Tạp chí Khoa học ĐHQGHN: Khoa học Tự nhiên và Công nghệ, Tập 30, Số 4 (2014) 48-57

Phân loại dữ liệu có liên kết sử dụng phương pháp đồng huấn luyện

Nguyễn Việt Tân[1], Hoàng Vũ[2,*], Đặng Vũ Tùng[3], Từ Minh Phương[4]

[1]*Đại học Công nghệ, ĐHQGHN, 144 Xuân Thủy, Cầu Giấy, Hà Nội, Việt Nam*
[2]*Viện Công nghệ thông tin, ĐHQGHN, 144 Xuân Thủy, Hà Nội, Việt Nam*
[3]*Học viện Thanh thiếu niên Việt Nam, 5 Chùa Láng, Đống Đa, Hà Nội, Việt Nam*
[4]*Học viện Công nghệ Bưu chính Viễn thông, 122 Hoàng Quốc Việt, Cầu Giấy, Hà Nội, Việt Nam*

Nhận ngày 10 tháng 10 năm 2014
Chỉnh sửa ngày 18 tháng 11 năm 2014; Chấp nhận đăng ngày 22 tháng 12 năm 2014

Tóm tắt: Trong một số ứng dụng phân loại tự động, bên cạnh các dữ liệu dạng vector còn có dữ liệu liên kết thể hiện quan hệ giữa các đối tượng như: trang web được nối bởi các siêu liên kết, bài báo khoa học được liên kết bởi các tài liệu tham khảo, các nút mạng được kết nối vật lý .v.v. Yêu cầu đặt ra với thuật toán phân loại là tận dụng và kết hợp dữ liệu liên kết với các thông tin khác để cho kết quả dự đoán chính xác hơn. Nhiều nghiên cứu trước đây đã giải quyết vấn đề này bằng cách sử dụng các thuật toán dựa trên đồ thị mà tiêu biểu là bộ phân lớp Gaussian-field, các mạng Hopfield và bộ phân lớp quan hệ láng giềng.v.v. Trong bài báo này, chúng tôi đề xuất giải quyết vấn đề kết hợp thông tin liên kết với các dữ liệu khác bằng cách sử dụng kỹ thuật đồng huấn luyện, trong đó các liên kết được coi là một góc nhìn (view) khác của dữ liệu. Phương pháp được thử nghiệm trên bộ dữ liệu WebKB. Kết quả thử nghiệm và so sánh cho thấy phương pháp đề xuất cho kết quả phân loại chính xác hơn phương pháp kết hợp dữ liệu liên kết dựa trên đồ thị.

Từ khóa: Đồng huấn luyện,dữ liệu liên kết

1. Giới thiệu

Phân loại hay phân lớp là kỹ thuật khai phá dữ liệu được nghiên cứu và sử dụng rộng rãi. Đây là phần quan trọng trong các dạng ứng dụng như phân loại văn bản, nhận dạng chữ viết, giọng nói, phân loại protein v.v.

Trên thực tế tồn tại một số bài toán trong đó giữa các đối tượng cần phân lớp có các liên kết với nhau. Chẳng hạn, khi phân loại trang web, ngoài nội dung trang có thể sử dụng như các đặc trưng dùng để phân loại, trong các trang lại có các siêu liên kết. Hay khi phân loại protein, các protein thường có các liên kết tương ứng với quan hệ tương tác giữa chúng. Các quan hệ liên kết cũng là dạng dữ liệu tiêu biểu với các ứng dụng cho mạng máy tính. Từ thực tế này, một vấn đề đặt ra là tận dụng dữ liệu có liên kết để tăng hiệu quả và độ chính xác cho thuật toán phân lớp.

Tác giả liên hệ. ĐT.: 84-903429148
Email: tannv@vnu.edu.vn

Nguyên tắc chung của việc phân lớp dữ liệu có liên kết là tạo ra các ràng buộc, theo đó những đối tượng được liên kết với nhau cần có nhãn phân lớp tương tự nhau. Dựa trên nguyên tắc chung này, nhiều thuật toán và kỹ thuật cụ thể đã được phát triển và ứng dụng.

