn luyện

Bước 1. Lựa chọn thuộc tính dữ liệu đầu vào và đầu ra

Bước 2. Phân cụm tập dữ liệu huấn luyện bằng SOM (n phân cụm)

Bước 3. Sử dụng thuật toán f-SVM hoặc SVM-IF để trích xuất ra các mô hình mờ TSK cho mỗi phân cụm dữ liệu

Bước 4. Thực nghiệm dự báo trên tập dữ liệu xác thực để chọn giá trị tối ưu cho các tham số ε, số phân cụm n

Bước 5. Trích xuất ra các mô hình mờ cho các phân cụm

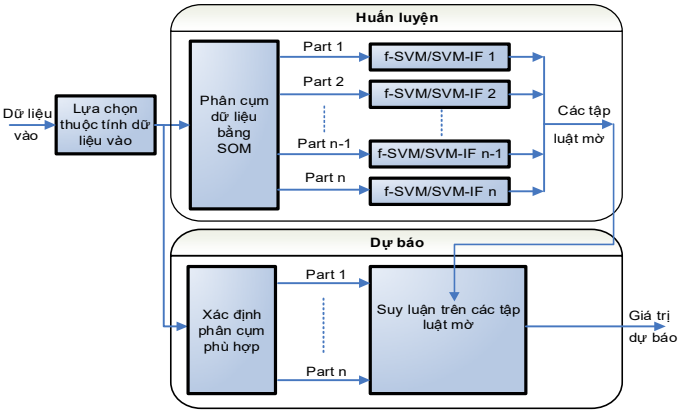

Fig. 1. Mô hình thực nghiệm cho bài toán dự báo dữ liệu chuỗi thời gian tài chính

Giai đoạn 2: Thực hiện dự báo trên tập dữ liệu thử nghiệm

Bước 1. Xác định phân cụm tương ứng với từng mẫu dữ liệu của tập thử nghiệm

Bước 2. Thực hiện dự báo trên tập dữ liệu thử nghiệm

Bước 3. Tính toán các sai số trên kết quả dự báo để đánh giá mô hình

Ở bước thứ 2 trong giai đoạn huấn luyện của mô hình dự báo, tập dữ liệu huấn luyện được phân thành n phân cụm. Tuy nhiên, qua nhiều thực nghiệm trên dữ liệu thật, kết quả phân cụm các tập dữ liệu cho kích thước phân cụm không cân đối. Điều này hoàn toàn có cơ sở, vì kết quả phân cụm bằng SOM phụ thuộc nhiều vào tính ngẫu nhiên của tập dữ liệu huấn luyện. Với mục tiêu của việc phân cụm là phân chia tập dữ liệu đầu vào thành nhiều phân cụm có sự tương đồng về phân bố thống kê và kích thước mỗi phân cụm là nhỏ so với tập dữ liệu đầu vào, nghiên cứu này đề xuất giải pháp điều chỉnh kết quả phân cụm sau khi thực hiện phân cụm bằng SOM. Phần tiếp theo của bài báo sẽ trình bày giải pháp điều chỉnh phân cụm này.

3 Điều chỉnh phân cụm tập dữ liệu chuỗi thời gian tài chính

Nghiên cứu này kế thừa sử dụng thuật toán phân cụm SOM đã được chuẩn hóa trong bộ công cụ SOM Toolbox 2.0, được phát triển bởi Juha Vesanto, Esa Alhoniemi và các đồng sự [7]. Theo đó, bản đồ phân cụm (sMap) được huấn luyện dựa vào tập dữ liệu huấn luyện, sử dụng hàm som_make(). Việc xác định một mẫu dữ liệu (nơ-ron) phù hợp nhất (bmu – Best-Matching Unit) với trung tâm của phân cụm (nơ-ron chiếm lĩnh) nào và thứ tự các phân cụm phù hợp được thực hiện bằng hàm som_bmus(). Trên cơ sở đó, nghiên cứu đề xuất thuật toán SOM*, cho phép điều chỉnh phân cụm sau khi huấn luyện SOM như ở hình vẽ Fig.2.

