ỨNG DỤNG THUẬT TOÁN TỐI ƯU TIẾN HÓA BẦY ĐÀN MỜ TRONG PHÂN TÍCH NHU CẦU KHÁCH HÀNG

Nguyễn Thị Như Na[a*]

[a]*Khoa Tự nhiên, Trường Cao đẳng Sư phạm Điện Biên, Lai Châu, Việt Nam*

Lịch sử bài báo
Nhận ngày 31 tháng 03 năm 2017 | Chỉnh sửa ngày 21 tháng 04 năm 2017
Chấp nhận đăng ngày 19 tháng 05 năm 2017

Tóm tắt

Bài báo này ứng dụng thuật toán tối ưu tiến hóa bầy đàn mờ cho bài toán phân tích nhu cầu khách hàng. Đây là bài toán có ý nghĩa ứng dụng lớn trong hoạt động sản xuất kinh doanh. Áp dụng thuật toán tối ưu tiến hóa bầy đàn mờ vào bài toán cụ thể là một công ty chuyên cung cấp thiết bị y tế của Mỹ muốn phân tích nhu cầu 500 bệnh viện trong khu vực về các thiết bị và vật tư y tế, hỗ trợ công ty đưa ra chiến lược kinh doanh phù hợp nhất với từng bệnh viện để đạt doanh thu cao.

Từ khóa: Bầy đàn; Di truyền; Phân cụm; Tập mờ; Tiến hóa; Tối ưu.

1. ĐẶT VẤN ĐỀ

Phân tích nhu cầu khách hàng là một việc rất quan trọng trong kinh doanh, phân tích nhu cầu khách hàng, phân khúc thị trường, xác định nhu cầu biến động của khách hang... Các thông tin thu được sử dụng để hỗ trợ các doanh nghiệp đưa ra chiến lược kinh doanh hiệu quả. Cụ thể, bài toán thực tế phân tích nhu cầu khách hàng là một công ty chuyên cung cấp các thiết bị y tế cho 500 bệnh viện (http://www.rci.rutgers.edu/~cabrera/sc/cs8/cs8.html) ở Mỹ muốn tìm giải pháp để tăng doanh số bán hàng. Phân tích nhu cầu tiêu thụ các thiết bị y tế của các bệnh viện ở Mỹ tìm ra bệnh viện có mức tiêu thụ loại thiết bị y tế nào cao nhằm đưa ra chiến lược kinh doanh phù hợp với từng bệnh viện để tăng hiệu quả kinh doanh. Dữ liệu đầu vào bài toán là 500 bệnh viện ở Mỹ được tổ chức trong tập tin *Customer.xls* bao gồm 19 trường và 4000 bản ghi.

Dữ liệu đầu ra bài toán là phân tích 500 bệnh viện thành 3 nhóm với mức tiêu thụ

* Tác giả liên hệ: Email: nhuna.cdsp@gmail.com

thiết bị y tế khác nhau: Bệnh viện có mức tiêu thụ thiết bị y tế thấp; Bệnh viện có mức tiêu thụ thiết bị y tế trung bình; Bệnh viện có mức tiêu thụ thiết bị y tế cao.

19 trường trong tập tin bao gồm các mã như sau:

1. ZIP: Mã bưu điện
2. HID: ID bệnh viện
3. CITY: Tên thành phố
4. STATE: Tên tiểu bang
5. BEDS: Số giường bệnh
6. RBEDS: Số giường chỉnh hình
7. OUT-V: số lượt khám ngoại trú
8. ADM: Chi phí hành chính (1000 $/năm)
9. SIR: Thu từ nội trú
10. SALESY: Bán trang thiết bị phục hồi chức năng từ ngày 1 tháng 1
11. SALES12: Bán trang thiết bị phục hồi chức năng cuối tháng 12
12. HIP95: Số hoạt động cho hông trong năm 1995
13. KNEE95: Số hoạt động đầu gối trong năm 1995
14. TH: Có hoạt động dạy học không? 0, 1
15. TRAUMA: Có chấn thương không? 0, 1
16. REHAB: Có chỉnh hình không? 0, 1
17. HIP96: Số hoạt động cho hông cho năm 1996
18. KNEE96: Số hoạt động đầu gối cho năm 1996
19. FEMUR96: Số hoạt động cho xương đùi cho năm 1996.

