Kỷ yếu Hội nghị Quốc gia lần thứ X về Nghiên cứu cơ bản và ứng dụng Công nghệ thông tin (FAIR), Đà Nẵng, ngày 17-18/08/2017
DOI: 10.15625/vap.2017.00090

PHÂN TÍCH SỰ ẢNH HƯỞNG CỦA MỘT SỐ ĐỘ ĐO LIÊN KẾT ÁP DỤNG VÀO BÀI TOÁN DỰ ĐOÁN LIÊN KẾT TRONG MẠNG ĐỒNG TÁC GIẢ

Phạm Minh Chuẩn[1], Trịnh Khắc Linh[2], Trần Đình Khang[2,*], Lê Hoàng Sơn[3]

[1] Trường Đại học Sư phạm Kỹ thuật Hưng Yên
[2] Trường Đại học Bách khoa Hà Nội, [*]Corresponding author
[3]Trường Đại học Khoa học tự nhiên, Đại học Quốc gia Hà Nội

chuanpm@gmail.com, linhtk.dhbk@gmail.com, khangtd@soict.hust.edu.vn, sonlh@vnu.edu.vn

TÓM TẮT: *Trong nghiên cứu khoa học, việc công bố ra các bài báo thường có sự tham gia và đóng góp một nhóm tác giả. Cũng như trong các mạng xã hội nói chung, sự liên kết đồng tác giả phụ thuộc vào nhiều yếu tố: sự quen biết, phối hợp, đồng tác giả trong quá khứ, hoặc lĩnh vực chuyên môn ... Bài báo hướng tới phân tích một số độ đo liên kết để xem xét sự ảnh hưởng của các độ đo đó trong dự báo về khả năng đồng tác giả của các ứng viên.*

Từ khóa: *mạng đồng tác giả, độ đo liên kết, dự báo liên kết, phân lớp.*

I. ĐẶT VẤN ĐỀ

Mạng đồng tác giả hay còn được gọi là mạng lưới học thuật, bao gồm những tác giả đã từng viết một hoặc nhiều bài báo, ấn phẩm được công khai về chủ đề, lĩnh vực nào đó. Trong mạng này, đỉnh là các nhà nghiên cứu, học giả, chuyên gia, ... và các cạnh thể hiện sự hợp tác khoa học giữa các chuyên gia đó. Hai tác giả được gọi là "đồng tác giả" nếu họ viết chung một hoặc nhiều bài báo.

Một mạng đồng tác giả có thể được định nghĩa như sau : $G^T = (V^T, E^T, P^T, T)$ trong đó,

- $T = \{t_1, t_2, \dots, t_k\}$ là tập các nhãn thời gian
- $V^T = \{v_1, v_2, \dots, v_N\}$ là tập các đỉnh được tạo trong thời gian T. Các nút đại diện cho các tác giả trong bài báo.
- $P^T = \{p_1, p_2, \dots, p_M\}$ là tập các bài báo trong thời gian$T$
- $E^T = \{(v_i, v_j, p_k, t_h)\}$ là tập các liên kết giữa các tác giả của bài báo trong thời gian T

Dự đoán liên kết trong mạng đồng tác giả là bài toán đưa ra dự đoán các tác giả trong mạng có khả năng hợp tác trong tương lai. Mục tiêu của bài toán là gợi ý cho các tác giả, nhà nghiên cứu tìm được cộng tác phù hợp với mình sau này. Đây là một vấn đề được quan tâm nghiên cứu bởi ý nghĩa thực tiễn, thiết thực. Có nhiều công trình nghiên cứu liên quan như [10, 11, 12, 13].

Để dự đoán liên kết trong tương lai, người ta thường dựa vào các thông tin về các liên kết trong quá khứ, như là thông tin về các nút có liên kết với nút đang xét, hay còn gọi là hàng xóm của nút. Dựa vào đó, định nghĩa một số độ đo liên kết mạng, như độ đo hàng xóm chung, hệ số Jaccard ... Các độ đo này khi tính toán cho một cặp ứng viên (u, v) có thể cho phép xác định khả năng u và v sẽ có liên kết trong tương lai, thường được xem là tham số đầu vào cho bài toán dự báo liên kết mạng. Có nhiều nghiên cứu về các độ đo liên kết mạng như [6-8, 14-20].