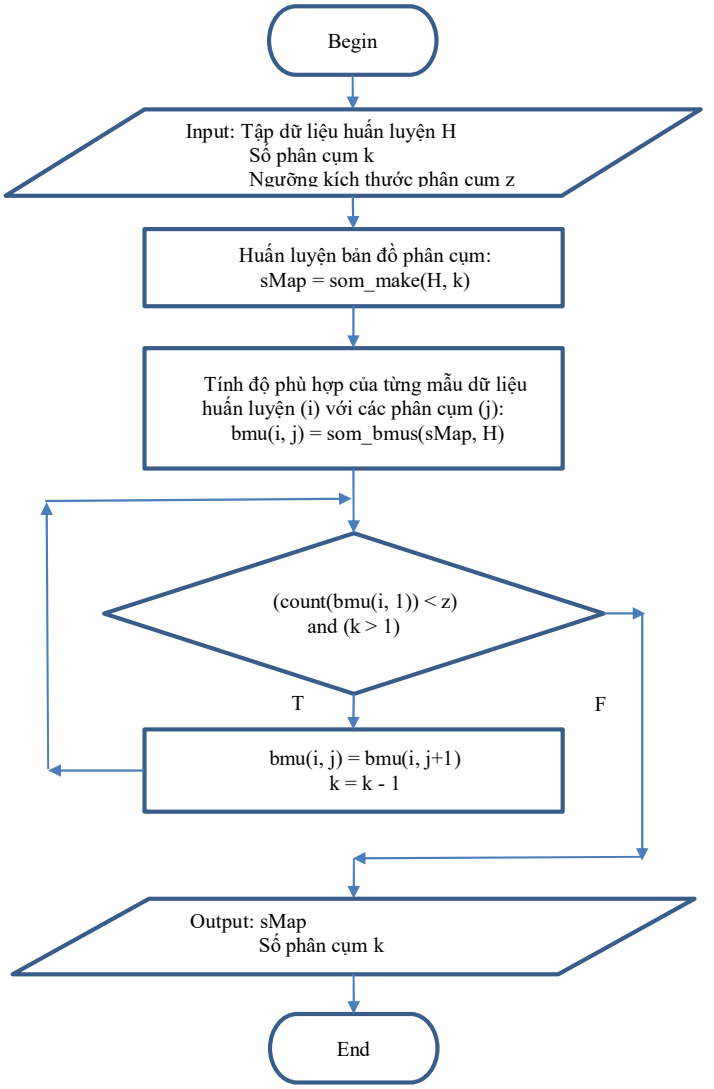

Fig. 2. Thuật toán phân cụm SOM* có điều chỉnh phân cụm

Đầu vào của thuật toán SOM* là tập dữ liệu huấn luyện H, có n mẫu dữ liệu. Ngoài ra thuật toán cũng cần xác định trước số phân cụm k và ngưỡng kích thước z của mỗi phân cụm. Ngưỡng kích thước phân cụm z được thiết lập để đảm bảo phân cụm dữ liệu đủ cho việc huấn luyện mô hình dự báo trong bước tiếp theo của mô hình ở Hình Fig.1. Trong thuật toán, ma trận bmu(i,j) cho biết thứ tự phân cụm phù hợp của mẫu dữ liệu thứ i là các phân cụm bmu(i, j). Có nghĩa rằng, nếu $j^1 < j^2$ thì mẫu dữ liệu i phù hợp với phân cụm bmu(i, j^1) hơn so với phân cụm bmu(i, j^2). Mặc nhiên, mỗi mẫu dữ liệu sẽ được phân vào phân

cụm phù hợp nhất (phân cụm j=bmu(i,1)). Tuy nhiên, nếu số mẫu dữ liệu phù hợp với phân cụm j bé hơn ngưỡng kích thước phân cụm z, thì mẫu dữ liệu tương ứng sẽ được điều chỉnh sang phân cụm có mức độ phù hợp kế tiếp (bmu(i,2)). Kết thúc thuật toán sẽ cho kết quả là một bản đồ phân cụm sMap và số phân cụm k đã được điều chỉnh.

4 Một số kết quả thực nghiệm và bàn luận

Nghiên cứu triển khai mô hình thực nghiệm trên công cụ Matlab. Nguồn dữ liệu thực nghiệm là mã cổ phiếu The Standard & Poor's stock index (S&P500) được thu thập trực tiếp từ kho dữ liệu lịch sử của sàn chứng khoán Yahoo Finance (http://finance.yahoo.com/). Dữ liệu được thu thập và sử dụng là giá đóng phiên của mã cổ phiếu, trong khoảng thời gian từ 03/01/2000 đến 23/12/2008 và được tiền xử và lựa chọn các biến đầu vào và đầu ra theo đề xuất của Tay và Cao trong [1] (công thức xác định các biến đầu vào và ra thể hiện trong bảng Tab. 1). Kết quả dữ liệu sau khi tiền xử lý sẽ được trích lập thành các tập 2016 mẫu dữ liệu dùng để huấn luyện, 200 mẫu dữ liệu dùng để xác thực mô hình và 200 mẫu dùng để kiểm thử đánh giá mô hình.

Tab. 1. Lựa chon các biến đầu vào và ra.

