A FUZZINNESS PARAMETER OPTIMIZATION METHOD TO EXTRACT THE OPTIMAL SET OF LINGUISTIC SUMMARIES FROM NUMERIC DATA

Pham Dinh Phong[1*], Pham Thi Lan[2], Tran Xuan Thanh[3,4]
[1]*University of Transport and Comminications,* [2]*Hanoi National University of Education*
[3]*East Asia University of Technology,* [4]*Graduate University of Science and Technology, VAST*

ARTICLE INFO

Received: 01/3/2024
Revised: 28/3/2024
Published: 29/3/2024

KEYWORDS

Linguistic summary
Hegde algebras
Interpretability
Multi-semantic structure
Particle swarm optimization

ABSTRACT

Extracting a set of linguistic summaries from numeric data aims to produce summary sentences expressed in natural language that describe the hidden knowledge in the numeric dataset. A number of genetic algorithm models have been proposed to extract the optimal set of linguistic summaries, in which the algorithm model for extracting the set of linguistic summaries ensures the interpretability of the content of the summary sentences by applying genetic algorithm with greedy strategy gives quite good results. However, the determination of fuzziness parameter values of the algorithm model depends on the expert's intuition. In this paper, a method to optimize the fuzziness parameter values to improve the quality of the set of linguistic summaries extracted from numeric data is proposed. Experimental results with the creep database show that with the optimized fuzziness parameter values, the quality of the extracted set of linguistic summaries is better on three measures: fitness function value, average truth value and number of sentences with linguistic quantifier greater than *a half*.

MỘT PHƯƠNG PHÁP TỐI ƯU THAM SỐ TÍNH MỜ TRÍCH RÚT TẬP CÂU TÓM TẮT TỐI ƯU TỪ DỮ LIỆU SỐ

Phạm Đình Phong[1*], Phạm Thị Lan[2], Trần Xuân Thanh[3,4]
[1]*Trường Đại học Giao thông vận tải,* [2]*Trường Đại học Sư phạm Hà Nội*
[3]*Trường Đại học Công nghệ Đông Á,* [4]*Học viện Khoa học và Công nghệ - Viện Hàn lâm Khoa học và Công nghệ Việt Nam*

THÔNG TIN BÀI BÁO

Ngày nhận bài: 01/3/2024
Ngày hoàn thiện: 28/3/2024
Ngày đăng: 29/3/2024

TỪ KHÓA

Tóm tắt ngôn ngữ
Đại số gia tử
Tính giải nghĩa được
Cấu trúc đa ngữ nghĩa
Tối ưu bầy đàn

TÓM TẮT

Trích rút tập câu tóm tắt bằng ngôn ngữ từ dữ liệu số giúp đưa ra các câu tóm tắt được diễn đạt bằng ngôn ngữ tự nhiên mô tả tri thức ẩn dấu trong tập dữ liệu số. Một số mô hình thuật toán di truyền được đề xuất nhằm trích rút tập câu tóm tắt tối ưu, trong đó, mô hình thuật toán trích rút tập câu tóm tắt đảm bảo tính giải nghĩa nội dung các câu tóm tắt trên cơ sở kết hợp thuật toán di truyền với chiến lược tham lam cho kết quả khá tốt. Tuy nhiên, việc xác định các tham số tính mờ của mô hình thuật toán phụ thuộc vào cảm nhận trực giác của chuyên gia. Trong bài báo này, chúng tôi đề xuất một thuật toán tối ưu các tham số tính mờ nhằm nâng cao chất lượng tập câu tóm tắt được trích xuất từ dữ liệu số. Kết quả thực nghiệm với cơ sở dữ liệu creep cho thấy, với bộ tham số tính mờ được tối ưu, chất lượng của tập câu tóm tắt được trích rút tốt hơn trên ba độ đo là giá trị hàm thích nghi, giá trị chân lý trung bình và số câu có từ lượng hóa lớn hơn *a half*.

DOI: https://doi.org/10.34238/tnu-jst.9824

* Corresponding author. *Email: phongpd@utc.edu.vn*

1. Giới thiệu

Ngày nay, dữ liệu nghiệp vụ trong các lĩnh vực của đời sống xã hội đang gia tăng nhanh chóng. Do đó, nhu cầu trích rút thông tin có ích ẩn chứa trong dữ liệu phục vụ công tác ra quyết định là vô cùng cấp thiết đòi hỏi các nhà nghiên cứu đề xuất các phương pháp khai phá dữ liệu một cách hiệu quả. Trong các phương pháp đó, trích rút tóm tắt dữ liệu dưới dạng các câu trong ngôn ngữ tự nhiên theo một cấu trúc cho trước, được gọi tắt là tóm tắt bằng ngôn ngữ (linguistic summary - LS), là phương pháp khai phá dữ liệu có ý nghĩa ứng dụng thực tế. Mỗi LS mô tả tri thức về các đối tượng trong thế giới thực được lưu trữ dưới dạng dữ liệu số trong tập dữ liệu. Tri thức được diễn đạt dưới dạng ngôn ngữ tự nhiên giúp người dùng dễ hiểu hơn so với những con số. Cấu trúc của LS được sử dụng trong nghiên cứu này là câu có từ lượng hóa của Yager [1] có dạng: "*Q y* are *S*" hoặc "*Q F y are S*" [1] - [11]. Ví dụ như "*Very few* (*Q*) sales of printers (*y*) is with *high* commission (*S*)" [7], "*Most* (*Q*) hospitals (*y*) with *very high* average hospital stay (*F*) have *very low* computer (*S*)" [5]. Người dùng đọc các câu tóm tắt để hiểu thông tin, tri thức trong tập dữ liệu thông qua ngữ nghĩa của các từ *'very few', 'most', 'high', 'very low', 'very high'* trong câu tóm tắt. Từ lượng hóa *Q* biểu diễn một tỷ lệ thỏa kết luận *S* so với tất cả đối tượng trong tập dữ liệu trong mẫu câu thứ nhất hoặc các đối tượng trong nhóm thỏa điều kiện lọc *F* trong mẫu câu thứ hai.