Với mạng đồng tác giả, cũng có thể áp dụng các độ đo đó để dự đoán liên kết. Bài báo này sẽ khảo sát một số độ đo thông dụng của mạng nói chung, xem mức độ ảnh hưởng của các độ đo này tới hiệu quả dự báo, bằng cách gán cho các độ đo bộ trọng số và tìm cách tính toán bộ trọng số phù hợp. Các trọng số này thể hiện mức độ quan trọng của độ đo ảnh hưởng tới hiệu quả dự báo. Việc tính toán bộ trọng số có thể được thực hiện qua thực nghiệm với một mạng đồng tác giả cụ thể được xây dựng từ bộ dữ liệu bài báo khoa học, dùng phương pháp phân lớp theo các nhãn có / không liên kết và dùng giải thuật di truyền để xác định bộ trọng số phù hợp.

Phần II của bài báo sẽ đưa ra năm độ đo liên kết mạng thông dụng được khảo sát trong bài báo, Phần III trình bày phương pháp phân lớp Weighted SVM áp dụng cho dự báo và giải thuật di truyền để tối ưu bộ trọng số. Phần IV nêu các kết quả thực nghiệm và đánh giá, Phần V là kết luận và khả năng phát triển các nghiên cứu tiếp theo.

II. MỘT SỐ ĐỘ ĐO LIÊN KẾT

Với mỗi nút x, ký hiệu $T(x)$là tập các hàng xóm của x trong đồ thị mạng đồng tác giả (G_{collab}).

Trong bài báo này sẽ khảo sát một số độ đo tiêu biểu dựa theo $T(x)$..

A. Độ đo Weighted Common Neighbours – WCN

Độ đo Common Neighbours CN [19] giữa hai nút u và v là tổng số hàng xóm chung giữa u và v. Số lượng hàng xóm chung càng cao thì độ tương đồng CN càng lớn, do dó khả năng (u, v) có liên kết trong tương lai càng cao.

$$SIM_{CN}(u,v) = |T(u) \cap T(v)|$$

Độ đo CN thể hiện được số lượng hàng xóm chung nhưng chưa tính đến mức độ liên kết giữa các hàng xóm, trong trường hợp này là số lượng các bài báo công bố cùng nhau. Với hai tác giả u và v, ký hiệu $w(u,v)$ là số lượng bài báo chung, được sử dụng làm trọng số liên kết giữa hai tác giả.

Theo đó, có thể mở rộng độ đo CN, tính toán thêm với các $w(u,v)$ giữa các nút. Với hai tác giả u và v, xét tất cả các hàng xóm chung z và trọng số liên kết giữa u và z, cũng như giữa v và z, ta có công thức của độ đo WCN [17]:

$$SIM_{WCN}(u,v) = \sum_{z \in T(u) \cap T(v)} \frac{w(u,z)+w(v,z)}{2} \quad (1)$$

B. Độ đo Weighted Adamic_Adar – WAA

Độ đo Adamic-Adar [18] quan sát thêm số lượng hàng xóm chung của hàng xóm chung. Với z là hàng xóm chung của cả u và v, thì độ đo Adamic-Adar tỷ lệ nghịch với số lượng hàng xóm chung của z. Tích lũy tất cả các hàng xóm chung, ta có công thức độ đo Adamic-Adar của hai nút u và v như sau:

$$SIM_{AA(u,v)} = \sum_{z \in T(u) \cap T(v)} \frac{1}{\log(|T(z)|)}$$

Cũng như với độ đo CN, xét thêm trọng số liên kết giữa các hàng xóm $w(u,v)$, ta có công thức cho độ đo WAA[17]:

$$SIM_{WAA}(u,v) = \sum_{z \in T(u) \cap T(v)} \frac{w(u,z)+w(v,z)}{2 * Log(\sum_{x \in T(z)} w(z,x))} \quad (2)$$

C. Độ đo Weighted Jaccard Coefficient – WJC

Độ đo Jaccard Coefficient JC [16] giữa hai nút u, v tỷ lệ thuận với số lượng hàng xóm chung của u, v, đồng thời tỷ lệ nghịch với tổng số hàng xóm của u và v. Độ đo JC cho tỉ lệ các đồng tác giả cùng làm việc với x cũng làm việc với y.

$$SIM_{JC}(u,v) = \frac{|T(u) \cap T(v)|}{|T(u) \cup T(v)|}$$

Cũng như với độ đo CN, xét thêm trọng số liên kết giữa các hàng xóm $w(u,v)$, ta có công thức cho độ đo WJC [15]:

$$SIM_{WJC}(u,v) = \frac{1}{\sum_{x \in T(u)} w(u,x) + \sum_{y \in T(v)} w(v,y))} \sum_{z \in T(u) \cap T(v)} \frac{w(u,z)+w(v,z)}{2} \quad (3)$$

