Kỷ yếu Hội nghị KHCN Quốc gia lần thứ XI về Nghiên cứu cơ bản và ứng dụng Công nghệ thông tin (FAIR); Hà Nội, ngày 09-10/8/2018
DOI: 10.15625/vap.2018.00034

MÔ HÌNH DỰ BÁO CHUỖI THỜI GIAN MỜ DỰA TRÊN ĐẠI SỐ GIA TỬ VÀ TỐI ƯU BẦY ĐÀN

Nghiêm Văn Tính[1*], Nguyễn Công Điều[2], Nguyễn Tiến Duy[1]

[1] Trường Đại học Kỹ thuật Công nghiệp - Đại học Thái Nguyên
[2] Trường Đại học Thăng Long

nghiemvantinh@tnut.edu.vn, ncdieu@yahoo.com, duy.infor@tnut.edu.vn

TÓM TẮT: *Những năm gần đây, nhiều mô hình dự báo dựa trên chuỗi thời gian mờ được đề xuất nhằm phân tích chuỗi thời gian. Trong mô hình dự báo, các yếu tố chính có thể ảnh hưởng đến độ chính xác dự báo của mô hình là độ dài của khoảng chia tập nền và nhóm quan hệ mờ. Trong bài báo này, chúng tôi đề xuất mô hình dự báo chuỗi thời gian mờ mới dựa trên đại số gia tử (ĐSGT) và thuật toán tối ưu bầy đàn (PSO). Trong đó, ĐSGT được sử dụng như một công cụ để chia tập nền thành các khoảng có độ dài khác nhau tương ứng với các khoảng ngữ nghĩa tính toán được của các hạng từ ngôn ngữ. Sau quá trình chia khoảng, các giá trị quan sát được biểu diễn bởi các tập mờ và sử dụng chúng để thiết lập các nhóm quan hệ mờ. Cuối cùng, mô hình đề xuất được kết hợp với kỹ thuật PSO để tìm ra khoảng chia phù hợp nhằm tăng độ chính dự báo của mô hình. Đánh giá hiệu quả của mô hình trên tập dữ liệu kinh điển về số lượng sinh viên nhập học tại Đại học Alabama. Thực nghiệm cho thấy mô hình đề xuất đưa ra kết quả dự báo chính xác hơn một số mô hình dự báo đã được công bố gần đây dựa vào chuỗi thời gian mờ bậc 1 và bậc cao.*

Từ khóa: *Chuỗi thời gian mờ, Nhóm quan hệ mờ phụ thuộc thời gian, Tối ưu bày đàn, PSO, Tuyển sinh.*

I. GIỚI THIỆU

Trong vài thập kỷ qua, nhiều mô hình dự báo đã được đề xuất nhằm giải quyết các bài toán dự báo khác nhau để giúp con người đưa ra các quyết định, như: dự báo tuyển sinh đại học cho năm tiếp theo, dự báo nhiệt độ cho các ngày tới, dự báo dân số hàng năm, dự báo tài chính,... Dựa trên lý thuyết tập mờ, Song và Chissom đã đưa ra hai mô hình chuỗi thời gian mờ không phụ thuộc thời gian [1] và phụ thuộc thời gian [2] bằng việc sử dụng các phép toán max - min trong quan hệ mờ để giải quyết bài toán dự báo tuyển sinh đại học của Trường Đại học Alabama. So sánh với các mô hình dự báo truyền thống trước đây như: Phân tích hồi huy, trung bình trượt, trung bình hàm mũ và mô hình ARIMA thì các mô hình [1], [2] có thể giải quyết tốt hơn đối với các bài toán dự báo có chuỗi số liệu được biểu diễn bởi giá trị ngữ nghĩa hay chuỗi dữ liệu không chắc chắn. Hơn nữa, các mô hình chuỗi thời gian mờ này, không yêu cầu số lượng quan sát lớn hay giả định tuyến tính như mô hình truyền thống. Tuy nhiên, các mô hình [1], [2] mất nhiều thời gian tính toán khi xử lý với ma trận mờ lớn. Do đó, để khắc phục hạn chế này, Chen [3] đã đưa ra phương pháp mới khá hiệu quả bằng việc sử dụng các phép tính số học đơn giản thay vì các phép tính kết hợp max-min phức tạp trong xử lý mối quan hệ mờ. Từ việc mở rộng của công trình [3] thành mô hình chuỗi thời gian mờ bậc cao [4] và mức ảnh hưởng của độ dài khoảng trong mô hình [5] cùng với việc phát triển từ các mô hình một nhân tố thành mô hình chuỗi thời gian mờ hai nhân tố [6] là nền tảng cho sự phát triển mạnh mẽ của mô hình chuỗi thời gian mờ trong những khoảng thời gian tiếp sau. Gần đây, nhiều tác giả đã sử dụng các kỹ thuật khác nhau vào từng pha (giai đoạn) trong mô hình chuỗi thời gian mờ nhằm nâng cao độ chính xác dự báo. Chen và Tanuwijaya [7] đã sử dụng phương pháp phân cụm tự động để chia tập nền thành các khoảng có độ dài khác nhau trong pha mờ hóa dữ liệu của mô hình. Một số tác giả khác dựa dựa trên kỹ thuật tối ưu kết hợp với các mô hình chuỗi thời gian mờ khác nhau nhằm điều chỉnh lại các khoảng chia từ tập nền [8]-[19]. Dựa trên tư tưởng tìm độ dài khoảng tối ưu, một số mô hình lại dùng kỹ thuật phân cụm để phân tập dữ liệu quan sát thành các cụm, sau đó điều chỉnh các cụm này thành các khoảng có độ dài khác nhau như: Phân cụm K-mean [20], [21] phân cụm C-mean [22], [23]. Một cách tiếp cận hoàn toàn khác biệt dựa trên lý thuyết đại số gia tử [24] để ngữ nghĩa hóa và giải ngữ phi tuyến [25] thay vì các phép mờ hóa dữ liệu và giải mờ dự báo trong mô hình chuỗi thời gian mờ. Cũng dựa trên đại số gia tử, trong công trình [26] sử dụng nó để phân chia tập nền thành các khoảng khác nhau bằng việc ánh xạ ngữ nghĩa của các hạng từ ngôn ngữ thành các khoảng mờ. Hai công trình theo hướng tiếp cận ĐSGT nêu trên chỉ tập trung vào xây dựng mô hình dự báo bậc 1 để dự báo số lượng sinh viên nhập học của Trường Đại học Alabama.

Dựa vào sự phân tích của các công trình trên cho thấy, độ dài khoảng và bậc của nhóm quan hệ mờ là các yếu tố ảnh hưởng rất lớn đến độ chính xác dự báo của mô hình. Bài báo này, chúng tôi đề xuất mô hình dự báo chuỗi thời gian mờ bậc một và bậc cao dựa trên ĐSGT và PSO cho bài toán tuyển sinh đại học trong [3]. Trong nghiên cứu này, trước tiên ĐSGT được sử dụng để phân chia tập nền thành các khoảng có độ dài khác nhau bằng cách định lượng chính các hạng từ ngôn ngữ dùng để biểu diễn chuỗi dữ liệu quan sát. Sau đó, tính giá trị đầu ra dự báo cho các nhóm quan hệ mờ bậc 1 và bậc cao đã được chúng tôi đề xuất trong công trình [14] bằng quy tắc giải mờ mới. Cuối cùng, mô hình đề xuất được kết hợp với thuật toán PSO để hiệu chỉnh lại độ dài khoảng chia ban đầu nhằm cải thiện độ chính xác dự báo hơn nữa.

Phần còn lại của bài báo được bố cục như sau: Phần II trình bày một số khái niệm liên quan đến chuỗi thời mờ và ĐSGT. Phần III giới thiệu từng bước của mô hình dự báo kết hợp giữa ĐSGT và PSO. Phần IV đánh giá hiệu quả dự báo của mô hình đề xuất so với các mô hình dự báo trước đây. Cuối cùng, các kết luận được đưa ra trong phần V.

II. MỘT SỐ KHÁI NIỆM CƠ BẢN VÀ THUẬT TOÁN LIÊN QUAN

Trong phần này tóm tắt một số khái niệm cơ bản về chuỗi thời gian mờ [1]- [3] và đại số gia tử [24] để làm cơ sở cho nghiên cứu này.

2. 1. Khái niệm cơ bản về chuỗi thời gian mờ (FTS)

Điểm khác chủ yếu giữa chuỗi thời gian mờ và khái niệm chuỗi thời gian truyền thống là giá trị của chuỗi thời gian được biểu diễn bởi các tập mờ (hay các nhãn ngôn ngữ), trong khi chuỗi thời gian truyền thống được biểu diễn bởi các giá trị số. Một số định nghĩa cơ bản về chuỗi thời gian mờ được đưa ra như sau:

Định nghĩa 1: Chuỗi thời gian mờ [1]

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 $Y(t)$.

