Tạp chí Khoa học Trường Đại học Cần Thơ
Phần A: Khoa học Tự nhiên, Công nghệ và Môi trường
website: sj.ctu.edu.vn

DOI:10.22144/ctu.jvn.2018.163

MỜ HÓA CHUỖI THỜI GIAN DỰA VÀO BÀI TOÁN PHÂN TÍCH CHÙM

Võ Văn Tài*, Phạm Bích Như, Nguyễn Văn Pha và Nguyễn Thu Hiền
Khoa Khoa học Tự nhiên, Trường Đại học Cần Thơ
**Người chịu trách nhiệm về bài viết: Võ Văn Tài (email: vvtai@ctu.edu.vn)*

Thông tin chung:
Ngày nhận bài: 06/03/2018
Ngày nhận bài sửa: 02/05/2018
Ngày duyệt đăng: 27/12/2018

Title:
Interpolate time series based on cluster analysis problem

Từ khóa:
Dự báo, chùm, chuỗi thời gian mờ, mờ hóa, thuật toán

Keywords:
Algorithm, cluster, forecast, fuzzy time series, interpolate

ABSTRACT

This article proposes a time series model to interpolate historical data and use them to forecast for future. This model is built based on the automatic algorithms in cluster analysis and performed by Matlab procedures. They are the algorithm to determine the suitable number of clusters, the elements in each cluster and the relationship of each element with established clusters. The proposed model's convenience and efficiency were tested by by many benchmarks and sets of real data. These numerical examples show the advantages of the proposed model in comparison with existing models and its effectiveness in practical applications.

TÓM TẮT

Bài báo này đề xuất mô hình chuỗi thời gian để mờ hoá dữ liệu lịch sử và sử dụng nó dự báo cho tương lai. Mô hình này được xây dựng dựa trên các thuật toán tự động trong phân tích chùm và được thực hiện bởi các chương trình viết trên Matlab. Chúng là thuật toán xác định số lượng chùm thích hợp, các phần tử cụ thể trong mỗi chùm và mối quan hệ của mỗi phần tử với các chùm đã được thiết lập. Tính hiệu quả và sự thuận lợi của mô hình đề nghị được kiểm tra bởi nhiều bộ dữ liệu chuẩn và thực tế. Các ví dụ số này đã thể hiện những ưu điểm của mô hình được đề xuất so với các mô hình hiện tại và sự hiệu quả của nó trong các ứng dụng thực tiễn.

Trích dẫn: Võ Văn Tài, Phạm Bích Như, Nguyễn Văn Pha và Nguyễn Thu Hiền, 2018. Mờ hóa chuỗi thời gian dựa vào bài toán phân tích chùm. Tạp chí Khoa học Trường Đại học Cần Thơ. 54(9A): 72-80.

1 GIỚI THIỆU

Dự báo là quá trình tiên đoán kết quả cho tương lai dựa vào kiến thức, kinh nghiệm, các hiện tượng và số liệu đã có của vấn đề được quan tâm. Dự báo có một vai trò rất quan trọng trong nhiều lĩnh vực khác nhau và được nhiều nhà khoa học quan tâm nghiên cứu. Tuy nhiên, cho đến nay dự báo vẫn là một bài toán chưa có lời giải cuối cùng. Khi thực hiện dự báo bằng các mô hình thống kê, điều kiện tiên quyết ban đầu là số liệu. Với số liệu kiểu chuỗi thời gian, một loại dữ liệu phổ biến có nhu cầu dự báo lớn trong thực tế hiện nay, hai mô hình chính được sử dụng để dự báo là hồi quy và chuỗi thời gian. Mô hình hồi quy có những ràng buộc về điều kiện của dữ liệu mà trong thực tế rất khó thỏa mãn, do đó nó có hạn chế trong nhiều trường hợp. Mô hình chuỗi thời gian (TSM) được đánh giá có nhiều ưu điểm hơn nên được sử dụng rất phổ biến ngày nay. Nhiều nhà nghiên cứu đã sử dụng các mô hình TSM như tự hồi qui (AR), mô hình tự hồi quy trung bình trượt (ARIMA) để ứng dụng trong kinh tế, môi trường và thuỷ văn (Box and Jenkins, 1970). Tuy nhiên, để xây dựng được mô hình TSM tốt thì dữ liệu phải thỏa một số điều kiện nhất định mà thực tế thường khó đáp ứng. Do đó, nhiều trường hợp cho

kết quả dự báo kém khi sử dụng các mô hình hồi quy không mờ. Mặc dù nhiều tác giả như Abreue *et al*. (2013), Oliveira và Ludermir (2014) đã cố gắng cải thiện mô hình ban đầu, nhưng họ vẫn còn gặp nhiều khó khăn để thực hiện dự báo hợp lý cho nhiều dữ liệu thực tế. Một mô hình có thể được đánh giá tốt hơn các mô hình khác dựa trên từng dữ liệu cụ thể mà không phải cho tất cả các trường hợp.

Một vấn đề khác là các mô hình TSM truyền thống không thể giải quyết các vấn đề dự báo, nơi các dữ liệu lịch sử được trình bày bằng các biến ngôn ngữ. Các mô hình chuỗi thời gian mờ (FTS) đã giải quyết nhược điểm này. Các mô hình FTS được phát triển theo hai hướng chính. Hướng thứ nhất là xây dựng các mô hình từ dữ liệu gốc và sử dụng nó để dự báo cho tương lai một cách trực tiếp. Abbasov và Mamedova (2003) đã có những đóng góp quan trọng theo hướng này. Hướng thứ hai là sự mờ hoá dữ liệu gốc để có được mối quan hệ giữa các phần tử, sau đó áp dụng các mô hình dự báo đã biết cho dữ liệu đã mờ hóa này. Hướng nghiên cứu này đã và đang được rất nhiều nhà thống kê quan tâm, trong đó Song và Chissom (1993) là những người tiên phong.