D. Độ đo Weighted Preferential Attachment – WPA

Độ đo Preferential Attachment PA [14] thể hiện mức độ liên kết rộng rãi của cả nút u và nút v, được tính bằng tích số lượng hàng xóm của cả hai nút.

$$SIM_{PA}(u,v) = T(u) \times T(v)$$

Xét thêm trọng số liên kết giữa các hàng xóm, ta có công thức cho độ đo WPA[17] như sau:

$$SIM_{WPA}(u,v) = \sum_{x \in T(u)} w(u,x) \times \sum_{y \in T(v)} w(v,y) \quad (4)$$

E. Độ đo SimRank

Độ đo SimRank [20] thể hiện mức độ tương tự giữa hai nút. Ký hiệu mức độ tương tự giữa hai nút u, v là $SIM_{SimRank}(u, v) \in [0,1]$, độ tương tự SimRank có thể được viết dưới dạng công thức đệ quy như sau, nếu u ≡ v thì $SIM_{SimRank}(u, v) = 1$, ngược lại tính theo công thức (5)

$$SIM_{SimRank}(u,v) = C.\sum_{z\in T(u)}\sum_{z'\in T(v)}\frac{SIM_{SimRank}(z,z')}{|T(u)|\times|T(v)|} \quad (5)$$

Trong đó $C \in [0.1]$ là hằng số.

III. PHÂN TÍCH SỰ ẢNH HƯỞNG CỦA CÁC ĐỘ ĐO

Để phân tích sự ảnh hưởng của các độ đo trong việc dự báo liên kết đồng tác giả, ta có thể gán trọng số cho các độ đo và thực nghiệm để tính toán bộ trọng số phù hợp, thực hiện qua các công việc sau:

(i) Xây dựng bảng dữ liệu các độ đo cho các cặp ứng viên: Tính toán các độ đo cho tất cả các cặp ứng viên (u,v) trong chu kỳ thời gian t_i. Trong bài báo này sẽ xét các độ đo WCN, WAA, WJC, WPA và SimRank. Tiếp theo, gán nhãn cho các cặp ứng viên (u, v) bằng cách xét chu kỳ thời gian tiếp theo t_{i+1}, xem u và v thực tế có phải là đồng tác giả ở khoảng thời gian t_{i+1} hay không. Gán nhãn "1" nếu u và v có công bố chung, nhãn "0', nếu ngược lại.

(ii) Sau khi hoàn thiện bảng dữ liệu gồm 5 thuộc tính độ đo và 1 nhãn cho các cặp ứng viên, ta có thể áp dụng một phương pháp phân lớp, dùng các dữ liệu đó để huấn luyện và kiểm tra, để dự đoán liên kết. Do đặc thù bài toán có số nhãn "0" vượt trội so với nhãn "1", nên trong bài báo này sử dụng phương pháp SVM có trọng số cho các bộ dữ liệu (Weighted Support Vector Machine)

(iii) Để phân tích sự ảnh hưởng của các độ đo đến kết quả dự báo liên kết đồng tác giả, ta có thể gán bộ trọng số (w_1, w_2, w_3, w_4, w_5) tương ứng cho 5 độ đo và tối ưu bộ tham số bằng giải thuật di truyền, với độ thích nghi của các cá thể (bộ trọng số) được đo bằng hiệu quả phân lớp.

Sau đây sẽ trình bày các nội dung chính.

A. Phương pháp phân lớp Weighted Support Vector Machine

Support vector machine (SVM) [1] là một phương pháp phân lớp nhị phân, coi việc học như là một vấn đề tối ưu. Các mẫu huấn luyện và kiểm tra được biểu diễn dưới dạng các véc tơ số thực d–chiều trong không gian đặc trưng mô tả dữ liệu, mỗi véc tơ trong tập huấn luyện được gán bởi nhãn dương hoặc nhãn âm. Bởi vậy, tập huấn luyện bao gồm các cặp (x_i, y_i), i =1,2,…,*l*. Trong đó $x_i \in \mathbf{R}^d$ là véc tơ huấn luyện thứ i, $y_i \in \{+1, -1\}$ là nhãn của véc tơ thứ i. Quá trình học sẽ cố gắng phân tách các véc tơ mang nhãn dương và nhãn âm bằng một siêu phẳng có dạng $\mathbf{w}^T x = b$. Ở đây, **w** là véc tơ pháp tuyến của siêu phẳng, b là hằng số xác định khoảng cách giữa góc toạ độ và siêu phẳng theo hướng pháp tuyến **w**. Để chọn **w** và b, SVM cực tiểu hoá hàm mục tiêu sau:

$$min_{w,b,\xi}\ \frac{1}{2}\|w\|^2 + C\sum_{i=1}^{l}\xi_i \quad (6)$$

thoả mãn điều kiện với mọi $\forall i: y_i(w^T\Phi(x_i)+b) \geq 1-\xi_i, \xi_i \geq 0$.