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	$(\overline{P(i+5)} - \overline{P(i)})/\overline{P(i)} * 100$ $với\ \overline{P(i)} = \overline{EMA_3(i)}$

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*. Các thuộc tính tương ứng với x_1->x_5 là các biến đầu vào và y là biến đầu ra cần dự đoán.

Kết quả các trường hợp phân cụm tập dữ liệu huấn luyện của mã cổ phiếu S&P500 được thể hiện ở bảng Tab. 2. Số liệu thể hiện ở bảng Tab. 3 cho thấy, khi áp dụng thuật toán SOM để phân cụm dữ liệu thì tồn tại một số phân cụm có kích thước cụm dữ liệu nhỏ (<30). Khi áp dụng thuật toán phân cụm SOM* thì số các phân cụm có kích thước nhỏ hơn ngưỡng giá trị z (z=30) đã được loại bỏ. Cụ thể trong Bảng Tab.3, trong trường hợp chọn số phân cụm ban đầu k=10, kết quả phân cụm bằng SOM nguyên thủy có 2 phân cụm kích kích thước <30; trong khi đó kết quả phân cụm bằng SOM* đã giảm số phân cụm thành 8 và không có phân cụm <30. Tương tự trong trường hợp chọn số phân cụm ban đầu là k=8, kết quả phân cụm bằng SOM* đã giảm số phân cụm thành 6 so với 8 phân cụm ban đầu của SOM nguyên thủy.

Tab. 2. Kết quả số mẫu dữ liệu theo từng phân cụm của tập dữ liệu S&P500.

Thuật toán phân cụm	SOM		SOM* (z = 30)	
Số phân cụm	10	8	10	8
Cụm số 1	253	198	398	135
Cụm số 2	195	364	211	411
Cụm số 3	198	15	178	397
Cụm số 4	19	421	211	101
Cụm số 5	155	205	173	494
Cụm số 6	231	22	423	478
Cụm số 7	12	417	97	#

Cụm số 8	314	374	325	#
Cụm số 9	202		#	
Cụm số 10	437		#	

Xét về độ phức tạp tính toán, thuật toán phân cụm SOM với số phân cụm k, kích thước tập dữ liệu huấn luyện là n và số lần thực hiện lặp lại điều chỉnh cấu trúc mạng sMap là t, thì được đánh giá có độ phức tạp O(n. k. t) [7]. Trong khi đó, đoạn thuật toán điều chỉnh phân cụm cho tập dữ liệu huấn luyện có độ phức tạp O(n. k). Như vậy độ phức tạp tổng thể của thuật toán SOM* là O(n. k. t).

Tab. 3. Tập luật trong 1 phân cụm của mã cổ phiếu S&P500.

Luật	Chi tiết
Rule#1	*IF x1=Gaussmf(0.10,-0.02) and x2=Gaussmf(0.10,-0.08) and x3=Gaussmf(0.10,0.02) and x4=Gaussmf(0.10,0.04) and x5=Gaussmf(0.10,0.02) THEN y=-0.02*
Rule#2	*IF x1=Gaussmf(0.10,0.02) and x2=Gaussmf(0.09,-0.00) and x3=Gaussmf(0.10,0.06) and x4=Gaussmf(0.10,0.05) and x5=Gaussmf(0.09,0.00) THEN y=0.04*
Rule#3	*IF x1=Gaussmf(0.09,-0.04) and x2=Gaussmf(0.10,0.07) and x3=Gaussmf(0.09,-0.16) and x4=Gaussmf(0.09,-0.14) and x5=Gaussmf(0.11,-0.05) THEN y=0.16*
Rule#4	*IF x1=Gaussmf(0.09,0.01) and x2=Gaussmf(0.10,0.08) and x3=Gaussmf(0.09,-0.06) and x4=Gaussmf(0.09,-0.09) and x5=Gaussmf(0.09,-0.04) THEN y=0.01*
Rule#5	*IF x1=Gaussmf(0.09,-0.05) and x2=Gaussmf(0.09,0.04) and x3=Gaussmf(0.10,-0.13) and x4=Gaussmf(0.10,-0.08) and x5=Gaussmf(0.08,-0.04) THEN y=-0.18*

Hiệu quả của việc áp dụng thuật toán SOM* là tinh chỉnh được kích thước tập dữ liệu trong từng phân cụm theo giá trị ngưỡng z được thiết lập trước. Với các phân cụm có kích thước dữ liệu đủ lớn (theo giá trị ngưỡng z) và có sự tương đồng nhất định về phân bố thống kê, sẽ mang lai hiệu quả trong việc ứng dụng thuật toán f-SVM hoặc SVM-IF để trích xuất ra các tập luật dùng cho giai đoạn dự báo. Kết quả một trường hợp tập luật trích rút được từ mô hình đề xuất được thể hiện ở bảng Tab. 3.