Mô hình FTS thông thường bao gồm ba giai đoạn: (i) xác định tập nền từ dữ liệu gốc, chia khoảng cho tập nền và tìm số lượng các phần tử cho mỗi khoảng; (ii) xây dựng các mối quan hệ mờ, và (iii) giải mờ. Đối với (i), nhiều tác giả đã sử dụng giá trị nhỏ nhất và giá trị lớn nhất của dữ liệu ban đầu để xác định tập nền (Chen, 1996; Chen và Hsu, 2004). Ngoài ra, Huarng (2001), Huarng và Yu (2006) đã đề xuất hai kỹ thuật mới để xác định các khoảng của tập nền dựa trên giá trị trung bình và phân phối của toàn chuỗi. Một cách khác để xây dựng tập nền là dựa trên sự thay đổi dữ liệu giữa các khoảng thời gian liên tiếp hoặc tỷ lệ phần trăm thay đổi (Abbasov và Mamedova, 2003). Bên cạnh đó, vấn đề xác định số tập mờ và các phần tử trong mỗi tập mờ là rất quan trọng để thiết lập mô hình. Nhiều tác giả đã chia số tập mờ bằng cách kiểm tra trong nhiều trường hợp để có các thông số đánh giá thích hợp cho từng dữ liệu mà không phải là một quy tắc chung cho tất cả các trường hợp. Số lượng tập mờ và các phần tử của chúng cũng được đề xuất dựa trên các thuật toán k – trung bình và thuật toán di truyền (Zhiqiang, 2012). Mặc dù, đã có nhiều tác giả thảo luận về vấn đề này nhưng cho đến nay sự lựa chọn tối ưu vẫn chưa được tìm thấy. Đối với (ii), một số nghiên cứu đã được thực hiện, Song và Chissom (1993) đã sử dụng các phép toán ma trận, Chen (1996) và một số nhà nghiên cứu khác sử dụng bảng nhóm quan hệ mờ. Trong khi đó, Aladag (2012) đã sử dụng mạng thần kinh nhân tạo để xác định mối quan hệ mờ. Các nhà thống kê tin rằng việc xử lý các mối quan hệ mờ là bước rất quan trọng để có mô hình FTS phù hợp. Đối với (iii), hầu hết các nghiên cứu sử dụng phương pháp trọng tâm để thực hiện (Chen, 1996; Huarng, 2001; Huarng và Yu, 2006).

Bài báo này đóng góp cho ba giai đoạn (i), (ii) và (iii) đối với mô hình FTS. Đối với giai đoạn (i), sau khi chuẩn hóa dữ liệu, bài báo đề xuất thuật toán để xác định số lượng các tập mờ. Cho giai đoạn (ii), dựa trên thuật toán phân tích chùm mờ, bài báo xác định số các phần tử trong mỗi tập mờ. Thuật toán này cũng xác định mối quan hệ mờ giữa mỗi phần tử và các tập mờ. Một quy tắc giải mờ mới cũng được thiết lập trong bài báo này. Đây là sự đóng góp cho giai đoạn (iii) của mô hình chuỗi thời gian mờ. Kết hợp tất cả các cải tiến, bài viết này đề nghị mô hình chuỗi thời gian mờ mới (NFTS) tốt hơn so với các mô hình hiện có thông qua nhiều bộ dữ liệu khác nhau. Các chương trình tính toán trên Matlab để thực hiện các thuật toán được đề xuất cũng được thiết lập. Các chương trình này giúp mô hình NFTS được thực hiện một cách hiệu quả cho các ví dụ số. Ngoài ra, mô hình NFTS cũng được sử dụng để dự báo đỉnh lũ tại trạm đầu nguồn con sông Tiền.

Phần tiếp theo của bài báo được cấu trúc như sau: Phần 2 xem xét một số khái niệm cơ bản về mô hình FTS và đưa ra mô hình FTS mới; Phần 3 giải quyết các vấn đề liên quan đến việc áp dụng trong thực tế của mô hình NFTS (Đó là thuật toán tìm số lượng các tập mờ phù hợp, xác định các phần tử trong mỗi tập mờ và các mối quan hệ của mỗi phần tử đến các chùm. Vấn đề tính toán cũng được xem xét trong phần này); Phần 4 là những ví dụ số để minh họa cho các vấn đề lý thuyết được trình bày (Phần này cũng so sánh các kết quả của NFTS và với một số mô hình hiện tại); Một ứng dụng thực tế được trình bày trong Phần 5; Cuối cùng là kết luận của bài viết.

2 MỘT SỐ ĐỊNH NGHĨA VÀ THUẬT TOÁN ĐỀ NGHỊ

2.1 Các định nghĩa

Định nghĩa 1. Cho U là tập nền, $U = \{u_1, u_2, \ldots, u_n\}$. Tập mờ A của U được định nghĩa như sau:

$$A = \{\mu_A(u_1)/u_1, \mu_A(u_2)/u_2, \ldots, \mu_A(u_n)/u_n\}, \quad (1)$$

trong đó $\mu_A(u_i)$ là hàm thuộc, $\mu_A(u_i): U \to [0,1]$, $\mu_A(u_i)$ cho biết mức độ liên hệ của u_i trong A, $\mu_A(u_i) \in [0,1]$, $1 \le i \le n$.

