Kỷ yếu Hội nghị Khoa học Quốc gia lần thứ IX "Nghiên cứu cơ bản và ứng dụng Công nghệ thông tin (FAIR'9)"; Cần Thơ, ngày 4-5/8/2016
DOI: 10.15625/vap.2016.00053

MỘT PHƯƠNG PHÁP MỚI DỰ BÁO CHUỖI THỜI GIAN MỜ DỰA TRÊN NGỮ NGHĨA NGÔN NGỮ

Nguyễn Duy Hiếu[1], Nghiêm Văn Tính[2], Vũ Như Lân[3]
[1]Trường Đại học Tây Bắc
[2]Trường Đại học Kỹ thuật Công nghiệp, Đại học Thái Nguyên
[3]Trường Đại học Thăng Long
hieu3210@gmail.com, nghiemvantinh@tnut.edu.vn, vnlan@ioit.ac.vn

TÓM TẮT— *Dự báo chuỗi thời gian là bài toán đã được rất nhiều tác giả trong và ngoài nước quan tâm nghiên cứu trong những năm gần đây. Tuy nhiên, việc dự báo trên những dữ liệu có sự biến đổi lớn, những dữ liệu được ghi nhận bằng các nhãn ngôn ngữ đã tạo ra những khó khăn khi giải quyết bằng các phương pháp toán học, thống kê truyền thống. Vì vậy, Q. Song và B.S Chissom đã đề xuất mô hình dự báo chuỗi thời gian mờ. Kể từ đó tới nay đã có nhiều nghiên cứu theo hướng này, nhằm đưa ra những phương pháp mới và cải tiến những phương pháp đã có nhằm tăng độ chính xác của dự báo. Trong nội dung bài báo này, chúng tôi đề xuất một phương pháp mới, sử dụng phép ngữ nghĩa hóa và giải ngữ nghĩa của đại số gia tử trong bài toán dự báo số lượng sinh viên nhập học của Trường Đại học Alabama. Mô hình dự báo, các kết quả và so sánh cũng sẽ được đưa ra thảo luận.*

Từ khóa— *Dự báo, chuỗi thời gian, chuỗi thời gian mờ, đại số gia tử, ngữ nghĩa, ngôn ngữ.*

I. MỞ ĐẦU

Vấn đề dự báo tương lai luôn là mong muốn, mơ ước của con người từ khi xuất hiện tới nay. Dự báo trước được những sự việc, hiện tượng xảy ra trong tương lai giúp cho con người hoạch định tốt hơn công việc của mình, giúp nâng cao hiệu quả, tiết kiệm thời gian và giảm bớt công sức trong công việc.

Trong vài chục năm gần đây, đã có rất nhiều nghiên cứu trong và ngoài nước được đưa ra nhằm giải quyết bài toán dự báo. Những nghiên cứu đó dù đi theo hướng nào đi chăng nữa thì mục tiêu cũng nhằm nâng cao độ chính xác của kết quả dự báo và giảm bớt khối lượng tính toán của bài toán này.

Những dữ liệu con người thu thập được có rất nhiều loại, và dữ liệu mà con người dùng cho bài toán dự báo cũng vậy. Dữ liệu dạng số liệu, rõ ràng và chính xác thường được dự báo bằng các phương pháp toán học, thống kê với các mô hình điển hình như ARMA, ARIMA,… Tuy nhiên, với những dữ liệu có sự biến động lớn (như thị trường chứng khoán) hay những dữ liệu trong thực tế được ghi nhận bằng các nhãn ngôn ngữ thì khó có thể giải quyết được bằng các phương pháp toán học, thống kê truyền thống. Chính vì vậy, trong nghiên cứu của mình Q. Song và B.S Chissom [1, 2, 3] đã đề xuất mô hình dự báo chuỗi thời gian mờ nhằm dự báo cho những dữ liệu có đặc điểm nói ở trên. Sau đó, S.M Chen [4] đã cải tiến phương pháp dự báo chuỗi thời gian mờ nhằm đưa ra cách tính toán đơn giản và hiệu quả hơn. Trong những nghiên cứu sau này, nhiều tác giả đã sử dụng các kỹ thuật khác nhau như phân cụm, tối ưu… làm cho độ chính xác của kết quả dự báo ngày càng được nâng cao hơn.

Trong bài toán dự báo chuỗi thời gian mờ, có hai yếu tố chính ảnh hưởng tới kết quả dự báo, đó chính là phép mờ hóa dữ liệu và giải mờ. Đối với việc mờ hóa dữ liệu, nhiều nghiên cứu đã chỉ ra rằng: số lượng khoảng, độ dài khoảng và bậc của chuỗi thời gian mờ ảnh hưởng nhiều tới độ chính xác dự báo. Nhiều tác giả cũng đã đưa ra các kết quả nghiên cứu việc tối ưu các tham số này với những kết quả dự báo khá tốt. Vấn đề giải mờ cũng ảnh hưởng lớn tới kết quả dự báo nếu lựa chọn khoảng giải mờ khác nhau. Việc tối ưu khoảng giải mờ này cũng cần được nghiên cứu để nâng cao độ chính xác dự báo.

Trong phép mờ hóa dữ liệu cần lựa chọn số lượng khoảng chia phù hợp. Nếu chọn số khoảng chia quá ít, dự báo có thể cho độ chính xác thấp do thiếu thông tin; còn nếu chọn số khoảng chia quá nhiều có thể làm mất hết ý nghĩa về tính mờ của giá trị ngôn ngữ.

Đại số gia tử [12] (ĐSGT) được N. Cat Ho và W. Wechler giới thiệu năm 1990 nhằm đưa ra một mô hình toán học phù hợp với dữ liệu không chắc chắn, theo đó các giá trị ngữ nghĩa của ngôn ngữ nằm trong một trật tự nhất định và chính thứ tự đó tạo nên giá trị ngữ nghĩa của từ ngôn ngữ. Đại số gia tử đã được ứng dụng trong các bài toán điều khiển, hồi quy, trích rút tri thức, tính toán trên từ[15, 16, 17, 18, 19, 20],… và cho nhiều kết quả tốt đẹp.

