A FUZZY TIME SERIES FORECASTING MODEL USING GRAPH – BASED CLUSTERING

Le Thi Luong
Industrial Economic Technology College

ARTICLE INFO

Received: 01/7/2021
Revised: 18/7/2021
Published: 21/7/2021

KEYWORDS

Forecasting
Fuzzy time series
Clustering
Fuzzy relation group
Enrolments
Salt peak

ABSTRACT

The fuzzy time series forecasting model is one of the tools which is used to deal with the complexity and uncertainty process. In the establishing of fuzzy time series model, the predictive accuracy depends on two main issues: (1) Partitioning and determining the effective lengths of intervals (2) Establishing the fuzzy relationships for prediction reasonably. In this study, a new fuzzy time series forecasting model that uses graph-based clustering to determine the different interval lengths is proposed. The proposed model is applied to two time series data sets, the historical data on the number of enrolments of university at the University of Alabama and the data set of salt peak for a coastal province in Vietnam. Computational results show that the proposed model has higher forecasting accuracy than the existing models when applied to two specifically datasets.

MÔ HÌNH DỰ BÁO CHUỖI THỜI GIAN MỜ SỬ DỤNG KỸ THUẬT PHÂN CỤM DỰA TRÊN ĐỒ THỊ

Lê Thị Lương
Trường Cao đẳng Công nghệ và Kinh tế Công nghiệp

THÔNG TIN BÀI BÁO

Ngày nhận bài: 01/7/2021
Ngày hoàn thiện: 18/7/2021
Ngày đăng: 21/7/2021

TỪ KHÓA

Dự báo
Chuỗi thời gian mờ
Phân cụm
Nhóm quan hệ mờ
Tuyển sinh
Đinh mặn

TÓM TẮT

Mô hình chuỗi thời gian mờ là một trong những công cụ được sử dụng để giải quyết quá trình phức tạp và không chắc chắn. Trong quá trình thiết lập mô hình chuỗi thời gian mờ, độ chính xác dự báo phụ thuộc vào hai vấn đề chính: (1) Phân khoảng và xác định độ dài khoảng dữ liệu hiệu quả, (2) Thiết lập các mối quan hệ mờ hợp lý cho dự báo. Trong nghiên cứu này, một mô hình dự báo chuỗi thời gian mờ mới sử dụng kỹ thuật phân cụm dựa trên đồ thị để xác định độ dài khoảng khác nhau được đề xuất. Mô hình đề xuất được áp dụng trên hai tập dữ liệu chuỗi thời gian, dữ liệu lịch sử về số lượng tuyển sinh đại học tại Đại học Alabama và dữ liệu về đỉnh muối của một tỉnh ven biển Việt Nam. Kết quả tính toán cho thấy, mô hình đề xuất có độ chính xác dự báo cao hơn các mô hình hiện có khi áp dụng cho hai tập dữ liệu cụ thể.

