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.2017.00059

MỘT GIẢI PHÁP QUẢN LÝ DỮ LIỆU THAM GIA PHÂN LỚP TRONG MÔ HÌNH HỌC BÁN GIÁM SÁT

Phạm Anh Phương[1], Quách Hải Thọ[2]

[1]Khoa Tin học, Trường Đại học Sư phạm, Đại học Đà Nẵng
[2] Tổ Cơ sở ngành, Trường Đại học Nghệ thuật, Đại học Huế

paphuong@yahoo.com,haitho37@gmail.com

TÓM TẮT: *Mô hình học máy tiên tiến với phương pháp phân lớp bán giám sát đã thu hút sự quan tâm của nhiều nhà nghiên cứu, một vài nghiên cứu đã chỉ ra rằng trong một số trường hợp các phương pháp phân lớp bán giám sát thực hiện không hiệu quả bằng các phương pháp phân lớp có giám sát, từ đó làm giảm độ tin cậy trong các ứng dụng thực tế. Với mong muốn phát triển một phương pháp phân lớp bán giám sát có tính đến độ an toàn của tập dữ liệu huấn luyện để thực hiện phân lớp tốt hơn so với phương pháp có giám sát, bài báo này đề xuất một giải pháp quản lý dữ liệu tham gia phân lớp đối với mô hình học bán giám sát bằng cách kiểm soát sự cân bằng giữa phân lớp bán giám sát và phân lớp có giám sát khi có sự tham gia của các dữ liệu không gán nhãn trong việc phân lớp. Các kết quả thực nghiệm cho thấy hiệu suất tổng thể của giải pháp chúng tôi đề xuất có khả năng cạnh tranh để áp dụng với mô hình học bán giám sát.*

Từ khóa: *Semi-supervised classification; Semi-supervised improvement; Manifold; Clustering; least-square support vector machine (LS-SVM).*

I. GIỚI THIỆU

Ngày nay, hầu hết mọi ngành công nghiệp đang làm việc với hàm lượng lớn dữ liệu và đều nhận ra tầm quan trọng của công nghệ học máy, học máy gây nên cơn sốt công nghệ trên toàn thế giới trong vài thập niên gần đây. Trong lĩnh vực học máy có 2 hướng tiếp cận được chấp nhận rộng rãi đó là học có giám sát và học không giám sát, nhưng cũng có những hướng tiếp cận khác như học bán giám sát, học tăng cường. Trong đó mô hình học có giám sát thì các dữ liệu dùng để huấn luyện bắt buộc phải được gán nhãn trước, đây chính là một trong những nhược điểm của phương pháp này, bởi vì không phải lúc nào việc gán nhãn chính xác cho dữ liệu cũng dễ dàng. Còn với mô hình học không có giám sát thì ngược lại, các dữ liệu huấn luyện không được gán nhãn, do đó kết quả thu được có độ chính xác không cao. Công việc gán nhãn cho dữ liệu sẽ tốn nhiều công sức và thời gian, việc tạo ra nhãn cho những dữ liệu đòi hỏi sự nỗ lực lớn của con người. Học bán giám sát đã khắc phục được các nhược điểm và phát huy được ưu điểm của học có giám sát và học không có giám sát, bằng cách kết hợp giữa học có giám sát và học không có giám sát, với một lượng lớn dữ liệu chưa gán nhãn và một lượng hạn chế những dữ liệu đã được gán nhãn, bằng các giải thuật học bán giám sát sẽ thu được kết quả vừa có độ chính xác cao vừa mất ít thời gian và công sức. Do đó, học bán giám sát là một phương pháp học đạt được hiệu quả tốt trong lĩnh vực học máy và là phương pháp học thu hút sự quan tâm nghiên cứu mạnh mẽ trong suốt các thập kỷ qua.

Các phương pháp phân lớp bán giám sát cố gắng khai thác các thông tin phân phối dữ liệu nội tại được đưa ra bởi các dữ liệu không gán nhãn. Hầu hết các phương pháp áp dụng một hoặc hai giả định phổ biến là giả định cụm (cluster) và các giả định đa dạng (manifold) [7, 8, 11, 15] để khai thác các dữ liệu không gán nhãn.

Một vài nghiên cứu chỉ ra rằng, có phương pháp phân lớp bán giám sát mang lại hiệu suất thực hiện kém hơn khi phát triển từ phương pháp phân lớp có giám sát bằng cách sử dụng dữ liệu không có gán nhãn [9,12,13], như trong [13] đã phát triển S3VM_us từ S3VM với giải pháp phân cấp cụm bằng việc lựa chọn các dữ liệu không gán nhãn có độ tin cậy cao được TSVM dự báo và phần còn lại được dự báo với SVM tham gia vào tập huấn luyện cho việc phân lớp. Còn trong [12] đã phát triển phương pháp S4VM từ S3VM, kết quả cuối cùng thì S4VM có tính cạnh tranh cao so với TSVM và không bao giờ thấp hơn đáng kể so với SVM. Cả 2 phương pháp S3VM_us và S4VM thực sự được cải thiện dựa trên quy nạp S3VM. Trong nghiên cứu [14] tác giả đã đưa ra giả thuyết sửa đổi phân cụm bằng cách chia sẻ các thành viên lớp tương tự hơn là một nhãn lớp rõ ràng, từ đó đã phát triển một phương pháp phân lớp bán giám sát mới dựa trên các thành viên lớp.