Việc ứng dụng đại số gia tử trong bài toán dự báo cũng đã được các tác giả trong [21] giới thiệu với mô hình dự báo cho kết quả khá tốt. Trong nghiên cứu này, chúng tôi sẽ sử dụng phép ngữ nghĩa hóa và giải ngữ nghĩa của đại số gia tử trên cơ sở việc chia tập nền thành 13 khoảng của S.M Chen trong [6]. Kết quả dự báo cũng sẽ được so sánh trực tiếp với nghiên cứu này nhằm chỉ ra tính ưu việt của phương pháp đề xuất.

Nội dung bài báo được trình bày thành ba mục: sau mục 1 là phần MỞ ĐẦU, trong mục 2 chúng tôi trình bày tổng quan về MÔ HÌNH DỰ BÁO CHUỖI THỜI GIAN MỜ theo các cách tiếp cận của Q. Song, B.S Chissom và S.M

Chen; mục 3 sẽ giới thiệu tóm tắt về LÝ THUYẾT ĐẠI SỐ GIA TỬ và mục 4 đề xuất phương pháp DỰ BÁO CHUỖI THỜI GIAN MỜ SỬ DỤNG ĐSGT VỚI 13 KHOẢNG CHIA. Các phương pháp tính toán và kết quả trong bài toán dự báo số lượng sinh viên nhập học tại Trường Đại học Alabama sẽ được đưa ra so sánh và thảo luận. Cuối cùng là phần KẾT LUẬN.

II. MÔ HÌNH DỰ BÁO CHUỖI THỜI GIAN MỜ

2.1. Một số khái niệm cơ bản của mô hình dự báo chuỗi thời gian mờ

Mô hình chuỗi thời gian mờ lần đầu tiên được Q. Song và B.S Chissom đưa ra [1, 2, 3] và được S.M Chen cải tiến [4, 5, 6, 7] để có thể xử lý bằng các phép tính số học đơn giản hơn nhưng chính xác, phù hợp với các ứng dụng dự báo chuỗi thời gian mờ. Có thể tóm lược qua một số khái niệm cơ bản sau đây:

Định nghĩa 2.1: Chuỗi thời gian mờ

Giả sử $Y(t), (t=... , 0,1,2,. ..)$, là tập các số thực và cũng là tập nền trên đó xác định các tập mờ $f_i(t), (i=1,2 , ...)$. Biến t là thời gian. Nếu $F(t)$ là một chuỗi các tập mờ của $f_i(t), (i=1,2,...)$, thì $F(t)$ được gọi là chuỗi thời gian mờ trên $Y(t), (t=... , 0,1,2,. ..)$.

Định nghĩa 2.2: Quan hệ mờ

Nếu tồn tại quan hệ mờ $R(t-1, t)$, sao cho $F(t)=F(t-1)*R(t-1, t)$, trong đó dấu * ký hiệu toán tử nào đó, thì $F(t)$ được suy ra từ $F(t-1)$. Quan hệ giữa $F(t)$ và $F(t-1)$ được xác định bằng ký hiệu:

$$F(t-1) \rightarrow F(t) \tag{2.1}$$

Ví dụ về toán tử * có thể là phép kết hợp *MaxMin*[2] hoặc *MinMax*[3] hay phép tính số học [4] .

Nếu $F(t-1)=A_i$ and $F(t)=A_j$, quan hệ logic giữa $F(t)$ and $F(t-1)$ được ký hiệu bằng $A_i \rightarrow Aj$, trong đó A_i là vế trái và A_j là vế phải của quan hệ mờ mô tả tập mờ dự báo.

Định nghĩa 2.3: Quan hệ mờ bậc n

Giả sử $F(t)$ là chuỗi thời gian mờ. Nếu $F(t)$ được suy ra từ $F(t-1), F(t-2),..., F(t-n)$, thì quan hệ mờ này được biểu diễn bằng biểu thức:

$$F(t-n),...,F(t-2), F(t-1) \rightarrow F(t) \tag{2.2}$$

và được gọi là chuỗi thời gian mờ bậc n.

Định nghĩa 2.4: Chuỗi thời gian mờ dừng

Giả sử $F(t)$ được suy ra từ $F(t-1)$ và được ký hiệu bằng $F(t-1) \rightarrow F(t)$, khi đó quan hệ mờ giữa $F(t)$ và $F(t-1)$ được mô tả bằng phương trình:

$$F(t)=F(t-1)*R(t-1, t) \tag{2.3}$$

Quan hệ mờ R thể hiện mô hình bậc nhất của $F(t)$. Nếu $R(t-1, t)$ không phụ thuộc t, sao cho với mọi t_1 và t_2 khác nhau, $R(t_1 , t_1-1)=R(t_2 ,t_2-1)$, thì $F(t)$ được gọi là chuỗi thời gian mờ dừng, còn lại được gọi là chuỗi thời gian mờ không dừng.

Định nghĩa 2.5: Nhóm quan hệ mờ (NQHM)

Các quan hệ mờ với cùng một tập mờ bên vế trái có thể đưa vào một nhóm gọi là nhóm quan hệ mờ hay nhóm quan hệ logic mờ.

Giả sử có các quan hệ mờ: $A_i \rightarrow A_{j1}; A_i \rightarrow A_{j2};....; A_i \rightarrow A_{jn}$ thì các quan hệ mờ trên có thể đưa vào một nhóm được ký hiệu như sau:

$$A_i \rightarrow A_{j1}, A_{j2}, , ..., A_{jn} \tag{2.4}$$

Tập mờ A_{jk} $(k=1,2,.., n)$ chỉ được xuất hiện 1 lần bên vế phải.

2.2. Mô hình dự báo của Q. Song và B.S Chissom

Mô hình dự báo chuỗi thời gian mờ lần đầu tiên được Q. Song và B.S Chissom đưa ra vào năm 1993 [1, 2, 3] và được ứng dụng để dự báo số sinh viên nhập học tại Trường Đại học Alabama với dữ liệu lịch sử qua 22 năm kể từ năm 1971 đến 1992.