DOI: https://doi.org/10.34238/tnu-jst.4720

Email: lethiluong88@gmail.com

1. Giới thiệu

Dự báo là quá trình đưa ra dự đoán dựa trên các dữ kiện quá khứ và các sự kiện liên quan, nhằm trợ giúp con người đưa ra quyết định tốt hơn trong những tình huống không chắc chắn. Tuy nhiên, dự báo giá trị tương lai của các sự kiện này với độ chính xác 100% là rất khó, nhưng hiệu quả dự báo và tốc độ của quá trình dự báo có thể được nâng cao. Trước đây, các mô hình hồi quy đã ảnh hưởng đáng kể đến vai trò trong dự báo bằng việc sử dụng phương pháp thống kê, nhưng chúng phải đối mặt trong thực tế với điều kiện dữ liệu không thể đáp ứng được. Các mô hình chuỗi thời gian không mờ như: mô hình trung bình trượt, trung bình hàm mũ và mô hình trung bình trượt tích hợp tự hồi quy (ARIMA) đã phần nào khắc phục được yếu điểm của mô hình hồi quy, tuy nhiên lại hoạt động kém khi có những thay đổi bất thường về dữ liệu hoặc chuỗi thời gian không ổn định. Để khắc phục những nhược điểm của các mô hình tuyến tính này, các mô hình tiên tiến đã được đề xuất, chẳng hạn như hồi quy đáp ứng đa biến [1], mạng nơron nhân tạo [2]... Tuy nhiên, các mô hình độc lập nêu trên vẫn còn nhiều hạn chế trong việc thực hiện các bài toán dự báo với tình huống thực tế. Chẳng hạn, các phương pháp truyền thống không thể xử lý các vấn đề dự báo trong đó dữ liệu lịch sử được biểu diễn dưới dạng ngôn ngữ hay các mô hình sử dụng mạng nơron cần số lượng lớn các quan sát để có được độ chính xác cao. Để khắc phục các hạn chế này, Song và Chissom [3] dựa trên lý thuyết tập mờ [4] đã đề xuất một mô hình dự báo chuỗi thời gian mờ (FTS) để giải quyết bài toán tuyển sinh đại học. Nối tiếp nghiên cứu này, Chen [5] đã phát triển mô hình FTS bậc 1và thu được các kết quả dự báo bằng các phép toán số học đơn giản thay vì các phép toán kết nhập max-min phức tạp [3]. Kết quả dự báo của Chen [5] tốt hơn nhiều so với các mô hình do Song và Chissom đề xuất [3]. Gần đây, nhiều nghiên cứu đã cung cấp một số cải tiến ở các giai đoạn khác nhau trong mô hình [5] như việc xác định độ dài khoảng hiệu quả bằng các kỹ thuật khác nhau [6], mờ hoá dữ liệu chuỗi thời gian [7], thiết lập quan hệ mờ [8], nhóm quan hệ mờ [9] và giải mờ [10]. Để tiếp tục nâng cao độ chính xác dự báo, nhiều nhà nghiên cứu đã đề xuất các mô hình FTS khác nhau để áp dụng dự báo vào các bài toán thực tế. Ví dụ, Chen et al. [11] đã giới thiệu một mô hình FTS mới để dự báo giá cổ phiếu bằng cách sử dụng lý thuyết trong dãy Fibonacci. Mô hình này dựa trên nền tảng của các mô hình FTS thông thường, có độ chính xác dự báo tốt hơn mô hình [5]. Thêm nữa, các công trình nghiên cứu trong [12] đã đề xuất các các mô hình FTS bậc cao nhằm khắc phục các hạn chế của các mô hình FTS bậc nhất [3], [5]. Để giảm thiểu thời gian tính toán phức tạp trong ma trận quan hệ mờ, Singh [13] đã đề xuất một phương pháp mới trong cách tiếp cận mô hình FTS. Li và Cheng [14] đã đưa ra mô hình FTS mới dựa trên số mờ hình thang để giải quyết ba vấn đề chính như hạn chế sự mơ hồ trong dự báo, phân khoảng một cách hợp lý và đảm bảo độ chính xác dự báo tốt với các độ dài khoảng khác nhau. Panigrahi và Bahera [15] đề xuất mô hình FTS kết hợp với kỹ thuật học máy (SVM) để giải quyết vấn đề liên quan đến việc xác định quan hệ mờ. Các phân tích so sánh cho thấy mô hình của họ đưa ra độ chính xác cao hơn so với các mô hình trong [3], [5], [16].

Như đã đề cập ở trên, việc xác định độ dài khoảng phù hợp và thiết lập các mối quan hệ mờ được coi là nhiệm vụ thách thức và ảnh hưởng đáng kể đến độ chính xác dự báo của mô hình FTS. Trong nghiên cứu này, chúng tôi trình bày một mô hình dự báo mới sử dụng kỹ thuật phân cụm dựa trên đồ thị dạng cây để xác định độ dài khoảng khác nhau khi áp dụng trên tập dữ liệu tuyển sinh Đại học Alabama và độ mặn đo được tại các Trạm quan trắc tỉnh Cà Mau.

2. Một số khái niệm cơ bản và thuật toán liên quan

Phần này tóm tắt một số khái niệm cơ bản về chuỗi thời gian mờ [3] và thuật toán phân cụm để làm cơ sở cho việc thiết lập mô hình dự báo.

2.1. Các khái niệm về chuỗi thời gian mờ [3]

Cho $Y(t)(t = \ldots 0,1,2,\ldots)$ là một tập con của tập số thực và cũng là tập nền, trên đó xác định các tập mờ $f_i(t)$. $F(t)$ là tập chứa các tập $f_i(t)$ $(i = 1,2,\ldots)$. Khi đó ta gọi $F(t)$ là chuỗi thời gian mờ xác định trên tập nền $Y(t)$.

Giả sử đặt $F(t-1) = A_i$ và $F(t) = A_j$, trong đó $F(t)$ được suy ra bởi $F(t-1)$. Quan hệ mờ giữa chúng được thay bởi quan hệ là: $A_i \rightarrow A_j$ và được gọi là mối quan hệ mờ bậc 1.

$F(t)$ là một chuỗi thời gian mờ. Nếu $F(t)$ được suy ra đồng thời bởi $F(t-1), F(t-2),\ldots, F(t-m)$, thì quan hệ giữa chúng được biểu diễn bởi $F(t-1), F(t-2), \ldots, F(t-m) \rightarrow F(t)$ và nó được gọi là mô hình chuỗi thời gian mờ bậc m một nhân tố.

2.2. Thuật toán phân cụm dựa trên đồ thị

Trong phần này, một phương pháp phân cụm dữ liệu thuộc lớp phân cụm dựa trên đồ thị để biểu diễn tập dữ liệu chuỗi thời gian thành các cụm được đề xuất. Phương pháp phân cụm đề xuất hiển thị tập dữ liệu dưới dạng cây nhị phân và tự động tạo các cụm thay vì số cụm cho trước. Cụ thể, trong bài báo này, phương pháp phân cụm dựa trên đồ thị được giới thiệu bằng một thuật toán bao gồm bốn thủ tục như sau:

(1) Thủ tục tìm nút gốc (Procedure of Finding Root Node - PFRN). Dựa trên chuỗi dữ liệu đầu vào, thủ tục này chỉ ra nút gốc.