Định nghĩa 2. Cho $X(t)$, (t = 1, 2, …) là tập nền với tập mờ $\mu_A(u_i)$, (i = 1, 2, …) và $F(t)$ là tập hợp các giá trị $\mu_A(u_i)$, (i = 1, 2, …). Khi đó $F(t)$ được gọi là chuỗi thời gian mờ (FTS) trên $X(t)$.

Định nghĩa 3. Cho một chuỗi dữ liệu thực tế $\{X_i\}$ và giá trị dự đoán tương ứng $\{\hat{X}_i\}$, $i = 1, 2, \ldots, n$, khi đó ta có các tiêu chuẩn sau để đánh giá các mô hình FTS:

Bình phương sai số trung bình:

$$MSE=\frac{1}{n}\sum_{i=1}^{n}\left(\hat{X}_i-X_i\right)^2. \quad (2)$$

Sai số tuyệt đối trung bình:

$$MAE=\frac{1}{n}\sum_{i=1}^{n}\left(\frac{\left|\hat{X}_i-X_i\right|}{X_i}\right). \quad (3)$$

Sai số phần trăm tuyệt đối trung bình:

$$MAPE=\frac{1}{n}\sum_{i=1}^{n}\left(\frac{\left|\hat{X}_i-X_i\right|}{X_i}.100\right). \quad (4)$$

Khi thực hiện dự báo, mô hình nào có các tiêu chuẩn trên càng nhỏ thì nó càng tốt.

2.2 Thuật toán đề nghị

Cho chuỗi thời gian $\{X_i\}$, $i = 1, 2, \ldots, n$. Dựa vào bài toán phân tích chùm, mô hình FTS được đề nghị với 6 bước như sau:

Bước 1. Chuẩn hóa dữ liệu về thang đo 100, $Y_i = 100X_i/\max\{X_i\}$, $i = 1, 2, \ldots, n$. Khi đó ta có tập nền

$$U = \{Y_i, i = 1, 2, \ldots, n\}.$$

Bước 2. Chia tập nền U thành c chùm $w_1, w_2, \ldots, w_c$ một cách thích hợp.

Bước 3. Xác định cụ thể các phần tử trong mỗi chùm w_i và tính trọng tâm m_i của mỗi chùm, $i = 1, 2, \ldots, c$.

Bước 4. Xác định mối quan hệ mờ μ_{ij} từ mỗi phần tử Y_i đến các chùm w_j; $i = 1, 2, \ldots, n$; $j = 1,2,\ldots, c$.

Bước 5. Dự báo cho Y_i theo nguyên tắc sau:

$$Y_i=\Sigma_{j=1}^{c}\mu_{ij}c_j, i=1,2,...,n. \quad (5)$$

Bước 6. Từ kết quả của Y_i, dự báo cho X_i theo công thức:

$$X_i = Y_i.\max\{X_i\}/100. \quad (6)$$

3 MỘT SỐ VẤN ĐỀ LIÊN QUAN ĐẾN THUẬT TOÁN ĐỀ NGHỊ

Trong mô hình đề nghị, việc xác định số chùm thích hợp c được chia ở Bước 2, tìm số phần tử trong mỗi chùm ở Bước 3 và việc tìm mối quan hệ mờ μ_{ij} của Bước 4 là những vấn đề rất quan trọng vì chúng quyết định đến hiệu quả của mô hình. Dựa vào các thuật toán phân tích chùm mờ của Chen và Hung (2015), thuật toán sau được đề nghị để giải quyết các vấn đề trên.

3.1 Thuật toán xác định số chùm thích hợp cho tập nền

Từ tập nền U của Bước 1, thuật toán để tìm số chùm thích hợp (the Suitable Numbers of Cluster (SNC)) được trình bày như sau:

Bước 1. Thiết lập dãy trọng tâm $Z^{(0)} = \{z_1^{(0)}, z_2^{(0)}, \ldots, z_n^{(0)}\} = \{y_1, y_2, \ldots, y_n\}$ và cho trước là số dương rất nhỏ ε.

Bước 2. Cập nhật dãy trọng tâm mới theo công thức

$$z_i^{(t+1)} = \frac{\sum_{i'=1}^{n} K_\lambda(z_i^{(t)}, z_{i'}^{(t)}) z_{i'}^{(t)}}{\sum_{i'=1}^{n} K_\lambda(z_i^{(t)}, z_{i'}^{(t)})}, \quad (7)$$

trong đó $K_\lambda(.)$ là hàm hạt nhân dạng chuẩn:

$$K_\lambda(z_i^{(t)}, z_{i'}^{(t)}) = \begin{cases} \exp(-d/\lambda) & \text{khi } d(z_i^{(t)}, z_{i'}^{(t)}) \le d_s \\ 0 & \text{khi } d(z_i^{(t)}, z_{i'}^{(t)}) > d_S \end{cases}$$

với λ là hằng số, $d(z_i^{(t)}, z_{i'}^{(t)})$ là khoảng cách giữa $z_i^{(t)}$ và $z_{i'}^{(t)}$, và d_s là trung bình khoảng cách của tất cả các phần tử trong tập dữ liệu. Nó được tính theo công thức sau:

$$d_S = \frac{2}{n(n-1)}\sum_{i<i'} d(z_i^{(t)}, z_{i'}^{(t)}),$$

Bước 3. Lặp lại bước 2 cho đến khi điều kiện sau được thỏa:

$$\max_i\{d(z_i^{(t+1)}, z_i^{(t)})\} < \varepsilon.$$

Trong thuật toán trên, các vấn đề sau cần được chú ý:

i) Sau khi vòng lặp kết thúc, mỗi phần tử trong bộ dữ liệu sẽ hội tụ đến phần tử đại diện của chùm chứa nó. Khi thuật toán dừng, c phần tử đại diện của các chùm sẽ được nhận. Từ đây ta cũng được c là số lượng các chùm của tập nền.