Chuỗi thời gian lần đầu tiên được xem xét dưới góc độ biến ngôn ngữ và bài toán dự báo đã có được một cách nhìn hoàn toàn mới trên quan điểm lý thuyết tập mờ. Mô hình dự báo đầu tiên là mô hình dự báo chuỗi thời gian mờ dừng [2, 3] và được triển khai qua các bước sau đây:

Bước 1. Xác định tập nền

Bước 2. Chia miền xác định của tập nền thành những khoảng bằng nhau
Bước 3. Xây dựng các tập mờ trên tập nền
Bước 4. Mờ hóa chuỗi dữ liệu
Bước 5. Xác định các quan hệ mờ
Bước 6. Dự báo bằng phương trình $A_i=A_{i-1}* R$, ở đây ký hiệu * là toán tử *MaxMin*
Bước 7. Giải mờ các kết quả dự báo.
Trong bước 5, quan hệ mờ R được xác định bằng biểu thức $R_i = A_s^T$ x A_q , với mọi quan hệ mờ $A_s \rightarrow A_q$,

$$R = \bigcup_{i=1}^{k} R_i \quad (2.5)$$

Ở đây x là toán tử *min*, *T* là phép chuyển vị và ∪ là phép hợp.

2.3. Mô hình dự báo S.M Chen

Do mô hình dự báo chuỗi thời gian mờ của Q. Song và B.S Chissom khá phức tạp trong bước 5 và bước 6, vì vậy S.M Chen [4] đã cải tiến cách tính toán chính xác hơn cho các mô hình dự báo chuỗi thời gian chỉ sử dụng các phép tính số học đơn giản trên cơ sở thông tin từ các quan hệ mờ và nhóm quan hệ mờ theo các bước sau đây:

Bước 1. Chia miền xác định của tập nền thành những khoảng bằng nhau.
Bước 2. Xây dựng các tập mờ trên tập nền.
Bước 3. Mờ hóa chuỗi dữ liệu.
Bước 4. Xác định các quan hệ mờ.
Bước 5. Tạo lập nhóm quan hệ mờ.
Bước 6. Giải mờ đầu ra dự báo.

2.4. Luật dự báo chuỗi thời gian mờ [4]

Luật dự báo cũng chính là phép giải mờ các kết quả đầu ra dự báo như ở bước 6 của mô hình dự báo [4].

Giả sử dữ liệu của chuỗi thời gian *F(t-1)* được mờ hóa bằng A_j, khi đó. Đầu ra dự báo của *F (t)* được xác định theo những luật (nguyên tắc) sau đây:

1. Nếu tồn tại quan hệ một - một trong nhóm quan hệ của A_j, ký hiệu là $A_j \rightarrow A_k$, và mức độ thuộc cao nhất của A_k tại khoảng u_k, thì đầu ra dự báo của *F (t)* là điểm giữa của u_k.

2. Nếu A_k là trống, có nghĩa là $A_j \rightarrow \varnothing$ và A_j có mức độ thuộc cao nhất tại khoảng u_j, thì đầu ra dự báo là điểm giữa của u_j.

3. Nếu tồn tại quan hệ một - nhiều trong nhóm quan hệ mờ của A_j , ký hiệu là $A_j \rightarrow A_1$, A_2 ,…, A_n, và mức độ thuộc cao nhất của A_1 , A_2 ,…, A_n tại các khoảng u_1 , u_2 ,…, u_n tương ứng, thì đầu ra dự báo được tính bằng trung bình các điểm giữa m_1 , m_2,…,m_n của u_1 , u_2,…, u_n. Đầu ra dự báo khi này có dạng: $(m_1+m_2+\ldots+m_n)/n$.

III. LÝ THUYẾT ĐẠI SỐ GIA TỬ

Đại số gia tử được giới thiệu bởi N.C. Ho và W. Wechler [12, 13] nhằm giải quyết vấn đề phương pháp luận và cung cấp một mô hình toán học mềm dẻo, linh hoạt và hiệu quả cho việc xử lý dữ liệu mờ. Đại số gia tử đã được ứng dụng thành công trong nhiều bài toán khác nhau như: điều khiển mờ, tính toán trên từ, phân lớp, hồi quy [15, 16, 17, 18, 19]… với nhiều kết quả tốt hơn những cách tiếp cận trước đó.

Trong phần này chúng tôi sẽ trình bày tóm tắt những lý thuyết của đại số gia tử, trong đó quan trọng nhất là các công thức để xác định mô hình tính toán trong đại số gia tử bao gồm cả phép ngữ nghĩa hóa và giải ngữ nghĩa hóa. Đây cũng là cơ sở để ứng dụng đại số gia tử trong bài toán dự báo chuỗi thời gian mờ

Trước hết, chúng ta xem xét một số khái niệm cơ bản của đại số gia tử sau đây:

Gọi $AX = (X, G, C, H, \leq)$ là một cấu trúc đại số, với X là tập nền của AX; $G = \{c-, c+\}$ là tập các phần tử sinh; $C = \{0, W, 1\}$, trong đó *0*, *W* và *1* tương ứng là những phần tử đặc trưng cận trái (tuyệt đối nhỏ), trung hòa và cận phải (tuyệt đối lớn); H là tập các toán tử một ngôi được gọi là các gia tử; $\leq$ là biểu thị quan hệ thứ tự trên các giá trị ngôn ngữ. Gọi H^- là tập hợp các gia tử âm và H^+ là tập hợp các gia tử dương của AX.

Ký hiệu $H^- = \{h_{-1}, h_{-2}, \ldots h_{-q}\}$, trong đó $h_{-1} < h_{-2} < \ldots < h_{-q}$ và $H^+ = \{h_1, h_2, \ldots, h_p\}$, trong đó $h_1 < h_2 < \ldots < h_p$.