(2) Thủ tục tạo cây (Tree Creation Procedure - TCP). Từ tập dữ liệu đầu vào và nút gốc, thủ tục này hiển thị cây.

(3) Thủ tục chèn nút vào cây (Node Insertion Procedure - NIP). Thủ tục này đưa các giá trị dữ liệu của chuỗi thời gian và nút gốc vào vị trí thích hợp trong cây.

(4) Thủ tục tạo các cụm (Node Clustering Procedure - NCP). Thủ tục này nhập vào cây được tạo bởi TCP và tạo ra các cụm dựa vào giá trị trên các nút.

Thuật toán phân cụm dữ liệu dựa trên đồ thị

Input: S ($x_1, x_2, \ldots x_n$)
Output: Clusters C ($c_1, c_2, \ldots c_k$)
BEGIN
(1) PROCEDURE_**PFRN** (S)
BEGIN
// Tinh (Rg) dựa vào giá trị lớn nhất và nhỏ nhất của S
Rg = $MAX_{value} - MIN_{value}$
For each i=1 to N
{ Mean = average(X_i)
$SD = Sqrt(\frac{1}{i} sum(X_i - Mean)^2)$
} $w = \frac{Rg}{SD*N}$
// Xác định tập nền U và giá trị gốc trên cây
U = [$MIN_{value} - w$, $MAX_{value} + w$];
Mid_u = (MIN_{value} + MAX_{value}) / 2 ;
Root = Mid_u
END;
---
(2) PROCEDURE_**TCP** (Root, S)
BEGIN
For each i = 1 to N
NIP(Root, X_i)
END;
---
(3) PROCEDURE_**NIP** (Root, S)
BEGIN
if (X_i < Root) **then**
if (Root.LEFT < > NULL) **then**
Call: NIP(Root. LEFT, X_i) **else**
Root.LEFT = NULL
end if

makeCluster(Root, minDiffnode**)**
}
if (minDiffnode == Root.RIGHT) **then**
if ((Root.RIGHT).LEFT < > NULL) **then**
add (Root.RIGHT).LEFT ; // chèn nút con này vào cụm
end if
if ((Root.RIGHT). RIGHT < >NULL) **then**
Call: NCP((Root.RIGHT).RIGHT)
end if Call: NCP(Root.LEFT)
else
if ((Root.LEFT). LEFT < >NULL) **then**
Call: NCP((Root.LEFT).LEFT)
end if
if ((Root.LEFT).RIGHT < > NULL) **then**
add ((Root.LEFT). RIGHT)
end if Call: NCP(Root. RIGHT)
end if
end if
else if (Root. RIGHT < > NULL && Root. LEFT == NULL) **then**
if Root is not presented in Cluster **then**
makeCluster(Root, Root.RIGHT)
if ((Root. RIGHT). LEFT < > NULL) **then**
add (Root. RIGHT). LEFT
end if
if ((Root.RIGHT). RIGHT < > NULL) **then**
Call: NCP((Root.RIGHT). RIGHT)
end if

```
else if (Xi> Root) then
    if (Root. RIGHT < > NULL) then
   Call: NIP(Root. RIGHT, Xi)
    Else   Root. RIGHT = NULL
end if
END;
-------------------------------------------------------
(4) PROCEDURE_NCP (Root)
BEGIN
if (Root == NULL) then
{
   "Nút gốc không tồn tại";
return
}
else if (Root.RIGHT < > NULL && Root.LEFT < > NULL) then
if (Root is not presented in Cluster) then
{
minDiffnode=makeDiff(Root,Root.RIGHT,Root.LEFT);
end if
else if (Root.RIGHT == NULL && Root.LEFT < > NULL) then
if Root is not presented in Cluster then makeCluster(Root, Root.LEFT)
if ((Root.LEFT). LEFT < > NULL) then Call: NCP((Root. LEFT). LEFT)
end if
if ((Root.LEFT). RIGHT < > NULL) then add ((Root. LEFT). RIGHT); // chèn nút con vào cụm
end if
end if
else if Root is not presented in the Cluster then  makeCluster(Root) end if
return
end if
END;
END.
```

3. Mô hình dự báo chuỗi thời gian mờ sử dụng kỹ thuật phân cụm dựa trên đồ thị

Trong phần này, mô hình dự báo chuỗi thời gian mờ kết hợp với kỹ thuật phân cụm dựa trên đồ thị được giới thiệu. Mô hình đề xuất được tổ chức thành hai giai đoạn chính: (1) Giai đoạn phân vùng dữ liệu dựa trên đồ thị được đề cập ở Bước 1; (2) Giai đoạn xây dựng mô hình dự báo FTS được đề cập từ Bước 2 đến Bước 7. Để thực hiện các bước trong mô hình dự báo đề xuất, tất cả dữ liệu tuyển sinh lịch sử [5] được sử dụng để minh họa quá trình phân cụm và xây dựng mô hình dự báo.