Trong đó, $\Phi(x_i)$ ánh xạ x_i vào không gian nhiều chiều, và C > 0 là một tham số chuẩn hoá.

Bởi vì **w** thường được xác định trong không gian nhiều chiều, do đó chúng ta sẽ giải quyết thông qua bài toán đối ngẫu:

$$min_{\alpha}\ \frac{1}{2}\alpha^T Q\alpha - e^T\alpha \quad (7)$$

thoả mãn, $y^T\alpha = 0$, $0 \leq \alpha_i \leq C$, i = 1,…,*l*

Trong đó, $e = [1,\ldots,1]^T$ là véc tơ tất cả giá trị bằng 1, Q là một ma trận nửa xác định dương (positive semidenite) với số chiều *l*×*l*. $Q_{ij} \equiv y_i y_j K(\mathbf{x_i}, \mathbf{x_j})$, và $K(\mathbf{x_i}, \mathbf{x_j}) \equiv \phi(x_i)^T\phi(x_j)$ là một hàm kernel.

Sau khi bài toán (2) được giải quyết, biến w tối ưu sẽ thoả mãn biểu thức sau.

$$w = \sum_{i=1}^{l} y_i\alpha_i\Phi(x_i)$$

Sau khi đã tìm được **w** và b, bộ phân lớp SVM sử dụng như một tiêu chuẩn để dự báo nhãn của một véc tơ mới (trong tập kiểm tra) như sau:

$$\text{Prediction}(x) := \text{sgn}(\mathbf{w}^T\phi(x) + b) = \text{sgn}\left(\sum_{i=1}^{l} y_i\alpha_i K(x_i, x) + b\right).$$

Đối với bài toán phân lớp nhị phân mà dữ liệu trong hai lớp là không cân bằng, khi đó một số tác giả trong [3, 4, 5] đã đề xuất sử dụng các tham số C khác nhau trong công thức SVM. Phương pháp Weighted SVM cực tiểu hóa hàm mục tiêu sau:

$$min_{w,b,\xi} \ \frac{1}{2}\|w\|^2 + C^+ \sum_{i:y_i=1} \xi_i + C^- \sum_{i:y_i=-1} \xi_i \tag{8}$$

thỏa mãn $\text{điều kiện với mọi } \forall i: y_i(w^T \Phi(x_i) + b) \geq 1 - \xi_i, \xi_i \geq 0$.

Trong đó, C^+, C^- là các tham số chuẩn hoá đối với lớp dương và âm tương ứng.

Bài toán đối ngẫu của (3) sẽ được giải quyết thông qua (4) như sau:

$$min_{\alpha} \ \frac{1}{2}\alpha^T Q\alpha - e^T\alpha \tag{9}$$

thoả mãn, $y^T\alpha = 0$, $0 \leq \alpha_i \leq C^+$, nếu $y_i = 1$, $0 \leq \alpha_i \leq C^-$, nếu $y_i = -1$, $i = 1,\ldots,l$

Các độ đo đánh giá hiệu quả phân lớp:

Một độ đo đánh giá hiệu quả của phương pháp phân lớp thường áp dụng cho bài toán phân lớp không cân bằng là độ đo AUC (Area Under the Curve). AUC có thể được định nghĩa như là xác suất chọn ngẫu nhiên cặp ứng viên có liên kết lớn hơn chọn cặp nút không có kết nối. Nếu AUC = 1 tương ứng với việc dự báo là tốt nhất, trong khi phương pháp lựa chọn ngẫu nhiên thì AUC = 0.5. Nếu giữa n phép so sánh độc lập, n' là số lần xác suất chọn cặp nút có liên kết cao hơn cặp nút không có liên kết, và n'' là số lần chọn cặp nút có liên kết có xác suất bằng với chọn cặp nút không có liên kết, khi đó giá trị của AUC được xác định bởi biểu thức (10) sau đây.

$$AUC = \frac{n' + 2n''}{n} \tag{10}$$

Ngoài ra, ta cũng xem xét độ chính xác của phương pháp dự báo dựa trên các độ đo Recall, Precision và F_measure.