Có nhiều cách tiếp cận để giải quyết bài toán phân tích nhu cầu khách hàng như là cách tiếp cận thống trị dựa trên tập rõ (DRSA) được phát triển bởi Greco, Matarazzo, và Slowinski (1998) và Greco, Matarazzo, và Slowinski (2000) rất hữu ích để giảm dữ

liệu trong phân tích định tính. Cách tiếp cận phân tích giỏ hàng (MBA) của Giudici và Passerone (2002) xác định mối liên hệ giữa khách hàng và các sản phẩm khác nhau trong một đơn vị đặc biệt như là bên trong một siêu thị. Các dữ liệu được phân tích trong MBA thường bao gồm tất cả các giao dịch mua hàng được thực hiện trong một khoảng thời gian nhất định để phân tích cấu trúc kết hợp giữa việc bán sản phẩm khác nhau có sẵn dựa vào đó doanh nghiệp lên kế hoạch cho các chính sách tiếp thị tốt hơn. Cách tiếp cận cho bài toán phân tích nhu cầu khách hàng mua bán trên Internet (Song, Kyeong, & Kim, 2001) sử dụng độ đo tượng tự và độ đo khác nhau cho các thay đổi của khách hàng, sau đó đánh giá độ đo thay đổi để đưa ra quy tắc thay đổi giúp các nhà quản lý đưa ra chiến lược kinh doanh phù hợp cho từng đối tượng khách hàng dựa trên số liệu kinh doanh cụ thể.

Phân cụm được coi như một công cụ độc lập để xem xét phân bố dữ liệu, làm bước tiền xử lý cho các thuật toán khác. Phân cụm ứng dụng rất nhiều trong các lĩnh vực như phân tích hình ảnh (Pappas, 1992), thông tin địa lý (Aksoy, 2006), khai phá web (Runkler & Bezdek, 2003) … Phương pháp phân cụm mờ là sự kết hợp của kỹ thuật phân cụm với lý thuyết mờ của Zadeh (1965) đang phát triển và được ứng dụng rộng rãi trong thực tiễn, ví dụ như phân tích rủi ro, dự báo nguy cơ phá sản cho ngân hàng và nhiều bài toán khác. Nhưng những vấn đề được quan tâm nhiều vẫn là nâng cao chất lượng phân cụm (Chen & Ludwig, 2014), tính toán thông qua một số độ đo chất lượng cụ thể… Và một số nghiên cứu ứng dụng thuật toán tối ưu tiến hóa phân cụm mờ như là nghiên cứu phân cụm mờ bằng PSO (Runkler & Katz, 2006) giải quyết bài toán phân cụm mờ bằng cực tiểu hóa mô hình *Fuzzy C – Means* (FCM), ứng dựng trên bộ dữ liệu bệnh ung thư phổi. Nghiên cứu về vấn đề phát hiện, định vị trực quan các tín hiệu không cố định của Biswal, Dash, và Panigrahi (2009). Các nghiên cứu dựa trên tối ưu tiến hóa mờ để giải bài toán phân tích nhu cầu khách hàng như trên là rất ít.

Bài báo này sẽ áp dụng thuật toán tối ưu tiến hóa mờ do Pang, Wang, Zhou, và Dong (2004) đề xuất cho bài toán phân cụm mờ để xác định nghiệm tối ưu toàn cục cho bài toán phân tích nhu cầu khách hàng. Với dữ liệu thực tế một công ty chuyên cung cấp các thiết bị y tế Mỹ muốn phân tích nhu cầu khách hàng là 500 bệnh viện trong khu vực để có kế hoạch và chiến lược kinh doanh phù hợp nhất cho từng đối tượng khách hàng, thích ứng với nhu cầu khách hàng mà đạt được doanh thu cao. Thuật toán tối ưu hóa bầy

đàn mờ (FPSO) (Pang và ctg., 2004) dựa trên thuật toán tối ưu bầy đàn cho bài toán người bán hàng. Thuật toán tối ưu bầy đàn (PSO) được Eberhart và Kennenday (1995) giới thiệu thuộc về lớp các bài toán tiến hóa, dựa trên khái niệm trí tuệ bầy đàn để giải bài toán tối ưu tiến hóa. PSO được áp dụng rộng rãi để cải tiến hiệu suất các thuật toán khác. Các ứng dụng như là bài toán lập kế hoạch (Weijun, Zhiming, Wei, & Genke, 2004), người bán hàng (Wang, Huang, Zhou, & Pang, 2003).