Một trong những tiếp cận sớm nhất chú ý tới mối quan hệ giữa các đối tượng được đề xuất bởi Chakrabarti và cộng sự [1]. Họ đề xuất một mô hình xác suất cho phân loại trang web bằng cách sử dụng kết hợp giữa nội dung của trang đã phân lớp, nhãn phân lớp của các trang liên kết và nội dung của các trang liên kết. Cũng thời gian này, Blum và Mitchell [2] đưa ra kỹ thuật Co-training với thử nghiệm phân lớp cho dữ liệu WebKB. Tuy nhiên 2 tập con đặc trưng đều dưới dạng text và 2 bộ phân lớp được sử dụng đều là loại truyền thống - Naïve Bayes. Gần đây, Macskassy và Provost [3] đã thử nghiệm phân lớp tập hợp cho dữ liệu liên kết bằng cách kết hợp một bộ phân lớp liên kết với một phương thức suy luận tập hợp (collective inferencing). Sen và cộng sự [4] so sánh bốn phương pháp phân loại tập hợp cho dữ liệu có liên kết. Bên cạnh các phương pháp phân loại tập hợp, một hướng tiếp cận được sử dụng rộng rãi khác là phương pháp học bán giám sát (semi-supervised learning) dựa trên đồ thị, trong đó tiêu biểu phải kể đến phương pháp Gaussian random field [5], phương pháp nhất quán địa phương và toàn cục (Local and global consistency) [6].

Trong bài báo này, chúng tôi đề xuất giải quyết vấn đề phân lớp cho dữ liệu có liên kết bằng cách kết hợp một bộ phân lớp liên kết (relation classifier) với một bộ phân lớp truyền thống (non-relation hay local classifier). Hai bộ phân lớp này sẽ học đồng thời trên hai tập đặc trưng con được trích chọn từ tập dữ liệu gốc. Phương pháp đồng huấn luyện (co-training) sẽ được sử dụng để gắn kết 2 bộ phân lớp nói trên. Hiệu quả của thuật toán được thử nghiệm và so sánh với một số phương pháp khác trên bộ dữ liệu WebKB. Đây là bộ dữ liệu thường được sử dụng để đánh giá các thuật toán phân loại cho dữ liệu có liên kết. Kết quả thử nghiệm cho thấy hiệu quả của phương pháp đề xuất.

2. Bài toán phân lớp cho dữ liệu có liên kết

Dữ liệu có liên kết, được gọi là Networked data hay Linked data, là trường hợp đặc biệt của dữ liệu quan hệ khi mà các phần tử trong đó có các kết nối với nhau. Ví dụ, các trang web được kết nối với nhau bằng các siêu liên kết, tài liệu được kết nối bằng các trích dẫn, tham khảo v.v. Các phương pháp phân lớp cho dữ liệu liên kết về cơ bản dựa trên giả thiết về Homophily (nguyên lý đồng đẳng): "*các đối tượng liên quan với nhau có xu hướng thuộc cùng một lớp*". Đây là một nguyên lý dựa trên các nghiên cứu và phân tích trên mạng xã hội cho rằng: sự giao tiếp giữa các đối tượng giống nhau xảy ra với tỉ lệ cao hơn so với giao tiếp giữa các đối tượng không giống nhau. Các đối tượng thường tìm kiếm, lựa chọn và kết bạn với những người giống với họ, có thể là về giới tính, về tuổi tác, về địa vị xã hội, về tầng lớp, về đặc điểm hành vi cá nhân, về niềm tin, lý tưởng.v.v.

So với phân lớp truyền thống, một trong những vấn đề chính cần lưu ý khi phân lớp dữ liệu có liên kết xuất phát từ bản chất quan hệ tự nhiên của của dữ liệu. Vì vậy, việc phân lớp của một nút có thể có ảnh hưởng đến các nút liên quan, và ngược lại. Để khắc phục vấn đề này, một kỹ thuật đã được công nhận rộng rãi là: các nút cần được ước tính và suy ra cùng một lúc

thay vì từng nút một. Kỹ thuật này được gọi là phân lớp tập hợp (collective classification).

Bài toán phân lớp cho dữ liệu có liên kết được phát biểu như sau: Cho đồ thị $G = (V, E, X)$ trong đó: V là tập nút (đỉnh) gồm n nút tương ứng với n đối tượng; E là tập các cạnh: $e_{i,j} \in E$ biểu thị một cạnh nối giữa 2 nút v_i và v_j ; X_i là thuộc tính phân lớp của nút v_i có thể nhận giá trị c ∈ X. Cho trước các giá trị x_i thuộc X_i cho tập con $V^K \in V$. Khi đó, phân lớp tập hợp là một tiến trình kết hợp một thuật toán phân lớp liên kết với một phép suy luận tập hợp để suy luận đồng thời các giá trị x_i thuộc X_i cho các đỉnh còn lại, $V^U = V - V^K$.