Định nghĩa 2: Quan hệ mờ (FLR) [1]

Tại các thời điểm t và t-1 có tồn tại một mối quan hệ mờ giữa $F(t)$ và $F(t-1)$ sao cho $F(t) = F(t-1) * R(t-1, t)$; trong đó * là toán tử max-min xác định trên tập mờ. $R(t-1, t)$ là mối quan hệ mờ. Ta cũng có thể ký hiệu mối quan hệ mờ giữa $F(t)$ và $F(t-1)$ bởi $F(t-1) \rightarrow F(t)$. Nếu đặt $F(t-1) = A_i$ và $F(t) = A_j$ thì mối quan hệ logic mờ giữa chúng được thay bởi quan hệ là: $A_i \rightarrow A_j$. Viết như thế này có thể hiểu là tập mờ A_j được suy ra từ tập mờ A_i.

Định nghĩa 3: Nhóm quan hệ mờ (FLRGs) [3]

Các quan hệ mờ trong tập luyện có thể gom thành một nhóm nếu các tập mờ bên vế phải của quan hệ có cùng các tập mờ bên vế trái thì gộp chúng thành một nhóm theo vế trái của quan hệ. Giả sử có các quan hệ logic mờ bậc một có cùng các tập mờ bên vế trái như sau:

$$A_i \rightarrow A_{k1}\ ; A_i \rightarrow A_{k2}; \ldots .; A_i \rightarrow A_{km}\ .$$

Theo Chen [3], các quan hệ này được gom thành một nhóm như sau: $A_i \rightarrow A_{k1}A_{k2}, \ldots, A_{km}$. Các quan hệ giống nhau (lặp lại) chỉ được tính duy nhất một lần khi tham gia vào nhóm quan hệ mờ.

Định nghĩa 4: Nhóm quan hệ mờ phụ thuộc thời gian [14]

Quan hệ mờ giữa hai quan sát liên tiếp F(t-1) và $F(t)$ được biểu diễn bởi $F(t\text{-}1) \rightarrow F(t)$. Nếu, đặt $F(t) = A_i(t)$ và $F(t-1) = A_j(t-1)$, thì quan hệ tại thời điểm t này được biểu diễn thành $A_j(t-1) \rightarrow A_i(t)$.

Nếu cũng tại thời điểm t, tồn tại các quan hệ sau: $A_j(t1-1) \rightarrow A_{i1}(t1); \ldots; A_j(tp-1) \rightarrow A_{ip}(tp)$ và $A_j(t-1) \rightarrow A_i(t)$ với $t1, t2, .., tp \leq t$. Nghĩa là các quan hệ tại thời điểm t1, t2, …, tp xảy ra trước quan hệ mờ tại thời điểm t, nhưng có cùng tập mờ bên vế trái là $A_j(t-1)$. Khi đó các quan hệ này được nhóm thành một nhóm quan hệ mờ là $A_j(t-1) \rightarrow A_{i1}(t1), A_{i2}(t2), A_{ip}(tp), A_i(t)$ và được gọi là nhóm quan hệ mờ phụ thuộc vào thời gian.

Ví dụ sau đây có thể hiểu rõ hơn về nhóm quan hệ mờ phụ thuộc thời gian và các nhóm quan hệ thông thường [3]. Giả sử tồn tại các quan hệ mờ tại các thời điểm khác nhau như sau:

$t = 1$ có quan hệ logic mờ $A_i \rightarrow A_j$;
$t = 2$ có quan hệ logic mờ $A_i \rightarrow A_k$;
$t = 3$ có quan hệ logic mờ $A_i \rightarrow A_j$.

Trong các quan hệ trên có hai quan hệ mờ giống nhau xuất hiện tại các thời điểm $F(t = 1)$ và $F(t = 3)$. Theo Chen [3], thì các quan hệ mờ giống nhau chỉ được tính một lần khi tham gia vào nhóm quan hệ mờ. Khi đó các quan hệ nói trên được gộp thành một nhóm quan hệ có dạng: $A_i \rightarrow A_j, A_k$. Điều đó có nghĩa rằng các quan hệ trùng nhau không được xem xét và dẫn đến thiếu thông tin trong quá trình dự báo. Do vậy, trong nhóm quan hệ đề xuất, chúng tôi xem xét đến thời điểm xuất hiện của các quan hệ mờ bên phải ở tại thời điểm dự báo t nào đó. Cùng ví dụ trên, giả sử thời điểm t =2, chúng tôi chỉ xét đến các quan hệ có cùng trạng thái bên trái mà có tập mờ bên phải xuất hiện từ thời điểm dự báo trở về trước thì được gộp thành một nhóm quan hệ có dạng: $A_i \rightarrow A_j, A_k$. Tương tự tại thời điểm dự báo t=3 thì nhóm quan hệ khác được thiết lập là $A_i \rightarrow A_j, A_k, A_j$.

Tư tưởng này cũng được áp dụng tương tự cho quan hệ bậc cao và được gọi là quan hệ mờ phụ thuộc thời gian bậc cao.

2.2. Cơ bản về đại số gia tử (ĐSGT) [24]

Giả sử ta có một tập các giá trị ngôn ngữ của biến ngôn ngữ $\mathcal{X} = \{Very\ Very\ small < Very\ small < small < Little\ small < Very\ Little\ small < medium < Very\ Little\ high < Little\ big < high < Very\ high < \cdots\}$. Các giá trị ngôn ngữ này được sử dụng trong các bài toán lập luận xấp xi dựa trên tri thức bằng luật. Một vấn đề đặt ra là cần có một cấu trúc đủ mạnh dựa trên tính thứ tự vốn có của giá trị ngôn ngữ trong miền của biến ngôn ngữ. Từ đó, có thể tính toán được ngữ nghĩa trên giá trị ngôn ngữ của biến ngôn ngữ trong các bài toán suy luận xấp xỉ.

Mỗi biến ngôn ngữ $\mathcal{X}$ được biểu thị như một cấu trúc đại số $\mathcal{AX} = (X, G, C, H, \leq)$, gọi là đại số gia tử, trong đó X là tập các hạng từ trong X; $\leq$ biểu thị mối quan hệ thứ tự ngữ nghĩa tự nhiên của các hạng từ trên X; $G = \{c^-, c^+\}$, $c^- \leq c^+$, được gọi là các phần tử sinh (ví dụ: $G = \{small < big\}$); $C = \{\mathbf{0}, \boldsymbol{W}, \mathbf{1}\}$ là tập các hằng, với $\mathbf{0} \leq c^- \leq \boldsymbol{W} \leq c^+ \leq \mathbf{1}$, để chỉ các phần tử có ngữ nghĩa nhỏ nhất, lớn nhất và phần tử trung hoà (ví dụ: $\boldsymbol{W} = medium$); $H = H^- \cup H^+$, với $H^- = \{h_{-q} \geq \cdots \geq h_{-2} \geq h_{-1}\}$ là tập các gia tử âm, $\forall h \in H^-$ thì $hc^+ \leq c^+$ và $H^+ = \{h_1 \leq h_2 \leq \cdots \leq h_p\}$ là các gia tử dương, $\forall h \in H^+$ thì $hc^+ \geq c^+$. Ví dụ $H^- = \{Little > Rather\}$, $H^+ = \{More < Very\}$. Với $x \in X$, $x = h_n h_{n-1} \ldots h_1 c$, $h_j \in H$, $c \in G$. Với quan hệ thứ tự giữa các phần tử sinh, các gia tử và chiều tác động của các gia tử như trên, có thể được biểu thị bằng dấu của chúng như sau:

Hàm dấu: $sgn: X \to \{-1, 0, 1\}$ được định nghĩa một cách đệ quy như sau: Với $k, h \in H$, $c \in G$, $x \in X$

$sgn(c^+) = +1$ *và* $sgn(c^-) = -1$ (2. 1)

$\{h \in H^+ | sgn(h) = +1\}$ *và* $\{h \in H^- | sgn(h) = -1\}$ (2. 2)

$sgn(hc^+) = +sgn(c^+)$ *nếu* $hc^+ \geq c^+$ *hoặc* $sgn(hc^-) = +sgn(c^-)$ *nếu* $hc^- \leq c^-$*và* $sgn(hc^+) = -sgn(c^+)$ *nếu* $hc^+ \leq c^+$ *hoặc* $sgn(hc^-) = -sgn(c^-)$ *nếu* $hc^- \geq c^-$. *Hay* $sgn(hc) = sgn(h)sgn(c)$. (2. 3)

$sgn(khx) = +sgn(hx)$ nếu k là dương đối với h $(sgn(k, h) = +1)$ và $sgn(khx) = -sgn(hx)$ nếu k là âm đối với h $(sgn(k, h) = -1)$. (2. 4)

$sgn(khx) = 0$ *nếu* $khx = hx$. (2. 5)

Tổng quát: $\forall x \in H(G)$, có thể được viết là: $x = h_n h_{n-1} \ldots h_1 c$, $h_j \in H$, $c \in G$.