Bài báo này ứng dụng thuật toán tối ưu tiến hóa phân cụm mờ cho công ty ở Mỹ chuyên cung cấp thiết bị y tế cho 500 bệnh viện khu vực, muốn phân tích nhu cầu về vật tư và thiết bị y tế của các bệnh viện trên. Kết quả thực nghiệm thu được 3 cụm khách hàng với các mức tiêu thụ vật tư và thiết bị y tế khác nhau, dựa vào đó công ty có cơ sở khoa học đưa ra giải pháp chiến lược kinh doanh phù hợp nhất cho từng nhóm đối tượng và từng đối tượng khách hàng làm tăng hiệu quả hoạt động kinh doanh. Ngoài ra, ứng dụng của thuật toán này có thể mở rộng không chỉ cho công ty chuyên cung cấp các thiết bị y tế mà còn có thể ứng dụng cho các doanh nghiệp kinh doanh sản xuất các mặt hàng khác.

2. THUẬT TOÁN TỐI ƯU TIẾN HÓA CHO PHÂN CỤM MỜ

2.1. Phương pháp phân cụm mờ

Bài toán phân cụm N vector $X=\{x_1,x_2,...,x_N\}$ thành c cụm dựa trên tính toán tối thiểu hóa hàm mục tiêu để đo chất lượng của cụm và tìm tâm cụm sao cho hàm độ đo không tương tự là nhỏ nhất. Một phân cụm mờ vector $X=\{x_1,x_2,...,x_N\}$ được biểu diễn bởi ma trận $U=[U_{ki}]_{N\times c}$ sao cho một điểm dữ liệu có thể thuộc về nhiều nhóm và được xác định bằng giá trị hàm thuộc u. Ma trận giá trị hàm thuộc có dạng như sau:

$$U=\begin{bmatrix} u_{11} & \cdots & u_{1c} \\ \vdots & \ddots & \vdots \\ u_{N1} & \cdots & u_{Nc} \end{bmatrix} \quad (1)$$

Thuật toán phân cụm mờ đã được xuất phát từ việc cực tiểu giá trị hàm mục tiêu:

$$J_m = \sum_{k=1}^{N}\sum_{j=1}^{c} u_{kj}^{m} d(x_k, z_j) \quad (2)$$

Trong đó: $d(x_k, z_j)$ là một độ đo không tương tự.

Giải bài toán $J_m(u,z) \to \min$ với ràng buộc sau:

$$\begin{cases} 0 \le u_{kj} \le 1 \\ \sum_{j=1}^{c} u_{kj} = 1 \\ 0 \le \sum_{k=1}^{N} u_{kj} \le N \end{cases} \quad (3)$$

2.2. Áp dụng thuật toán tối ưu bầy đàn cho phân cụm mờ (FPSO)

Pang và ctg. (2004) đã đề xuất thuật toán tối ưu bầy đàn mờ. Trong thuật toán này vị trí và vận tốc của các cá thể xác định lại tương ứng với các biến mờ. Phương pháp này mô tả cho bài toán phân cụm mờ.

Với X là vị trí các cá thể, thể hiện mối quan hệ mờ của các đối tượng dữ liệu. Tâm cụm $Z = \{z_1, z_2, ..., z_c\}$, X được biểu diễn như sau:

$$X = \begin{bmatrix} \mu_{11} & \cdots & \mu_{1c} \\ \vdots & \ddots & \vdots \\ \mu_{N1} & \cdots & \mu_{Nc} \end{bmatrix} \quad (4)$$

Trong đó: μ_{ij} là một hàm thuộc của cá thể i thuộc cụm j:

$$\mu_{ij} \in [0,1], \forall i = 1,2,...,N; \forall j = 1,2,...,c$$
$$\sum_{j=1}^{c} \mu_{ij} = 1, \forall i = 1,2,...,N \quad (5)$$

Ma trận vị trí của mỗi cá thể giống ma trận mờ μ trong thuật toán FCM. Ngoài ra, vận tốc của mỗi cá thể là ma trận N dòng và c cột, các cá thể của ma trận trong phạm vi [-1, 1].

$$V = \begin{bmatrix} v_{11} & \cdots & v_{1c} \\ \vdots & \ddots & \vdots \\ v_{N1} & \cdots & v_{Nc} \end{bmatrix} \quad (6)$$