$$\text{Re}call = \frac{|TP|}{|TP| + |FN|} \tag{11}$$

$$\text{Pr}ecision = \frac{|TP|}{|TP| + |FP|} \tag{12}$$

$$F - measure = \frac{2 * \text{Re}call * \text{Pr}ecision}{\text{Re}call + \text{Pr}ecision} \tag{13}$$

Trong đó |TP|, |FP| và |FN| lần lượt là số lượng véc tơ mang nhãn dương được dự đoán đúng (True Positives), số nhãn dương được dự đoán sai (False Positives) và số nhãn âm được dự đoán sai (False Negatives).

B. Áp dụng Giải thuật di truyền tính bộ trọng số các độ đo liên kết

Quần thể được dùng có kích thước N gồm các cá thể, mỗi cá thể là một bộ trọng số gồm 5 giá trị không âm có tổng bằng 1. Với mỗi cá thể, ta xây dựng lại bảng dữ liệu ứng viên với các giá trị độ đo nhân thêm với trọng số tương ứng. Độ thích nghi của mỗi cá thể được xác định bằng hiệu quả phân lớp khi sử dụng phương pháp Weighted SVM để phân lớp với bảng dữ liệu sau khi nhân với bộ trọng số tương ứng với cá thể đó. Quá trình di truyền sẽ cho ra kết quả các bộ trọng số phù hợp.

Cụ thể, các bước của giải thuật như sau:

Bước 1: Khởi tạo quần thể, kích thước N.

Bước 2: Tính độ thích nghi cho các cá thể, bằng cách thực hiện phân lớp Weighted SVM với bảng dữ liệu được nhân với bộ trọng số tương ứng với các thể đó. Độ đo hiệu quả phân lớp cho ta độ thích nghi của cá thể.

Bước 3: Kiểm tra điều kiện kết thúc.

Bước 4: Chọn lọc các cá thể cho thế hệ tiếp theo, có thể sử dụng bánh xe Roulette kết hợp với giữ lại các cá thể tốt nhất.

Bước 5: Lai ghép hai cá thể bố mẹ (theo xác suất lai ghép) để tạo ra các cá thể mới. Tính lại bộ trọng số tương ứng cho của cá thể mới.

Bước 6: Đột biến theo xác suất đột biến. Tính lại bộ trọng số tương ứng của cá thể đột biến. Quay lại Bước 2.

IV. THỰC NGHIỆM

Thực nghiệm được tiến hành trong môi trường Matlab, sử dụng thư viện LIBSVM của Chang & Lin [2] (được đăng tải tại địa chỉ https://www.csie.ntu.edu.tw/~cjlin/ libsvm/), cài đặt Weighted SVM gọi hàm *svmpredict*, với các tham số như sau: s = 0, w_1= tỷ lệ nhãn âm, w_{-1} = tỷ lệ nhãn dương, và h = 0.

A. Dữ liệu thực nghiệm

Để chuẩn bị thực nghiệm, chúng tôi thu thập dữ liệu về các tác giả trong lĩnh vực vật lý công bố bài báo khoa học về chủ đề High Energy Phisics – Lattice [9] từ năm 1992 đến năm 2000 , có 3.555 tác giả tham gia đóng góp 4.111 bài báo. Từ dữ liệu này, tính được mạng liên kết đồng tác giả với số lượng 19069 liên kết.

Tiếp theo là các bước tạo bảng dữ liệu.

(i) Từ dữ liệu đồng tác giả, tính tập các cặp ứng viên cho từng năm, ký hiệu:

- Tập C1 lưu trữ các cặp ứng viên của năm 1992,

- Tập C2 lưu trữ các cặp ứng viên của năm 1993,

- Tiếp tục như vậy, cho đến tập C9 lưu trữ các cặp ứng viên của năm 2000.

(ii) Từ các tập C1 đến C9, xây dựng các bảng dữ liệu để huấn luyện và kiểm tra cho phương pháp phân lớp.

- Bảng D1 có 5 thuộc tính độ đo và 1 thuộc tính nhãn, tính cho các cặp ứng viên đồng tác giả của các năm 1992, 1993, 1994 (dữ liệu trong các bảng C1, C2, C3), và gán nhãn "1" / "0" bằng dữ liệu liên kết đồng tác giả của năm 1995.

- Tương tự như vậy, có bảng D2 cho các cặp ứng viên của các năm 1993, 1994, 1995, và gán nhãn bởi dữ liệu liên kết đồng tác giả năm 1996.