❖ **Giai đoạn phân vùng dữ liệu dựa trên đồ thị**

Bước 1: Phân tập dữ liệu lịch sử S thành các khoảng sử dụng thuật toán phân cụm đề xuất trong Phần 2.2.

Bước này, thuật toán phân cụm được áp dụng để biểu diễn tập dữ liệu chuỗi thời gian thành các cụm. Dựa trên các cụm đạt được, điều chỉnh các cụm thành các khoảng với độ dài khác nhau.

Bước 1.1: Áp dụng thuật toán phân cụm để phân dữ liệu thành các cụm .

Để phân vùng dữ liệu chuỗi thời gian thành các cụm, bốn thủ thục của thuật toán phân cụm dựa trên đồ thị trong Phần 2.2 được sử dụng. Kết quả của bốn thủ tục này trên tập dữ liệu tuyển sinh được giải thích ngắn gọn như sau:

1) Tạo nút gốc và tìm giá trị của nút gốc (PFRN)

Input: Chuỗi dữ liệu tuyển sinh : S (13055, 13563, 13867, . . . , 19328, 19337, 18876).

Tính Rg = $MAX_{value} - MIN_{value}$ = 6282;

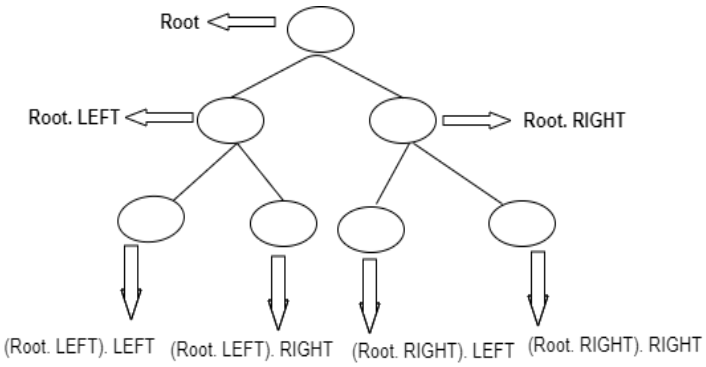
Tính độ lệch chuẩn SD = 1774.72; w = $\frac{R_g}{SD*N}$ = 0.16;

Tập nền được xác định: U = [MIN_{value} – w, MAX_{value} + w] = [13054.84, 19337.16];

Gái trị của nút gốc bằng điểm giữa của tập nền U: Mid_u = (MIN_{value}+ MAX_{value}) / 2 =16196; root = Mid_u =16196

2) Tạo cây phân cụm và chèn nút vào cây

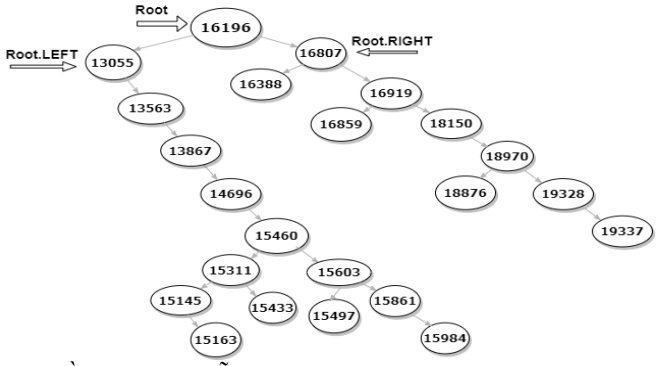
Từ tập dữ liệu đầu vào S và Root. Chúng tôi sử dụng hai thủ tục TCP và NIP để tạo cây và chèn các nút vào cây. Kết quả của hai thủ tục này được thể hiện trong Hình 1 và Hình 2 tương ứng.



Hình 1. *Đồ thị biểu diễn hình dạng cây được thực hiện bởi thủ tục TCP và NIP*

3) Tạo các cụm từ cây dựa vào thủ tục NCP

Sau khi có được cây dữ liệu trong Hình 2, quá trình tạo các cụm được giải thích ngắn gọn theo các điều kiện như sau:



Hình 2. *Cây biểu diễn dữ liệu đầu vào của chuỗi thời gian dựa trên hai thủ tục TCP và NIP với nút gốc là 16196*

1. Ban đầu, kiểm tra xem Root có tồn tại hay không và Root có chứa cây con trái hay con phải hay không.

2. Nếu cả hai con tồn tại cho mỗi Root thì tính toán sự khác biệt giữa các giá trị của Root và (Root. RIGHT), Root và (Root. LEFT). Sau đó, tạo cụm với các nút con tương ứng (Root. LEFT hoặc Root. RIGHT) với sự khác biệt so với Root là nhỏ hơn.