Khi đó: $sgn(x) = sgn(h_n, h_{n-1}) \ldots sgn(h_2, h_1) sgn(h_1) sgn(c)$ (2. 6)

$sgn(hx) = +1 \Rightarrow (hx \geq x)$ *và* $sgn(hx) = -1 \Rightarrow (hx \leq x)$

Độ đo tính mờ: Khái niệm "mờ" của thông tin ngôn ngữ mờ là rất quan trọng trong việc tính toán giá trị ngữ nghĩa của từ ngữ. Ngữ nghĩa của giá trị ngôn ngữ trong AX được xây dựng từ các tập $H(x) = \{x = h_n h_{n-1} \ldots h_1 c, h_j \in H, c \in G\} \cup \{x\}, x \in X$, có thể coi như một mô hình mờ của x. Tập $H(x), x \in X$, xác định độ đo tính mờ fm của X, chính bằng "bán kính" của $H(x)$ và có thể được tính toán một cách đệ quy từ độ đo tính mờ của các phần tử sinh, $fm(c^-)$, $fm(c^+)$ và độ đo tính mờ của gia tử $\mu(h)$, $h \in H$. Chúng được gọi là các tham số mờ của X.

$fm: X \to [0, 1]$ gọi là độ đo tính mờ nếu thỏa mãn các điều kiện sau:

$fm(c^-) + fm(c^+) = 1$ và

$\sum_{h \in H} fm(hx) = fm(x)$, *với* $\forall x \in X$ (2. 7)

Với các phần tử $\mathbf{0}$, $\boldsymbol{W}$ và $\mathbf{1}$,

$fm(\mathbf{0}) = fm(\boldsymbol{W}) = fm(\mathbf{1}) = 0$ (2. 8)

Và với $\forall x, y \in X, \forall h \in H, \frac{fm(hx)}{fm(x)} = \frac{fm(hy)}{fm(y)}$ (2. 9)

Đẳng thức (2. 9) không phụ thuộc vào các phần tử x, y, nó đặc trưng cho gia tử h, gọi là độ đo tính mờ của h, ký hiệu là $\mu(h)$. Tính chất của $fm(x)$ và $\mu(h)$ như sau:

Ta có $x \in X$, $x = h_n h_{n-1} \ldots h_1 c$,

$fm(hx) = \mu(h) fm(x), \forall x \in X$ (2. 10)

$fm(h_n h_{n-1} \ldots h_1 c) = \mu(h_n)\mu(h_{n-1}) \ldots \mu(h_1) fm(c), c \in G$ (2. 11)

$\sum_{i=-1}^{-q} \mu(h_i) = \alpha$ *và* $\sum_{i=1}^{p} \mu(h_i) = \beta$, *với* $\alpha, \beta > 0$ *và* $\alpha + \beta = 1$ (2. 12)

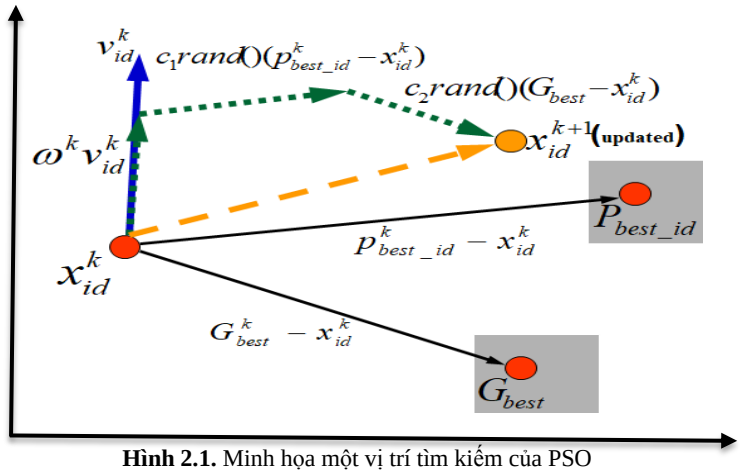
2.3. Thuật toán PSO [27]

PSO là thuật toán tìm kiếm ngẫu nhiên dựa trên việc mô phỏng hành vi tương tác của bầy chim hay đàn cá tìm nguồn thức ăn. Mỗi con chim (hay cá thể, phần tử) trong đàn (quần thể) được đặc trưng bởi hai tham số là véctơ vị trí x_{id} và véctơ vận tốc (dịch chuyển) v_{id}. Ban đầu PSO được khởi tạo bởi vị trí và vận tốc một cách ngẫu nhiên. Sau mỗi bước dịch chuyển (lặp) mỗi cá thể đánh giá khả năng tìm kiếm bằng hàm đo độ thích nghi (fitness function). Đồng thời mỗi cá thể cập nhật vận tốc v_{id} và vị trí x_{id} của mình theo công thức (2.13) và (2.14). Cũng tại mỗi bước lặp, mỗi cá thể phản ánh bởi hai thông tin: Thông tin thứ nhất là vị trí tốt nhất mà nó đạt được cho tới thời điểm hiện tại, gọi là $\mathrm{P_{best\_id}}$. Thông tin thứ hai là vị trí tốt nhất trong quá trình tìm kiếm của quần thể từ trước cho tới thời điểm hiện tại, gọi là $\mathrm{G_{best}}$. Mô hình hóa việc cập nhật vị trí của mỗi cá thể theo vị trí tốt nhất của nó và của tất cả các cá thể trong quần thể tính tới thời điểm hiện tại được minh họa trong Hình 2.1.

$$\mathrm{V_{id}^{k+1}} = \omega^k * \mathrm{V_{id}^k} + \mathrm{C_1} * \mathrm{Rand}() * (\mathrm{P_{best\_id}} - \mathrm{x_{id}^k}) + \mathrm{C_2} * \mathrm{Rand}() * (\mathrm{G_{best}} - \mathrm{x_{id}^k}) \quad (2.13)$$

$$\mathrm{x_{id}^{k+1}} = \mathrm{x_{id}^k} + \mathrm{v_{id}^{k+1}} \quad (2.14)$$

$$\omega^k = \omega_{max} - \frac{k*(\omega_{max} - \omega_{min})}{iter\_max} \quad (2.15)$$



Hình 2.1. Minh họa một vị trí tìm kiếm của PSO

Vị trí tốt nhất của các cá thể được đặc trưng bởi một véctơ $P_{best\_id} = [p_{id,1}, p_{id,2}, \dots, p_{id,n-1}]$ và giá trị P_{best_id} của mỗi cá thể id, $x_{id} = [x_{id,1}, x_{id,2}, \dots, x_{id,n-1}]$ được tính như sau:

$$P_{best\_id}^{k+1} = f(x) = \begin{cases} P_{best\_id}^k, & if\ fitness(x_{id}^{k+1}) > P_{best\_id}^k \\ fitness(x_{id}^{k+1}), & if\ fitness(x_{id}^{k+1}) \le P_{best\_id}^k \end{cases} \quad (2.16)$$

Giá trị G_{best} tại lần lặp thứ k là: $G_{best} = \min(P_{best_{id}}^k)$ (2.17)

III. MÔ HÌNH DỰ BÁO ĐỀ XUẤT KẾT HỢP GIỮA ĐSGT VÀ PSO

Trong mục này, chúng tôi giới thiệu mô hình dự báo chuỗi thời gian mờ dựa trên việc kết hợp giữa đại số gia tử và tối ưu bày đàn cho dự báo tuyển sinh đại học. Trước tiên ĐSGT được áp dụng để chia tập nền thành các khoảng có độ dài khác nhau bằng việc ánh xạ định lượng các hạng từ ngôn ngữ thành các khoảng mờ. Dựa trên các khoảng mới đạt được này, chúng tôi xác định các tập mờ và mờ hóa dữ liệu trên mỗi khoảng đã chia. Sau đó, thực hiện nhóm quan hệ mờ theo [14] và tính giá trị đầu ra dự báo bằng quy tắc dự báo đề xuất. Cuối cùng, để tăng độ chính xác dự báo hơn nữa, mô hình đề xuất được kết hợp với PSO trong việc hiệu chỉnh lại độ dài khoảng nhằm tìm ra các khoảng chia tối ưu từ tập nền. Để xác minh tính hiệu của mô hình đề xuất, toàn bộ dữ liệu liệu về số lượng sinh viên nhập học trong tài liệu [3] được sử dụng làm minh chứng cho quá trình dự báo dự trên chuỗi thời gian mờ bậc 1 và bậc cao. Mô hình dự báo đề xuất bao gồm các bước sau:

Bước 1: Xác định tập nền của chuỗi dữ liệu quan sát

Giả sử tập nền hay miền trị tham chiếu $U = [D_{min}, D_{max}] = [I_{min} - N_1, I_{max} + N_2]$, trong đó I_{min}, I_{max} là giá trị nhỏ nhất và lớn nhất của chuỗi dữ liệu quan sát và N_1, N_2 là hai số dương được chọn sao cho tập nền U bao trọn vẹn chuỗi dữ liệu lịch sử và đảm bảo nhiễu của dữ liệu kiểm thử. Không mất tính tổng quát, chúng tôi xác định tập nền U giống như trong công trình [3] là $U = [13000, 20000]$. Trong đó I_{min} =13055, I_{max} = 19337 và N_1= 55, N_2 = 663.