Như vậy, bài toán phân lớp tập hợp cho dữ liệu liên kết được thực hiện nhờ hai thủ tục. Thủ tục thứ nhất là phân lớp liên kết (relational classification), theo đó nhãn phân lớp được xác định dựa trên các hàng xóm. Một số thuật toán tiêu biểu cho bước này bao gồm: *Weighted-Vote Relational Neighbor Classifier (wvRN), Class-Distribution Relational Neighbor Classifier (cdRN), Network-Only Bayes Classifier (nBC)* hay *Network-Only Link-Base Classifier (nLB)* [7][4]. Thủ tục thứ hai là suy luận tập hợp (collective inference). Bản chất của bước này là xác định nhãn phân lớp đồng thời cho các nút trên mạng. Một số thuật toán suy luận tập hợp bao gồm: *Iterative Classification (IC), Gibbs Sampling (GS) hay Relaxation Labeling (RL)* [7].

Trong bài này, chúng tôi lựa chọn và sử dụng bộ phân lớp liên kết *wvRN* và phương pháp suy luận tập hợp *RL* do tính đơn giản và hiệu quả phân lớp đã được đánh giá là tốt đối với dạng bài toán phân loại trang web. Chúng tôi xin trình bày tóm tắt hai thuật toán này.

Bộ phân lớp liên kết *wvRN* ước tính xác suất đối tượng thuộc một lớp dựa trên 2 giả định: (i) nhãn của một nút chỉ phụ thuộc vào hàng xóm trực tiếp của nó và (ii) sự tồn tại của Homophily.

Thuật toán *wvRN*:

Cho một nút i và một tập các nút hàng xóm N. Bộ phân lớp *wvRN* tính xác suất nút i thuộc lớp c bằng trung bình trọng số các xác suất của tất cả các nút hàng xóm.

$$P(x_i = c|N_i) = \frac{1}{Z}\sum_{j\in N} w_{i,j}.P(x_j = c|N_j) \quad (1)$$

Trong đó $w_{i,j}$ là trọng số giữa i và j, thường tính bằng số lượng liên kết xuất hiện giữa 2 nút; Z là hệ số chuẩn hóa để đảm bào giá trị nằm trong khoảng [0,1], và được tính bằng số lượng các liên kết xuất hiện giữa i với các nút đã được dán nhãn.

Vì đây là một định nghĩa đệ quy (cho đồ thị vô hướng, $v_j \in N_i \Leftrightarrow v_i \in N_j$) nên bộ phân lớp sẽ sử dụng ước tính "hiện tại" cho xác suất $P(x_j = c|N_j)$

Phương pháp suy luận tập hợp *RL* dùng để lưu giữ các nhãn tạm thời , theo dõi các ước tính xác suất "hiện tại" cho x^U . Hơn nữa, thay vì ước tính mỗi lần một nút và ghi giá trị ngay vào đồ thị, *RL* "đóng băng" ước tính "hiện tại" sao cho tại bước $t+1$, tất cả các đỉnh sẽ được cập nhật dựa trên ước tính từ bước t. Tuy nhiên, làm như vậy sẽ dẫn tới sự dao động giữa các trạng thái. Có thể sử dụng một phương pháp tiếp cận của *giải thuật luyện kim* (Simulated Annealing – SA) để giải quyết vấn đề này. Sau mỗi bước lặp, trọng số cho nút hiện tại sẽ được tăng lên và ảnh hưởng của các nút láng giềng sẽ bị giảm xuống.

Suy luận tập hợp *RL* được định nghĩa như sau:

$$\hat{c}_i^{(t+1)} = \beta^{(t+1)}.wvRN(v_i^{(t)}) + (1-\beta^{(t+1)}).\hat{c}_i^{(t)} \quad (2)$$

Trong đó $\hat{c}_i^t$ là vector các xác suất (phân bố xác suất) biểu diễn ước tính của $P(x_i|N_i)$ tại bước t và $wvRN(v_i^{(t)})$ biểu thị áp dụng *wvRN*

với mọi ước tính từ thời điểm bước *t*. Người ta xác định các tham số của giải thuật luyện kim như sau:

$$\beta^0 = k$$

$$\beta^{(t+1)} = \beta^{(t)}.\alpha,$$

Với *k* là hằng số giữa 0 và 1 thường được chọn là 1; α là hệ số suy giảm thường được chọn giữa 0.9 và 0.99.