3. Nếu chỉ có một con tồn tại cho mỗi Root thì tạo cụm theo Root và (Root. LEFT) hoặc Root và (Root. RIGHT).

4. Lặp lại các điều kiện 2-3, cho đến khi tất cả giá trị của các nút trong cây được thêm vào các cụm.

Dựa trên các thủ tục của thuật toán phân cụm trên, chúng tôi đạt được 10 cụm và các phần tử tương ứng của chúng. Kết quả phân cụm đạt được chỉ ra trong Bảng 1 như sau:

Bảng 1. *Các phần tử trong cụm và tâm cụm tương ứng*

Số cụm	Các phần tử trong cụm
C1	(16196, 16807, 16388)
C2	(16919, 16859)
C3	(18150, 18970, 18876)
--	----------
C9	(15311, 15433)
C10	(15145, 15163)

Bước 1.2: Điều chỉnh các cụm thành các khoảng với độ dài khác nhau.

Để đạt được các khoảng từ các cụm trong Bước 1.1, chúng tôi lấy giá trị nhỏ nhất và lớn nhất của các cụm C_i là giá trị cận trên và cận dưới của khoảng u_i. Các khoảng thu được chỉ ra trong trong Bảng 2.

Bảng 2. *Kết quả các khoảng thu được từ thuật toán phân cụm*

Số khoảng	Khoảng	Giá trị điểm giữa
1	u_1= [16196, 16807]	16292
2	u_2= [16859, 16919]	16889
--	----------	-----
9	u_9= [15311, 15433]	15372
10	u_{10}= [15145, 15163]	15154

❖ **Giai đoạn xây dựng mô hình dự báo chuỗi thời gian mờ**

Trong giai đoạn này, sử dụng các bước dự báo được đề xuất bởi công trình [17] làm cơ sở để thiết lập mô hình dự báo FTS. Các bước tiếp theo của mô hình đề xuất được tóm tắt như sau:

Bước 2. Xác định các tập mờ cho các quan sát trên mỗi khoảng thu được ở Bước 1.

Bước 3: Mờ hóa dữ liệu lịch sử dựa trên các tập mờ đã xác định.

Bước 4: Xác định các quan hệ mờ.

Bước 5: Thiết lập nhóm quan hệ mờ phụ thuộc thời gian.

Bước 6: Giải mờ và tính giá trị dự báo đầu ra.

Bước 7: Tính độ chính xác dự báo của mô hình.

Hai tiêu chí như: sai số trung bình bình phương MSE (*mean square error*) và MAPE *(mean absolute percentage error*) được sử dụng để so sánh độ chính xác dự báo giữa mô hình đề xuất và các mô hình khác. Giá trị của hàm MSE và MAPE được tính theo công thức (1) và (2) sau:

$$MSE = \frac{1}{n}\sum_{i=\lambda}^{n}(F_i - R_i)^2 \quad (1)$$

$$MAPE = \frac{1}{n}\sum_{i=\lambda}^{n}\left|\frac{F_i - R_i}{R_i}\right| * 100\% \quad (2)$$

Trong đó: F_i giá trị dự báo tại thời điểm i, R_i là giá trị thực tại thời điểm i, n là tổng số dữ liệu tham gia dự báo, λ là bậc của quan hệ.

4. Tổ chức thực nghiệm và đánh giá kết quả

Trong bài báo này, mô hình dự báo đề xuất được áp dụng trên hai chuỗi dữ liệu, đó là dữ liệu tuyển sinh của Đại học Alabama [5] và dữ liệu về độ mặn đo được tại các trạm quan trắc tỉnh Cà Mau. Trước khi triển khai mô hình dự báo đề xuất, các tập dữ liệu chuỗi thời gian được mô tả ngắn gọn. Sau đó, các kết quả mô phỏng và phân tích liên quan đến các tập dữ liệu này được đưa ra. Các đặc điểm thống kê của hai chuỗi thời gian này được thể hiện như sau.

4.1. Mô tả chuỗi dữ liệu thời gian

(1) Chuỗi dữ liệu tuyển sinh của trường Đại học Alabama: Tập dữ liệu tuyển sinh chứa 22 quan sát trong khoảng thời gian từ 1971 đến 1992. Tập dữ liệu kinh điển này đã được số lượng lớn các công trình nghiên cứu [3], [5], [6], [9], [10] sử dụng làm mô phỏng và đưa ra kết quả dự báo tin cậy. Một trong số kết quả thu được trong các công trình này cũng được sử dụng để so sánh với mô hình đề xuất.

(2) Dữ liệu đỉnh mặn trên địa bàn tỉnh Cà Mau, Việt Nam bao gồm ba trạm đo chính là: Sông Cửa Lớn (CL), sông Gành Hào (GH) và Ông Đốc (OĐ). Dữ liệu này được cung cấp bởi Đài Khí tượng Thủy văn khu vực Nam Bộ, đặt tại Thành phố Hồ Chí Minh, giai đoạn 2000 – 2017 bao gồm 17 quan sát trên mỗi trạm.