- Bảng D3 cho các năm 1994, 1995, 1996 và gán nhãn bởi dữ liệu 1997.

- Bảng D4 cho các năm 1995, 1996, 1997 và gán nhãn bởi dữ liệu 1998.

- Bảng D5 cho các năm 1996, 1997, 1998 và gán nhãn bởi dữ liệu 1999.

- Bảng D6 cho các năm 1997, 1998, 1999 và gán nhãn bởi dữ liệu 2000.

(iii) Dùng các tập dữ liệu D1 đến D4 để tính độ thích nghi của các cá thể theo giải thuật di truyền.

- Từ D1 tính E1 bằng cách nhân các độ đo với trọng số tương ứng với các thể đang xét. Tương tự, tính được E2 từ D2, E3 từ D3, và E4 từ D4.

- Thực hiện thủ tục Weighted SVM ba lần, lần 1 dùng E1 để huấn luyện, E2 để kiểm tra; lần 2 dùng E2 để huấn luyện, E3 để kiểm tra và lần 3 dùng E3 để huấn luyện , E4 để kiểm tra. Sau đó tổng hợp các độ đo hiệu quả phân lớp của 3 lần đó cho độ thích nghi của cá thể.

(iv) Dùng các tập dữ liệu D5, D6 để kiểm tra lại bộ trọng số tốt nhất được tính bởi giải thuật di truyền

- Tính E5 từ D5, E6 từ D6 bằng cách nhân với bộ trọng số đó

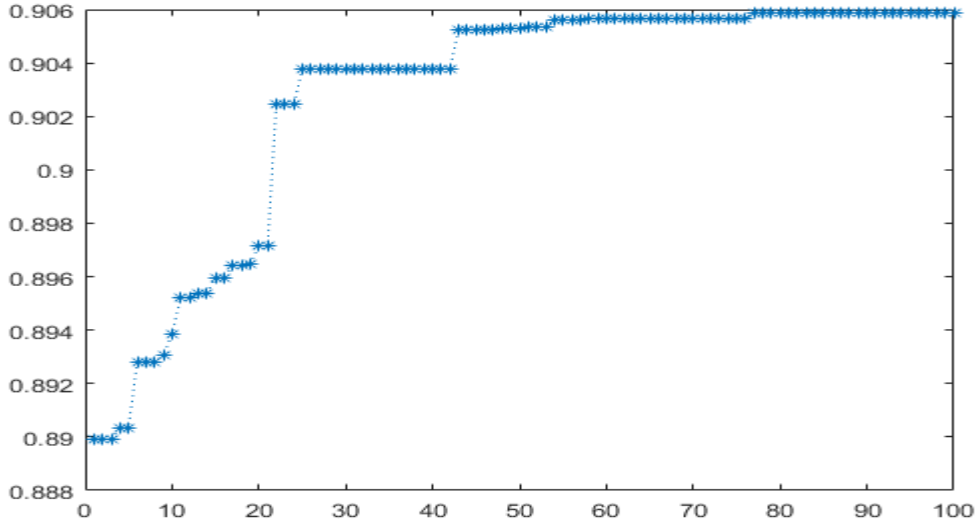
- Thực hiện thủ tục Weighted SVM với E5 là tập huấn luyện và E6 là tập kiểm tra.

B. Kết quả thực nghiệm

Áp dụng giải thuật di truyền với quần thể có 40 cá thể, chạy 100 thế hệ, xác suất lai ghép 25%, đột biến 5%, chọn lọc dùng bánh xe Roulette, độ thích nghi của các cá thể tính bằng độ đo AUC và độ đo F_measure

Kết quả bộ trọng số với độ thích nghi AUC:

Cho kết quả bộ trọng số tốt nhất: (0.0470943, 0.0581799, 0.145314, 0.175576, 0.573835) có độ thích nghi AUC = **0.905895**.