Các bộ phân lớp liên kết chỉ quan tâm tới cấu trúc liên kết của một nút. Nếu tất cả các nút trong tập kiểm tra được kết nối tới ít nhất một nút trong tập huấn luyện thì không có vấn đề gì, nhưng trên thực tế có rất nhiều dữ liệu không thỏa mãn điều kiện này. Khi đó, bộ phân lớp liên kết sẽ không thể phân lớp cho những nút không có nút hàng xóm trong tập huấn luyện. Để bù đắp những thiếu hụt này, bộ phân lớp tập hợp có thể kết hợp một bộ phân lớp liên kết với một bộ phân lớp truyền thống nhằm cố gắng tăng độ chính xác khi phân lớp. Với cách sử dụng bộ phân lớp truyền thống trong bước lặp đầu tiên (t=1), bộ phân lớp tập hợp bảo đảm rằng tất cả các nút sẽ có xác suất phân lớp ban đầu. Bộ suy luận tập hợp sau đó sẽ sử dụng bộ phân lớp liên kết và dựa vào các xác suất ban đầu đó để tiếp tục phân lớp.

3. Phân lớp dữ liệu liên kết dựa trên kỹ thuật đồng huấn luyện

Trong phần này, chúng tôi trình bày phương pháp đề xuất, trong đó vấn đề phân lớp cho dữ liệu liên kết được thực hiện theo nguyên lý đồng huấn luyện. Để tiện cho việc trình bày, trước hết chúng tôi tóm tắt nguyên lý đồng huấn luyện, sau đó sẽ mô tả chi tiết cách sử dụng kỹ thuật này cho phân loại tập hợp đối với dữ liệu liên kết.

3.1. Đồng huấn luyện

Đồng huấn luyện là kỹ thuật học bán giám sát được giới thiệu lần đầu bởi Blum và Mitchell vào năm 1998 [2]. Mục đích của đồng huấn luyện là cung cấp khả năng phân loại một cách chính xác và hiệu quả một tập lớn dữ liệu không gán nhãn dựa vào một tập nhỏ ban đầu các dữ liệu gán nhãn. Trong kỹ thuật đồng huấn luyện, giả sử rằng (i) các đặc trưng có thể phân chia thành hai tập riêng biệt; (ii) mỗi tập đặc trưng con là đủ để huấn luyện một bộ phân lớp tốt; (iii) hai tập con phải thỏa mãn tính chất độc lập có điều kiện khi cho trước lớp. Ban đầu, hai bộ phân lớp được học với các dữ liệu đã gán nhãn trên hai tập đặc trưng tương ứng. Mỗi bộ phân lớp sau đó lại phân lớp dữ liệu chưa gán nhãn rồi chọn các nhãn dự đoán mà nó cảm thấy có độ tin cậy cao nhất để đưa thêm vào tập huấn luyện. Tiếp theo, mỗi bộ phân lớp học lại trên tập huấn luyện vừa được bổ sung bởi bộ phân lớp còn lại. Quá trình được lặp lại cho tới khi hết dữ liệu không gán nhãn hoặc sau một số bước thiết lập trước.

3.2. Phân lớp cho dữ liệu liên kết theo phương pháp đồng huấn luyện

Chúng tôi chia dữ liệu gốc thành 2 tập đặc trưng gọi là *Content* và *Link*. Tập *Content* chứa thông tin về các đặc trưng nội dung của từng đối tượng. Ví dụ, trong trường hợp đối tượng cần phân lớp là trang web, thông tin nội dung sẽ là các từ xuất hiện trong trang. Đối với đối tượng là protein, thông tin nội dung có thể là mức độ biểu hiện gen tương ứng với protein đó. Tập *Link* chứa thông tin về liên kết giữa các đối tượng. Ví dụ, thông tin liên kết được tạo thành từ các siêu liên kết trong trang đối với dữ liệu web hay thông tin tương tác giữa protein trong trường hợp phân lớp protein.

Một bộ phân lớp truyền thống sử dụng vector đặc trưng sẽ được huấn luyện trên đặc trưng nội dung của phần *Content*. Trong thực nghiệm ở đây, chúng tôi sử dụng bộ phân lớp Naïve Bayes [8] để phân lớp trên dữ liệu nội dung. Đây là phương pháp phân loại được sử dụng rộng rãi cho dữ liệu văn bản.

Một bộ phân lớp liên kết được sử dụng để dự đoán trên dữ liệu liên kết của phần *Link*. Trong nghiên cứu này, chúng tôi sử dụng bộ phân lớp liên kết *wvRN* cùng phương thức suy luận tập hợp *RL* để học và phân lớp trên tập *Link*.