Bước 2: Chia tập nền U thành n khoảng khác nhau dựa vào ĐSGT

Như đã biết, chuỗi thời gian là một tập các dữ liệu quan sát được diễn biến theo thứ tự thời gian. Các dữ liệu quan sát này, theo tiếp cận mờ thì chúng được biểu diễn bởi các tập mờ (hạng từ ngôn ngữ) và gọi đó là chuỗi thời gian mờ. Nếu nhìn trên phương diện ĐSGT thì mỗi hạng từ ngôn ngữ đại diện cho một vài giá trị quan sát thuộc vào khoảng mờ nào đó, mà các hạng từ này luôn đảm bảo về thứ tự ngữ nghĩa. Điều đó có thể thấy rằng, khi sử dụng ĐSGT để ánh xạ định lượng ngữ nghĩa các hạng từ ngôn ngữ thành các giá trị trên miền mờ luôn thỏa mãn tính chất chia khoảng trên miền thực. Vì vậy, trong phần này chúng tôi áp dụng ĐSGT để chia tập nền U hay miền trị tham chiếu thành các khoảng tương ứng với các hạng từ ngôn ngữ dùng để định tính giá trị quan sát trong chuỗi thời gian mờ. Trong phần này, bài báo sử dụng ĐSGT có cấu trúc như sau: ĐSGT $\mathcal{AX} = (X, G, C, H, \le)$ với X là tập các hạng từ của biến ngôn ngữ "enrollment"; $G = \{c^-, c^+\} = \{Low, High\}$, Low $\le$ *High là tập các phần tử sinh;* Tập các hằng $C = \{0, 1, W\}$, hai gia tử là $H = \{Very, Little\}$. Để so sánh kết quả dự báo của mô hình đề xuất với các mô hình khác. Trong bài báo này chúng tôi sử dụng số khoảng chia bằng với số lượng hạng từ ngôn ngữ dùng để định tính các giá trị quan sát. Cụ thể, xuất phát từ số lượng hạng từ ngôn ngữ cho trước là 7 và 14 được đưa ra trong Bảng 3.1, chúng tôi xác định được các số khoảng tương ứng là 7 và 14 khoảng.

Bảng 3.1. Số lượng hạng từ ngôn ngữ

Số lượng hạng từ	Các hạng từ có thứ tự
7	A_1 = Very Very Low (VVL) < A_2= Little Verry Low (LVL) < A_3= Little Little Low (LLL) < A_4=Very Little Low (VLL) < A_5 =Verry Little High (VLH) < A_6= Little Little High (LLH) < A_7= Very High (VH).
14	A_1 = VVS < A_2 = LLVS < A_3 = VLVS < A_4 = VLLL< A_5 = LLLS < A_6 = LVLS < A_7 = VVLL < A_8 = VVLH < A_9 = LVLH < A_{10} = LLLH < A_{11} = VLLH < A_{12} = VLVH <A_{13} = LLVH < A_{14} = VVH

Bước này, sử dụng 7 khoảng chia để minh họa từng bước cho việc xác định khoảng mờ dựa trên ĐSGT như sau:

Bước 2.1: Miền trị tham chiếu U = [13000, 20000] được ánh xạ sang miền [0, 1]

Giả sử trong tập dữ liệu lịch sử chọn giá trị 16807 là giá trị trung bình khi đó khoảng tính mờ của các phần tử sinh được thiết lập là $fm(low) = \frac{16807-13000}{20000-13000} = 0{,}544$, $fm(high) = 1 - 0{,}544 = 0{,}456$.

Từ đây, có thể tính được khoảng mờ của các từ ngôn ngữ trên miền [0,1] là: $fm(VVL)$ = 0,1471, $fm(LVL)$ = 0,1358, $fm(LLL)$ = 0,1253, $fm(VLL)$ = 0,1358, $fm(VLH)$ = 0,11138, $fm(LLH)$ = 0,1051, $fm(VH)$ = 0,2371.

Bước 2.2: Ánh xạ ngược lại miền U

Giả sử gọi $cofm(G)$ là độ rộng của hai phần tử sinh là $cofm(Low) = fm(Low) \times LU = 0{,}544 \times 7000 = 3808$ và $cofm(High) = fm(High) \times LU = 0{,}456 \times 7000 = 3192$, trong đó độ dài của miền U ký hiệu là LU = 20000-13000 =7000.

Bước 2.3: Xác định khoảng mờ của nhãn ngôn ngữ trên tập nền U

Trong bài báo này, chúng tôi chọn độ đo tính mờ của các gia tử âm và gia tử dương tương ứng là $\mu(Little) = 0.48$ và $\mu(Very) = 1 - \mu(little) = 0{,}52$.

Kết hợp Bước 2.2, ta có thể tính được giá trị cho các hạng từ thuộc vào các khoảng mờ như sau:

$cofm(A_1)$ = μ(*Verry*) × μ(*Very*) × *cofm*(Low) = 0,52×0,52×3808 = 1029,683;

$cofm(A_2)$ = μ(*Little*) × μ(*Very*) × *cofm*(*Low*) = 0,48×0,52×3808 = 950,477;

Một cách tương tự cho các hạng khác chúng tôi xác định được 7 khoảng chia trên miền thực U như sau:
u_1= [13000, 14029,68), u_2= [14029.68, 14980), u_3= [14980, 15858), u_4= [15858, 16808), u_5= [16808, 17605), u_6= [17605, 18340), u_7= [18340, 20000].

Thực hiện tương tự các bước trên đối với số hạng từ là 14, chúng tôi đưa ra 14 khoảng chia tương ứng với các hạng từ trong tập nền U như sau:
u_1= [13000, 13539,5), u_2= [13539,5, 14079) , u_3= [14079, 14438,5), u_4= [14438,5, 14798), u_5= [14798, 15157,5), u_6= [15157,5, 15517), u_7= [15517, 15756,5), u_8= [15756,5, 15996), u_9= [15996, 16316,5) , u_{10}= [16316,5, 16637), u_{11}= [16637, 17117,5), u_{12}= [17117,5, 17598), u_{13}= [17598, 18799), u_{14}= [18799, 20000].

Bước 3: Khởi tạo ngẫu nhiên m các cá thể trong quần thể

Theo thuật toán 2, mỗi cá thể trong PSO được đặc trưng bởi hai thành phần là vị trí và vận tốc; Giả sử id là một cá thể trong quần thể. Khi đó:

- Vị trí X_{id} và vận tốc V_{id} là các véctơ gồm n-1 (n=7) phần tử được biểu diễn như hình 3.1.

$x_{id,1}$	$x_{id,2}$	...	$x_{id,k}$	...	$x_{id,6-1}$

$v_{id,1}$	$v_{id,2}$	...	$v_{id,k}$	...	$v_{id,6}$

Hình 3.1. Cấu trúc vị trí và vận tốc của cá thể id, (1≤ k ≤ n-1)

Trong đó, các phần tử X_{id} được khởi tạo một cách ngẫu nhiên trong tập nền và được sắp xếp theo thứ tự tăng dần như sau: $D_{min} < x_{id,1} < \cdots < x_{id,6-1} < D_{max}$ và các phần tử V_{id} được khởi tạo ngẫu nhiên trong miền [$-V_{max}, V_{max}$].

- Vị trí tốt nhất của cá thể id ghi nhận được là một véctơ $P_{best\_id} = [p_{id,1}, p_{id,2}, \ldots, p_{id,6}]$ và ban đầu được khởi tạo giống như khởi tạo vị trí của cá thể id.

Bước 4: Hiệu chỉnh lại độ dài các khoảng chia tại bước 2 và tính giá trị hàm mục tiêu cho mỗi cá thể trong PSO.