Theo tiếp cận lý thuyết tập mờ, độ đúng đắn của mỗi câu tóm tắt được tính toán dựa trên giá trị của hàm thuộc của các tập mờ biểu diễn ngữ nghĩa của các từ ngôn ngữ trong câu như *'very few', 'most', 'high', 'very low', 'very high'* trong các ví dụ trên. Kết quả của thuật toán trích rút tóm tắt từ một tập dữ liệu cụ thể là tập các câu tóm tắt có độ đúng đắn (*T*) lớn hơn một ngưỡng cho trước. Khi cả ba thành phần *Q, F* và *S* hoàn toàn chưa được xác định về thuộc tính cũng như từ ngôn ngữ thì số lượng câu tóm tắt được trích rút là rất lớn, đặt ra thách thức lớn về khối lượng tính toán. Tuy nhiên, đó là mức độ tổng quát nhất nên người dùng có thể phát hiện được những tri thức hữu ích và thú vị chưa được khai khá trong tập dữ liệu. Trong thực tế, người dùng không thể đọc hết một số lượng khổng lồ các câu tóm tắt được trích rút mà chỉ cần đọc một số câu hữu ích nào đó. Do đó, các nghiên cứu trong [5], [12] – [14] đã ứng dụng thuật toán di truyền để trích rút một tập câu tóm tắt tối ưu dựa trên các điều kiện ràng buộc và hàm đánh giá chất lượng cho tập câu tóm tắt.

Các mô hình thuật toán di truyền trích rút tập câu tóm tắt tối ưu được đề xuất trong [13] và [14] chưa loại bỏ hết câu có độ đúng đắn $T = 0$ và vẫn còn ba câu có độ đúng đắn $T < 0,8$. Kết quả này có thể do tập từ lượng hóa được sử dụng chỉ có năm từ ngôn ngữ '*none', 'few', 'half', 'much', 'most'* là quá ít nên không mô tả đầy đủ các phần tử dữ liệu.

Để khắc phục hạn chế của tiếp cận lý thuyết tập mờ với các tập mờ được thiết kế theo cảm nhận trực giác của các chuyên gia và các mô hình thuật toán di truyền trong [13] và [14], Phạm Thị Lan và các cộng sự đã đề xuất mô hình thuật toán di truyền kết hợp chiến lược tham lam trích rút tập câu tóm tắt [15]. Trong mô hình được đề xuất, các tác giả đã ứng dụng đại số gia tử (ĐSGT) mở rộng [16] để sinh cấu trúc đa ngữ nghĩa cho các từ ngôn ngữ của các biến ngôn ngữ đảm bảo tính chung riêng của tập từ ngôn ngữ giúp tăng cơ hội thu được các câu tóm tắt có giá trị đúng đắn gần với 1 [17]. Sau đó, thuật toán chỉ sinh ngẫu nhiên thành phần lọc *F* và cấu trúc của thành phần kết luận *S*. Các từ ngôn ngữ trong các thành phần *S* và *Q* được xác định theo chiến lược tham lam với giá trị đúng đắn *T* và thứ tự ngữ nghĩa của *Q* càng lớn càng tốt. Tuy nhiên, tập giá trị của các tham số tính mờ được sử dụng để sinh các phân hoạch mờ trên miền giá trị của các thuộc tính của tập dữ liệu được xác định dựa trên kinh nghiệm của các chuyên gia nên có thể chưa đủ tốt dẫn đến tập câu tóm tắt được trích rút chưa tối ưu. Trong bài báo này, chúng tôi đề xuất một phương pháp tối ưu tập giá trị của các tham số tính mờ của ĐSGT nhằm nâng cao chất lượng tập câu tóm tắt được trích rút từ tập dữ liệu số, trong đó thuật toán tối ưu bầy đàn (PSO) [18] kết hợp với thuật toán di truyền và chiến lược tham lam để tối ưu đồng thời tập giá trị của

các tham số tính mờ và trích xuất tập câu tối ưu. Kết quả thực nghiệm với cơ sở dữ liệu creep đã chứng tỏ tính hiệu quả của phương pháp tối ưu được đề xuất.