Hai bộ phân lớp nói trên sẽ được sử dụng cùng nhau theo kiểu đồng huấn luyện. Tại mỗi bước, từng bộ phân loại được huấn luyện trên dữ liệu có nhãn hiện có, sau đó dự đoán nhãn cho những nút còn lại. Các dự đoán có độ tin cậy cao nhất của mỗi phương pháp được thêm vào tập nhãn huấn luyện của phương pháp kia. Thuật toán lặp lại cho tới khi toàn bộ các nút được gán nhãn.

Thuật toán đề xuất được thể hiện trên Hình 1.

Input:

L: Tập các mẫu đã gán nhãn, ***U***: Tập các mẫu chưa gán nhãn;

$\boldsymbol{F_1}$ là tập *Content*, $\boldsymbol{F_2}$ là tập *Link*;

$\boldsymbol{C_1}$ là bộ phân lớp *Naïve Bayes*, $\boldsymbol{C_2}$ là bộ phân lớp $wvRN_{\mathrm{RL}}$.

n là số nhãn phân loại mới sau mỗi bước

Output:

Tập nhãn cho toàn bộ các mẫu

Thuật toán:

Lặp cho đến khi $\boldsymbol{U}=\emptyset$:

- Học bộ phân lớp $\boldsymbol{C_1}$ bằng dữ liệu ***L*** trên tập $\boldsymbol{F_1}$
- Học bộ phân lớp $\boldsymbol{C_2}$ bằng dữ liệu ***L*** trên tập $\boldsymbol{F_2}$
- Với mỗi bộ phân lớp $\boldsymbol{C_i}$ (i = 1,2): thực hiện:
  - $\boldsymbol{C_i}$ dự đoán nhãn cho các mẫu trong tập ***U***
  - $\boldsymbol{C_i}$ chọn *n* mẫu có độ tin cậy cao nhất trong ***U***
  - $\boldsymbol{C_i}$ thêm *n* mẫu đã chọn vào ***L***
  - $\boldsymbol{C_i}$ loại bỏ *n* mẫu đã chọn khỏi ***U***

Hình 1. Thuật toán đồng huấn luyện áp dụng cho bài toán phân lớp dữ liệu liên kết

Trong thuật toán ở Hình 1, tại mỗi bước, thuật toán lựa chọn và thêm *n* nhãn mới dự đoán vào tập *L*. Các nhãn được chọn là nhãn có độ tin cậy phân lớp cao nhất. Trong cả hai trường hợp phân loại *Naïve Bayes* và *wvRN*, độ tin cậy được xác định bằng xác suất hậu nghiệm, ví dụ, xác suất $P(x_j = c| N_i)$ trong trường hợp *wvRN*. Cụ thể, thuật toán sắp xếp các nhãn mới dự đoán theo thứ tự giảm dần của xác suất hậu nghiệm, sau đó lựa chọn *n* nhãn

đứng đầu danh sách. Số lượng *n* được lựa chọn cố định và là tham số của thuật toán.

4. Thực nghiệm và kết quả

4.1. Dữ liệu

Dữ liệu thực nghiệm là bộ dữ liệu được sử dụng rộng rãi WebKB (http://www.cs.cmu.edu/~WebKB/). Bộ này bao gồm hơn 8000 trang web lấy từ 4 website bộ môn Khoa học máy tính của các trường đại học: Cornell, Texas, Washington và Wisconsin. Mỗi trang web được lưu vào một tệp tin dạng *.html* với tên chính là URL thực của trang web đó. Người ta đã thực hiện việc phân lớp thủ công cho từng trang web vào 1 trong 7 lớp: *course, department, faculty, project, staff, student, other* bằng cách chia vào các thư mục có tên tương ứng. Để tương thích và tiện so sánh với các kết quả nghiên cứu trước đây, chúng tôi loại bỏ các trang web trong lớp *other* và thực hiện việc phân chia dữ liệu vào 6 lớp còn lại.

4.2. Công cụ

Trong quá trình thực nghiệm học và phân lớp, chúng tôi sử dụng 2 bộ công cụ mã nguồn mở:

- Network Learning Toolkit **(Netkit-SRL** http://sourceforge.net/projects/netkit-srl/). Đây là một trong số rất ít công cụ mã nguồn mở có khả năng thực hiện các thuật toán phân lớp cho dữ liệu liên kết như: *WVRN, CDRN, NBC*, *NLB*. Mỗi thuật toán phân lớp lại có thể kết hợp với một phương thức suy luận tập hợp như: *GS, RL, IC*.