Định nghĩa 3.1: Độ đo tính mờ

fm: $X \rightarrow [0, 1]$ 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 $\forall x \in X$ (3.1)

+) Với các phần tử *0*, *W* và *1*, $fm(0) = fm(W) = fm(1) = 0$ (3.2)

+) Và với $\forall x,y \in X$, $\forall h \in H$, $\frac{fm(hx)}{fm(x)} = \frac{fm(hy)}{fm(y)}$ (3.3)

Đẳng thức (3.3) không phụ thuộc vào các phần tử *x, y* và do đó ta có thể ký hiệu là $\mu(h)$ và đây là độ đo tính mờ của gia tử *h*. Tính chất của *fm(x)* và $\mu(h)$ như sau:

+) $fm(hx) = \mu(h) fm(x), \ \forall x \in X$ (3.4)

+) $\sum_{i=-q, i\neq 0}^{p} fm(h_i c) = fm(c)$, với $c \in \{c-, c+\}$ (3.5)

+) $\sum_{i=-q, i\neq 0}^{p} fm(h_i x) = fm(x)$ (3.6)

+) $\sum_{i=-1}^{-q} \mu(h_i) = \alpha$ và $\sum_{i=1}^{p} \mu(h_i) = \beta$, với $\alpha, \beta > 0$ và $\alpha+\beta = 1$ (3.7)

Định nghĩa 3.2: Hàm dấu

Hàm *sign*: $X \to \{-1, 0, 1\}$ là một ánh xạ được gọi là hàm dấu với $h, h' \in H$ và $c \in \{c-, c+\}$ trong đó:

+) $sign(c-) = -1$, $sign(c+) = +1$; (3.8)

+) $sign(hc) = - sign(c)$, nếu *h* là âm đối với *c*; (3.9)

+) $sign(hc) = + sign(c)$, nếu *h* là dương đối với *c*; (3.10)

+) $sign(h'hx) = - sign(hx)$, nếu $h'hx \neq hx$ và *h'* là âm đối với *h*; (3.11)

+) $sign(h'hx) = + sign(hx)$, nếu $h'hx \neq hx$ và *h'* là dương đối với *h*; (3.12)

+) $sign(h'hx) = 0$ nếu $h'hx = hx$. (3.13)

Gọi *fm* là một độ đo tính mờ trên *X*, ánh xạ ngữ nghĩa định lượng $v: X \to [0, 1]$, được sinh ra bởi *fm* trên *X*, được xác định như sau:

+) $v(W) = \theta = fm(c^-)$, (3.14)

+) $v(c^-) = \theta - \alpha\, fm(c^-) = \beta\, fm(c^-)$, (3.15)

+) $v(c^+) = \theta + \alpha\, fm(c^+) = 1 - \beta\, fm(c^+)$ (3.16)

+) $v(h_j x) = v(x) + sign(h_j x)\left[\sum_{i=sign(j)}^{j} fm(h_i x) - \omega(h_j x) fm(h_j x)\right]$ (3.17)

với $\omega(h_j x) = \frac{1}{2}\left[1 + sign(h_j x).sign(h_p h_j x)(\beta - \alpha)\right] \in \{\alpha, \beta\}, j \in [-q, p], j \neq 0.$ (3.18)

Ngữ nghĩa hóa (Semantization) và giải ngữ nghĩa hóa (Desemantization)

Để thuận tiện cho việc biểu diễn ngữ nghĩa của các giá trị ngôn ngữ, giả sử rằng miền tham chiếu thông thường của các biến ngôn ngữ *X* là đoạn [a, b] còn miền tham chiếu ngữ nghĩa X_s là đoạn $[a_s,b_s]$ ($0 \leq a_s < b_s \leq 1$). Việc chuyển đổi tuyến tính từ [a, b] sang $[a_s,b_s]$ được gọi là phép ngữ nghĩa hóa tuyến tính (Linear Semantization) còn việc chuyển ngược lại từ đoạn $[a_s,b_s]$ sang [a, b] được gọi là phép giải nghĩa tuyến tính (Linear Desemantization). Trong nhiều ứng dụng của ĐSGT đã sử dụng miền ngữ nghĩa là đoạn $[a_s=0, b_s=1]$, khi đó phép ngữ nghĩa hóa tuyến tính được gọi là phép chuẩn hóa (Linear Semantization = Normalization) và phép giải nghĩa tuyến tính được gọi là phép giải chuẩn (Linear Desemantization = Denormalization). Như vậy có thể biểu diễn phép ngữ nghĩa hóa tuyến tính và phép giải nghĩa tuyến tính đơn giản như sau:

- *Linear Semantization* $(x) = x_s = a_s + (b_s - a_s)(x - a)/(b - a)$ (3.19a)
- *Linear Desemantization* $(x_s) = x = a + (b - a)(x_s - a_s)/(b_s - a_s)$ (3.20a)
- *Normalization* $(x) = x_s = (x - a)/(b - a)$ (3.19b)
- *Denormalization* $(x_s) = x = a + (b - a)x_s$ (3.20b)

trong đó *a, b* là các số thực.

Nhiều ứng dụng của ĐSGT trong những lĩnh vực khoa học đòi hỏi mở rộng không gian tham số trong các phép ngữ nghĩa hóa và phép giải nghĩa để có nhiều tham số lựa chọn mềm dẻo hơn nữa. Điều này chỉ có thể có được khi mở rộng phép ngữ nghĩa hóa và phép giải nghĩa từ tuyến tính đến phi tuyến. Tương tự trên, phép ngữ nghĩa hóa phi tuyến và phép giải nghĩa phi tuyến có thể được biểu diễn như sau:

- *Nonlinear Semantization* $(x) = f(x_s,sp)$ (3.19c)

  Với điều kiện: $0 \leq f(x_s,sp) \leq 1$ và $f(x_s=0,sp) = 0$ và $f(x_s=1,sp) = 1$

- *Nonlinear Desemantization* $(x_s) = g(x,dp)$ (3.20c)