3.1. Xác định các tập mờ và mờ hóa dữ liệu quan sát

Dựa trên véctơ vị trí của mỗi cá thể id gồm n-1 phần tử này, chúng tôi xác định n khoảng chia từ tập nền. Để thuận tiện và không mất tính tổng quát, chúng tôi chọn số lượng khoảng giống như số khoảng đã chia ở bước 2. Giả sử

số khoảng chia là n =7, khi đó các khoảng đạt được là: $u_1 = (D_{min}, x_{id,1}]$, $u_2 = (x_{id,1}, x_{id,2}]$, …, $u_7 = (x_{id,6}, D_{max}]$. Từ các khoảng mới đạt được, chúng tôi xác định các tập mờ dựa theo [1] đưa ra trong công thức (3.1) như sau:

$$A_i = {}^{a_{i1}}/_{u_1} + {}^{a_{i2}}/_{u_2} + \ldots + {}^{a_{ij}}/_{u_j} + \cdots + {}^{a_{i7}}/_{u_7} \quad (3.1)$$

$$a_{ij} = \begin{cases} 1 & j = i \\ 0{,}5 & j = i-1, j = i+1 \\ 0 & \text{otherwise} \end{cases} \quad (3.2)$$

Trong đó, $a_{ij} \in [0,1]$, $(1 \le i \le 7, 1 \le j \le 7)$ chỉ cấp độ của khoảng u_j vào tập mờ A_i , u_j là khoảng thứ j của tập nền U. Để đơn giản, mỗi giá trị độ thuộc của tập mờ A_i được lựa chọn theo công thức (3.2), có dạng hàm thuộc tam giác với cấp độ thuộc tương ứng là 1, 0,5 và 0.

3.2. Mờ hóa dữ liệu lịch sử thành các tập mờ

Mờ hóa các dữ liệu rõ thành dữ liệu mờ được biểu diễn bằng tập mờ, trước tiên cần gán giá trị ngôn ngữ liên quan đến mỗi tập mờ đã xác định trong bước 4.2 vào mỗi khoảng tương ứng. Cách đơn giản là tìm ra một khoảng u_i mà giá trị lịch sử của biến chuỗi thời gian tại thời điểm nào đó thuộc vào khoảng này mà có cấp độ thuộc cao nhất của khoảng u_i tại tập mờ A_i, khi đó dữ liệu lịch của biến chuỗi thời gian được mờ hóa là A_i.

Ví dụ: Giá trị lịch sử của năm 1972, Y(1972) là 13563 thuộc vào khoảng u_1= (13000, 14029.68] mà cấp độ thuộc lớn nhất của khoảng này xảy ra tại A_1 là 1, vậy giá trị mờ hóa tại thời điểm t =1972, F(t)=F(1972) là A_1 có nhãn ngôn ngữ là "*not many*". Bằng cách tương tự cho các giá trị quan sát khác trong chuỗi thời gian.

3.3. Xác định các quan hệ mờ

Dựa trên định nghĩa 2 và 3 về khái niệm quan hệ logic mờ bậc 1 và bậc cao, một quan hệ mờ được xác định bởi một hay nhiều tập mờ liên tiếp trong chuỗi thời gian. Để xác định các quan hệ logic mờ với bậc khác nhau, chúng ta tìm ra các quan hệ có dạng: $F(t-\lambda), F(t-\lambda+1), , \ldots, F(t-1) \to F(t)$; trong đó, $F(t-\lambda), F(t-\lambda+1), , \ldots, F(t-1)$ và $F(t)$ được gọi là trạng thái hiện tại và trạng thái tương lai của quan hệ. Sau đó quan hệ này được thay thế bởi quan hệ giữa các tập mờ là: $A_{i\lambda}, A_{i(\lambda-1)}, \ldots, A_{i2}, A_{i1} \to A_k$. Hai ví dụ minh họa cho quan hệ mờ bậc 1 và bậc 2 được trình bày như sau:

- Trong trường hợp quan hệ mờ bậc 1(λ=1), hai tập mờ liên tiếp được sử dụng để xác định quan hệ mờ bậc 1. Giả sử ở bước 4.2, dữ liệu tại năm F(1973) được mờ hóa là A_2và dữ liệu tại năm F(1974) được mờ hóa là A_1. Khi đó quan hệ giữa thời điểm F(1993) với F(1994) là $F(1973) \to F(1974)$ và được thay bởi quan hệ mờ là $A_2 \to A_1$. Hoàn toàn tương tự có thể thiết lập được các quan hệ mờ bậc 1 khác.
- Trong trường hợp quan hệ mờ bậc cao (giả λ=2), ba tập mờ liên tiếp theo thứ tự thời gian được sử để tạo thành quan hệ mờ bậc 2. Giả sử ba năm liên tiếp F(1973), F(1974), F(1975) được mờ hóa tương ứng với các tập mờ là A_2, A_1, A_2. Khi đó quan hệ mờ bậc hai tại thời điểm t =1975 được biểu diễn là: $A_2, A_1 \to A_2$. Một cách tương tự để xác định các quan hệ mờ bậc hai khác tại thời điểm khác nhau.

3.4. Thiết lập nhóm quan hệ mờ phụ thuộc thời gian

Trong bước này, chúng tôi tạo nhóm quan hệ mờ dựa trên định nghĩa 4 về nhóm quan hệ mờ phụ thuộc thời gian bậc 1 và bậc cao. Giả sử tồn tại các quan hệ mờ tại các thời điểm khác nhau như sau:

Tại $t = 1973$ ta có mối quan hệ mờ $A_1 \to A_2$

Tại $t = 1974$ ta có mối quan hệ mờ $A_2 \to A_1$

Tại $t = 1975$ ta có mối quan hệ mờ $A_1 \to A_2$

Khi đó, tại các thời điểm t lần lượt là 1973, 1974, 1975 chúng ta nhận được ba nhóm quan hệ theo thứ tự thời gian trên là G_1: $A_1 \to A_2$; G_2: $A_2 \to A_1$ và G_3: $A_1 \to A_2, A_2$. Một cách tương tự cho nhóm quan hệ mờ bậc cao.

3.5. Giải mờ và tính giá trị dự báo đầu ra

Để giải mờ dữ liệu đã mờ hóa và tính toán giá trị cho nhóm quan hệ mờ bậc 1 và bậc cao. Thứ nhất, chúng tôi đề xuất các kỹ thuật giải mờ mới để tính toán giá trị dự báo cho các nhóm quan hệ mờ với các bậc khác nhau trong giai đoạn huấn luyện. Thứ hai, sử dụng quy tắc giải mờ được đề xuất trong [11] để tính toán giá trị dự báo cho các nhóm quan hệ mờ trong giai đoạn thử nghiệm. Các giá trị dự báo cho các nhóm quan hệ mờ dựa vào chuỗi thời gian mờ bậc 1 và bậc cao được tính theo các quy tắc sau:

Quy tắc 1: Trong trường hợp nhóm quan hệ mờ bậc 1(bậc λ =1)

Để tính toán giá trị dự báo cho tất cả các nhóm quan hệ mờ bậc 1, chúng tôi xem xét thứ tự xuất hiện của các tập mờ bên vế phải trong cùng nhóm quan hệ kể cả các tập mờ lặp lại, sau đó gán các trọng số có tầm quan trọng khác nhau cho các tập mờ này theo thứ tự xuất hiện. Tức là các quan hệ xuất hiện gần đây hơn thì được gán với trọng số cao hơn. Điều này đã

thể hiện rõ sự khác biệt so với các quan hệ mờ được xây dựng trong các công trình trước đây trong [3], [11]. Giả sử có nhóm quan hệ mờ bậc 1 xuất hiện cùng vế trái là A_j như sau: $A_j(t-1) \rightarrow A_{i1}(t1), A_{i2}(t2), \dots, A_{ik}(tk)\dots$; Khi đó, giải mờ dự báo cho năm t có nhóm này được tính theo công thức (3.3) sau đây:

$$forecasted = \frac{1*m_{i1}+2*m_{i2}+\cdots+k*m_{ik}+\cdots+p*m_{ip}}{1+2+\cdots+k+\cdots+p} \quad (3.3)$$

Trong đó:

- ✓ m_{i1}, m_{i2} và m_{ik} là điểm giữ của các khoảng u_{i1}, u_{i2} và u_{ik} tương ứng, mà cấp độ thuộc cao nhất của các tập mờ A_{i1}, A_{i2} , . .. ,A_{ik} xảy ra tại các khoảng này.
- ✓ $k(1 \le k \le p)$ là các trọng số được xác định theo thứ tự thời gian

***Quy tắc 2*:** Trường hợp nhóm quan hệ mờ bậc cao ($\lambda \ge 2$)

Để thiết lập giá trị dự báo cho các nhóm quan hệ mờ phụ thuộc thời gian bậc cao, chúng tôi xem xét thêm thông tin của các tập mờ xuất hiện bên vế phải của các quan hệ mờ trong cùng nhóm. Cụ thể của quy tắc được tính như sau:

Đối với mỗi nhóm quan hệ mờ bậc cao, chúng tôi chia mỗi khoảng tương ứng với các tập mờ bên vế phải trong cùng nhóm thành 4 khoảng con có độ dài bằng nhau và giải mờ dự báo cho mỗi nhóm này trong giai đoạn huấn luyện được tính theo công thức (3.4):

$$\text{forecasted}_{\text{output}} = \frac{1}{n}\sum_{j=1}^{n} subm_{kj} \quad (3.4)$$

Trong đó, ($1 \le j \le n$, $1 \le k \le 4$)

- ✓ n là tổng số tập mờ bên vế phải của nhóm;
- ✓ $subm_{kj}$ là điểm giữa của một trong 4 khoảng con (điểm giữa của khoảng con thứ k) tương ứng với tập mờ thứ j bên vế phải của nhóm quan hệ.