4.2. Thử nghiệm và áp dụng dự báo trên các tập dữ liệu khác nhau

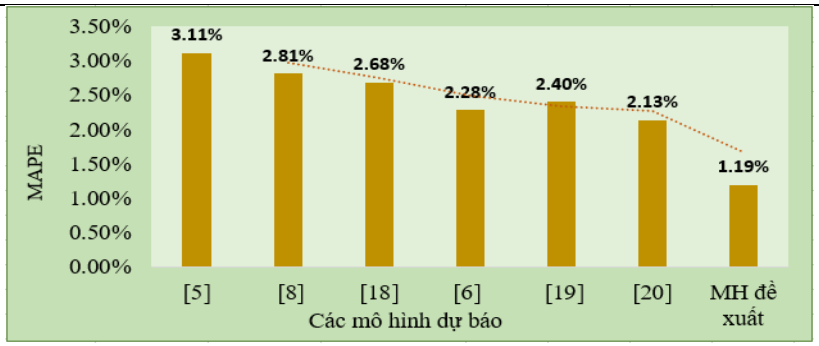
4.2.1. Áp dụng dự báo tuyển sinh đại học

Để xác minh quả dự báo của mô hình dựa trên quan hệ mờ bậc nhất với số khoảng chia khác nhau, kết quả dự báo thu được từ mô hình đề xuất được so sánh với kết quả dự báo của các mô

hình trong các nghiên cứu [5], [6], [8], [18]-[20]. Kết quả dự báo và độ chính xác MSE (1) giữa mô hình đề xuất và các mô hình khác được đưa ra trong Bảng 3. Trong đó, cột thứ 1 và cột thứ 2 thể hiện dữ liệu năm dự báo và dữ liệu tuyển sinh thực tế. Các cột còn lại là kết quả dự báo tương ứng với các mô hình được chọn để so sánh.

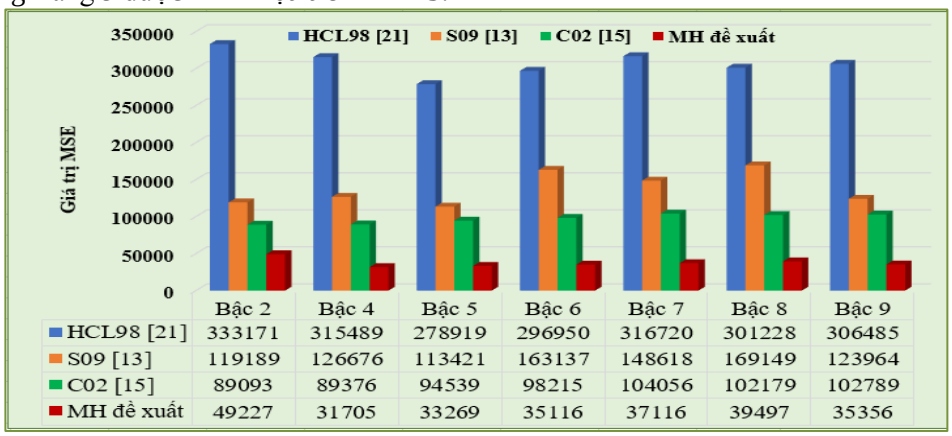
Bảng 3. *So sánh mô hình đề xuất với các mô hình khác dựa trên chuỗi thời gian bậc 1 với 10 khoảng chia*

Year	Actual	[5]	[8]	[18]	[6]	[19]	[20]	MH đề xuất
1971	13055	-	-	-	-	-	-	
1972	13563	14000	13486	13944	14279	14242	13820	13309
1973	13867	14000	14156	13944	14279	14242	13820	13957.33
---	---	---	---	---	---	---	---	---
1991	19337	19000	18808	18933	19257	19144	19135	19332.5
1992	18876	19000	18808	18933	19257	19144	19135	18817.5
MSE		**407707**	**334431**	**255959**	**198203**	**228920**	**194746**	**57473**



Hình 3. *Đồ thị biểu diễn độ chính xác MAPE giữa mô hình đề xuất với các mô hình khác*

Kết quả trong Bảng 3 cho thấy, mô hình đề xuất có sai số dự báo (MSE = 57473) nhỏ nhất trong số tất cả các mô hình so sánh dựa trên quan hệ mờ bậc nhất với số khoảng chia bằng 10. Điểm khác biệt chủ yếu giữa mô hình đề xuất và các mô hình so sánh là cách thức nhóm quan hệ mờ và kỹ thuật chia khoảng được sử dụng. Điểm khác biệt này chứng tỏ rằng, mô hình dự báo đề xuất hiệu quả hơn so với mô hình được so sánh khi thử nghiệm trên tập dữ liệu tuyển sinh Đại học Alabama. Trực quan hơn có thể thấy, độ chính xác phần trăm MAPE của các mô hình so sánh trong Bảng 3 được minh họa trên Hình 3.