  Với điều kiện: $a \le g(x,dp) \le b$, $g(x = a,dp) = a$ và $g(x = b,dp) = b$

Các hàm *f(.)* và *g(.)* được chọn tùy theo từng ứng dụng và là các hàm liên tục, đồng biến, trong đó $sp \in [-1\ 1]$ là tham số ngữ nghĩa hóa, $dp \in [-1\ 1]$ là tham số giải nghĩa. Ví dụ có thể chọn *f(.)* phi tuyến theo x_s thể hiện qua $f(x_s,sp)$ và *g(.)* phi tuyến theo *x* thể hiện qua *Denormalization* $(f(x_s,sp))$ như sau:

$$f(x_s,sp) = sp*x_s*(1-x_s)+x_s \quad (3.19d)$$

$$g(x,dp) = dp*((\text{Denormalization } (f(x_s,sp))–a)*(b – \text{Denormalization } (f(x_s,sp))) / (b-a) +$$

$$+ \text{Denormalization } (f(x_s,sp)) \quad (3.20d)$$

$$\text{trong đó Denormalization } (f(x_s,sp)) = (sp*x*(1-x)+x)*(b-a) + a \quad (3.20d1)$$

Hàm $f(x_s,sp)$ là hàm biểu diễn ngữ nghĩa phi tuyến trong phép giải nghĩa phi tuyến *g(x.dp)* chưa được sử dụng trong các ứng dụng của ĐSGT. Lưu ý rằng: có thể chọn các hàm $f(x_s,sp)$ và $g(x,dp)$ độc lập với nhau.

Khi *sp=dp=0* tính phi tuyến bị loại bỏ và biểu thức (3.19d) trở thành (3.19b) và (3.20d) trở thành (3.20b).

Cho trước độ đo tính mờ của các gia tử $\mu(h)$ và các giá trị độ đo tính mờ của các phần tử sinh $fm(c^-)$, $fm(c^+)$ và θ là phần tử trung hoà (neutral). Khi đó mô hình tính toán của ĐSGT được xây dựng trên cơ sở các biểu thức từ (3.1) đến (3.20) được kích hoạt và thực tế đã được sử dụng hiệu quả trong rất nhiều ứng dụng. Phép mờ hóa và phép giải mờ trong tiếp cận mờ được thay thế tương ứng bằng phép ngữ nghĩa hóa và phép giải nghĩa trong tiếp cận ĐSGT. Hệ luật được thể hiện bằng siêu mặt làm cơ sở cho quá trình suy luận xấp xỉ. Một lưu ý quan trọng của quá trình tính toán trong tiếp cận ĐSGT là cần xác định các tham số ban đầu như độ đo tính mờ của các phần tử sinh và độ đo tính mờ của các gia tử trong biến ngôn ngữ một cách thích hợp dựa trên cơ sở phân tích ngữ nghĩa của miền ngôn ngữ trong từng bài toán ứng dụng cụ thể. Khi đó mô hình tính toán của tiếp cận ĐSGT sẽ cho các kết quả hợp lý trong các ứng dụng.

IV. DỰ BÁO CHUỖI THỜI GIAN MỜ SỬ DỤNG ĐSGT VỚI 13 KHOẢNG CHIA

Trong phần này, chúng tôi sẽ đề xuất việc sử dụng lý thuyết của đại số gia tử, cụ thể là sử dụng phép ngữ nghĩa hóa và giải ngữ nghĩa hóa trong bài toán dự báo chuỗi thời gian mờ theo cách chia khoảng của S.M Chen [6].

Về việc ứng dụng đại số gia tử trong mô hình chuỗi thời gian mờ cho bài toán dự báo số sinh viên nhập học trên đã được Nguyễn Duy Hiếu đề xuất trong [21]. Trong nghiên cứu đó đã chỉ rõ việc sử dụng các công thức tính toán của đại số gia tử để đưa ra mô hình dự báo theo 6 bước cơ bản. Trong nghiên cứu này, chúng tôi muốn thử nghiệm tính hiệu quả của mô hình trên với cách chia khoảng mới của S.M Chen [6] đối với bài toán dự báo số lượng sinh viên nhập học tại Trường Đại học Alabama theo số liệu ghi nhận được như bảng sau:

Bảng 4.1 Số sinh viên nhập học tại Trường Đại học Alabama từ 1971 đến 1992 [2]

Năm	Số sinh viên nhập học	Năm	Số sinh viên nhập học
1971	13055	1982	15433
1972	13563	1983	15497
1973	13867	1984	15145
1974	14696	1985	15163
1975	15460	1986	15984
1976	15311	1987	16859
1977	15603	1988	18150
1978	15861	1989	18970
1979	16807	1990	19328
1980	16919	1991	19337
1981	16388	1992	18876

Theo S.M Chen [6], có thể chia lại tập nền thành 13 khoảng (không đều nhau) từ 7 khoảng như cách chia trước đó [2, 3, 4] trên cơ sở thống kê số lượng các điểm dữ liệu thuộc về các khoảng đó. Theo đó, những khoảng nào có nhiều dữ liệu lịch sử thuộc vào hơn thì chia thành nhiều khoảng hơn và ngược lại. Cá biệt có khoảng không có dữ liệu lịch sử thuộc vào thì có thể bỏ đi.

Bảng 4.2 Thống kê lịch sử dữ liệu của cách chia 7 khoảng

Khoảng	[13000,14000]	[14000,15000]	[15000,16000]	[16000,17000]	[17000,18000]	[18000,19000]	[19000,20000]
Số dữ liệu	3	1	9	4	0	3	2

Theo S.M Chen [6], ta chia khoảng có 9 dữ liệu lịch sử thành 4 khoảng con, khoảng có 4 dữ liệu lịch sử thành 3 khoảng con, khoảng có 3 dữ liệu lịch sử thành 2 khoảng con, khoảng không có dữ liệu lịch sử thuộc vào thì bỏ đi, còn lại giữ nguyên. Các nhãn giá trị ngôn ngữ được Chen dùng ở đây gồm: A_1=very very very very few, A_2=very very very few, A_3=very very few, A_4 =very few, A_5 =few, A_6 = moderate, A_7=many, A_8=many many, A_9= very many, A_{10}=too many, A_{11}=too many many, A_{12}=too many many many và A_{13}= too many many many many.