Quy tắc 3: Trường hợp nhóm quan hệ rỗng (Nhóm quan hệ mờ có vế phải chưa xác định tập mờ).

Để tính toán giá trị dự báo cho nhóm quan hệ trong giai đoạn thử nghiệm, chúng tôi sử dụng lược đồ đề xuất trong [11]. Ý tưởng của lược đồ như sau:

Đối với nhóm quan hệ chưa có mẫu luyện, tức là nhóm quan hệ chưa có tập mờ hóa bên vế phải của quan hệ. Giả xử xuất hiện nhóm quan hệ mờ bậc λ như sau: $A_{i\lambda}, A_{i(\lambda-1)}, \dots, A_{i1} \rightarrow \#$. Khi đó, lược đồ gắn trọng số cao nhất w_h đối với tập mờ xuất hiện gần nhất về tương lai A_{i1} và trọng số bằng 1 cho các tập xuất hiện trước đó nằm bên vế trái của nhóm quan hệ mờ và giải mờ dự báo được tính theo công thức (3.5) sau:

$$Forecated_{for\#} = \frac{(M_{i1}*w_h)+M_{i2}+\cdots+M_{i\lambda}}{w_h+(\lambda-1)} \quad (3.5)$$

Trong đó, w_h là phiếu bầu cử cao nhất (trọng số lớn nhất) được cho trước bởi người dùng. Trong báo cáo này, để so sánh với các mô hình dự báo trước đây, chúng tôi chọn w_h=15 giống như công trình được công bố trong [11]. $M_{i1}, M_{i2}, \dots, M_{i\lambda}$ là giá trị điểm giữa của các khoảng $u_{i1}, u_{i2}, \dots, u_{i\lambda}$, với ($1 \le i \le \lambda$).

3.6. Tính giá trị hàm mục tiêu cho mỗi cá thể trong PSO

Mỗi cá thể đạt một giải pháp tối ưu thông qua giá trị hàm mục tiêu MSE (mean square error) or RMSE (root mean square error) như sau:

$$MSE = \frac{1}{n}\sum_{id=\lambda}^{n}(F_{id} - R_{id})^2 \quad (3.6)$$

$$RMSE = \sqrt{\frac{1}{n}\sum_{id=\lambda}^{n}(F_{id} - R_{id})^2} \quad (3.7)$$

Trong đó; F_{id} giá trị dự báo tại thời điểm id, R_{id} là giá trị thực tại thời điểm id, n là tổng số dữ liệu tham gia dự báo, λ là bậc của quan hệ.

Bước 5: Cập nhật vị trí tốt nhất P_{best_id} của mỗi cá thể và G_{best} của quần thể

Trong PSO mỗi cá thể trong quần thể được đặc trưng bởi hai thành phần tốt nhất đó là: Véctơ vị trí tốt nhất của từng cá thể id P_{best_id} = [$P_{id,1}$, $P_{id,2}$, $P_{id,3}$, …, $P_{id,n-1}$] và vị trí tốt nhất trong cả quần thể. Mỗi cá thể id được cập nhật theo giá trị của hàm mục tiêu MSE (3.6). Nếu giá trị của MSE ở thời điểm hiện tại k nhỏ hơn giá trị MSE ở thời điểm trước đó k-1 thì P_{best_id}=MSE (x_{id}^{k}) và G_{best} = min (P_{best_id})

Bước 6: Cập nhật lại các thành phần của mỗi cá thể trong PSO

- Cập nhật trọng số ω theo công thức (2.15);
- Cập nhật vận tốc của mỗi cá thể theo công thức (2.13);
- Cập nhật vị trí của mỗi cá thể theo công thức (2.14).

Bước 7: Kiểm tra điều kiện dừng

Nếu số lượng lần lặp hiện tại còn nhỏ hơn số lần lặp tối đa (k < iter_max) hay (*chưa tìm thấy giải pháp tối ưu*), quay lại bước Bước 4. Trái lại đưa ra kết quả dự báo và độ chính xác dự báo của mô hình bằng giá trị MSE.

IV. KẾT QUẢ THỰC NGHIỆM

Mục này thảo luận về một số kết quả dự báo đạt được từ tập dữ liệu huấn luyện và dữ liệu kiểm thử. Chúng tôi thực hiện mô hình dự báo bằng việc sử dụng ngôn ngữ lập trình C# trên máy tính Intel Core i7 PC, 8 GB RAM để dự báo tuyển sinh đại học với bộ dữ liệu về số lượng sinh viên nhập học [3] từ giai đoạn 1971 đến 1992. Các tham số để thực hiện mô hình dự báo được đưa ra trong Bảng 4.1.

Bảng 4.1. Các tham số sử dụng trong mô hình dự báo trên hai tập dữ liêu

Các tham số	Giá trị cho dữ liệu tuyển sinh Đại học
Số lượng cá thể trong quần thể: $\boldsymbol{N}$ =	50
Số lần lặp tối đa (số thế hệ): ***iter_max*** =	150
Trọng số quán tính $\boldsymbol{\omega}$ *(Giảm tuyến tính)*	0,9 to 0,4
Các hệ số tự tin cậy và hệ số xã hội $\boldsymbol{C_1 = C_2}$	2
Miền giới hạn vận tốc của mỗi cá thể: $\boldsymbol{V}$ =	[-100,100]
Miền giới hạn vị trí của mỗi cá thể: $\boldsymbol{X}$ =	[13000, 20000]

Để đánh giá hiệu quả của mô hình đề xuất dựa trên chuỗi thời gian mờ bậc 1 với số lượng khoảng chia bằng 7, các mô hình trong các công trình sau được lựa chọn cho việc so sánh: mô hình [28], mô hình [29], mô hình của Wei Lu [30], mô hình [26] và mô hình [25]. Từ các tham số thiết lập cho dữ liệu tuyển sinh trong Bảng 4.1, mô hình đề xuất thực hiện 20 lần chạy, kết quả của lần chạy có giá trị MSE (3.6) hoặc RMSE (3.7) nhỏ nhất được chọn là giá trị dự báo cuối cùng. Hiệu quả của mô hình dự báo đề xuất được so sánh với các mô hình trước đây chỉ ra trong Bảng 4.2. Trong đó, cột thứ 1, cột thứ 2 và cột thứ 3 thể hiện dữ liệu năm dự báo, dữ liệu tuyển sinh, các tập mờ biểu diễn dữ liệu tuyển sinh. Các cột còn lại là kết quả dự báo tương ứng với các mô hình được chọn để so sánh trong giai đoạn huấn luyện.

Bảng 4.2. So sánh mô hình đề xuất với các mô hình khác dựa trên chuôi thời gian bậc 1 với 7 khoảng chia

Năm	Dữ liệu thực	Tập mờ	Mô hình [28]	Mô hình [29]	Mô hình [30]	Mô hình [26]	Mô hình [25]	Mô hình đề xuất
1971		A1	-	-	-	-	-	-
1972	13563	A1	13486	13944	14279	13820	13865	13848
1973	13867	A1	14156	13944	14279	13820	14082	13848
1974	14696	A2	15215	13944	14279	13820	14514	14426
1975	15460	A3	15906	15328	15392	15402	15391	15420
1976	15311	A3	15906	15753	15392	15536	15219	15420
1977	15603	A4	15906	15753	15392	15536	15219	15644
1978	15861	A4	15906	15753	16467	16461	16219	15757
1979	16807	A6	16559	16279	16467	16461	16625	16765
1980	16919	A6	16559	17270	17161	17444	16951	17270
1981	16388	A5	16559	17270	17161	17444	16439	16548
1982	15433	A3	16559	16279	14916	15402	15219	15420
1983	15497	A3	15906	15753	15392	15536	15219	15532
1984	15145	A2	15906	15753	15392	15536	15219	15321
1985	15163	A2	15906	15753	15392	15536	16219	15142
1986	15984	A5	15906	15753	15470	15536	15812	15664
1987	16859	A6	16559	16279	16467	16461	17439	16653
1988	18150	A7	16559	17270	17161	17444	19165	17811
1989	18970	A7	19451	19466	19257	19135	19165	19075
1990	19328	A7	18808	18933	19257	19135	19165	19075
1991	19337	A7	18808	18933	19257	19135	19165	19075
1992	18876	A7	18808	18933	19257	19135	15219	19075
1993	N/A	N/A						19170
RMSE			**578.3**	**506**	**445.2**	**441.3**	**210.9**	**196.1**
MSE			**334430.9**	**256036**	**198203**	**194745.7**	**44507**	**38422.7**