ii) Có nhiều khoảng cách khác nhau để đánh giá mức độ tương tự giữa hai phần tử (Webb, 2002). Trong bài viết này, $d(z_i^{(t)}, z_{i'}^{(t)})$ là khoảng cách Euclide.

iii) Trong thực tế, giá trị λ quyết định số khoảng của tập dữ liệu. Khi $\lambda \to 0$ thì dữ liệu có n khoảng và khi $\lambda \to \infty$ thì dữ liệu có một khoảng duy nhất. Tùy theo bộ dữ liệu, giá trị λ được chọn sao cho phù hợp.

Để có thể chọn tham số này một cách thống nhất cho các mô hình FTS, trong Bước 1 của thuật toán SNC dữ liệu được chuẩn hóa về thang đo 100. Bên cạnh đó, theo kinh nghiệm cho nhiều bộ dữ liệu khác nhau, bài báo này đã chọn $\lambda = d_S/16$ cho các ví dụ số.

iv) $\mathcal{E}$ là số dương rất nhỏ, được chọn tuỳ ý. Giá trị này càng nhỏ thì số vòng lặp của thuật toán càng cao và thời gian tính toán càng nhiều. Trong bài báo ε=0,001 được chọn.

3.2 Thuật toán xác định số phần tử của mỗi chùm và các mối quan hệ mờ

Sau khi xác định số chùm thích hợp c để chia tập nền U, các phần tử đại diện trong mỗi chùm {wi} , i = 1, 2, …, c sẽ được tìm và các mối quan hệ mờ của mỗi phần tử yi, i = 1, 2, …, n với các chùm được thiết lập. Các vấn đề này được thực hiện bởi thuật toán DFR (Determining the Fuzzy Relation):

Bước 1. Chia tập nền U thành c chùm $w_1, w_2, \ldots, w_c$ một cách ngẫu nhiên. Thiết lập ma trận phân vùng ban đầu $U^{(0)} = [\mu_{ij}]_{k\times n}$, với $\mu_{ij} = 1$ nếu phần tử thứ j thuộc chùm w_i và $\mu_{ij} = 0$ đối với trường hợp ngược lại.

Bước 2. Xác định phần tử đại diện v_i của mỗi tập mờ bằng công thức:

$$v_i = \frac{\sum_{j}^{n} \mu_{ij}^{2} y_j}{\sum_{j}^{n} \mu_{ij}^{2}}, \qquad (8)$$

trong đó $1 \le i \le c$, μ_{ij} là xác suất để phần tử thứ j được xếp vào chùm w_i.

Bước 3. Cập nhật ma trận phân vùng mới $U^{(1)}$ bởi qui tắc sau:

$$\mu_{ij}^{(1)} = \begin{cases} \dfrac{1}{\sum_{l=1}^{c} \left(d_{ij}/d_{lj}\right)^2} & \text{khi} \quad d_{ij} > 0, \\ 0 & \text{khi} \quad d_{ij} = 0, \end{cases} \qquad (9)$$

trong đó d_{ij} là khoảng cách từ y_j đến v_i.

Bước 4. Tính $S = \max_{ij}\left(\left|\mu_{ij}^{(1)} - \mu_{ij}^{(0)}\right|\right)$. Nếu $S<\varepsilon$ thuật toán dừng lại, ngược lại lặp lại Bước 2 và Bước 3.

Trong thuật toán DFR, các vấn đề sau cần được lưu ý:

i) Khoảng cách Euclide được sử dụng để đánh giá sự tương tự của các phần tử.

ii) Khi thuật toán DFR kết thúc, ma trận cỡ ($c \times n$) sẽ được thiết lập. Trong ma trận này, tổng của mỗi cột bằng 1 ($\sum_{j=1}^{c} \mu_{ij} = 1$). Nếu max $\{\mu_{ij}\} = \mu_{im}, 1 \le m \le c\}$ thì phần tử $y_i, 1 \le i \le n$ được xếp vào chùm w_m.

Hai thuật toán SNC và DFR đã được lập trình hoàn chỉnh trên phần mềm Matlab. Chúng được áp dụng một cách hiệu quả cho các ví dụ số trong Phần 4.

4 VÍ DỤ MINH HỌA CHO THUẬT CÁC TOÁN ĐỀ NGHỊ

Trong phần này tập dữ liệu sinh viên của trường đại học Alabama được sử dụng để minh họa cho các thuật toán được đề nghị. Đây là tập dữ liệu được trình bày trong nhiều nghiên cứu (Chen, 1996; Ming, 2002; Chen and Hsu, 2004). Nó thường được sử dụng khi so sánh hiệu quả của các mô hình chuỗi thời gian mờ với nhau. Các bước thực hiện được tính cụ thể như sau:

Bước 1. Từ tập dữ liệu đã cho X_i, chuẩn hóa dữ liệu về thang đo 100 ta được các giá trị Y_i.