Trong phần tiếp theo của bài báo này, chúng tôi tập trung vào một số phương pháp phân lớp bán giám sát đã được phát triển trong suốt các thập kỷ qua [7, 8, 10, 11] và đề xuất một cơ chế quản lý dữ liệu tham gia phân lớp bằng cách kiểm soát sự cân bằng giữa phân lớp bán giám sát và phân lớp có giám sát khi có sự tham gia của các dữ liệu không gán nhãn trong việc phân lớp.

Phần còn lại của bài báo này được cấu trúc như sau: Phần II giới thiệu phương pháp phân lớp bán giám sát dựa vào thành viên lớp. Phần III đề xuất giải pháp quản lý dữ liệu tham gia phân lớp trong mô hình học bán giám sát dựa vào thành viên lớp. Phần IV trình bày các kết quả thực nghiệm và đánh giá kết quả. Cuối cùng là phần kết luận và hướng phát triển.

II. PHƯƠNG PHÁP PHÂN LỚP BÁN GIÁM SÁT DỰA VÀO THÀNH VIÊN LỚP

Giả thiết phân cụm giả định là các trường hợp tương tự nên chia sẻ nhãn cùng lớp, điều này ngầm giả định rằng mỗi trường hợp nên có nhãn rõ ràng. Tuy nhiên, trong thực tế có một số trường hợp khó gán nhãn cho một lớp đơn, ví dụ như các trường hợp nằm ở biên. Trong những trường hợp như thế này thì giả thiết cụm không thể phản ánh đầy đủ phân bố dữ liệu thực và sẽ dẫn đến dự đoán không hiệu quả trong phân lớp bán giám sát. Do đó, trong nghiên cứu [14] tác giả đã đưa ra giả thiết sửa đổi phân cụm bằng cách chia sẻ các thành viên lớp tương tự hơn là một nhãn lớp rõ ràng. Đối với mỗi trường hợp, các thành viên lớp được biểu diễn như một vector, giá trị của từng yếu tố thể hiện khả năng của các trường hợp liên quan thuộc về lớp.

Áp dụng thông qua giả thiết sửa đổi cụm cũng như nguyên tắc học địa phương (hạn chế mỗi trường hợp và trọng số địa phương của nó là chia sẻ các vector thành viên cùng một nhãn), [14] đã phát triển một phương pháp phân lớp bán giám sát mới gọi là phân lớp bán giám sát dựa vào thành viên lớp.

Cụ thể như sau: Cho tập dữ liệu gán nhãn $X_l = \{x_i\}_{i=1}^{nl}$ với các nhãn tương ứng $Y = \{y_i\}_{i=1}^{nl}$ và tập dữ liệu không gán nhãn $X_u = \{x_j\}_{j=nl+1}^{n}$, trong đó $x_i \in R^d, y_i \in R^C$ cho phân lớp lớp C và $n_u = n - n_l$. Với hàm quyết định f(x) và một hàm thành viên nhãn v(x), thì phương pháp phân lớp bán giám sát dựa trên thành viên lớp được xây dựng như sau:

$$\min_{f,v_k(x_j)} \sum_{i=1}^{nl} \|f(x_i) - y_i\|^2 + \lambda_s \sum_{i=1}^{nl} \|f(\hat{x}_i) - y_i\|^2 + \sum_{k=1}^{C} \sum_{j=n_l+1}^{n} v_k(x_j)^2 \|f(x_j) - r_k\|^2 + \lambda_s \sum_{k=1}^{C} \sum_{j=n_l+1}^{n} v_k(x_j)^2 \|f(\hat{x}_j) - r_k\|^2 + \lambda \|f\|_{\mathcal{H}}^2 \quad (1)$$

với ràng buộc: $\sum_{k=1}^{C} v_k(x_j) = 1, \quad 0 \le v_k(x_j) \le 1, k = 1 \dots C, j = n_l + 1 \dots n$

Trong đó $\|.\|_{\mathcal{H}}$ là một chuẩn trong không gian Hilbert nhân tự sinh, $\{r_k\}_{k=1}^{C}$ là kiểu mã hóa cho các lớp C, $y_i \in R^C$ và $r_k \in R^C$ được mã hóa bởi một trong những luật C, tức là yếu tố k của y_i là 1 và yếu tố còn lại là 0 nếu x_i thuộc về lớp k, yếu tố k của r_k là 1 và các yếu tố còn lại là 0, $f(x_i) \in R^C$ và $v(x_i) \in R^C$ cho mỗi x_i, và $v_k(x_i)$ thể hiện xác suất của x_i thuộc lớp k, $\hat{x}_j$ là giá trị trung bình trọng số địa phương của x_i được xác định bởi:

$$\hat{x}_i = \frac{\sum_{x_j \in Ne(x_i)} S_{ij} x_j}{\sum_{x_j \in Ne(x_i)} S_{ij}} \quad (2)$$

Trong đó $Ne(x_i)$ là k láng giềng gần nhất của x_i được đo bằng khoảng cách Euclide và S_{ij} là một đại lượng tỉ lệ nghịch với khoảng cách giữa x_i và $x_j \forall k = 1 \dots C, i = 1 \dots n$.