Thêm nữa, mô hình đề xuất được so sánh với các mô hình trước đây dựa trên chuỗi thời gian mờ bậc 1 với số lượng khoảng chia là 14 khoảng. Các mô hình sau được lựa chọn cho việc so sánh là: C96 [3], H01[5], CC06a [8], HPSO [11], AFPSO [12], VGPSO [14], Wei Lu [30]. Từ các kết quả thực nghiệm cho thấy mô hình dự báo đề xuất hiệu quả hơn so với các mô hình trước đây dựa trên chuỗi thời gian mờ bậc 1. Cụ thể với số khoảng chia bằng 7 mô

hình đề xuất đưa ra sai số dự báo (***MSE =38422.7)*** nhỏ nhất trong số mô hình đưa ra so sánh trong Bảng 4.2, trong Bảng 4.3 với số khoảng chia bằng 14 đưa ra sai số dự báo (***MSE = 5249.9)*** cũng tốt hơn các mô hình hiện có trong bảng.

Bảng 4.3. So sánh mô hình đề xuất với các mô hình khác dựa trên chuỗi thời gian bậc 1 với 14 khoảng chia

Năm	Dữ liệu thực	C96	H01	CC06a	HPSO	Wei Lu	AFPSO	VGPSO	MH đề xuất
1971	13055	---	---	---	---	---	---	---	---
1972	13563	14000	14000	13714	13555	13512	13579	13434	13433
1973	13867	14000	14000	13714	13994	13998	13812	13841	13851
----	----	----	----	----	----	-----	----	----	----
1990	19328	19000	19000	19300	19340	19241	19418	19340	19486
1991	19337	19000	19500	19149	19340	19666	19260	19340	19486
1992	18876	19000	19149	19014	19014	18718	19031	18820	18869
MSE		**407507**	**226611**	**35324**	**22965**	**14534**	**8224**	**7475**	**5249.9**

Hơn thế, trong bài báo này chúng tôi thực hiện mô hình dự báo dựa trên quan hệ mờ bậc cao từ bậc 2 đến bậc 9 với số khoảng chia được cố định là **7** khoảng. Kết quả dự báo dựa trên mô hình chuỗi thời gian mờ bậc cao được thể hiện trong Bảng 4.4 sau đây. Quan sát Bảng 4.4 cho thấy mô hình dự báo đề xuất càng hiệu quả khi bậc của quan hệ mờ tăng lên theo sự tăng số lượng quan sát của chuỗi thời gian.

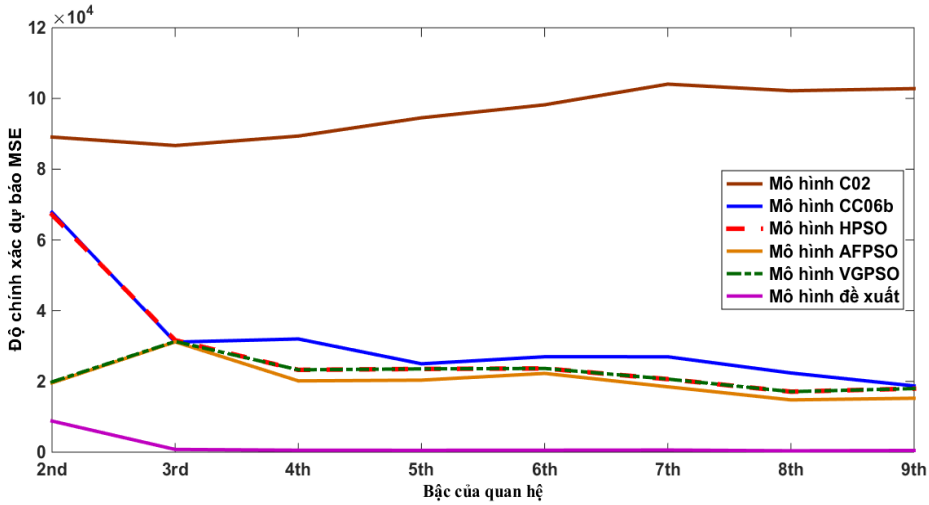
Bảng 4.4. Kết quả dự báo của của mô hình đề xuất dựa trên chuỗi thời gian mờ bậc cao với số khoảng chia bằng 7

Năm	Giá trị thực	Bậc 2	Bậc 3	Bậc 4	Bậc 5	Bậc 6	Bậc 7	Bậc 8	Bậc 9
1973	13867	13874							
1974	14696	14678	14694						
1975	15460	15488	15457	15468					
1976	15311	15310	15333	15311	15310				
1977	15603	15595	15580	15591	15606	15612			
1978	15861	15906	15852	15852	15856	15851	15873		
1979	16807	16724	16814	16829	16862	16798	16863	16830	
1980	16919	17066	16951	16926	16862	16927	16863	16919	16886
1981	16388	16390	16381	16387	16394	16396	16386	16394	16388
----	----	----	----	----	----	----	----	----	----
1991	19337	19298	19356	19354	19331	19358	19329	19356	19326
1992	18876	18958	18927	18924	18878	18930	18879	18873	18847
MSE		**8802.55**	**774.05**	**550.17**	**526.29**	**554.94**	**603.37**	**396.29**	**491**

Để xác minh tính hiệu quả của mô hình dự báo đề xuất dựa trên chuỗi thời gian mờ bậc cao, bốn mô hình liệt kê trong Bảng 4.5 được lựa chọn cho việc so sánh: Trong Bảng 4.5 mô hình CC06b [9] sử dụng giải thuật di truyền để tối ưu khoảng, các mô hình HPSO [11], AFPSO [12], VGPSO [14] và mô hình đề xuất đều sử dụng PSO để chia khoảng. Nhưng điểm khác biệt chính giữa các mô hình cùng sử dụng PSO là kỹ thuật giải mờ đầu ra và cách nhóm quan hệ mờ. Ngoài việc sử dụng PSO để tìm khoảng chia phù hợp, mô hình đề xuất được kết hợp thêm ĐSGT để chia các khoảng ban đầu có độ dài khác nhau thay vì các khoảng có độ dài bằng nhau. Từ kết quả so sánh về độ chính xác dự báo MSE (3.6) liệt kê trong Bảng 4.5 cho thấy mô hình đề xuất đưa ra độ chính xác dự báo với giá trị MSE nhỏ hơn so với các mô hình được chọn để so sánh dựa trên quan hệ mờ bậc cao (từ bậc 2 đến bậc 9) với cùng số khoảng chia bằng 7. Đặc biệt mô hình đề xuất đưa sai số dự báo tốt nhất thông qua giá trị (***MSE = 396.29***) trong trường hợp quan hệ mờ bậc 8. Điều đó, chứng tỏ rằng mô hình dự báo đề xuất hiệu quả hơn so với mô hình dự báo trước đây khi thử nghiệm trên tập dữ liệu tuyển sinh Đại học Alabama. Để trực quan hơn, thiên hướng dự báo của mô hình đề xuất với các mô hình trước đây cũng được minh họa trên Hình 4.1.