Ta có công thức (7) cập nhật vị trí và vận tốc của các cá thể:

$$V(t+1) = \mathrm{w}.\mathrm{V}(\mathrm{t}) + (\mathrm{c}_1 r_1)(pbest(t) - X(t)) + (\mathrm{c}_2 r_2)(gbest(t) - X(t))$$
$$\mathrm{X}(\mathrm{t}+1) = X(t) \oplus V(t+1) \quad (7)$$

Sau khi cập nhật ma trận vị trí của cá thể, có thể không thỏa mãn điều kiện như trong (8). Vì vậy phải chuẩn hóa ma trận vị trí.

$$\mu_{\mathrm{ij}} \in [0,1], \forall i = 1,2,...,N; \forall j = 1,2,...,\mathrm{c}$$
$$\sum_{j=1}^{c} \mu_{\mathrm{ij}} = 1, \forall i = 1,2,...,N \quad (8)$$

Để chuẩn hóa ta chuyển các số âm trong ma trận thành 0. Nếu các cá thể trên một hàng của ma trận là 0, chúng cần chuẩn hóa lại bằng cách lấy ngẫu nhiên trong [0,1], ta có ma trận được chuẩn hóa được biểu diễn như trong (9).

$$Xnormal = \begin{bmatrix} \mu_{11} \Big/ \sum_{j=1}^{c} \mu_{1j} & \cdots & \mu_{1c} \Big/ \sum_{j=1}^{c} \mu_{1j} \\ \vdots & \ddots & \vdots \\ \mu_{N1} \Big/ \sum_{j=1}^{c} \mu_{Nj} & \cdots & \mu_{Nc} \Big/ \sum_{j=1}^{c} \mu_{Nj} \end{bmatrix} \quad (9)$$

Thuật toán FPSO cũng tương tự như các thuật toán cải tiến khác, đều cần một hàm đánh giá kết quả tổng quát gọi là hàm độ đo thích nghi. Công thức (10) là hàm độ đo thích nghi sử dụng để đánh giá các kết quả.

$$f(X) = \frac{K}{J_m} \quad (10)$$

Trong đó K là hằng số; J_m là hàm mục tiêu $J_m(\mu, \mathrm{Z}) = \sum_{i=1}^{N} \sum_{j=1}^{c} \mu_{\mathrm{ij}}^m \left\| x_i - z_j \right\|^2$

Từ các công thức trên, ta có thuật toán FPSO như sau:

Bước 1. Khởi tạo tham số P, c_1, c_2, w và tham số đếm vòng lặp tối đa Maxinter.

Bước 2. Tạo P cá thể trong quần thể (X, *pbest*, *gbest* , V ma trận $N \times c$ cá thể).

Bước 3. Khởi tạo X, V, *pbest* cho mỗi cá thể và *gbest* cho quần thể.

Bước 4. Tính các tâm cụm cho mỗi cá thể bằng công thức (11):

$$z_j = \frac{\sum_{i=1}^{N} \mu_{ij}^{m} x_i}{\sum_{i=1}^{N} \mu_{ij}^{m}} \quad (11)$$

Bước 5. Tính giá trị hàm độ đo thích nghi cho mỗi cá thể bằng công thức (10).

Bước 6. Tính *pbest* cho mỗi cá thể.

Bước 7. Tính *gbest* cho quần thể.

Bước 8. Cập nhật ma trận vận tốc cho mỗi cá thể bằng công thức (9).

Bước 9. Cập nhật ma trận vị trí cho mỗi cá thể bằng công thức (7).

Bước 10. Nếu thỏa gặp điều kiện kết thúc, quay lại Bước 4.

Điều kiện kết thúc là số lần lặp tối đa hoặc giá trị *gbest* trong vòng lặp không cải thiện nữa.

2.3. Ưu điểm và nhược điểm của bài toán

Thuật toán FPSO là sự kết hợp ưu điểm của thuật toán FCM, dễ dàng giải quyết các bài toán tối ưu hàm mục tiêu khác nhau của thuật toán PSO để phân cụm trong môi trường mờ. Hiệu suất tốt hơn so với một số thuật toán phân cụm mờ như FCM. Sử dụng FPSO dễ dàng giải quyết các bài toán phân cụm mờ, tối ưu hóa tổ hợp khó giải quyết trong phạm vi lớn trong môi trường mờ (Mehdizadeh & Moghaddam, 2008). Tuy nhiên, nhược điểm là thuật toán FPSO hội tụ chậm hơn so với thuật toán phân cụm mờ FCM.