2. Phương pháp nghiên cứu

2.1. Trích rút tập câu tóm tắt từ dữ liệu

Để tóm tắt tập dữ liệu số bằng các câu trong ngôn ngữ tự nhiên, Yager [1] đã đề xuất cấu trúc câu được trích xuất dưới dạng mệnh đề mờ có từ lượng hóa. Bài toán trích rút tập câu tóm tắt từ tập dữ liệu số được phát biểu như sau:

Cho $Y = \{y_1, y_2, \ldots, y_n\}$ là tập các đối tượng (bản ghi) trong cơ sở dữ liệu như tập các khách hàng của một ngân hàng; $A = \{A_1, A_2, \ldots, A_m\}$ là tập các thuộc tính cần xem xét của các đối tượng trong tập Y như AGE, SALARY, MARITAL,… Ký hiệu $A_i(y_j)$ là giá trị thuộc tính A_i của đối tượng y_j. Cơ sở dữ liệu được cho bởi tập D = $\{\{A_1(y_1), A_2(y_1), \ldots, A_m(y_1)\}, \ldots, \{A_1(y_n), A_2(y_n), \ldots, A_m(y_n)\}\}$ là đầu vào của bài toán trích rút tóm tắt bằng ngôn ngữ. Đầu ra của bài toán là tập câu tóm tắt bằng ngôn ngữ chứa từ lượng hóa có một trong hai dạng cấu trúc tổng quát sau:

$$Q\ y \text{ are } S \tag{1}$$

$$Q\ F\ y \text{ are } S \tag{2}$$

trong đó, S là thành phần *Kết luận* của câu tóm tắt được diễn đạt bằng một từ trong miền giá trị của biến ngôn ngữ. Q là *Từ lượng hóa* với ngữ nghĩa thể hiện tỷ lệ các đối tượng thỏa *Kết luận* S trong toàn bộ tập dữ liệu D như trong câu dạng (1) hoặc trong nhóm đối tượng thỏa điều kiện lọc F như trong câu dạng (2). Điều kiện lọc F là tùy chọn để xác định một nhóm các đối tượng trong tập đối tượng Y được xem xét trong câu tóm tắt. Ví dụ, một điều kiện lọc mờ có dạng như AGE = '*young*' tức là chỉ xét các đối tượng trong nhóm tuổi '*young*'. Giá trị đúng đắn T là một giá trị trong khoảng [0, 1] đánh giá mức độ đúng đắn của câu tóm tắt. Giá trị T được coi là giá trị chân lý của mệnh đề mờ có từ lượng hóa và được tính theo một trong hai công thức sau [14], [15], [17]:

$$T(Q \text{ y are } S) = \mu_Q\left[\frac{1}{n}\sum_{i=1}^{n}\mu_S\left(y_i\right)\right] \tag{3}$$

$$T(Q\ F \text{ y are } S) = \mu_Q\left[\frac{\sum_{i=1}^{n}\left(\mu_F\left(y_i\right)\wedge\mu_S\left(y_i\right)\right)}{\sum_{i=1}^{n}\mu_F\left(y_i\right)}\right] \tag{4}$$

trong đó, μ_Q, μ_F và μ_S tương ứng là giá trị hàm thuộc của các tập mờ biểu diễn ngữ nghĩa của các từ ngôn ngữ trong các thành phần Q, F và S.

Để đảm bảo chất lượng của tập câu tóm tắt được trích rút, chỉ có các câu có giá trị chân lý T lớn hơn một ngưỡng δ cho trước (chẳng hạn $\delta = 0{,}8$ [14]) mới được đưa vào tập câu. Ngoài ra, một số độ đo khác như độ đo tính mờ (imprecision), độ đo bao phủ (covering), độ đo tập trung (focus), độ đo sự phù hợp (appropriateness) [11], [14] cũng được sử dụng để đánh giá độ tốt của câu tóm tắt.

Mặc dù đã đặt ngưỡng δ để giới hạn số câu tóm tắt nhưng số lượng câu tóm tắt được trích rút vẫn rất lớn. Do đó, Donis-Diaz và cộng sự trong [13], [14] đã ứng dụng thuật toán di truyền để trích rút tập câu tóm tắt tối ưu dựa trên độ tốt (goodness) và độ đa dạng (deversity).

Trong [14], độ tốt Gn của một câu tóm tắt được đánh giá theo công thức (5), trong đó $St(Q)$ là trọng số của từ lượng hóa Q được chọn trước dựa trên mức độ ưu tiên của các từ lượng hóa. Trong [13], [14], các giá trị của $St(Q)$ là $St(Most) = 1, St(Much) = 0{,}75, St(Half) = 0{,}20, St(Some) = 0{,}15, St(Few) = 0{,}05$. Trong cả hai nghiên cứu này, độ tốt Gd của một tập câu tóm tắt là trung bình cộng độ tốt của các câu tóm tắt trong tập câu theo công thức (6) với l là số câu tóm tắt trong tập câu.

$$Gn = T \cdot St(Q) \tag{5}$$

$$Gd = \frac{\sum_{i=1}^{l} Gn_i}{l} \quad (6)$$

Trong [13], [14], độ đa dạng của một tập câu tóm tắt được xác định bằng công thức (7) sau:

$$De = \frac{C}{l} \quad (7)$$

trong đó, l là số câu trong tập câu tóm tắt và C là số cụm khi thực hiện phân cụm tập câu tóm tắt được xác định dựa trên hàm tính độ tương tự L như sau:

$$L(p1, p2) = \begin{cases} Yes & if \ \sum_{k=0}^{m} H(p1_k, p2_k) < 2 \\ No & trong \ trường \ hợp \ khác \end{cases} \quad (8)$$