- Waikato Environment for Knowledge Analysis: WEKA. Đây là công cụ rất tiện dụng trong xây dựng các mô hình khai phá dữ liệu. WEKA triển khai hầu hết các kỹ thuật khai phá dữ liệu như Classification, Clustering, Association Rule,.. Trong mỗi kỹ thuật, WEKA triển khai rất nhiều thuật toán cho phép lựa chọn thuật toán phù hợp với yêu cầu và dữ liệu trong việc khai phá dữ liệu.

4.3. Phương pháp thực nghiệm

Chúng tôi sử dụng công cụ WEKA với bộ phân lớp *Naïve Bayes* để tiến hành học và phân lớp trên dữ liệu *Content*. Bộ công cụ Netkit-SRL với bộ phân lớp *wvRN* kết hợp với phương thức suy luận tập hợp *RL* sẽ được dùng để học và phân lớp trên dữ liệu *Link*. Phương pháp đồng huấn luyện như mô tả trong Hình 1 sẽ được sử dụng để kết hợp hai bộ phân lớp trên.

Khi có kết quả phân lớp áp dụng phương pháp Đồng huấn luyện, chúng tôi sẽ đánh giá và so sánh với hai phương pháp phân lớp ban đầu cũng như so sánh với phương pháp phân lớp tập hợp kết hợp bộ phân lớp liên kết với bộ phân lớp truyền thống.

4.4. Quá trình và kết quả thực nghiệm

4.4.1. Xây dựng và trích chọn các đặc trưng

Đầu tiên, chúng tôi tiến hành trích chọn đặc trưng của các trang web và chia thành 2 tập chứa các đặc trưng riêng biệt. Đặc trưng thứ nhất của trang web chính là các từ xuất hiện trong trang web đó. Mỗi trang web sẽ được biểu diễn dưới dạng vector theo mô hình không gian vector (Vector Space Model). Mỗi thành phần của vector là một từ khóa riêng biệt xuất hiện trong website và được gán một giá trị gọi là hàm f chỉ mật độ xuất hiện của từ khóa đó. Chúng tôi gọi tập *Content* là tập chứa các vector này.

Một đặc trưng nữa của trang web là các siêu liên kết có trong mỗi trang. Chúng tôi xây dựng một tập tên là *Link* chứa các thông tin bao gồm:

"*x*", "*y*" và "*Trọng số liên kết giữa x và y*"; trong đó *x*, *y* là 2 trang web có liên kết với nhau và cùng nằm trong một website.

Thông tin siêu liên kết lại được chia làm 2 loại là Direct Link và Cocite. Direct Link là kiểu liên kết trực tiếp giữa 2 trang web (*x* có chứa siêu liên kết tới *y*). Khi đó, trọng số liên kết dạng Direct Link giữa 2 trang *x* và *y* là tổng số lần xuất hiệu siêu liên kết từ trang *x* tới trang *y*. Cocite là một kiểu liên kết khác. Hai trang *x* và *y* gọi là liên kết dạng Cocite (theo z) khi *x* liên kết trực tiếp với *z* và *y* cũng liên kết trực tiếp tới *z*. Để tính trọng số liên kết kiểu Cocite giữa *x* và *y*, ta lấy tổng số lần xuất hiện siêu liên kết từ trang *x* tới trang *z* rồi nhân với tổng số lần xuất hiện siêu liên kết từ trang *y* tới trang *z*.

4.4.2. Tiến hành phân lớp

Trước khi tiến hành phân lớp bằng phương pháp đồng huấn luyện, chúng tôi thực hiện phân lớp trên 2 bộ dữ liệu và 2 bộ phân lớp riêng lẻ để kiểm tra việc tiền xử lý dữ liệu và đánh giá độ chính xác của các bộ phân lớp.

Đầu tiên, chúng tôi sử dụng phần mềm WEKA để tiến hành học và phân lớp trên tập *Content*. Bảng 1 biểu diễn kết quả phân lớp dựa trên *Content* với bộ phân lớp Naive Bayes, tùy chọn thử nghiệm là 5 fold cross validation.

Bảng 1. Tỷ lệ chính xác khi phân lớp dựa trên tập Content và bộ phân lớp Naive Bayes

	Cornell	**Texas**	**Washington**	**Wisconsin**
Course	0.649	0.795	0.781	0.792
Department	0	0	0	0
Faculty	0.444	0.615	0.406	0.54
Project	0.12	0.2	0.125	0.421
Staff	0.417	0.057	0.04	0.061
Student	0.757	0.811	0.664	0.78
Trung bình	**0.612**	**0.714**	**0.599**	**0.694**

Dựa vào kết quả ở Bảng 1 ta thấy độ chính xác của bộ phân lớp *Naive Bayes* là ở mức tin cậy được với độ chính xác trung bình cao nhất lên tới 71.4% và thấp nhất là 61.2%.