5 Kết luận

Thuật toán phân cụm SOM* đề xuất có độ phức tạp tính toán tương đương với thuật toán SOM nguyên thủy. Kết quả thực nghiệm trên tập dữ liệu S&P500 cho thấy thuật toán áp dụng có hiệu quả trong việc điều chỉnh kết quả phân cụm tập dữ liệu huấn luyện. Tuy nhiên, việc hiệu chỉnh này còn phụ thuộc vào việc lựa chọn giá trị ngưỡng z của kích thước các phân cụm dữ liệu. Ứng với mỗi tập dữ liệu nhất định, cần thiết phải tiến hành nhiều thực nghiệm để có một kết quả thống kê đủ tin cậy; từ đó mới có cơ sở đề xuất một giá trị ngưỡng z phù hợp cho từng trường hợp.

Tập luật trích xuất được từ tập dữ liệu huấn luyện cần phải được thử nghiệm dự báo trên tập dữ liệu thử nghiệm để đánh giá hiệu quả dự báo của mô hình. Ngoài ra nghiên cứu cũng cần được tiếp tục thử nghiệm trên nhiều tập dữ liệu (các mã cổ phiếu khác, chuỗi thời gian khác), để từ đó có những số liệu thống kê, đánh giá hiệu quả sử dụng mô hình.

Tài liệu tham khảo

1. Li Yuan Cao, Francis Eng Hock Tay (2001), Improved financial time series forecasting by combining Support Vector Machines with self-organizing feature map. Intelligent Data Analysis 5, IOS press, 339-354.
2. Kamalpreet Kaur Jassar, Kanwalvir Singh Dhindsa (2016), Comparative Study and Performance Analysis of Clustering Algorithms, IJCA - Proceedings on International Conference on ICT for Healthcare ICTHC 2015(1), 1-6.
3. Meizhen Liu, Chunmei Duan (2018), A Review of Using Support Vector Machine Theory to Do Stock Forecasting, 2018 International Conference on Network, Communication, Computer Engineering.

4. 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.
5. Zhe Gao, and Jianjun Yang (2014), Financial Time Series Forecasting with Grouped Predictors using Hierarchical Clustering and Support Vector Regression, International Journal of Grid Distribution Computing Vol.7, No.5, 53-64.
6. O. Maimon, L. Rokach (2010), Chapter 14 & 56, Data mining and knowledge discovery handbook, 2nd edition, Springer, New York.
7. Juha Vesanto, Johan Himberg, Esa Alhoniemi, and Jaha Parhankangas (1999), Self-organizing map in Matlab: the SOM Toolbox, Proceedings of the Matlab DSP Conference 1999, 35-40. Toolbox available at http://www.cis.hut.fi/projects/somtoolbox/.
8. Juan C. Figueroa-García, Cynthia M. Ochoa-Rey, José A. Avellaneda-González (2015), Rule generation of fuzzy logic systems using a self-organized fuzzy neural network, Neurocomputing– ELSEVIER, 151, 955–962.
9. Teuvo Kohonen (1998), The self-organizing map, Elsevier, Neurocomputing 21, 1-6.
10. Duc-Hien Nguyen, Manh-Thanh Le (2014), 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.
11. Nguyễn Đức Hiển, Lê Mạnh Thạnh (2015), Mô hình mờ TSK dự đoán giá cổ phiếu dựa trên máy học véc-tơ hỗ trợ hồi quy, Tạp chí khoa học Trường Đại học Cần Thơ, Số chuyên đề Công nghệ thông tin, 144-151.
12. Nguyễn Đức Hiển, Lê Mạnh Thạnh (2015), Tối ưu hóa mô hình mờ TSK trích xuất từ máy học véc-tơ hỗ trợ hồi qui với tham số epsilon, Tạp chí Khoa học và Công nghệ Đại học Đà Nẵng, Số 12(97), Quyển 2, 15-19.
13. Nguyễn Đức Hiển, Lê Mạnh Thạnh (2018), Một số giải pháp tối ưu tập luật mờ TSK trích xuất từ máy học véc-tơ hỗ trợ hồi quy. Kỷ yếu Hội nghị FAIR'2018.
14. Nguyễn Đức Hiển (2019), Tối ưu hóa tập luật mờ hướng dữ liệu bằng giải pháp rút gọn tập thuộc tính dữ liệu vào. Kỷ yếu Hội thảo khoa học quốc gia – CITA2019 – ISBN: 987-604-84-4453-2
15. Vạn Duy Thanh Long, Lê Minh Duy, Nguyễn Hoàng Tú Anh (2011), Phương pháp dự đoán xu hướng cổ phiếu dựa trên việc kết hợp K-means và SVM với ước lượng xác suất lớp, Đại học quốc gia – Tp HCM.