Hình 4. *So sánh độ chính xác dự báo MSE giữa mô hình đề xuất và các mô hình khác dựa trên quan hệ mờ bậc cao với số khoảng chia khác nhau*

Thêm nữa, mô hình đề xuất cũng được mô phỏng dựa trên quan hệ mờ bậc cao khác nhau từ bậc 2 đến bậc 9 với số khoảng chia được cố định là 10 khoảng. Để xác minh tính hiệu quả của mô hình dự báo dựa trên chuỗi thời gian mờ bậc cao, ba mô hình có tên là HCL [21], S09 [13] và C02 [15] được lựa chọn cho việc so sánh với mô hình đề xuất. Từ kết quả so sánh về độ chính

xác dự báo MSE (1) liệt kê trong Hình 4 cho thấy, mô hình đề xuất đưa ra sai số dự báo nhỏ hơn so với các mô hình được chọn để so sánh trong tất cả các bậc với số khoảng chia bằng 10, đặc biệt nhận được giá trị (MSE = 31705) nhỏ nhất trong trường hợp quan hệ mờ bậc 4.

4.2.2. Áp dụng dự báo đỉnh mặn tại tỉnh Cà Mau

Trong phần này, mô hình dự báo đề xuất được áp dụng để dự báo đỉnh mặn tại ba trạm đo trên địa bàn tỉnh Cà Mau. Từ số liệu trích dẫn bởi công trình [22], chúng tôi lần lượt dự báo độ mặn tại trạm Cửa Lớn, Gành Hào và Ông Đốc. Kết quả dự báo tại các trạm thu được từ mô hình đề xuất được ghi trong Bảng 4.

Bảng 4. *Kết quả và độ chính xác dự báo của mô hình đề xuất dựa trên quan hệ mờ bậc 1*

Năm	Cửa Lớn		Gành Hào		Ông Đốc	
	DL thực	DL dự báo	DL thực	DL dự báo	DL thực	DL dự báo
2000	29,6		31,5		30,8	
2001	29,4	29,67	30,8	30,75	31,8	32,2
2002	34,4	32,31	30,5	30,75	34,7	34,55
---	---	---	---	---	---	---
2016	35,9	33,56	32,9	32,21	37,9	35,93
2017	36,5	34,62	33,7	33,08	38,8	35,58
MSE	**2,217**		**0,260**		**2,279**	

Quan sát Bảng 4 thấy rằng, dữ liệu dự báo được từ mô hình đề xuất khá bám sát với dữ liệu thực tế tương ứng với từng trạm đo trên địa bàn tỉnh Cà Mau. Dựa vào độ chính xác MSE trên Bảng 4 cho thấy sự tác động rất lớn của độ dài khoảng chia từ thuật toán phân cụm trong mô hình đề xuất trên mỗi tập dữ liệu khác nhau.

Để chứng minh tính ưu việt của mô hình dự báo đề xuất trên tập dữ liệu về độ mặn, độ chính xác của mô hình tham chiếu trong công trình [22] được lựa chọn để so sánh. Kết quả so sánh giữa mô hình đề xuất và mô hình này dựa trên hai tiêu chí đánh giá MSE (1) và MAPE (2) đưa ra trong Bảng 5. Quan sát các giá trị MSE và MAPE cho thấy hiệu quả dự báo của mô hình đề xuất vượt trội hơn mô hình [22].

Bảng 5. *Kết quả so sánh độ chính xác dự báo giữa mô hình đề xuất với MH [22]*

Dữ liệu	Mô hình	MSE	MAPE
Cửa Lớn	MH [22]	38,928	5,167
	MH đề xuất	**2,217**	**3,700**
Gành Hào	MH [22]	8,376	2,509
	MH đề xuất	**0,260**	**1,114**
Ông Đốc	MH [22]	47,096	5,854
	MH đề xuất	**2,279**	**3,075**

5. Kết luận

Nghiên cứu này đề xuất một mô hình dự báo chuỗi thời gian mờ mới sử dụng kỹ thuật phân cụm dựa trên đồ thị nhằm cải thiện hiệu suất dự báo trong các ứng dụng khác nhau. Trong mô hình dự báo đề cập đến hai vấn đề chính được xem là ảnh hưởng lớn đến độ chính xác dự báo, đó là vấn đề xác định khoảng chia từ tập nền và cách thiết lập nhóm quan hệ mờ. Để khắc phục những hạn chế của các mô hình chuỗi thời gian mờ cùng sử dụng nhóm quan hệ mờ, mô hình đề xuất sử dụng khái niệm nhóm quan hệ mờ phụ thuộc thời gian được chứng minh là hiệu quả và phù hợp với điều kiện thực tế hơn. Thêm nữa, thuật toán phân cụm mới dựa trên đồ thị được đề xuất để xác định độ dài khoảng chia khác nhau trong mô hình chuỗi thời gian mờ nhằm khắc phục những nhược điểm của các mô hình sử dụng độ dài khoảng bằng nhau. Từ kết quả thu được trong các Bảng 3, 4 và 5 cho thấy, việc sử dụng phương pháp phân khoảng có kích thước khác nhau có thể tạo ra độ chính xác dự báo tốt hơn so với các khoảng có kích thước bằng nhau, dẫn đến hiệu quả dự báo vượt trội hơn so với một số mô hình dự báo trước đây. Tuy nhiên, mô hình

dự báo hiện tại chỉ được áp dụng đối với chuỗi thời gian mờ một nhân tố. Kỳ vọng trong thời gian tới, mô hình đề xuất sẽ được mở rộng và phát triển trên các tập dữ liệu có nhiều nhân tố hơn.