Trong phân lớp bán giám sát dựa vào thành viên lớp thì mỗi trường hợp có thể thuộc đa lớp với các thành viên lớp tương ứng, ngoài ra mỗi trường hợp và giá trị trung bình trọng số địa phương của nó chia sẻ cùng một vector thành viên nhãn. Vấn đề tối ưu có thể được giải quyết hiệu quả bằng chiến lược lặp đi lặp lại luân phiên, trong đó mỗi bước tạo ra một phương pháp giải có nghiệm kín. Sự hội tụ của quá trình giải được lặp đi lặp lại có thể được đảm bảo về mặt lý thuyết. Cuối cùng, phương pháp phân lớp bán giám sát dựa trên thành viên lớp đạt được giá trị cực đại cạnh tranh so với một số phương pháp phân lớp bán giám sát tiên tiến như TSVM, LapSVM và S3VM.

Cũng giống như các phương pháp phân lớp bán giám sát khác, phương pháp phân lớp bán giám sát dựa trên thành viên lớp có thể mang lại hiệu suất kém hơn so với phương pháp có giám sát LS-SVM. Vì vậy, trong bài báo này, chúng tôi thảo luận và đề xuất một giải pháp quản lý các thành viên tham gia phân lớp đối với phương pháp phân lớp bán giám sát dựa trên thành viên lớp. Tuy nhiên, trong tính toán của mỗi giá trị trung bình trọng số địa phương có rủi ro tiềm ẩn [2] rằng các trường hợp từ lớp ngược lại có thể lựa chọn trong k láng giếng gần nhất và do đó giá trị trung bình trọng số địa phương thu được có thể rơi vào lớp ngược lại. Để tránh rủi ro như vậy ngay từ đầu của mô hình mà chúng tôi đưa ra, chúng tôi loại bỏ các số hạng liên quan đến giá trị này từ (1) và đơn giản hóa vấn đề tối ưu của phương pháp phân lớp bán giám sát dựa trên thành viên lớp như sau:

$$\min_{f,v_k(x_j)} \|f\|_{\mathcal{H}}^2 + \lambda_1 \sum_{i=1}^{nl} \|f(x_i) - y_i\|^2 + \lambda_2 \sum_{k=1}^{C} \sum_{j=n_l+1}^{n} v_k(x_j)^2 \|f(x_j) - r_k\|^2 \quad (3)$$

với ràng buộc: $\sum_{k=1}^{C} v_k(x_j) = 1, 0 \le v_k(x_j) \le 1, k = 1 \dots C, j = n_l + 1 \dots n$

III. GIẢI PHÁP QUẢN LÝ DỮ LIỆU THAM GIA PHÂN LỚP

Trong phần này, chúng tôi đề xuất giải pháp quản lý dữ liệu tham gia phân lớp (gọi là OPTMEM), bao gồm: Tối ưu thuật toán và giá trị của tham số λ.

Kết quả mà chúng tôi đề xuất sẽ có dự báo cuối cùng là sự kết hợp giữa phương pháp phân lớp bán giám sát dựa trên thành viên lớp khi các dữ liệu không gán nhãn là có lợi cho việc học và gần với dự báo của LS-SVM khi dữ liệu không có gán nhãn không có lợi cho việc học. Để sử dụng các dự báo của LS-SVM, chúng tôi gọi hàm quyết định của LS-SVM là $g(x)$, sau đó với hàm quyết định $f(x)$ và hàm nhãn thành viên $v(x)$ thì thiết lập đối với OPTMEM như sau:

$$\min_{f,v_k(x_j)} \|f\|_{\mathcal{H}}^2 + \lambda_1 \sum_{i=1}^{n} \|f(x_i) - y_i\|^2 + \lambda_2 \sum_{k=1}^{C} \sum_{j=n_l+1}^{n} v_k(x_j)^2 \left\|f(x_j) - r_k\right\|^2 + \left(\frac{1}{\lambda} - 1\right) \sum_{j=n_l+1}^{n} \left\|f(x_j) - g(x_j)\right\|^2 \quad (4)$$

với ràng buộc: $\sum_{k=1}^{C} v_k(x_j) = 1, 0 \le v_k(x_j) \le 1, k = 1 \dots C, j = n_l + 1 \dots n$

Ba số hạng đầu tiên trong hàm mục tiêu (4) là tìm kiếm hàm quyết định $f(x)$ và hàm nhãn thành viên $v(x)$, đồng thời khai thác cả dữ liệu có gán nhãn và dữ liệu không gán nhãn như phương pháp dựa vào thành viên lớp thực hiện, trong khi đó số hạng cuối cùng được sử dụng để kiểm soát sự khác biệt của các dự báo (khi sử dụng dữ liệu không gán nhãn) đưa ra với LS-SVM tương ứng. Điều này dẫn đến dự báo của (4) đưa ra trở thành cân bằng giữa những dự báo của (3) và LS-SVM tương ứng. Giá trị tham số λ được điều chỉnh trong khoảng [0,1] theo các dữ liệu không có gán nhãn có sẵn, khi giá trị λ gần đến 0 để giá trị $1/\lambda$ là vô cực thì dự báo của (4) sẽ suy biến cũng như dự báo của LS-SVM, và khi λ gần đến 1 thì dự báo của (4) sẽ suy biến như dự báo của (3). Do đó, giá trị của tham số λ là quan trọng trong việc kiểm soát sự cân bằng của OPTMEM, chúng tôi sẽ thảo luận việc xác định tham số λ như thế nào? trong phần tiếp theo.

Trong OPTMEM, vấn đề tối ưu là hai mặt lồi trong (f, v) và chúng tôi sử dụng giải pháp lặp đi lặp lại xen kẽ có đảm bảo hội tụ [3]. Mỗi bước lặp tối ưu mang lại một giải pháp có hình thức đóng cho cả $f(x)$ và $v(x)$.