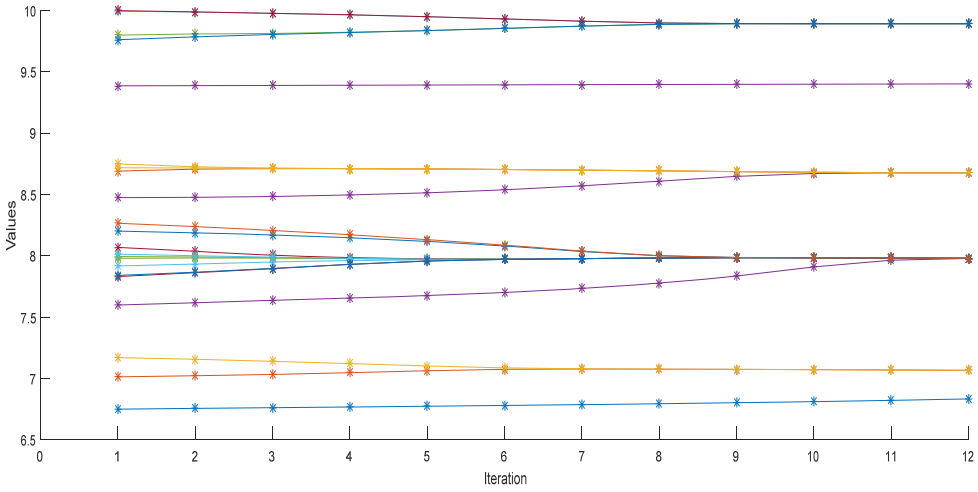
Bảng 1: Số liệu sinh viên thực tế và chuẩn hóa

Năm	X_i	Y_i	Năm	X_i	Y_i
1971	13055	6,751	1982	15433	7,981
1972	13563	7,014	1983	15497	8,014
1973	13867	7,171	1984	15145	7,832
1974	14696	7,600	1985	15163	7,841
1975	15460	7,995	1986	15984	8,266
1976	15311	7,918	1987	16859	8,719
1977	15603	8,069	1988	18150	9,386
1978	15861	8,202	1989	18970	9,810
1979	16807	8,692	1990	19328	9,995
1980	16919	8,750	1991	19337	10,00
1981	16388	8,475	1992	18876	9,762

Bước 2. Áp dụng thuật toán SNC với 12 bước lặp, ta được các giá trị sau:

7,0112	7,0113	7,0113	7,9804	7,9804	7,9804	7,9804	7,9804
8,6780	8,6780	8,6780	7,9804	7,9804	7,9804	7,9804	7,9804
8,6780	9,4139	9,8920	9,8920	9,8920	9,8920		

Các vòng lặp của thuật toán này được minh họa bởi Hình 1.



Hình 1: Sự hội tụ của thuật toán SNC cho dữ liệu sinh viên qua 12 vòng lặp

Như vậy, thuật toán SNC cho ta 6 phần tử đại diện, do đó ta chia tập dữ liệu thành 6 chùm.

Bước 3. Sử dụng thuật toán DFR với số chùm là 6, chúng ta có các chùm cụ thể

$w_1=\{Y_1\}$, $w_2=\{Y_2;Y_3\}$, $w_3=\{Y_4;Y_5;Y_6;Y_7;Y_8;Y_{12};Y_{13};Y_{14};Y_{15};Y_{16}\}$, $w_4=\{Y_9;Y_{10};Y_{11};Y_{17}\}$, $w_5=\{Y_{18}\}$, $w_6=\{Y_{19};Y_{20};Y_{21};Y_{22}\}$

Trọng tâm của các chùm trên lần lượt là 6,7510, 7,0925, 7,9718, 8,6590, 9,3860 và 9,8918.

Bước 4. Thuật toán DFR cũng cho ta mối quan hệ μ_{ij} từ mỗi phần tử y_i đến các chùm w_j; i = 1, 2, …, 22; j = 1, 2, …, 6 bởi ma trận phân chia sau:

$$[\mu_{ij}]_{6x22}=\begin{bmatrix} 0{,}9956 & 0{,}2446 & 0{,}0023 & \cdots & 0{,}0009 & 0{,}0018 \\ 0{,}0037 & 0{,}7310 & 0{,}9969 & \cdots & 0{,}0012 & 0{,}0024 \\ 0{,}0004 & 0{,}0151 & 0{,}0006 & \cdots & 0{,}0024 & 0{,}0051 \\ 0{,}0002 & 0{,}0052 & 0{,}0002 & \cdots & 0{,}0054 & 0{,}0132 \\ 0{,}0001 & 0{,}0025 & 0{,}0001 & \cdots & 0{,}0262 & 0{,}1156 \\ 0{,}0001 & 0{,}0017 & 0{,}0000 & \cdots & 0{,}9638 & 0{,}8619 \end{bmatrix}$$

Bước 5. Dự báo cho Y_i theo (4), ta có các giá trị dự báo $\hat{Y}_i$ được cho bởi Bảng 2.