Trong công thức (8), nếu hàm $L(p1, p2)$ có giá trị là '*Yes*' thì hai câu tóm tắt $p1$ và $p2$ là tương tự nhau. $p1$ và $p2$ là hai véctơ có $m + 1$ thành phần, trong đó thành phần thứ $p1_0$ và $p2_0$ là chỉ số của từ lượng hóa Q trong $Dom(Q)$ và các thành phần $p1_i$ và $p2_i$ lần lượt là chỉ số của từ ngôn ngữ trong miền giá trị ngôn ngữ $Dom(A_i)$ của biến ngôn ngữ ứng với thuộc tính A_i của vectơ biểu diễn câu tóm tắt $p1$ và $p2$. Nếu thuộc tính A_i không có trong câu tóm tắt thì thành phần thứ i trong vectơ biểu diễn câu tóm tắt nhận giá trị 0. Hàm $H(p1_k, p2_k)$ được tính theo công thức (9) để so sánh thành phần thứ k trong hai vectơ có khác biệt nhau không.

$$H(p1_k, p2_k) = \begin{cases} 1 & if \ \ |p1_k - p2_k| > round\left(20\% * size\left(Dom(A_k)\right)\right) or \\ & if \quad p1_k = 0 \ and \ p2_k \neq 0 \quad or \\ & if \quad p1_k \neq 0 \ and \ p2_k = 0 \\ 0 & trong \ trường \ hợp \ khác \end{cases} \quad (9)$$

Dựa trên độ tốt Gd và độ đa dạng De của tập câu, hàm thích nghi Fit cho một cá thể biểu diễn một tập câu tóm tắt được tính theo công thức sau:

$$Fit = m_g Gd + m_d De \quad (10)$$

trong đó, m_g và m_d lần lượt là trọng số của ứng với Gd và De thỏa điều kiện $m_g + m_d = 1$. Các tác giả trong [14] đã chọn $m_g = 0{,}7$ và $m_d = 0{,}3$.

2.2. Trích rút tập câu tóm tắt tối ưu đảm bảo tính giải nghĩa nội dung của câu tóm tắt

Với các đề xuất trong [13], [14], các phân hoạch mờ ứng với các biến ngôn ngữ được xây dựng theo kinh nghiệm của các chuyên gia và số từ ngôn ngữ ứng với mỗi biến bị giới hạn bởi 7 ± 2, là lượng thông tin mà bộ não con người có thể xử lý tại một thời điểm. Do chưa có cầu nối hình thức giữa ngữ nghĩa của các từ ngôn ngữ và các tập mờ tương ứng của chúng nên các từ ngôn ngữ chỉ là các nhãn ngôn ngữ được gán cho các tập mờ và không đảm bảo người sử dụng giải nghĩa một cách đúng đắn nội dung của câu tóm tắt nhận được từ thuật toán trích rút tóm tắt.

Nhằm khắc phục các hạn chế của tiếp cận trong [13], [14], trong nghiên cứu [17], các tác giả đã đề xuất một phương pháp xây dựng các phân hoạch mờ dưới dạng cấu trúc đa ngữ nghĩa có khả năng mở rộng và đảm bảo hai quan hệ dựa trên ngữ nghĩa vốn có của các từ ngôn ngữ là quan hệ thứ tự và quan hệ chung - riêng (khái quát - cụ thể) của các từ ngôn ngữ dựa trên phương pháp hình thức hóa của lý thuyết đại số gia tử. Các quan hệ này được bảo toàn khi ánh xạ từ tập các từ ngôn ngữ sang tập các tập mờ biểu diễn ngữ nghĩa của chúng. Khi người dùng muốn mở rộng số từ ngôn ngữ được sử dụng để xây dựng phân hoạch mờ, họ chỉ cần bổ sung các từ ngôn ngữ có tính riêng hơn ở mức $k + 1$ trong khi không làm thay đổi ngữ nghĩa của các từ ngôn ngữ đang được sử dụng. Hình 1 thể hiện cấu trúc đa ngữ nghĩa dựa trên tập mờ hình thang ba mức (từ mức 1 đến mức 3). Tiếp cận này đảm bảo tính giải nghĩa nội dung của các câu tóm tắt được trích rút. Để hạn chế số câu tóm tắt có có giá trị đúng đắn $T = 0$ và với mẫu câu tóm tắt có cấu trúc như trong [17] như sau:

"*Q*os are $o(E_s)$," and "*Q*os that are $o(F_q)$ is $o(E_s)$" (11)

trong đó, $o(E_s)$ là thành phần kết luận, $o(F_q)$ là thành phần lọc, Thuật toán sinh câu tóm tắt theo chiến lược tham lam **Random-Greedy-LS** được đề xuất trong [15] được tóm tắt như sau:

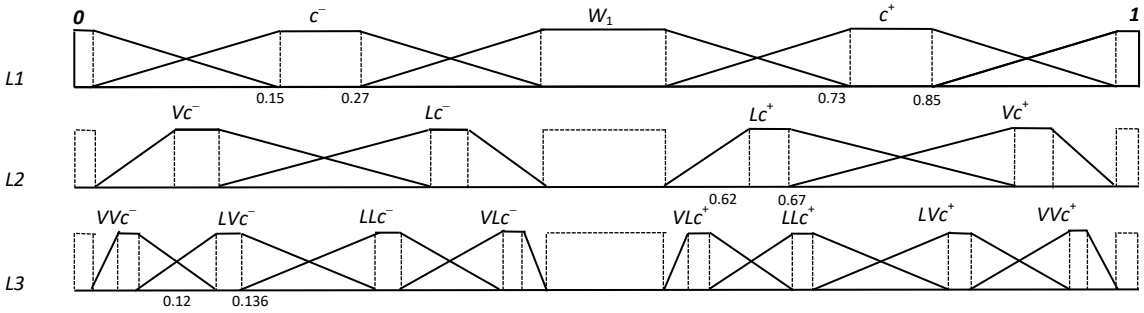
- *Bước 1*: Sinh ngẫu nhiên thành phần lọc $o(F_q)$ bao gồm cả thuộc tính và từ ngôn ngữ tương ứng. Tính độ hỗ trợ cho $o(F_q)$ theo công thức $supp(o(F_q)) = \sum_{i=1}^{n} \mu_{F_q}(y_i) / n$, trong đó n là số bản

ghi trong cơ sở dữ liệu. Nếu $supp(o(F_q)) > \beta$ cho trước thì thành phần $o(F_q)$ này được chấp nhận và chuyển sang bước 2, ngược lại thì sinh ngẫu nhiên thành phần lọc $o(F_q)$ khác.

- *Bước 2*: Chọn ngẫu nhiên thuộc tính trong thành phần $o(E_s)$ theo số lượng đã cho, duyệt các tổ hợp từ ngôn ngữ được sử dụng của các thuộc tính trong $o(E_s)$ để tìm một một tổ hợp từ ngôn ngữ mà biểu thức $r = \frac{\sum \mu_{o(Fq)} \wedge \mu_{o(Es)}}{\sum \mu_{o(Fq)}}$ đạt giá trị lớn nhất.

- *Bước 3:* Chọn một từ lượng hóa Q^* trong tập từ ngôn ngữ được sử dụng sao cho giá trị $T = \mu_{Q^*}(r)$ đạt giá trị lớn nhất. Nếu có nhiều từ ngôn ngữ Q^* để T đạt giá trị lớn nhất thì chọn từ ngôn ngữ Q^* có thứ tự ngữ nghĩa lớn nhất.

Kết quả thực hiện của thuật toán tham lam trên là một câu tóm tắt hướng đến độ đo tốt *Gn* lớn trong các câu tóm tắt cùng thành phần $o(F_q)$ và cùng cấu trúc $o(E_s)$.



Hình 1. *Cấu trúc đa ngữ nghĩa hình thang của các từ ngôn ngữ 3 mức* [17]

Để trích rút được tập câu tóm tắt có độ tốt và độ đa dạng càng cao càng tốt, các tác giả trong [15] đã đề xuất một thuật toán di truyền kết hợp chiến lược tham lam được trình bày ở trên tìm kiếm tập câu tóm tắt tối ưu. Thủ tục di truyền **Greedy-GA** kết hợp chiến lược tham lam **Random-Greedy-LS** được đề xuất trong [15] được thực hiện thông qua ba phép toán di truyền như sau: Toán tử chọn lọc lựa chọn một tỷ lệ nhất định các cá thể tốt nhất (dựa vào giá trị hàm thích nghi *Fit* theo công thức (10)) sang thế hệ tiếp theo, *Toán tử lai ghép* thực hiện hoán đổi các câu giữa hai cá thể, Toán tử đột biến thực hiện thay thế một số câu tóm tắt của một cá thể bằng các câu tóm tắt mới được sinh ra từ thủ tục **Random-Greedy-LS**. Như vậy, *Toán tử lai ghép* không làm thay đổi độ tốt của từng câu tóm tắt nhưng có thay đổi độ tốt của cả tập câu tóm tắt và *Toán tử đột biến* làm thay đổi cả độ tốt *Gd* và độ phong phú *De* của tập câu tóm tắt. Kết quả thực nghiệm cho thấy, so với các phương pháp **Hybrid-GA** được đề xuất trong [14] thì phương pháp **Greedy-GA** được đề xuất trong [15] có giá trị hàm mục tiêu lớn hơn, có số lượng câu có từ lượng hóa có thứ tự ngữ nghĩa lớn hơn 'a half' nhiều hơn, có số lượng câu có giá trị chân lý $T >$ 0,8 đạt tối đa 30 câu và không có câu tóm tắt nào có giá trị chân lý $T = 0$.