Khác với cách tiếp cận của S.M Chen, chúng tôi đề xuất mô hình đại số gia tử được xây dựng bởi các phần tử sinh c^- (small) và c^+ (large) với tác động của hai gia tử (Little, Very) thuộc H. Việc lựa chọn các giá trị ngữ nghĩa tương ứng với cách chia khoảng của Chen cụ thể như bảng 4.3.

Bảng 4.3 Nhãn ngữ nghĩa của các khoảng

Stt	Phân đoạn	Kí hiệu	Giá trị ngữ nghĩa
1	$u_{1,1}$ = [13000, 13500]	A_1	Very Very Small
2	$u_{1,2}$ = [13500, 14000]	A_2	Little Very Small
3	u_2 = [14000, 15000]	A_3	Small
4	$u_{3,1}$ = [15000, 15250]	A_4	Very Very Little Small
5	$u_{3,2}$ = [15250, 15500]	A_5	Little Very Little Small
6	$u_{3,3}$ = [15500, 15750]	A_6	Very Little Little Small
7	$u_{3,4}$ = [15750, 16000]	A_7	Little Little Little Small
8	$u_{4,1}$ = [16000, 16333]	A_8	Little Little Little Large
9	$u_{4,2}$ = [16333, 16667]	A_9	Little Little Large
10	$u_{4,3}$ = [16667, 17000]	A_{10}	Very Little Little Large
11	$u_{6,1}$ = [18000, 18500]	A_{11}	Very Little Large
12	$u_{6,2}$ = [18500, 19000]	A_{12}	Little Very Large
13	u_7 = [19000, 20000]	A_{13}	Very Large

Trong đó, các A_i, i=1..13 là các kí hiệu (nhãn ngữ nghĩa) tương ứng các giá trị ngữ nghĩa được chọn của đại số gia tử. Việc lựa chọn các giá trị ngữ nghĩa này đảm bảo tỉ lệ, mật độ chia khoảng.

Đối với các giá trị ngữ nghĩa được chọn, giá trị ngữ ngữ nghĩa định lượng của 13 nhãn ngữ nghĩa A_1, A_2,… A_{13} được tính toán cụ thể theo các công thức sau:

- SA_1 = v(Very Very Small) = $\theta-3\theta\alpha+3\theta\alpha^2-\theta\alpha^3$;
- SA_2 = v(Little Very Small) = $\theta-2\theta\alpha+2\theta\alpha^2-\theta\alpha^3$;
- SA_3 = v(Small) = $\theta-\theta\alpha$;
- SA_4 = v(Very Very Little Small) = $\theta-\theta\alpha+\theta\alpha^2-2\theta\alpha^3+\theta\alpha^4$;
- SA_5 = v(Little Very Little Small) = $\theta-\theta\alpha+\theta\alpha^2-\theta\alpha^3+\theta\alpha^4$;
- SA_6 = v(Very Little Little Small) = $\theta-\theta\alpha+2\theta\alpha^2-2\theta\alpha^3+\theta\alpha^4$;
- SA_7 = v(Little Little Little Small) = $\theta-\theta\alpha+3\theta\alpha^2-3\theta\alpha^3+\theta\alpha^4$;
- SA_8 = v(Little Little Little Large) = $\theta+\alpha-3\alpha^2+3\alpha^3-\alpha^4-\theta\alpha+3\theta\alpha^2-3\theta\alpha^3+\theta\alpha^4$;
- SA_9 = v(Little Little Large) = $\theta+\alpha-2\alpha^2+\alpha^3-\theta\alpha+2\theta\alpha^2-\theta\alpha^3$;
- SA_{10} = v(Very Little Little Large) = $\theta+\alpha-2\alpha^2+2\alpha^3-\alpha^4-\theta\alpha+2\theta\alpha^2-2\theta\alpha^3+\theta\alpha^4$;
- SA_{11} = v(Very Little Large) = $\theta+\alpha-\alpha^2+\alpha^3-\theta\alpha+\theta\alpha^2-\theta\alpha^3$;
- SA_{12} = v(Little Very Large) = $\theta+2\alpha-2\alpha^2+\alpha^3-2\theta\alpha+2\theta\alpha^2-\theta\alpha^3$;
- SA_{13} = v(Very Large) = $\theta+2\alpha-\alpha^2-2\theta\alpha+\theta\alpha^2$;

Trong đó kí hiệu SA_i = Semantization(A_i) là giá trị ngữ nghĩa định lượng của nhãn ngữ nghĩa A_i.

Nếu chọn trước α=0.5 và θ=0.5 thì giá trị ngữ nghĩa định lượng tính được như sau:

- SA_1 = v(Very Very Small) = 0.0625;
- SA_2 = v(Little Very Small) = 0.1875;
- SA_3 = v(Small) = 0.25;
- SA_4 = v(Very Very Little Small) = 0.28125;
- SA_5 = v(Little Very Little Small) = 0.34375;
- SA_6 = v(Very Little Little Small) = 0.40625;
- SA_7 = v(Little Little Little Small) = 0.46875;
- SA_8 = v(Little Little Little Large) = 0.53125;
- SA_9 = v(Little Little Large) = 0.5625;
- SA_{10} = v(Very Little Little Large) = 0.59375;
- SA_{11} = v(Very Little Large) = 0.6875;
- SA_{12} = v(Little Very Large) = 0.8125;
- SA_{13} = v(Very Large) = 0.875;

Chúng ta dễ thấy rằng các giá trị ngữ nghĩa định lượng luôn được theo thứ tự:

$$SA_1 < SA_2 < \ldots < SA_{13}$$

hay nói cách khác thứ tự ngữ nghĩa luôn được đảm bảo. Đây cũng chính là điểm khác biệt quan trọng của lý thuyết đại số gia tử so với lý thuyết mờ khi đặt các giá trị ngôn ngữ trong thứ tự của nó, và cũng chính thứ tự đó tạo nên cấu trúc của ngôn ngữ.