Tiếp theo, chúng tôi sử dụng phần mềm Netkit-SRL để học và phân lớp trên tập *Link-Cocite*. Trong quá trình tiền xử lý dữ liệu chúng tôi phát hiện ra việc dùng dữ liệu dạng Direct link trong bài toán phân loại trang web sẽ cho kết quả kém chính xác hơn nhiều so với việc sử dụng dữ liệu dạng Cocite. Chính vì vậy trong các phần tiếp theo chúng tôi chỉ sử dụng dữ liệu liên kết dạng Cocite. **Bảng 2** chứa kết quả phân lớp dựa trên *Link-Cocite* với thuật toán phân lớp quan hệ *wvRN* và phương thức suy luận tập hợp *RL*.

Bảng 2. Tỷ lệ chính xác khi phân lớp dựa trên tập Link-Cocite và bộ phân lớp $wvRN_{RL}$

	Cornell	**Texas**	**Washington**	**Wisconsin**
Course	0.37621	0.53169	0.7538	0.83226
Department	0.15254	0.35204	0.09302	0.61446
Faculty	0.34564	0.48413	0.24038	0.09412
Project	0.37786	0.0979	0.06742	0.2459
Staff	0	0.17647	0	0.08
Student	0.87263	0.96415	0.87762	0.99003
Trung bình	**0.56057**	**0.65976**	**0.61379**	**0.70845**

Kết quả phân lớp ở Bảng 2 cho ta thấy độ chính của bộ phân lớp *wvRN* với dạng dữ liệu Cocite ở mức chấp nhận được với độ chính xác trung bình trong khoảng 56.057% đến 70.845%.

Tiếp đó chúng tôi dùng phần mềm Netkit-SRL để phân lớp tập hợp kết hợp bộ phân lớp liên kết $wvRN_{RL}$ với bộ phân lớp truyền thống *Naive Bayes.*

Bảng 3. Tỷ lệ chính xác khi phân lớp tập hợp $wvRN_{RL}$+Naive Bayes

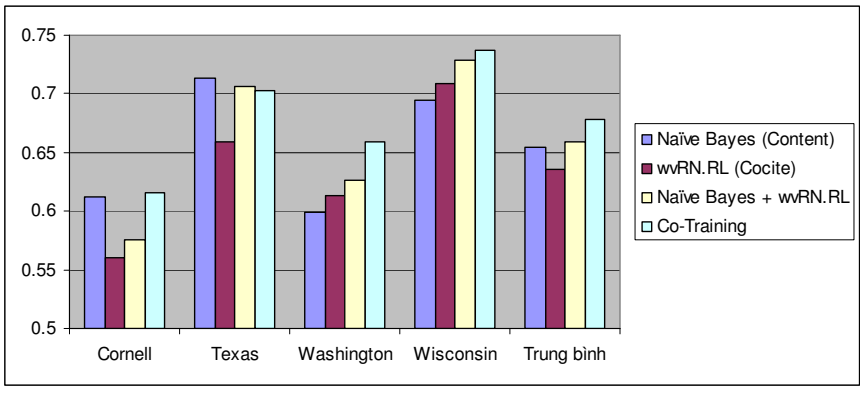
	Cornell	**Texas**	**Washington**	**Wisconsin**
Course	0.48438	0.56881	0.73143	0.77652
Department	0.23913	0.42697	0.03704	0.72222
Faculty	0.26562	0.55682	0.26872	0.03175
Project	0.51786	0.22642	0.06349	0.18487
Staff	0	0.33333	0	0.03922
Student	0.86232	0.96319	0.88446	0.99619
Trung bình	**0.57571**	**0.70588**	**0.62673**	**0.72825**

Cuối cùng, chúng tôi thử nghiệm học và phân lớp bằng phương pháp đồng huấn luyện. Các bước thực hiện phân lớp theo phương pháp này đã được mô tả tại **Hình 1** với các tham số cho mô hình được thiết lập như sau :

- **L***:* chứa 20% số trang được chọn ngẫu nhiên trên mỗi website *;*
- **n**=10% số mẫu ban đầu trong **U**

Bảng 4. Tỷ lệ chính xác khi phân lớp bằng phương pháp đồng huấn luyện

	Cornell	**Texas**	**Washington**	**Wisconsin**
Course	0.53846	0.5641	0.77709	0.79755
Department	0.30189	0.58571	0.11905	0.66667
Faculty	0.41129	0.53	0.36538	0.02817
Project	0.43011	0.12727	0.08642	0.20455
Staff	0	0.26667	0	0
Student	0.87584	0.95046	0.88312	1
Trung bình	**0.61571**	**0.70294**	**0.65862**	**0.73662**



Hình 2. Biểu đồ so sánh độ chính xác của 4 bộ phân lớp.