Bảng 2: Kết quả mờ hóa dữ liệu sinh viên

Năm	$\hat{Y}_i$	$\hat{X}_i$	Năm	$\hat{Y}_i$	$\hat{X}_i$
1971	6,7535	13059	1982	7,9718	15415
1972	7,0407	13615	1983	7,9733	15418
1973	7,0928	13715	1984	7,9542	15381
1974	7,6540	14801	1985	7,9579	15388
1975	7,9721	15416	1986	8,2170	15889
1976	7,9717	15415	1987	8,6595	16745
1977	7,9853	15441	1988	9,3860	18150
1978	8,1069	15676	1989	9,8536	19054
1979	8,6588	16743	1990	9,8635	19073
1980	8,6630	16752	1991	9,8608	19068
1981	8,5829	16597	1992	9,7949	18940

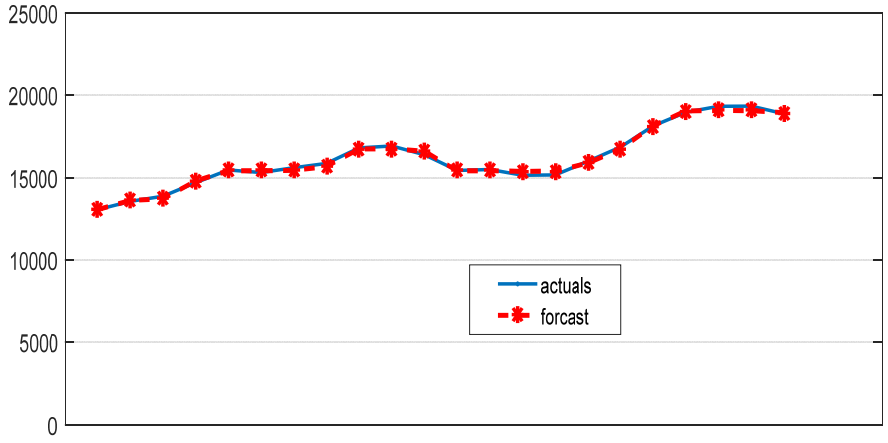
Bước 6. Từ kết quả của $\hat{Y}_i$, theo (5) ta có kết quả dự báo của X_i là $\hat{X}_i$ (xem Bảng 2).

Kết quả dự báo này cho MSE = 21292, MAE = 121,96 và MAPE = 0,748. Giá trị thực và dự báo cho số liệu sinh viên được cho bởi Hình 2.

So sánh 3 tham số MAE, MAPE và MSE của phương pháp đề nghị và một số phương pháp khác, ta có kết quả sau:

Bảng 3: So sánh mô hình đề nghị và các phương pháp khác

Phương pháp	MAE	MAPE	MSE
AM (Abbasov – Mamedova)	459,34	2,78	301599,00
Chen	502,38	3,08	413980,98
Singh	254,64	1,53	84880,31
Heuristic	419,62	2,59	235891,26
Chen – Hsu	293,45	1,76	138366,80
Mô hình đề nghị	**121,96**	**0,75**	**21292,00**



Hình 2: Đồ thị cho giá trị thực và giá trị dự báo của sinh viên

Hình 2 thể hiện các giá trị thực sự và dự báo gần như trùng khớp nhau. Các tham số của Bảng 3 cho thấy phương pháp đề nghị tốt hơn các phương pháp khác.

5 ÁP DỤNG TRONG DỰ BÁO ĐỈNH LŨ SÔNG TIỀN

Đồng bằng sông Cửu Long của Việt Nam chịu tác động mạnh mẽ của sông Tiền và sông Hậu, đây là hai con sông chính trong khu vực này. Bên cạnh việc đem đến sự màu mỡ cho đất đai, sự dồi dào về nước ngọt và thủy sản cho khu vực thì chế độ thủy văn phức tạp trên hai con sông, đặc biệt là lũ lụt đã gây ra nhiều thiệt hại cho người dân. Việc dự báo lũ cho chúng là vấn đề quan trọng ảnh hưởng đến nhiều chính sách phát triển kinh tế xã hội của vùng. Vì thế, vấn đề này nhận được sự quan tâm đặc biệt của các cấp chính quyền và các nhà khoa học. Cũng như nhiều vấn đề dự báo khác, dự báo lũ lụt là một bài toán chưa có lời giải cuối cùng. Trong phần này, mô hình đề nghị được sử dụng để thực hiện dự báo đỉnh lũ tại các trạm đo được đặt đầu nguồn trên sông Tiền. Số liệu dự báo sẽ là cơ sở quan trọng cho các trạm đo khác cũng như cho các cấp chính quyền trong việc hoạch định các chính sách vĩ mô.

Số liệu về đỉnh lũ trong giai đoạn 1990 – 2012 được cho trong bảng sau:

Bảng 4: Đỉnh lũ trên sông Tiền giai đoạn 1990 – 2012

Năm	Đỉnh lũ (cm)	Năm	Đỉnh lũ (cm)
1990	418	2005	436
1991	463	2006	417
1992	343	2007	408
1993	344	2008	377
1994	453	2009	412
1995	430	2010	320
1996	486	2011	486
1997	418	2012	432
1998	281	2011	486
1999	420	2012	432
2000	506	2013	435
2001	479	2014	396
2002	482	2015	251
2003	406	2016	307
2004	440	2017	343