Kết hợp dữ liệu của bảng 4.1 với cách gán nhãn ngôn ngữ theo bảng 4.3 ta được bảng dữ liệu sinh viên nhập học với nhãn ngữ nghĩa theo bảng 4.4 bên dưới.

Bảng 4.4 Dữ liệu sinh viên nhập học với nhãn ngữ nghĩa tương ứng

Năm	Số SVNH	Kí hiệu	Năm	Số SVNH	Kí hiệu
1971	13055	A_1	1982	15433	A_5
1972	13563	A_2	1983	15497	A_5
1973	13867	A_2	1984	15145	A_4
1974	14696	A_3	1985	15163	A_4
1975	15460	A_5	1986	15984	A_7
1976	15311	A_5	1987	16859	A_{10}
1977	15603	A_6	1988	18150	A_{11}
1978	15861	A_7	1989	18970	A_{12}
1979	16807	A_{10}	1990	19328	A_{13}
1980	16919	A_{10}	1991	19337	A_{13}
1981	16388	A_9	1992	18876	

Từ bảng 4.4 ta tìm được các nhóm quan hệ ngữ nghĩa như sau:

Bảng 4.5 Các nhóm quan hệ ngữ nghĩa

Nhãn	Nhóm quan hệ
A_1	$A_1 \rightarrow A_2$
A_2	$A_2 \rightarrow A_2, A_3$
A_3	$A_3 \rightarrow A_5$
A_4	$A_4 \rightarrow A_4, A_7$
A_5	$A_5 \rightarrow A_4, A_5$ (2 lần), A_6
A_6	$A_6 \rightarrow A_7$
A_7	$A_7 \rightarrow A_{10}$ (2 lần)
A_8	không có quan hệ
A_9	$A_9 \rightarrow A_5$
A_{10}	$A_{10} \rightarrow A_9, A_{10}, A_{11}$
A_{11}	$A_{11} \rightarrow A_{12}$
A_{12}	$A_{12} \rightarrow A_{13}$
A_{13}	$A_{13} \rightarrow A_{13}$

Về việc lựa chọn khoảng giải nghĩa, chúng tôi lựa chọn sao cho các khoảng giải nghĩa với từng điểm dự báo bên trái của quan hệ ngữ nghĩa sẽ được chọn sao cho hai đầu khoảng bao được các giá trị nằm ở bên phải của nhóm quan hệ ngữ nghĩa.

Bảng 4.6 Khoảng giải nghĩa cho các điểm dự báo

Khoảng giải nghĩa cho các điểm dự báo	Giá trị đầu khoảng	Giá trị cuối khoảng	Khoảng giải nghĩa cho các điểm dự báo	Giá trị đầu khoảng	Giá trị cuối khoảng
1 (1972)	13100	15500	12 (1983)	13400	20000
2 (1973)	13300	16000	13 (1984)	13200	19300
3 (1974)	13400	19900	14 (1985)	14300	16700
4 (1975)	14000	18600	15 (1986)	14100	19500
5 (1976)	14000	18200	16 (1987)	15300	18000
6 (1977)	14600	17700	17 (1988)	15500	20000
7 (1978)	13900	18300	18 (1989)	16800	19500
8 (1979)	14200	18800	19 (1990)	15600	20000
9 (1980)	13500	19300	20 (1991)	15000	20000
10 (1981)	13000	18700	21 (1992)	13000	20000
11 (1982)	14600	17200			

Sử dụng mô hình dự báo chuỗi thời gian mờ bằng đại số gia tử với 6 bước được giới thiệu trong [21] với cách chia khoảng theo Chen [6], cách tính toán ngữ nghĩa định lượng, các nhóm quan hệ mờ, khoảng giải nghĩa như trên chúng tôi thu được kết quả dự báo như sau:

Bảng 4.7 So sánh kết quả dự báo

Stt	Năm	Số sinh viên nhập học	Phương pháp Chen [21]	Phương pháp đề xuất
1	1971	13055		
2	1972	13563	13750	13500
3	1973	13867	13875	13830
4	1974	14696	14750	14676
5	1975	15460	15375	15461
6	1976	15311	15313	15334
7	1977	15603	15625	15584
8	1978	15861	15813	15852
9	1979	16807	16834	16836
10	1980	16919	16834	16950
11	1981	16388	16416	16391
12	1982	15433	15375	15426
13	1983	15497	15375	15496
14	1984	15145	15125	15137
15	1985	15163	15125	15137
16	1986	15984	15938	15983
17	1987	16859	16834	16847
18	1988	18150	18250	18177
19	1989	18970	18875	18969
20	1990	19328	19250	19424
21	1991	19337	19250	19346
22	1992	18876	18875	19084
MSE			**5344**	**2988**

Chú ý rằng bảng tổng hợp số liệu trên sử dụng kết quả của S.M Chen [6] tuy nhiên đã làm tròn đến phần nguyên theo quy tắc làm tròn cho hợp lý hơn về số lượng sinh viên (của Chen vẫn để số lẻ). Kết quả tính toán theo phương pháp đề xuất cũng được làm tròn tương tự.

Tham số ngữ nghĩa hóa (sp) và tham số giải nghĩa (dp) dùng để tính toán trong mô hình dự báo theo đại số gia tử đã giới thiệu ở công thức (3.19c) và (3.20c) được chọn tương ứng là 0.2 và -0.3.

Công thức xác định sai số bình phương trung bình (MSE) là:

$$MSE = \frac{\sum_{i=2}^{22}(\text{Dữ\_liệu\_thực}_i - \text{Dữ\_liệu\_dự\_báo}_i)}{21}$$

(chỉ dự báo 21 năm từ 1972 tới 1992).