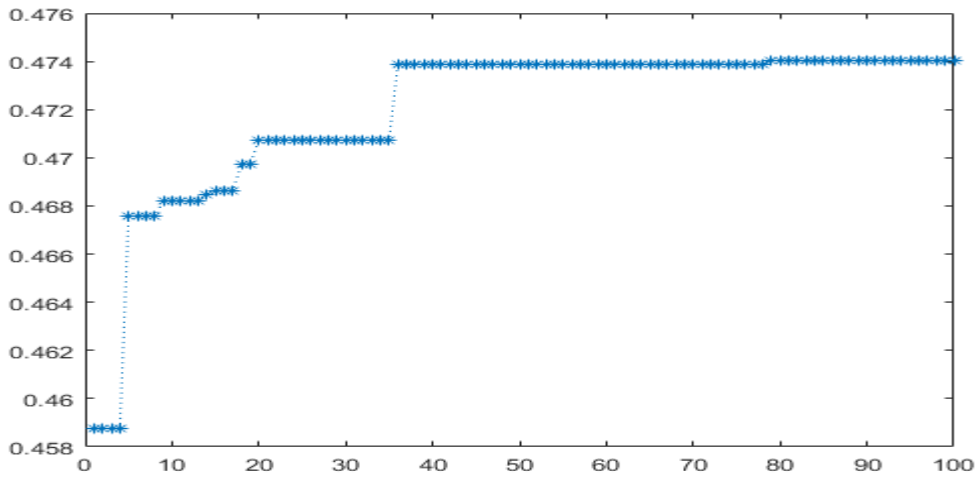


Hình 1. Độ thích nghi AUC của cá thể tốt nhất qua các thế hệ

Kiểm tra lại phân lớp Weighted SVM dùng bộ dữ liệu E5 để huấn luyện và E6 để kiểm tra, cho ta độ đo AUC với bộ trọng số trên là **0.89114**, trong khi nếu không dùng trọng số, khi phân lớp Weighted SVM với D5, D6 thì cho kết quả AUC là **0.86441**.

Kết quả bộ trọng số với độ thích nghi F_measure

Cho kết quả bộ trọng số tốt nhất: (0.400487, 0.0534822, 0.377146, 0.118194, 0.0506908) có độ thích nghi F-measure = **0.474008**.



Hình 2. Độ thích nghi F_measure của cá thể tốt nhất qua các thế hệ

Kiểm tra lại phân lớp Weighted SVM dùng bộ dữ liệu E5 để huấn luyện và E6 để kiểm tra, cho ta độ đo F_measure với bộ trọng số trên là **0.29895**, trong khi nếu không dùng trọng số, khi phân lớp Weighted SVM với D5, D6 thì cho kết quả AUC là **0.27171.**

Với bộ trọng số này thì độ đo Precision cũng được cải thiện: độ chính xác **0.20952**, so với **0.16667** nếu không dùng bộ trọng số.

C. Đánh giá

Độ đo AUC thể hiện hiệu quả phân lớp theo xác suất chọn ngẫu nhiên cặp ứng viên, độ đo SimRank có trọng số vượt trội, thể hiện sự "tương tự" giữa các nút. Trong mạng đồng tác giả, có thể là sự gần gũi về lĩnh vực chuyên môn, đồng nghiệp … là các đặc trưng cần lưu ý.

Độ đo F_measure liên quan đến độ chính xác và độ bao phủ, các độ đo liên kết WCN, WJC có trọng số lớn hơn, thể hiện vai trò của "hàng xóm" chung của các nút.

Việc đưa thêm vào bộ trọng số cho các độ đo, về bản chất là tăng thêm các tham số điều chỉnh. Trong khi trường hợp không dùng trọng số có thể hiểu là một trường hợp riêng khi các trọng số bằng nhau = (0.2, 0.2, 0.2, 0.2, 0.2). Vì vậy, xét thêm trọng số cho cơ hội cải thiện hiệu quả phân lớp. Điều này được minh chứng khi kiểm tra lại với bộ dữ liệu D5+D6, dùng D5 cho training, D6 cho testing với thủ tục Weighted-SVM, các giá trị độ đo hiệu quả phân lớp đều được cải thiện.

V. KẾT LUẬN

Bài báo đã đưa ra phương pháp phân tích sự ảnh hưởng một số độ đo liên kết đến hiệu quả dự báo liên kết đồng tác giả, hiện qua bộ trọng số gán cho các độ đo. Mạng đồng tác giả có đặc thù là mạng thưa, số cặp nút có liên kết ít hơn hẳn so với số cặp nút không có liên kết, vì vậy độ chính xác dự báo không cao. Việc tính toán thêm bộ trọng sô áp dụng váo phân lớp đã làm tăng hiệu quả dự báo, đã được trình bày qua thực nghiệm trong bài báo này.

Hiện tại, chúng tôi mới đang xét đến các độ đo liên kết thông dụng cho các mạng nói chung, chưa xét đến các đặc thù riêng của mạng đồng tác giả, như đặc trưng nhóm nghiên cứu, lĩnh vực nghiên cứu, địa chỉ, … Việc xây dựng các độ đo mới cho loại mạng này có thể là hướng phát triển của bài báo, cũng như thử nghiệm các phương pháp phân lớp và độ thích nghi phù hợp để tăng hiệu quả phương pháp dự báo.