Bảng 5: So sánh các mô hình trên tập huấn luyện cho đỉnh lũ

Năm	Thực tế	Chen	Singh	Heuristic	Chen-Hsu	AM	NFTS
1990	418	-	-	-	-	-	415,51
1991	463	412,79	-	457,79	457,79	-	458,67
1992	343	377,43	-	377,43	341,27	-	343,51
1993	344	393,50	327,99	393,50	337,25	344,00	343,51
1994	453	393,50	460,39	393,50	457,79	345,00	456,42
1995	430	377,43	428,88	377,43	425,64	454,00	434,05
1996	486	412,79	489,93	457,79	481,89	431,00	482,14
1997	418	436,36	422,82	436,36	425,64	487,00	415,51
1998	281	412,79	297,07	372,07	297,07	419,00	281,00
1999	420	425,64	422,32	425,64	433,68	282,00	417,50
2000	506	412,79	489,93	457,79	489,93	421,00	506,00
2001	479	436,36	489,84	436,36	481,89	507,00	481,57
2002	482	436,36	490,84	489,93	489,93	480,00	482,33
2003	406	436,36	393,50	436,36	393,50	483,00	416,79
2004	440	425,64	422,69	425,64	425,64	407,00	436,52
2005	436	412,79	426,79	372,07	425,64	441,00	435,34
2006	417	412,79	429,43	372,07	417,61	437,00	414,94
2007	408	412,79	398,00	372,07	385,46	418,00	416,25
2008	377	399,53	381,81	351,31	386,47	422,86	377,00
2009	412	407,56	409,19	407,56	418,11	391,17	420,47
2010	320	399,53	323,19	351,31	323,19	426,36	320,00
2011	486	491,94	491,46	491,94	391,94	334,12	483,39
MAE		41,59	6,62	30,38	6,42	61,95	**5,10**
MAPE		10,69	1,63	7,82	1,61	15,58	**1,19**
MSE		264362	69,87	1334,57	61,82	5879,53	**59,99**

(Kí hiệu "–" cho dữ liệu bị khuyết)

Dựa vào số liệu được trình bày trong bảng trên, bài viết sẽ tiến hành dự báo cho một số năm tiếp theo bằng nhiều mô hình khác nhau. Trong ứng dụng này, hai trường hợp sau được xem xét:

i) Trường hợp 1: Chia dữ liệu thành hai phần: Tập huấn luyện và tập kiểm tra với tỉ lệ 80% (22 năm) và 20% (6 năm) để đánh giá mô hình đề nghị khi so sánh với các mô hình khác bởi các tham số MAE, MAPE và MSE. Kết quả thực hiện trên tập huấn luyện được tổng kết như Bảng 5.

Mô hình đề nghị cho kết quả mờ hóa tốt nhất đối với tập huấn luyện nên nó được sử dụng để tiến hành dự báo cho 6 năm tiếp theo bằng các mô hình ARIMA (ARIMAP) và AM (AMP) để so sánh với các mô hình được xây dựng từ dữ liệu gốc (ARIMAR và AMR). Các kết quả so sánh được cho bởi Bảng 6.

Bảng 6: So sánh các mô hình của tập kiểm tra cho đỉnh lũ

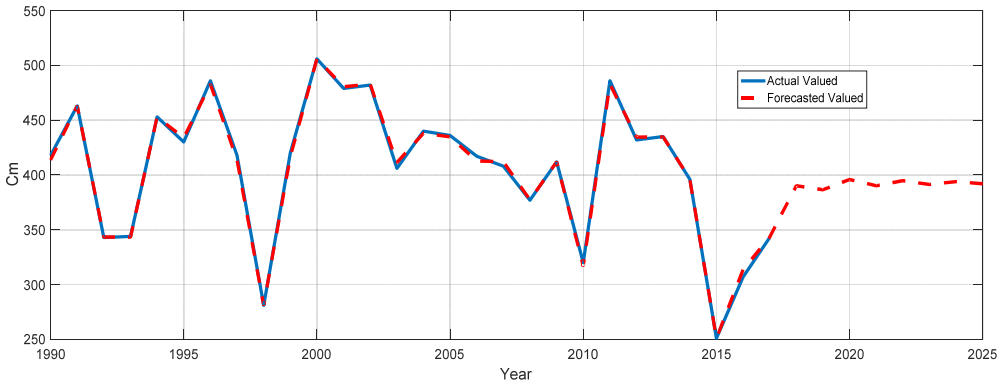
Tham số	ARIMAR	AMR	AMP	ARIMAP
MAE	91,23	174,33	162,07	**79,44**
MAPE	28,63	55,08	51,37	**25,81**
MSE	10370,37	37750,26	32975,05	**8735,04**

Bảng 6 cho thấy các mô hình được xây dựng từ dữ liệu mờ hóa ARIMAP cho kết quả tốt nhất.

ii) Trường hợp 2: Sử dụng toàn bộ dữ liệu để mờ hóa bằng các mô hình khác nhau. Sử dụng mô hình tốt nhất trong những trường hợp này để dự báo cho một số năm tiếp theo. Kết quả mờ hóa được cho bởi Bảng 7.