V. KẾT LUẬN

Trong nghiên cứu này, chúng tôi đã sử dụng đại số gia tử trong bài toán dự báo chuỗi thời gian mờ theo cách chia tập nền của bài toán dự báo sinh viên nhập học Alabama thành 13 khoảng theo S.M Chen. Qua kết quả dự báo, ta dễ dàng thấy được phương pháp đề xuất có kết quả dự báo tốt hơn nhiều so với kết quả của Chen.

Trong [21] đã so sánh kết quả dự báo theo cách chia truyền thống 7 đoạn giữa phương pháp sử dụng đại số gia tử và các phương pháp khác sử dụng lý thuyết mờ, thêm kết quả của nghiên cứu này cho thấy khả năng ứng dụng của đại số gia tử trong bài toán dự báo chuỗi thời gian mờ là một hướng đi tốt, có thể tiếp tục mở rộng nghiên cứu.

Chúng ta có thể nghiên cứu việc sử dụng đại số gia tử trong bài toán dự báo chuỗi thời gian mờ với việc tối ưu các tham số của đại số gia tử, tối ưu khoảng chia và áp dụng phương pháp này cho các tập dữ liệu khác để có được cái nhìn khách quan, toàn diện hơn độ chính xác và hiệu quả dự báo.

TÀI LIỆU THAM KHẢO

[1] Q. Song, B.S Chissom. Fuzzy time series and its models. Fuzzy Sets and Syst. 54 269–277, 1993

[2] Q. Song, B.S Chissom, Forecasting enrollments with fuzzy time series – part 1. Fuzzy Sets and Syst. 54, 1–9, 1993

[3] Q. Song, B.S Chissom, Forecasting enrollments with fuzzy time series – part 2. Fuzzy Sets and Syst. 62, 1–8, 1994.

[4] S.M Chen, Forecasting Enrollments Based on Fuzzy Time Series. Fuzzy Sets and Syst. 81, 311–319, 1996

[5] S.M Chen, Forecasting Enrollments based on High Order Fuzzy Time Series. Cybernetics and Systems: An International Journal. 33,1-16, 2002.

[6] S.M Chen, C.C Hsu, A New Method to Forecast Enrollments using Fuzzy Time Series. Int. Journal Applied Science and Engineering 2, 234-244, 2004.

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

[8] S.M Chen, K. Tanuwijaya, Multivariate fuzzy forecasting based on fuzzy time series and automatic clustering techniques. Expert Systems with Applications 38, 10594–10605, 2011

[9] K. Huarng, Heuristic models of fuzzy time series for forecasting. Fuzzy Sets and Systems, 123: 369-386, 2001.

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

[11] M. H Lee, R. Efendi, Z. Ismad, Modified Weighted for Enrollments Forecasting Based on Fuzzy Time Series. MATEMATIKA, 25(1), 67-78, 2009.

[12] N. Cat Ho and W. Wechler, Hedge algebras: An algebraic approach to structures of sets of linguistic domains of linguistic truth variable, Fuzzy Sets and Systems, Vol. 35,3, pp.281-293, 1990

[13] N. Cat Ho and W. Wechler, Extended hedge algebras and their application to Fuzzy logic, Fuzzy Sets and Systems 52, 259-281, 1992.

[14] Cat Ho N. and H. Van Nam: An algebraic approach to linguistic hedges in Zadeh's fuzzy logic, Fuzzy Set and System, 129, 229-254, 2002.

[15] Nguyen Cat Ho, Vu Nhu Lan, Le Xuan Viet, Optimal hedge-algebras-based controller: Design and Application, Fuzzy Sets and Systems 159, 968– 989, 2008

[16] Dinko Vukadinović, Mateo Bašić, Cat Ho Nguyen, Nhu Lan Vu, Tien Duy Nguyen Hedge-Algebra-Based Voltage Controller for a Self-Excited Induction Generator, Control Engineering Practice, 30, 78–90, 2014.

[17] Nguyen Dong Anh, Bui Hai Le, Vu Nhu Lan and Tran Duc Trung, Application of hedgealgebras-based fuzzy controller to active control of a structure against earthquake Struct. Control Health Monit 20, 483–495, 2013

[18] Hai Le Bui, Duc Trung Tran, Lan Nhu Vu, Optimal fuzzy control of inverted pendulum. Journal of Vibration and Control, 18 (14), 2097-2110, 2012

[19] Nguyen Dinh Duc, Vu Nhu Lan, Tran Duc Trung and Bui Hai Le A study on the application of hedge algebras to active fuzzy control of a seism-excited structure, Journal of Vibration and Control,18 (14), 2186–2200, 2012

[20] Nguyễn Công Điều, Một thuật toán mới cho mô hình chuỗi thời gian mờ, Tạp chí Khoa học và Công nghệ, Tập 49, Số 4, 11-25, 2011

[21] Nguyễn Duy Hiếu, Vũ Như Lân, Nguyễn Cát Hồ, Dự báo chuỗi thời gian mờ dựa trên ngữ nghĩa, Kỷ yếu Hội nghị Quốc gia lần thứ 8 về Nghiên cứu cơ bản và ứng dụng CNTT (FAIR), 232-243, 2015.

A NEW METHOD TO FORECAST USING FUZZY TIME SERIES BASE ON LINGISTIC SEMANTICS

Nguyen Duy Hieu, Nghiem Van Tinh, Vu Nhu Lan

ABSTRACT*— The time series forcasting problem has reseached by many authors in recent years. But forcasting on data with large changes by time or data recorded by the linguistic labels caused many difficulties when solving it with traditional mathematical and statistical methods. So Q. Song and B.S Chissom proposed the fuzzy time series forcasting model. Since then, there are many studies in this direction, in order to provide new methods or improve existing methods to increase the forcasting accuracy. In this paper, we proposed a new method using hegde algebra semantization and desemantization to Alabama enrollments forcasting problem. The forcasting model, the results and the comparisons will also be discussed.*

Keywords*— Forecasting, prediction, times series, fuzzy time series, hedge algebra, semantic, linguistic.*