2.3. Phương pháp tối ưu tham số tính mờ trích rút tập câu tóm tắt tối ưu được đề xuất

Tuy nhiên, kết quả của phương pháp **Greedy-GA** [15] cho kết quả tốt hơn so với phương pháp **Hybrid-GA** [14] nhưng giá trị của các tham số tính mờ của đại số gia tử được sử dụng để sinh các phân hoạch mờ trên miền giá trị của các thuộc tính của tập dữ liệu được xác định trước dựa trên cảm nhận trực giác của các chuyên gia nên có thể chưa tối ưu do nhận thức của họ về dữ liệu có thể chưa đầy đủ. Trong nghiên cứu này, chúng tôi đề xuất một phương pháp tối ưu tập giá trị của các tham số tính mờ trên cơ sở của sự tương tác giữa ngữ nghĩa của các từ ngôn ngữ với tập dữ liệu, nhờ đó nâng cao chất lượng của tập câu tóm tắt được trích rút. Cụ thể, thuật toán tối ưu bầy đàn (PSO) [18] được sử dụng để hiệu chỉnh tối ưu tập giá trị của các tham số tính mờ trong miền ràng buộc do thuật toán này chỉ lựa chọn các giá trị trong miền tham chiếu theo xác suất mà không cần đến các phép toán di truyền. Do ứng dụng ĐSGT mở rộng [16] nên mỗi phần

tử thứ i ở thế hệ thứ t của bầy đàn biểu diễn một tập giá trị các tham số tính mờ $X_i^t = \{m(0), m(c^-), m(W), m(1), \mu(Little), \mu(h_0)\}$, trong đó $m(0)$, $m(c^-)$, $m(W)$, $m(1)$, $\mu(L)$ và $\mu(h_0)$ lần lượt là độ đo tính mờ của hằng tử nhỏ nhất 0, phần tử sinh âm c^-, phần tử trung hòa W, hằng tử lớn nhất 1, gia tử âm *Little* (L) và gia tử nhân tạo h_0. Khi đó, độ đo tính mờ của phần tử sinh dương c^+ được tính theo công thức $m(c^+) = 1 - m(0) - m(c^-) - m(W) - m(1)$ và của gia tử dương *Very* (V) được tính theo công thức $\mu(Very) = 1 - \mu(Little) - \mu(h_0)$ vì chỉ sử dụng hai gia tử. Với mỗi tập giá trị của các tham số tính mờ X_i^t được chọn sẽ được sử dụng làm đầu vào của thuật toán **Greedy-GA**. Dựa trên các giá trị cụ thể của X_i^t, các phân hoạch mờ đa ngữ nghĩa dựa trên tập mờ hình thang tương ứng được sinh ra một cách tự động phục vụ cho quá trình lập luận. Trong quá trình lập luận, ngữ nghĩa tính toán dựa trên tập mờ hình thang được tương tác với dữ liệu để tính toán các tham số của mô hình thuật toán. Đầu ra của thuật toán **Greedy-GA** là tập câu tóm tắt có giá trị hàm thích nghi *Fit* lớn nhất sẽ được so sánh với giá trị của biến *bestGlobalFit* lưu giá trị hàm thích nghi toàn cục tốt nhất của PSO để cập nhật giá trị tốt nhất cho biến này. Như vậy, tập giá trị của các tham số tính mờ được hiệu chỉnh thích nghi qua các lần lặp của thuật toán PSO với hàm mục tiêu là giá trị *Fit* thu được từ thuật toán **Greedy-GA**. Mô hình thuật toán được đề xuất như sau:

Procedure FPO_GreadyGA.
Đầu vào: Tập dữ liệu D;
Các biến: G_{max}, N, ngữ nghĩa cú pháp của ĐSGT; //G_{max}: số thế hệ, N: Kích thước quần thể
Đầu ra: Bộ giá trị tham số tính mờ tối ưu *bestGlobalFit*;
Begin
Sinh ngẫu nhiên thế hệ ban đầu $S_0 = \{X_i^0 \mid i = 1, \ldots, N\}$ với $X_i^0 = \{m(c^-), \mu(Little)\}$;
Khởi tạo giá trị 0 cho PG^t và P_i^t lưu vị trí tốt nhất toàn cục và cá nhân;
bestGlobalFit = 0;
For each phần tử x_i **do begin**
F_i^0 = **Greedy-GA**; //Gọi thuật toán di truyền kết hợp tham lam
If F_i^0 > *bestGlobalFit* **then begin**
bestGlobalFit = F_i^0; $PG^0 = X_i^0$;
End;
Cập nhật vị trí cá nhân tốt nhất P_i^0 của phần tử x_i dựa trên F_i^0;
End;
t = 1;
Repeat
For each phần tử x_i **do begin**
Cập nhật tốc độ V_i^t của phần tử x_i;
Cập nhật vị trí mới X_i^t của phần tử x_i;
F_i^t = **Greedy-GA**; //Gọi thuật toán di truyền kết hợp tham lam
If F_i^t tốt hơn F_i^{t-1} **then begin**
Cập nhật vị trí cá nhân tốt nhất P_i^t của phần tử x_i dựa trên F_i^t;
If F_i^t > *bestGlobalFit* **then begin**
bestGlobalFit = F_i^t; $PG^t = X_i^t$;
End;
End;
End;
$t = t + 1$;
Until $t = G_{max}$;
Return *bestGlobalFit* và $PG^{G_{max}-1}$;
End.

Đầu ra của thuật toán **FPO_GreadyGA** trên là tập giá trị của các tham số tính mờ trong biến $PG^{G_{max}-1}$ ứng với giá trị hàm thích nghi tốt nhất *bestGlobalFit* ở thế hệ cuối cùng. Trong cài đặt thực tế, biến *bestGlobalFit* có thể là một biến đối tượng chứa tập câu tóm tắt tối ưu.