Bảng 7: Mờ hóa dữ liệu đỉnh lũ từ các mô hình

Năm	Thực tế	Chen	Singh	Heuristic	Chen-Hsu	AM	NFTS
1990	418	-	-	-	-	-	422,21
1991	463	399,53	-	449,75	463,81	-	467,39
1992	343	393,50	-	393,50	358,34	-	343,50
1993	344	407,56	356,19	407,56	351,31	357,50	343,50
1994	453	407,56	463,41	407,56	463,81	358,50	455,93
1995	430	393,50	432,23	393,50	435,69	467,50	449,63
1996	486	445,06	491,94	463,81	484,91	444,50	483,46
1997	418	445,06	413,78	445,06	407,56	500,50	422,21
1998	281	399,53	295,06	351,31	295,06	432,50	281,00
1999	420	407,56	413,28	407,56	414,59	295,50	423,87
2000	506	399,53	491,94	449,75	491,94	434,50	505,96
2001	479	445,06	490,34	445,06	484,91	520,50	483,09
2002	482	445,06	491,51	491,94	491,94	493,50	483,25
2003	406	445,06	408,78	445,06	407,56	496,50	419,04
2004	440	399,53	434,15	449,75	428,66	420,50	457,43
2005	436	445,06	436,30	421,63	435,69	454,50	457,81
2006	417	445,06	411,52	421,63	400,53	450,50	421,81
2007	408	399,53	412,44	351,31	400,53	431,50	420,54
2008	377	399,53	381,81	351,31	386,47	422,50	377,00
2009	412	407,56	409,19	407,56	414,59	391,50	421,58
2010	320	399,53	323,19	351,31	316,16	426,50	320,00
2011	486	491,94	491,47	491,94	491,94	334,50	483,46
2012	432	445,06	442,34	445,06	435,69	500,50	453,91
2013	435	415,59	419,10	463,50	434,28	442,16	434,63
2014	396	415,59	405,59	370,77	366,47	445,23	396,42
2015	251	349,52	262,59	316,68	266,94	406,15	251,00
2016	307	308,95	306,48	308,95	290,84	261,34	313,69
2017	343	413,27	327,74	413,27	330,69	317,56	343,34
MAE		38,88	7,52	28,64	7,83	60,16	**2,02**
MAPE		10,31	1,92	7,73	2,12	16,12	**0,50**
MSE		2834,42	78,23	1471,58	95,02	5775,76	**7,78**



Hình 3: Đồ thị cho đỉnh lũ thực tế và dự báo

Bảng 7 cho thấy mô hình đề nghị cho kết quả tốt nhất vì vậy ta sử dụng tập dữ liệu này để dự báo đến năm 2025. Kết quả thực hiện được cho bởi Bảng 8.

Hình 3 cho thấy mô hình đề nghị cho kết quả rất gần với số liệu thực và tình trạng lũ trong tương lai trên sông Tiền ở mức độ trung bình thấp.

Bảng 8: Dự báo đỉnh lũ đến năm 2020

Năm	Dự báo	Năm	Dự báo
2018	403,88	2022	406,01
2019	407,70	2023	406,46
2020	405,47	2024	406,19
2021	406,77	2025	406,35

6 KẾT LUẬN

Dựa vào bài toán phân tích chùm, bài viết đã thiết lập được mô hình chuỗi thời gian mờ mới. Mô hình này được bổ sung bởi hai thuật toán quan trọng là xác định số tập mờ và các mối quan hệ mờ dựa vào bài toán phân tích chùm. Mô hình đề nghị được đánh giá có ưu điểm hơn so với nhiều mô hình trước đó qua bộ số liệu đối chứng và thực tế. Việc áp dụng mô hình đề nghị trong thực tế được thực hiện nhanh chóng bởi các code được thiết lập trên phần mềm Matlab. Áp dụng thực tế cũng cho thấy tính hợp lý và tiềm năng trong việc áp dụng mô hình đề nghị cho nhiều lĩnh vực khác nhau. Tuy nhiên, trong bài viết này, sự hội tụ của thuật toán đề nghị vẫn chưa chứng minh. Đây sẽ là hướng nghiên cứu mở rộng của bài viết này.

TÀI LIỆU THAM KHẢO

Abbasov, A. and Mamedova, M., 2003. Application of fuzzy time series to population forecasting. Vienna University of Technology. 1(2): 545–552.

Abreu, P. H., Silva, D. C., Mendes-Moreira, J., Reis, L. P., and Garganta, J., 2013. Using multivariate adaptive regression splines in the construction of simulated soccer team's behavior models. International Journal of Computational Intelligence Systems. 6(5): 893–910.

Aladag, S., Aladag, C. H., Mentes, T., and Egrioglu, E., 2012. A new seasonal fuzzy time series method based on the multiplicative neuron model and SARIMA. Hacettepe Journal of Mathematics and Statistics. 41(3): 145–163.

Box, G. E. P. and Jenkins, G. M., 1970. Time series analysis: Forecasting and control. Holden-Day. San Fancisco, 546 pages.

Chen, S. M., 1996. Forecasting enrollments based on fuzzy time series. Fuzzy Sets and Systems. 81(3): 311–319.

Chen, S. M. and Hsu, C. C., 2004. A new method to forecast enrollments using fuzzy time series. International Journal of Applied Science and Engineering. 2(3): 234–244.

Chen, J. and Hung, W., 2015. An automatic clustering algorithm for probability density functions. J. Stat. Comput. Simul. 85(1): 3047–3063.

Huarng, K., 2001. Heuristic models of fuzzy time series for forecasting. Fuzzy Sets and Systems. 123(3): 369–386.

Huarng, K. and Yu, T., 2006. Ratio-based lengths of intervals toimprove fuzzy time series forecasting. IEEE Trans Syst ManCybern-Part B: Cybern. 36(2): 328–340.

Ming, C. S., 2002. Forecasting enrollments based on high-order fuzzy time series. Fuzzy Sets and Systems. 33(1): 1–16.

Oliveira, D. J and Ludermir, T. B., 2014. A distributed PSO-ARIMA-SVR hybrid system for time series forecasting. In 2014 IEEE International Conference on Systems, Man, and Cybernetics (SMC). 15(2): 3867–3872.

Song, Q. and Chissom, B. S., 1993. Fuzzy time series and its models. Fuzzy Sets and Systems. 54(3): 269–277.

Zhiqiang, Z. and Qiong, Z., 2012. Fuzzy time series forecasting based on k-means clustering. Open Journal of Applied Sciences. 25(1):100–103.

Webb, A., 2002. Statistical pattern recognition, 2nd Ed. John Wiley & Sons, London, 725 pages