TÀI LIỆU THAM KHẢO

[1] Cortes, C., & Vapnik, V., Support-vector networks. *Machine learning*, *20*(3) (1995), 273-297.

[2] Chang, C. C., Lin, C. J., LIBSVM: a library for support vector machines. ACM Transactions on Intelligent Systems and Technology (TIST) 2(3) (2011) 27.

[3] Osuna, E., Freund, R., & Girosi, F. (1997), Support vector machines: Training and applications.

[4] Brank, J., Grobelnik, M., Milic-Frayling, N., & Mladenic, D. (2003), *Training text classifiers with SVM on very few positive examples* (Vol. 486). Technical Report MSR-TR-2003-34, Microsoft Corp.

[5] Vapnik, V. N., & Vapnik, V. (1998), *Statistical learning theory* (Vol. 1). New York: Wiley.

[6] Adamic, L. A., & Adar, E. (2003). Friends and neighbors on the web. Social networks, 25(3), 211-230.

[7] Akcora, C. G., Carminati, B., & Ferrari, E. (2011). Network and profile based measures for user similarities on social networks. Proceedings of the 2011 IEEE International Conference on Information Reuse and Integration (IRI)(pp. 292-298).

[8] Akcora, C. G., Carminati, B., & Ferrari, E. (2013). User similarities on social networks. Social Network Analysis and Mining, 3(3), 475-495.

[9] Cornell University (2016). High Energy Physics Theory. Available at: https://arxiv.org/archive/hep-th/ (Accessed on: 17/10/2016).

[10] Fei Gao, Katarzyna Musial, Colin Cooper, Sophia Tsoka (2014), Link Prediction Methods and Their Accuracy for Different Social Networks and Network Metrics, http://eprints.bournemouth.ac.uk/22934/1/%5Bgamu15%5Dlink_prediction.pdf

[11] David Liben-Nowell, Jon Kleinberg (2004), The Link Prediction Problem for Social Networks, https://www.cs.cornell.edu/home/kleinber/link-pred.pdf

[12] Han, X., Wang, L., Farahbakhsh, R., Cuevas, Cuevas, R., Crespi, N., & He, L. (2016). CSD: A multi-user similarity metric for community recommendation in online social networks. Expert Systems withApplications, 53, 14-26.

[13] Bliss, C. A., Frank, M. R., Danforth, C. M., & Dodds, P. S. (2014). An evolutionary algorithm approach to link prediction in dynamic social networks. Journal of Computational Science, 5(5), 750-764.

[14] Mitzenmacher, M. (2004). A brief history of generative models for power law and lognormal distributions. Internet mathematics, 1(2), 226-251.

[15] Gne, Gndz-dc, & Ataltepe, Z. (2016). Link prediction using time series of neighborhood-based node similarity scores. Data Mining and Knowledge Discovery,30(1), 147-180.

[16] Salton, G. & Mc Gill, M.J. (1983). Introduction to Modern Information Retrieval. Mc Graw-Hill, NewYork.

[17] Murata, T., & Moriyasu, S. (2007). Link prediction of social networks based on weighted proximity measures. Proceedings of the IEEE/WIC/ACM international conference on In Web Intelligence, 85-88.

[18] Adamic, L. A., & Adar, E. (2003). Friends and neighbors on the web. Social networks, 25(3), 211-230.

[19] Newman, M. E. (2001). Clustering and preferential attachment in growing networks. Physical review E,64(2), 025102, 1-13.

[20] Jeh, G., & Widom, J. (2002, July). SimRank: a measure of structural-context similarity. In Proceedings of the eighth ACM SIGKDD international conference on Knowledge discovery and data mining (pp. 538-543). ACM.

ANALYSE THE EFFECT OF SOME METRICS TO APPLY TO LINK PREDICTION PROBLEM IN CO-AUTHORSHIP NETWORK

Pham Minh Chuan, Trinh Khac Linh, Tran Dinh Khang, Le Hoang Son

ABSTRACT: *In scientific research field, publishing papers often involves the participation and contribution from multiple authors. Similar to general social networking, co-authorship depends on various factors: acquaintance, collaboration, past relationship / co-authorship, or specific researching field, etc. This paper aims to analyse some linking metrics to observe the effect of those metrics in predicting the possibility of a co-authorship between certain candidates.*