Để cố định $v(x)$, vấn đề tối ưu của OPTMEM có thể được viết lại như sau:

$$\min_{f,v_k(x_j)} \|f\|_{\mathcal{H}}^2 + \lambda_1 \sum_{i=1}^{n_l} \|f(x_i) - y_i\|^2 + \lambda_2 \sum_{k=1}^{C} \sum_{j=n_l+1}^{n} v_k(x_j)^2 \left\|f(x_j) - r_k\right\|^2 + \left(\frac{1}{\lambda} - 1\right) \sum_{j=n_l+1}^{n} \left\|f(x_j) - g(x_j)\right\|^2 \quad (5)$$

Theo [5] thì cực tiểu của (5) có dạng $f(x) = \sum_{i=1}^{n} \alpha_i K(x_i, x)$, trong đó $\alpha_i \in R^{C \times 1}$

Từ đó (5) có thể được viết lại như sau:

$$min_\alpha M_1 = tr(\alpha K \alpha^T) + \lambda_1 tr((\alpha K_l - Y)(\alpha K_l - Y)^T) + \lambda_2 \sum_{k=1}^{C} tr\left((\alpha K_u - L_k)\hat{V}_k(\alpha K_u - L_k)^T\right) + \left(\frac{1}{\lambda} - 1\right) tr((\alpha K_u - \alpha_0 K_u)(\alpha K_u - \alpha_0 K_u)^T) \quad (6)$$

Trong đó $\alpha = [\alpha_1, \alpha_2, \dots, \alpha_n] \epsilon R^{C \times n}$ là ma trận nhân tử Lagrange và α_0 là ma trận nhân tử Lagrange cho LS-SVM. $K = [K_l K_u] = \begin{bmatrix} K_u & K_{lu} \\ K_{ul} & K_{uu} \end{bmatrix}$ là ma trận Kernel, trong đó $K_u = \langle \phi(X_l), \phi(X_l) \rangle_{\mathcal{H}}$, $K_{lu} = \langle \phi(X_l), \phi(X_u) \rangle_{\mathcal{H}}$ và $K_{uu} = \langle \phi(X_u), \phi(X_u) \rangle_{\mathcal{H}}$. L_k là ma trận $C \times n_u$ với dòng thứ k là một vector tất cả đều bằng 1 và những dòng còn lại là các vector bằng 0. Cho V_k là vector nhãn thành viên tương ứng với lớp thứ k, và $\hat{V}_k$ được định nghĩa là một ma trận đường chéo với các thành phần trên đường chéo là các giá trị bình phương của các phần tử trong V_k.

Bây giờ, cho đạo hàm của M_1 bằng 0 đối với α, chúng ta có như sau:

$$\frac{\partial M_1}{\partial \alpha} = \alpha K + \lambda_1 (\alpha K_l - Y) K_l^T + \lambda_2 \sum_{k=1}^{C} (\alpha K_u - L_k) \hat{V}_k K_u^T + \left(\frac{1}{\lambda} - 1\right)\left((\alpha K_u - \alpha_0 K_u) K_u^T\right) = 0 \quad (7)$$

Từ đó đưa đến giải pháp sau:

$$\alpha = \left(\lambda_1 K_l^T Y + \lambda_2 \sum_{k=1}^{C} K_u \hat{V}_k L_k + \left(\frac{1}{\lambda} - 1\right) \alpha_0 K_u K_u^T\right) x \left(\lambda_1 K_l^T K_l + \lambda_2 \sum_{k=1}^{C} K_u \hat{V}_k K_u^T + K + \left(\frac{1}{\lambda} - 1\right) K_u K_u^T\right)^{-1} \quad (8)$$

Để cố định $f(x)$, thì vấn đề tối ưu của OPTMEM được ghi lại như sau:

$$\min_{v_k(x_j)} \sum_{k=1}^{C} \sum_{j=n_l+1}^{n} v_k(x_j)^2 \left\|f(x_j) - r_k\right\|^2 \quad (9)$$

với ràng buộc: $\sum_{k=1}^{C} v_k(x_j) = 1$ và $0 \le v_k(x_j) \le 1, k = 1 \dots C, j = n_l + 1 \dots n$

Sử dụng phương pháp nhân tử Lagrange, chúng ta định nghĩa như sau:

$$M_2 = \sum_{k=1}^{C} \sum_{j=n_l+1}^{n} v_k(x_j)^2 \left\|f(x_j) - r_k\right\|^2 - \lambda_j \left(\sum_{k=1}^{C} v_k(x_j) - 1\right) \quad (10)$$

Tương tự như vậy, cho đạo hàm M_2 bằng 0 với mỗi $v_k(x_j), \forall k = 1 \dots C, j = 1 \dots n_u$, chúng ta có như sau:

$$\frac{\partial M_2}{\partial v_k(x_j)} = 2\left\|f(x_j) - r_k\right\|^2 - \lambda_j = 0 \quad (11)$$

Vì vậy
$$v_k(x_j) = \lambda_j / 2\left\|f(x_j) - r_k\right\|^2 \quad (12)$$

Hơn nữa, kết hợp các ràng buộc $\sum_{k=1}^{C} v_k(x_j) = 1$

chúng ta có như sau: $v_k(x_j) = \frac{1/\|f(x_j) - r_k\|^2}{\sum_{k=1}^{C} 1/\|f(x_j) - r_k\|^2}$ trong đó $k \in \{1 \dots C\}, i \in \{1 \dots n\}$ (13)