Kết quả trong *Bảng 4* và *Hình 2* cho thấy, về tổng thể, trong 4 phương pháp phân lớp thì độ chính xác của phương pháp đồng huấn luyện là cao hơn cả. Trong phần lớn các trường hợp, độ chính xác của phương pháp đồng huấn luyện là cao nhất. Không có trường hợp nào phương pháp đồng huấn luyện cho kết quả kém nhất.

5. Kết luận

Thông qua việc đề xuất và thử nghiệm phương pháp đồng huấn luyện để phân lớp cho dữ liệu có liên kết, chúng tôi muốn kiểm chứng đồng thời hai vấn đề. Thứ nhất, việc tận dụng thông tin của các đối tượng liên quan trong dữ liệu liên kết sẽ giúp nâng cao hiệu suất phân lớp. Thứ hai, chúng tôi muốn kiểm tra và củng cố tính đúng đắn của phương pháp đồng huấn luyện khi áp dụng cho một kiểu dữ liệu mới. Kết quả thử nghiệm đã cho thấy tính đúng đắn và ưu việt của phương pháp này khi áp dụng cho dạng dữ liệu có liên kết.

Tài liệu tham khảo

[1] S. Chakrabarti, B. Dom, and P. Indyk (1998). Enhanced hypertext categorization using hyperlinks. In Proceedings of the 1998 ACM SIGMOD International Conference on Management of Data, pp: 307–319, 1998

[2] Blum A., Mitchell T. (1998): Combining labeled and unlabeled data with co- training. In Proceedings of the 11th Annual Conference on Computational Learning Theory (COLT-98).

[3] Macskassy, S.A., Provost, F. (2005): Suspicion scoring based on guilt-by-association, collective inference, and focused data access. In: International Conference on Intelligence Analysis.

[4] Sen, P., Namata, G., Bilgic, M., Getoor, L., Gallagher, B., Eliassi-Rad, T. (2008): Collective Classification in Network Data. AI Magazine 93-106.

[5] Zhu, X.: Semi-supervised learning literature survey (2008): Technical Report 1530, Department of Computer Science, University of Wisconsin at Madison.

[6] Zhou, D., Bousquet, O., Lal, T., Weston, J., & Scholkopf, B. (2004): Learning with local and global consistency. Advances in Neural Information Processing Systems 16. MIT Press, Cambridge, MA.

[7] Macskassy, S.A., Provost, F. (2007): Classification in Networked Data: A toolkit and a univariate case study. Journal of machine learning research. Vol. 8. pp: 935-983.

[8] Bilgic, M., Getoor, L. (2010): Active inference for collective classification. Proceedings of 24-th AAAI conference on Artificial Intelligence.

A Co-training Method for Linked Data Classification

Nguyễn Việt Tân[1], Hoàng Vũ[2], Đặng Vũ Tùng[3], Từ Minh Phương[4]

[1]*VNU University of Engineering and Technology, E3 Building, 144 Xuân Thủy, Cầu Giấy, Hanoi, Vietnam*

[2]*VNU The Information Technology Institute, E3 Building, 144 Xuân Thủy, Cầu Giấy, Hanoi, Vietnam*

[3]*Vietnam Youth Academy, 5 Chua Lang Street, Dong Da District, Hanoi, Vietnam*

[4]*Posts and Telecommunications Institute of Technology, 122 Hoàng Quốc Việt, Cầu Giấy, Hanoi, Vietnam*

Abstract: In some automatic classification applications, data points can be represented not only by vectors but also by linked structures or linked data describing the relationship among objects such as: Hyperlinks-linked websites, references-cited scientific papers, physical networks and so on. A critical requirement for classification methods is to employ and combine linked data with other information to achieve more accurate prediction results. To solve this problem, graph-based methods have been

proposed such as the Gaussian-field classifier, Hopfield networks, neighbor-based classifiers and so on. In the paper, we propose a co-training method to solve the problem of combining linked data with other information. In the proposed method, links are considered as another view of data. The proposed method was tested on the WebKB dataset. Experimental results and the comparative evaluation shown that the proposed method achieves the better results and higher accuracy than graph-based methods when tested on linked datasets.

Keywords: Networked data, linked data, co-training.