3. Kết quả và bàn luận

3.1. Cài đặt thực nghiệm

Các thực nghiệm trong bài báo này được cài đặt bằng ngôn ngữ lập trình C#, chạy trên máy tính có cấu hình Intel Core i5-8250U 1.60GHz, 8GB RAM và hệ điều hành Windows 11 64 bit. Tập dữ liệu thực nghiệm là creep [14]. Các tham số của chương trình thực nghiệm như sau:

- Số lần lặp của thuật toán PSO là 5, số phần tử mỗi thế hệ là 10, hệ số Inertia là 0,7, hệ số nhận thức cá nhân và xã hội đều là 1,5.
- Số thế hệ của thuật toán di truyền là 100, số cá thể mỗi thế hệ là 20, tỷ lệ chọn lọc là 0,15, tỷ lệ lai ghép chéo là 0,8, tỷ lệ đột biến là 0,1.
- Số câu tóm tắt được trích rút trong mỗi tập câu là 30.
- Các ràng buộc về giá trị của các tham số tính mờ được thiết lập dựa trên kinh nghiệm trong [15] và qua quá trình thực nghiệm như sau:

+ Ràng buộc trên thuộc tính *creep* như sau: $0{,}001 \le m(0) \le 0{,}05;\ 0{,}2 \le m(c^-) \le 0{,}35;\ 0{,}02 \le m(W) \le 0{,}05; 0{,}35 \le m(1) \le 0{,}4; 0{,}2 \le \mu(L) \le 0{,}5;\ 0{,}05 \le \mu(h_0) \le 0{,}3.$

+ Ràng buộc cho tập từ lượng hóa Q: $0{,}001 \le m(0) \le 0{,}05;\ 0{,}2 \le m(c^-) \le 0{,}45;\ 0{,}001 \le m(W) \le 0{,}15; 0{,}002 \le m(1) \le 0{,}05; 0{,}2 \le \mu(L) \le 0{,}5;\ 0{,}05 \le \mu(h_0) \le 0{,}3.$

+ Ràng buộc trên các thuộc tính khác:

$0{,}001 \le m(0) \le 0{,}05;\ 0{,}2 \le m(c^-) \le 0{,}45;\ 0{,}002 \le m(W) \le 0{,}15; 0{,}001 \le m(1) \le 0{,}05; 0{,}2 \le \mu(L) \le 0{,}5;\ 0{,}05 \le \mu(h_0) \le 0{,}3.$

+ Mức đặc tả được thiết lập cho tất cả các thuộc tính và tập từ lượng khóa: $k = 3$.

3.2. Kết quả thực nghiệm và so sánh

Tập dữ liệu creep [14] được sử dụng để thực nghiệm. Mô hình tối ưu **FPO_GreadyGA** được đề xuất được chạy thực nghiệm 10 lần. Các kết quả của 10 lần chạy được tính trung bình và được thể hiện trong Bảng 1.

Bảng 1. *So sánh kết quả thực nghiệm của mô hình được đề xuất FPO_GreadyGA với các mô hình Greedy-GA trong* [15] *và Hybrid-GA trong* [14]

Mô hình	Giá trị hàm thích nghi *Fit*	Trung bình giá trị chân lý *T*	Số câu có Q > *a half*	Số câu có giá trị chân lý *T* > 0,8	Số câu có giá trị chân lý *T* = 0
Hybrid-GA [14]	0,6659	0,9139	17,8	27,0	1,0
Greedy-GA [15]	0,7905	0,9951	18,8	30,0	0,0
FPO_GreadyGA	0,8828	0,9970	21,9	30,0	0,0

Các kết quả trong Bảng 1 cho thấy, mô hình được đề xuất **FPO_GreadyGA** có các giá trị hàm thích nghi trung bình *Fit* là 0,8828, giá trị chân lý trung bình *T* là 0,9970 và số câu có $Q > a\ half$ là 21,9 đều lớn hơn các mô hình **Greedy-GA** trong [15] và **Hybrid-GA** trong [14]. Bên cạnh đó, mô hình được đề xuất có Số câu có giá trị chân lý $T > 0{,}8$ đạt giá trị tối đa là 30 câu và không có câu nào có giá trị chân lý $T = 0$, tương đương với kết quả của mô hình **Greedy-GA**. Ngoài ra, quan sát số liệu qua các thế hệ của thuật toán được đề xuất ta thấy rằng, giá trị hàm thích nghi *Fit* tốt lên chủ yếu do tăng độ tốt *Gd* trong khi độ đa dạng *De* của tập câu tăng khá ít. Nguyên nhân *De* không tăng nhiều là do các phân hoạch mờ đều được thiết lập với mức đặc tả (tính riêng) của các từ ngôn ngữ là $k = 3$.

Với kết quả so sánh trên, có thể kết luận rằng phương pháp được đề xuất cho kết quả tốt hơn so với hai phương pháp được đối sánh, đặc biệt là tốt hơn phương pháp **Greedy-GA** với cùng phương pháp luận dựa trên ĐSGT do giá trị của các tham số tính mờ được tối ưu.

4. Kết luận

Bài báo đề xuất một phương pháp tối ưu giá trị của các tham số tính mờ nhằm nâng cao chất lượng của tập câu tóm tắt được trích rút từ cơ sở dữ liệu. Cụ thể, thuật toán PSO được sử dụng để hiệu chỉnh tối ưu tập giá trị của các tham số tính mờ làm đầu vào cho thuật toán di truyền kết hợp chiến lược tham lam trích xuất tập câu tóm tắt có độ thích nghi tốt. Kết quả thực nghiệm với cơ sở dữ liệu creep cho thấy tính hiệu quả của phương pháp tối ưu được đề xuất so với các phương pháp được so sánh do bộ tham số ngữ nghĩa được tối ưu là kết quả của sự tương tác giữa ngữ nghĩa của các từ ngôn ngữ với dữ liệu.