Vì vậy, đối với mỗi phiên bản x tùy ý

$$v_k(x) = \frac{1/\|f(x)-r_k\|^2}{\sum_{k=1}^{C} 1/\|f(x)-r_k\|^2} \quad \text{trong đó} \quad k \in \{1 \dots C\} \tag{14}$$

Như trong phương pháp phân lớp dựa trên thành viên lớp, dự báo cho mỗi phiên bản đã cho trong OPTMEM có thể được thực hiện không chỉ bởi hàm quyết định $f(x)$, mà còn hàm thành viên lớp $v(x)$, phản ánh các khả năng có thể xảy ra của phiên bản đó đến các lớp cá nhân. Như chúng ta thấy ở mệnh đề sau đây, khác với các dự báo không nhất quán có thể được dẫn ra bởi 2 hàm $f(x)$, $v(x)$ như thế trong phương pháp phân lớp dựa trên thành viên lớp, 2 dự báo trong OPTMEM luôn luôn nhất quán.

Mệnh đề: Các dự báo cho mỗi mẫu được cho bởi hàm quyết định và hàm thành viên lớp luôn luôn nhất quán [14].

Chứng minh: Cho phiên bản x_i tùy ý, nhãn lớp của nó do hàm quyết định dự báo là $\hat{y}_i = max_{k=1\dots C} f_k(x_i)$, do đó $x_i \in X_k$ có nghĩa là $f_k(x_i) > f_j(x_i), \forall j = 1 \dots C, j \neq k$, trong đó X_k là tập hợp các phiên bản phụ thuộc vào cụm thứ k. Trong khi đó nhãn lớp của nó do hàm thành viên lớp dự báo $\tilde{y}_i = max_{k=1\dots C} v_k(x_i)$, vì vậy từ (14), $x_i \in X_k$ có nghĩa là $\|f(x_i) - r_k\|^2 < \left\|f(x_i) - r_j\right\|^2$, thì $f(x_i)^T r_k > f(x_i)^T r_j$, hoặc là tương đương $f_k(x_i) > f_j(x_i), \forall j = 1 \dots C, j \neq k$. Do đó, các điều kiện dự báo cho $x_i \in X_k$ bởi cả hai $f(x)$ và $v(x)$ là tương đương, vì vậy hai dự báo là không mâu thuẫn nhau.

III.1 Mô tả thuật toán

Vấn đề tối ưu của OPTMEM theo chiến lược luân phiên lặp đi lặp lại. Các giá trị ban đầu đối với các thành viên nhãn của các phiên bản không có gán nhãn trong OPTMEM có thể đạt được bởi một số chiến lược như gán ngẫu nhiên, một số kỹ thuật phân cụm mờ như FCM, hoặc đơn giản là thiết lập tất cả chúng bằng 0. Sự lặp đi lặp lại kết thúc khi $|M_k - M_{k-1}| < \varepsilon M_{k-1}$, trong đó M_k là giá trị hàm mục tiêu tại lần lặp k và ε là ngưỡng xác định trước. Thuật toán cụ thể của OPTMEM được trình bày như sau.

Input:

X_l: tập dữ liệu được gán nhãn

Y_l: nhãn của X_l

X_u: tập dữ liệu chưa gán nhãn

$\lambda, \lambda_1, \lambda_2$: tham số chính tắc

ε: tham số bước lặp

σ: tham số kernel

Maxiter: số lượng lần lặp tối đa

Output:

$f(x)$: hàm quyết định

$v(x)$: hàm nhãn thành viên

Method:

Khởi tạo các nhãn thành viên cho dữ liệu không có nhãn.

Thiết lập giá trị hàm mục tiêu ban đầu đến vô cực, nghĩa là: $M_0 = INF$

For k= 1…Maxiter

Cập nhật α bằng công thức (8), và $f(x)$ bởi định lý và nhận được α

Cập nhật $v(x)$ bởi (14);

Cập nhật giá trị hàm mục tiêu M_k

Nếu $|M_k - J_{k-1}| < \varepsilon M_{k-1}$

Dừng, trả về giá trị $f(x)$ và $v(x)$

End if

End for.

Quá trình lặp lại luân phiên đạt tối ưu của OPTMEM được đảm bảo về mặt lý thuyết để hội tụ và minh chứng chi tiết trong [14].

III.2 Giá trị λ.