Bảng 4.5. So sánh độ chính xác dự báo MSE giữa mô hình đề xuất và các mô hình C02, CC06b, HPSO, AFPSO dựa trên các bậc khác nhau và số khoảng chia bằng 7

Mô hình	Số bậc của quan hệ								
	2	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**Average**
C02 [4]	89093	86694	89376	94539	98215	104056	102179	102789	95867.63
CC06b [9]	67834	31123	32009	24984	26980	26969	22387	18734	31377.5
HPSO [11]	67123	31644	23271	23534	23671	20651	17106	17971	28121.38
AFPSO [12]	19594	31189	20155	20366	22276	18482	14778	15251	20261.38
VGPSO [14]	19868	31307	23288	23552	23684	20669	17116	17987	22183
MH đề xuất	**8802.55**	**774.05**	**550.17**	**526.29**	**554.94**	**603.37**	**396.36**	**491**	**1587.34**



Hình 4.1. Xu thế dự báo của mô hình đề xuất so với các mô hình trước đây dựa trên các quan hệ mờ bậc cao với 7 khoảng chia

V. KẾT LUẬN

Nghiên cứu này, chúng tôi đưa ra mô hình dự báo chuỗi thời gian mờ kết hợp giữa đại số gia tử và kỹ thuật tối ưu bầy đàn. Mô hình đề xuất đã đề cập đến ba vấn đề được xem là quan trọng và ảnh hưởng lớn đến độ chính xác dự báo, là vấn đề xác định khoảng chia từ tập nền, cách thiết lập nhóm quan hệ mờ và các quy tắc giải mờ dự báo đầu ra. 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 và được chứng minh là hiệu quả và phù hợp với điều kiện thực tế hơn. Thêm nữa, kỹ thuật tối ưu PSO được áp dụng trong việc tìm khoảng chia tối ưu từ tập nền nhằm nâng cao độ chính xác dự báo của mô hình. Trong số các kỹ thuật khai phá và tự tìm giải pháp tối ưu, PSO được xem là thực hiện tốt hơn với các kỹ thuật heuristic khác về tỷ lệ thành công cũng như chất lượng giải pháp. Bằng việc kết hợp giữa ĐSGT và kỹ thuật tối ưu PSO, hiệu quả dự báo của mô hình đề xuất được cải thiện một cách đáng kể. Từ việc thử nghiệm trên tập dữ liệu về tuyển sinh đại học của Trường Đại học Alabama, kết quả dự báo cho thấy mô hình đề xuất vượt trội hơn so với các mô hình trước đây dựa trên chuỗi thời gian mờ bậc một và bậc cao. Chi tiết cho sự so sánh được thể hiện trên các Bảng 4.2 - 4.5. 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.

VI. REFERENCES

[1] Song, Q., Chissom, B. S., 1993b.. Fuzzy time series and its models. Fuzzy Sets and Systems, vol.54, no.3, 269-277.

[2] Q. Song, B. S. Chissom. “Forecasting Enrollments with Fuzzy Time Series - Part I”. Fuzzy set and systems, vol. 54, pp.1-9. 1993b.

[3] S. M. Chen. “Forecasting Enrollments based on Fuzzy Time Series”. Fuzzy set and systems, vol. 81, pp. 311-319. 1996.

[4] S. M. Chen. “Forecasting Enrollments based on hight-order Fuzzy Time Series”. Int. Journal: Cybernetic and Systems, N.33, pp. 1-16, 2002.

[5] Huarng K. “Effective lengths of intervals to improve forecasting in fuzzy time series”. Fuzzy Sets and Systems, 123, (2001b), 387-394.

[6] Lee, L. W. et al.. Handling forecasting problems based on two-factors high-order fuzzy time series. IEEE Transactions on Fuzzy Systems, 14, 468-477, 2006.

[7] S. M. Chen, K Tanuwijaya. “Fuzzy forecasting based on high- order fuzzy logical relationships and automatic clustering techniques”. Expert Systems with Applications. 38, 15425-15437, 2011.

[8] Chen S. M., & Chung N. Y. “Forecasting enrollments of students by using fuzzy time series and genetic algorithms”. International Journal of Information and Management Sciences, 17, 1-17, 2006a.

[9] Chen S. M., Chung N. Y. Forecasting enrollments using high-order fuzzy time series and genetic algorithms. International of Intelligent Systems 21, 485-501, 2006b.

[10] Lee L. W. Wang L. H., & Chen, S. M.. “Temperature prediction and TAIFEX forecasting based on hight order fuzzy logical ralationship and genetic simulated annealing techniques”. Expert Systems with Applications, 34, 328-336, 2008.

[11] I. H. Kuo, et al.. “An improved method for forecasting enrollments based on fuzzy time series and particle swarm optimization”. Expert systems with applications, 36, 6108-6117, 2009.

[12] Huang Y. L. et al.. A hybrid forecasting model for enrollments based on aggregated fuzzy time series and particle swarm optimization. Expert Systems with Applications, 38, 8014-8023, 2011

[13] Ling-Yuan Hsu et al.. Temperature prediction and TAIFEX forecasting based on fuzzy relationships and MTPSO techniques, Expert Syst. Appl.37, 2756-2770, 2010.

[14] Nguyen Cong Dieu, Nghiem Van Tinh. Fuzzy time series forecasting based on time-depending fuzzy relationship groups and particle swarm optimization, In: Proceedings of the 9th National conference on Fundamental and Applied Information Technology Research (FAIR'9), pp.125-133, 2016

[15] Park J. I., Lee D. J., Song C. K., Chun M. G.. TAIFEX and KOSPI 200 forecasting based on two-factors high-order fuzzy time series and particle swarm optimization, Expert Systems with Applications 37, 959-967, 2010.

[16] Chen, S. M, Bui Dang H. P.. Fuzzy time series forecasting based on optimal partitions of intervals and optimal weighting véctơs. Knowledge-Based Systems 118, 204-216, 2017.

[17] Chen S. M., Jian W. S.. Fuzzy forecasting based on two-factors second-order fuzzy-trend logical relationship groups, similarity measures and PSO techniques. Information Sciences 391-392, 65-79, 2017.

[18] M. Bose, K. Mali. A novel data partitioning and rule selection technique for modelling high-order fuzzy time series. Applied Soft Computing, https://doi.org/10.1016/j.asoc.2017.11.011, 2017.

[19] Tian Z. H., Wang P., He T. Y.. Fuzzy time series based on K-means and particle swarm optimization algorithm. Man-Machine-Environement System Engineering. Lecture Note in Electrical Enginearing 406, 181-189, Springer 2016.

[20] Zhiqiang Zhang, Qiong Zhu. "Fuzzy time series forecasting based on k-means clustering". Open Journal of Applied Sciences, 2,100-103, 2012.

[21] Nghiem Van Tinh & Nguyen Cong Dieu. Improving the forecasted accuracy of model based on fuzzy time series and k-means clustering. Journal of science and technology: issue on information and communications technology, No.2, 51-60, 2017

[22] Bulut E., Duru O., & Yoshida, S. A.. Fuzzy time series forecasting model formulti-variate forecasting analysis with fuzzy c-means clustering. WorldAcademy of Science, Engineering and Technology, 63, 765-771, 2012.

[23] S. Askari, N. Montazerin. A high-order multi-variable. Fuzzy Time Series forecasting algorithm based on fuzzy clustering, Expert Systems with Applications ,42, 2121-2135, 2015.

[24] Ho N. C.,Wechler W.."Hedge algebra: An algebraic approach to structures of sets of linguistic truth values", fuzzy Sets and Systems, 35, pp. 281-293, 1990.

[25] Nguyễn Cát Hồ, Nguyễn Công Điều, Vũ Như Lân. "Ứng dụng đại số gia tử trong dự báo chuỗi thời gian mờ". Tạp chí Khoa học và Công nghệ, Vol 54, No.2, 2016.

[26] Hoang Tung, Nguyen Dinh Thuan, Vu Minh Loc. The partitioning method based on hedge algebras for fuzzy time series forecasting, Journal of Science and Technology, 54 (5), 571-583, 2016.

[27] Kennedy J., & Eberhart R.. Particle swarm optimization. Proceedings of IEEE international Conference on Neural Network, 1942-1948, 1995.

[28] Lizhu Wang, Xiaodong Liu, Witold Pedrycz. "Effective intervals determined by information granules to improve forecasting in fuzzy time series". Expert Systems withApplications, vol.40, pp.5673-5679, 2013.

[29] Lizhu Wang et al. "Determination of temporal information granules to improve forecasting in fuzzy time series". Expert Systems with Applications, vol.41, pp.3134-3142, 2014

[30] Wei 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.

A FUZZY TIME SERIES FORECASTING MODEL BASED ON THE HEDGE ALGEBRAS AND PARTICLE SWARM OPTIMIZATION

Nghiem Van Tinh, Nguyen Cong Dieu, Nguyen Tien Duy

ABSTRACT: *In recent years, many forecasting models based on fuzzy time series that have been proposed for the analysis of time series. In the forecasting model, the main factors that may affect the forecasted accuracy of model are partitioning the universe of discourse and determining fuzzy logical relationship groups. In this paper, we propose a new fuzzy time series forecasting model based on hedge algebra (HA) and particle swarm optimization (PSO). In that, HA is used as a tool to partition the universe of discourse into intervals with unequal length corresponding to the semantic intervals that calculated from the linguistic terms. After processing of generating the interval, the observation data of time are represented by fuzzy sets and use them to establish fuzzy logic relationship groups. Finally, the proposed model is combined with the PSO technique to find the appropriate divisor to increase the forecasting probability. Finally, the proposed model combined with the PSO technique to find the proper length of each interval for increasing forecast accuracy. Evaluating the performance of the proposed model based on historical data of enrolments at the University of Alabama. The experimental results show that the proposed model the achieves good forecasting results compared to other existing forecasting models based on the first - order and high-order fuzzy time series.*