Lời cám ơn

Nghiên cứu này được tài trợ bởi Trường Đại học Giao thông vận tải (ĐH GTVT) trong đề tài mã số T2023-CN-008TĐ.

TÀI LIỆU THAM KHẢO/ REFERENCES

[1] R. R. Yager, "A new approach to the summarization of data," *Information Sciences*, vol. 28, no. 1, pp. 69-86, 1982.

[2] J. Kacprzyk, R. R. Yager, and S. Zadrożny, "A fuzzy logic based approach to linguistic summaries of databases," *International Journal of Applied Mathematics and Computer Science*, vol. 10, no. 4, pp. 813-834, 2000.

[3] J. Kacprzyk and S. Zadrożny, "Linguistic database summaries and their protoforms: towards natural language based knowledge discovery tools," *Information Sciences*, vol. 173, no. 4, pp. 281-304, 2005.

[4] C. A. D. Díaz, R. B. Pérez, and E. V. Morales, "Using Linguistic Data Summarization in the study of creep data for the design of new steels," in *11th International Conference on Intelligent Systems Design and Applications (ISDA)*, 2011, pp. 160-165.

[5] T. Altintop, R. R. Yager, D. Akay, F. E. Boran, and M. Ünal, "Fuzzy Linguistic Summarization with Genetic Algorithm: An Application with Operational and Financial Healthcare Data," *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems*, vol. 25, no. 04, pp. 599-620, 2017.

[6] R. J. Almeida, M.-J. Lesot, B. Bouchon-Meunier, U. Kaymak, and G. Moyse, "Linguistic summaries of categorical time series for septic shock patient data," *Fuzz-IEEE 2013-IEEE International Conference on Fuzzy Systems*, Hyderabad, India. IEEE, Jul. 2013, pp.1-8.

[7] J. Kacprzyk and R. R. Yager, "Linguistic summaries of data using fuzzy logic," *International Journal of General System*, vol. 30, no. 2, pp. 133-154, 2001.

[8] M. D. Peláez-Aguilera, M. Espinilla, M. R. F. Olmo, and J. Medina, "Fuzzy linguistic protoforms to summarize heart rate streams of patients with ischemic heart disease," *Complexity*, vol. 2019, pp. 1-11, 2019.

[9] A. Duraj, P. S. Szczepaniak, and L. Chomatek, "Intelligent Detection of Information Outliers Using Linguistic Summaries with Non-monotonic Quantifiers," *International Conference on Information Processing and Management of Uncertainty in Knowledge-Based Systems*, 2020, pp. 787-799.

[10] A. Jain, M. Popescu, J. Keller, M. Rantz, and B. Markway, "Linguistic summarization of in-home sensor data," *Journal of biomedical informatics*, vol. 96, 2019, Art. no. 103240.

[11] A. Wilbik, I. Vanderfeesten, D. Bergmans, S. Heines, and W. van Mook, "Linguistic summaries for compliance analysis of a glucose management clinical protocol," *IEEE International Conference on Fuzzy Systems (FUZZ-IEEE)*, 2018, pp. 1-7.

[12] F. E. Boran and D. Akay, "A generic method for the evaluation of interval type-2 fuzzy linguistic summaries," *IEEE transactions on cybernetics*, vol. 44, no. 9, pp. 1632-1645, 2013.

[13] C. Donis-Diaz, R. Bello, and J. Kacprzyk, "Linguistic data summarization using an enhanced genetic algorithm," *Technical Transactions – Automatic Control*, vol. 2013, no. 2, pp. 3-12, 2013.

[14] C. Donis-Diaz, A. Muro, R. Bello-Pérez, and E. V. Morales, "A hybrid model of genetic algorithm with local search to discover linguistic data summaries from creep data," *Expert Systems with Applications*, vol. 41, no. 4, pp. 2035-2042, 2014.

[15] T. L. Pham, C. H. Nguyen, and D. P. Pham, "Extracting an optimal set of linguistic summaries using genetic algorithm combined with greedy strategy," *Journal on Information Technologies & Communications*, vol. 2020, no. 2, pp. 75-87, 2020.

[16] C. H. Nguyen, T. S. Tran, and D. P. Pham, "Modeling of a semantics core of linguistic terms based on an extension of hedge algebra semantics and its application," *Knowledge-Based Systems*, vol. 67, pp. 244-262, 2014.

[17] C. H. Nguyen, T. L. Pham, T. N. Nguyen, C. H. Ho, and T. A. Nguyen, "The linguistic summarization and the interpretability, scalability of fuzzy representations of multilevel semantic structures of word-domains," *Microprocessors and Microsystems*, vol. 81, 2021, Art. no. 103641.

[18] J. Kennedy and R. C. Eberhart, "Particle Swarm Optimization," *Proceedings of the IEEE International Conference on Neural Networks*, Piscataway, New Jersey. IEEE Service Center, 1995, pp. 1942-1948.