Giá trị λ trong OPTMEM điều khiển dự báo giữa những dự báo bởi phương pháp thành viên lớp và LS-SVM tương ứng. Tiếp theo sau đây, chúng ta lựa chọn giá trị của nó từ [0, 1], thực sự đây là một bài toán phụ thuộc dữ liệu. Khi có đủ các phiên bản có gán nhãn, thì chúng ta có thể chấp nhận các chiến lược lựa chọn tham số điển hình như kiểm chứng chéo. Khi không có đủ các phiên bản có gán nhãn thì kiểm chứng chéo đó không hợp lệ [11, 12, 13, 14] và vì vậy chúng ta chấp nhận chiến lược tập hợp như trong [16]. Cụ thể, bài toán lựa chọn λ có thể đơn giản hóa bằng cách chọn giá trị tốt nhất của nó từ $\{\lambda_1, \lambda_2, \dots, \lambda_m\}$, trong đó $\lambda_1 = 0$ và $\lambda_m = 1$ hoặc tương đương, chọn lựa hàm quyết định tốt nhất f từ $\{f_1, f_2, \dots, f_m\}$, trong đó f_1 và f_m là các hàm quyết định đạt được tương ứng từ LS-SVM và phương pháp thành viên lớp. Từ chiến lược tập hợp đối với học bán giám sát [4, 16], tối ưu hóa f được biểu diễn như một kết hợp tuyến tính của các hàm cơ bản $\{f_i\}_{i=1}^m$ nghĩa là $f = \sum_{i=1}^m \omega_i f_i, \sum_{i=1}^m \omega_i = 1, \omega_i \geq 0, i = 1 \dots m$. Cuối cùng, cho $\{\lambda_i\}_{i=1}^m$, chúng ta có thể đạt được $\{f_i\}_{i=1}^m$ tương ứng, sau đó bài toán được chuyển đổi để tìm kiếm tập hợp các trọng số kết hợp $\{\omega_1, \omega_2, \dots, \omega_m\}$, như sau:

$$\min_{\omega_k} \frac{1}{2}\eta(FR_\omega)^T L(FR_\omega) - \vec{1}_C^T(Y_l F_l R_\omega) \tag{15}$$

với ràng buộc: $\omega^{\mathrm{T}}\mathrm{e} = 1, \omega_{\mathrm{k}} \geq 0$

Trong đó $\omega = [\omega_1, \omega_2, \dots, \omega_m]^T, e \in R^{m\times 1}$ là một vector với các thành phần bằng 1. $F = [f^1(X)^T, f^2(X)^T, \dots, f^m(X)^T] \in R^{n\times(C\times m)}$ và $F_l = [f^1(X_l)^T, f^2(X_l)^T, \dots, f^m(X_l)^T] \in R^{l\times(C\times m)}$ là ma trận dự báo cho toàn thể tập dữ liệu và toàn thể tập dữ liệu có gán nhãn tương ứng. $Y_l \in R^{C\times l}$ là ma trận nhãn đối với dữ liệu có gán nhãn. $R = [R^1 R^2 \dots R^m]^T \in R^{(C\times m)\times m}$, mỗi $R^k \in R^{m\times C}$trong đó hàng k là một vector mà tất cả bằng 1 và các hàng còn lại là các vector bằng 0. $\vec{1}_C \in R^{C\times 1}$ với các yếu tố phân tử bằng 1. L là công thức đồ thị Laplace khi $\mathrm{L} = \mathrm{D} - \mathrm{W}$, trong đó $W = \left[\omega_{ij}\right]_{n\times n}$ là ma trận trọng số trên đồ thị và D là ma trận chéo với mục chéo con i là $D_{ii} = \sum_{j=1}^n W_{ij}$, η là tham số chính tắc.

Số hạng đầu tiên trong hàm mục tiêu (15) đảm bảo thông suốt trên đồ thị Laplace [5] và số hạng thứ 2 đảm bảo các dự báo đúng đối với dữ liệu có gán nhãn. Tối ưu hóa của (15) là một bài toán phương trình bậc 2, có thể được giải hiệu quả bằng bất kỳ phương trình bậc 2 nào. Thông qua việc chấp nhận chiến lược tập hợp, bài toán thiết lập λ được chuyển đổi để tối ưu hệ số kết hợp tuyến tính $\{\omega_1, \omega_2, \dots, \omega_m\}$ đối với tập hợp các hàm biệt thức $\{f_i\}_{i=1}^m$ đối với $\{\lambda_i\}_{i=1}^m$, mà phần nào tương tự với học nhiều kernel [6], trong đó lựa chọn cho các tham số tối ưu của một kernel đơn được phân bổ đến lựa chọn các hệ số kết hợp tuyến tính của nhiều kernel.

IV. THỰC NGHIỆM

Trong phần thực nghiệm, chúng tôi minh họa cho OPTMEM đối với phân lớp bán giám sát là như thế nào và chúng tôi tiến hành đánh giá hiệu suất của OPTMEM trên nhiều bộ dữ liệu thực với phương pháp có giám sát LS-SVM, TSVM và các phương pháp có sẵn như S3VM_us [13] và S4VMs [12].

Đối với phương pháp dựa vào thành viên lớp, chúng tôi sử dụng (3) để tiến hành so sánh trực tiếp. Còn đối với S4VM, chúng tôi áp dụng các phiên bản với mẫu đại diện đạt hiệu quả tính toán hơn [12]. Phương pháp bán giám sát mà chúng tôi đề xuất có thể dự đoán tốt các dữ liệu không gán nhãn bởi các hàm quyết định. Vì vậy, khi tiến hành thực nghiệm để so sánh với các phương pháp truyền dẫn S3VM_us và S4VM_us, thì chúng tôi thực hiện theo kiểu truyền dẫn, nghĩa là sẽ học trên cả dữ liệu gán nhãn, không gán nhãn và sau đó sẽ dự đoán trên những dữ liệu không gán nhãn nhất định.