TÀI LIỆU THAM KHẢO/ REFERENCES

[1] J. H. Friedman, "Multivariate adaptive regression splines," *Annals of Statistics*, vol. 19, no. 1, pp. 1–141, 1991.

[2] S. Aladag, C. H. Aladag, T. Mentes, and E. Egrioglu, "A new seasonal fuzzy time series method based on the multiplicative neuron model and SARIMA," *Hacettepe Journal of Mathematics and Statistics*, vol. 41, no. 3, pp. 145-163, 2012.

[3] Q. Song and B. S. Chissom, "Forecasting enrollments with fuzzy time series – Part I," *Fuzzy Sets and Systems*, vol. 54, no. 1, pp. 1-9, 1993.

[4] L. A. Zadeh, "Fuzzy sets," *Information and Control*, vol. 8, no. 3, pp. 338-353, 1965.

[5] S. M. Chen, "Forecasting enrollments based on fuzzy time series," *Fuzzy Sets and Systems*, vol. 81, pp. 311-319, 1996.

[6] W. Lu, et al., "Using interval information granules to improve forecasting in fuzzy time series," *International Journal of Approximate Reasoning*, vol. 57, pp. 1-18, 2015.

[7] J. R. Hwang, S. M. Chen, and C. H. Lee, "Handling forecasting problems using fuzzy time series," *Fuzzy Sets and Systems*, vol. 100, pp. 217-228, 1998.

[8] K.-H. Huarng and T. H.-K. Yu, "Modeling fuzzy time series with multiple observations," *International Journal of Innovative Computing, Information and Control*, vol. 8, no.10(B), pp. 7415-7426, 2012.

[9] N. Van Tinh and N. C. Dieu, "A new hybrid fuzzy time series forecasting model based on combining fuzzy c-means clustering and particle swam optimization," *Journal of Computer Science and Cybernetics*, vol. 35, no. 3, pp. 267-292, 2019.

[10] P. Singh and B. Borah, "An efficient time series forecasting model based on fuzzy time series," *Engineering Applications of Artificial Intelligence*, vol. 26, pp. 2443-2457, 2013.

[11] T.-L. Chen, C.-H. Cheng, and H. J. Teoh, "Fuzzy time-series based on Fibonacci sequence for stock price forecasting," *Physica A: Statistical Mechanics and its Applications*, vol. 380, pp. 377-390, 2007.

[12] R. M. Pattanayak, S. Panigrahi, H. S. Behera, "High order fuzzy time series forecasting by membership values along with data and support vector machine," *Arabian J. of Scien. and Engg*., vol. 45, pp. 7865-7867, 2020.

[13] S. R. Singh, "A robust method of forecasting based on fuzzy time series," *Applied Mathematics and Computation*, vol. 188, no. 1, pp. 472-484, 2007.

[14] S.-T. Li and Y.-C. Cheng, "Deterministic fuzzy time series model for forecasting enrollments," *Computers and Mathematics with Applications*, vol. 53, no. 12, pp. 1904-1920, 2007.

[15] S. Panigrahi and H. S. Behera, "A study on leading machine learning techniques for high order fuzzy time series forecasting," *Eng. Appl. Artif. Intell.*, vol. 87, pp. 1-10, 2020.

[16] N.-Y. Wang and S.-M. Chen, "Temperature prediction and TAIFEX forecasting based on automatic clustering techniques and two-factors high-order fuzzy time series," *Expert Systems with Applications*, vol. 36(2), Part 1, pp. 2143-2154, 2009.

[17] N. C. Dieu and N. V. Tinh, "Fuzzy time series forecasting based on time depending fuzzy relationship groups and particle swarm optimization," Proceedings of the 9th National Conference on Fundamental and Applied Information Technology Research (FAIR'9), Can Tho, Viet Nam, 2016, pp. 125-133.

[18] L. Wang, X. Liu, W. Pedrycz, and Y. Shao, "Determination of temporal information granules to improve forecasting in fuzzy time series," *Expert Syst. Appl.*, vol. 41, no. 6, pp. 3134-3142, 2014, doi: http://dx.doi.org/10.1016/j.eswa.2013.10.046.

[19] C. H. Cheng, G. W. Cheng, and J. W. Wang, "Multi-attribute fuzzy time series method based on fuzzy clustering," *Expert Systems with Applications*, vol. 34, pp. 1235-1242, 2008.

[20] T. Hoang, D. T. Nguyen, and M. L. Vu, "The partitioning method based on hedge algebras for fuzzy time series forecasting*," Journal of Science and Technology*, vol. 54, no. 5, pp. 571-583, 2016.

[21] J. R. Hwang, S. M. Chen, and C. H. Lee, "Handling forecasting problems using fuzzy time series," *Fuzzy Sets and Systems*, vol. 100, no. 1-3, pp. 217-228, 1998.

[22] V. V. Tai et al., "An improved fuzzy time series forecasting model," (in Vietnamese), *Can Tho University Journal of Science*, vol. 56(1A), pp. 68-94, 2020