IV.1 Dữ liệu thực nghiệm

Chúng tôi thực hiện đánh giá trên cả 2 bộ dữ liệu là UCI và bộ dữ liệu chuẩn, được thể hiện trong bảng sau:

Bảng 1. Các tập dữ liệu cho thực nghiệm

STT	Tập dữ liệu	Số mẫu (n)	Số chiều (d)
1	House	232	16
2	Heart	270	9
3	Vehicle	435	16
4	WDBC	569	14
5	Isolet	600	51
6	Austra	690	15
7	Optdigits	1143	42
8	BCI	400	117
9	Digitl	1500	241
10	USPS	1500	241

Trong đó, mỗi bộ dữ liệu UCI được chia ngẫu nhiên thành 2 phần, một phần dành cho việc huấn luyện và phần còn dùng để thử nghiệm, trong phần dành cho việc huấn luyện chỉ có 100 trường hợp được gán nhãn và phần còn lại

không gán nhãn. Quá trình này cùng với việc học phân lớp được tiến hành lặp đi lặp lại 30 lần, sau đó lấy giá trị trung bình để kiểm tra và lập báo cáo. Cả hai hàm Linear Kernel và Gauss Kernel đều được sử dụng ở đây, các tham số chính tắc C_1 và C_2 được thiết lập cố định là 1 và 0.1, tương ứng như vậy các giá trị λ_1, λ_2 và η có giá trị cố định lần lượt là 100,1 và 1, giá trị ε là 10^{-3}, thông số chiều rộng σ trong Gauss kernel được thiết lập là khoảng cách trung bình giữa các cặp thực thể.

Còn đối với các bộ dữ liệu chuẩn, chúng tôi lấy ra từ các thực nghiệm trong [7] và [14]. Hơn nữa, đối với mỗi bộ dữ liệu và mỗi thiết lập, có 15 tập con của dữ liệu được dán nhãn và cuối cùng giá trị thực hiện trung bình trên các dữ liệu không gán nhãn được dùng để báo cáo. Các tham số chính tắc C_1 và C_2 trong các phương pháp so sánh đưa ra được thiết lập giá trị từ 100 và 0.1, các giá trị λ_1, λ_2 và η có giá trị cố định là 100, 0.1 và 1, giá trị ε cũng được thiết lập là 10^{-3}. Cả hai hàm Linear Kernel và Gauss Kernel đều được sử dụng ở đây, thông số chiều rộng σ trong Gauss kernel được thiết lập là khoảng cách trung bình giữa các cặp thực thể và 100 trường hợp được gán nhãn thì ký hiệu là δ và lựa chọn chéo 5 giá trị $\{0.25, 0.5, 1, 2, 4\}$ cho δ trên các dữ liệu gán nhãn dùng để huấn luyện.

Ngoài ra, các giá trị của λ được chọn thống nhất từ khoảng [0,1] với khoảng cách là 0.2, tức là giá trị λ như sau: [0, 0.2, 0.4, 0.6, 0.8, 1].

IV.2 Đánh giá kết quả thực nghiệm

Bảng 2 và Bảng 3 thể hiện việc thực hiện 100 trường hợp có gán nhãn thông qua các hàm Linear Kernel và GaussKernel.

Bảng 2. So sánh hiệu suất với hàm Linear Kernel

STT	Tập dữ liệu	LS-SVM	TSVM	S3VM_us	S4VM_s	OPTMEM
1	House	73.1	71.5	71.6	71.5	76.8
2	Heart	85.8	85.1	85.1	85.5	86.2
3	Vehicle	83.8	81.9	82.9	82.0	84.7
4	WDBC	58.5	57.7	59.1	59.3	58.4
5	Isolet	88.9	86.7	89.6	90.6	89.5
6	Austra	76.3	80.0	75.0	75.2	75.6
7	Optdigits	91.4	87.3	91.7	91.4	93.3
8	BCI	63.5	58.2	69.5	70.1	70.3
9	Digitl	95.7	83.6	95.6	94.9	96.0
10	USPS	73.8	70.6	73.6	74.8	77.1
Hiệu suất trung bình		**79.08**	**76.26**	**79.37**	**79.53**	**80.79**

Bảng 3. So sánh hiệu suất với hàm Gauss Kernel

STT	Tập dữ liệu	LS-SVM	TSVM	S3VM_us	S4VM_us	OPTMEM
1	House	96.7	96.6	96.4	96.6	97.0
2	Heart	87.8	81.4	85.7	85.9	87.6
3	Vehicle	84.7	80.0	79.4	80.5	86.9
4	WDBC	65.3	66.5	66.0	66.7	65.9
5	Isolet	91.9	81.6	89.6	90.7	92.2
6	Austra	68.6	65.8	67.2	67.2	69.8
7	Optdigits	94.1	88.2	92.5	91.9	94.5
8	BCI	87.2	87.5	87.5	87.2	89.2
9	Digitl	95.1	89.8	94.9	94.8	96.5
10	USPS	75.1	71.2	72.6	73.4	75.2
Hiệu suất trung bình		**84.65**	**80.86**	**83.18**	**83.49**	**85.48**

Từ Bảng 2 và Bảng 3, chúng ta có thể có một số quan sát khi tính hiệu suất trung bình trên các bộ dữ liệu kiểm tra như sau:

a) Các trình diễn tổng thể của LS-SVM và TSVM là có thể so sánh được. Tuy nhiên, các trình diễn tổng thể của LS-SVM tốt hơn các trình diễn của TSVM (dựa trên SVM), vì giả định phân cụm được thay đổi có thể thu nạp lại phân phối dữ liệu thực tốt hơn giả định phân cụm. Đó cũng chính là lý do chúng tôi đề xuất giải pháp quản lý thành viên tham gia phân lớp trong phương pháp phân lớp bán giám sát dựa trên thành viên lớp, mặc dù cơ chế như vậy cũng có thể được áp dụng tương tự với các phương pháp phân lớp bán giám sát khác như TSVM.

b) Cơ hội thực hiện suy biến trong S3VM_us nhỏ hơn nhiều so với TSVM. Đồng thời, hiệu suất tổng thể của S4VM có tính cạnh tranh cao so với các hiệu suất tổng thể của TSVM. Do đó, S3VM_us và S4VM_us có thể được áp dụng cho phân lớp bán giám sát có tính đến giải pháp quản lý các thành viên tham gia phân lớp.

c) Dự kiến các phương pháp bán giám sát có tính giải pháp quản lý các thành viên tham gia phân lớp có thể thực hiện không kém hơn các phương pháp có giám sát và đồng thời thực hiện không kém hơn các phương pháp bán giám sát ban đầu.

d) Hiệu suất tổng thể của giải pháp chúng tôi đề xuất tốt hơn hiệu suất tổng thể của cả S3VM_us và S4VM_us, cho thấy khả năng cạnh tranh cao của giải pháp này để phân lớp bán giám sát có tính đến giải pháp quản lý các thành viên tham gia phân lớp.

V. KẾT LUẬN

Trong một số trường hợp, phương pháp phân lớp bán giám sát có thể mang lại hiệu suất kém hơn phương pháp phân lớp có giám sát khi sử dụng các dữ liệu không có nhãn. Do đó, sẽ làm giảm độ tin cậy khi áp dụng các phương pháp phân lớp bán giám sát vào thực tế. Để giải quyết vấn đề này mà không làm giảm hiệu suất khi sử dụng phương pháp phân lớp bán giám sát, chúng tôi đã đề xuất một giải pháp quản lý các dữ liệu tham gia phân lớp bằng cách kiểm soát sự cân bằng có khả năng thích nghi giữa học có giám sát và học bán giám sát đối với dữ liệu có sẵn không có nhãn. Cuối cùng, kết quả thực nghiệm trên nhiều bộ dữ liệu thực cho thấy hiệu suất tổng thể của giải pháp này không kém hơn hơn hiệu suất tổng thể của các phương pháp đưa ra đánh giá. Đồng thời, hiệu quả tính toán của giải pháp này cạnh tranh so với các phương pháp phân lớp bán giám sát có sẵn.

Giá trị của tham số λ là một vấn đề quan trọng trong giải pháp quản lý các dữ liệu tham gia phân lớp giữa phương pháp học có giám sát và phương pháp học bán giám sát. Trong bài báo này, nó có thể được lựa chọn bằng cách xác nhận chéo khi có đủ dữ liệu có gán nhãn. Tuy nhiên, tìm kiếm các giá trị tối ưu cho các tham số trong phương pháp học bán giám sát vẫn còn là một vấn đề mở đáng nghiên cứu và đó là vấn đề hoàn toàn cần thiết để tiếp tục nghiên cứu trong tương lai.

TÀI LIỆU THAM KHẢO

[1] B. Pfahringer (2006), “A semi-supervised spam mail detector” in *Proc. 17th Eur. Conf. Mach. Learn. 10th Eur. Conf. Principles Pract. Knowl.Discovery Databases*, pp. 1-5.

[2] C. J. Taylor (2013), “Towards fast and accurate segmentation”in *CVPR*, pages 1916-1922.

[3] J. Wang, J. Yang, K. Yu, F. Lv, T. Huang, and Y. Gong (2010), “Locality-constrained linear coding for image classification”in *CVPR*, pp. 3360-3367.

[4] K. Chen and S. Wang (2011), “Semi-supervised learning via regularized boosting working on multiple semi-supervised assumptions”,Published in: *IEEE Transactions on Pattern Analysis and Machine Intelligence* (Volume: 33, Issue: 1, pp. 129 - 143).

[5] M. Belkin, P. Niyogi, and V. Sindhwani (2006), “Manifold regularization:A geometric framework for learning from labeled and unlabeled examples,” *J. Mach. Learn. Res.*, vol. 7, no. 1, pp. 2399-2434.

[6] M. Gonen and E. Alpaydin (2011), “Multiple kernel learning algorithms,” *J.Mach. Learn. Res.*, vol. 12, pp. 2211-2268.

[7] O. Chapelle, B. Scholkopf, and A. Zien (2006), *Semi-Supervised Learning*. Cambridge, MA, USA: MIT Press.

[8] P. K. Mallapragada, R. Jin, A. K. Jain, and Y. Liu (2009), “Semi-boost:Boosting for semi-supervised learning”, *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 31, no. 11, pp. 2000-2014.

[9] T. Yang and C. E. Priebe (2011), “The effect of model misspecification on semi-supervised classification,” *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 33, no. 10, pp. 2093-2103.

[10] X. Liu, M. Song, D. Tao, Z. Liu, L. Zhang, C. Chen, and J. Bu (2013), *Semi-supervised node splitting for random forest construction*. In CVPR, pages 492-499.

[11] V. Jothi Prakash , Dr. L.M. Nithya (2014), "A Survey On Semi-Supervised Learning Techniques" in *International Journal of Computer Trends and Technology (IJCTT)* - vol.8, no.1, pp25-29.

[12] Y.-F. Li and Z.-H. Zhou (2011), “Towards making unlabeled data never hurt,” in *Proc. 28th Int. Conf. Mach. Learn*, pp. 1081-1088.

[13] Y.-F. Li and Z.-H. Zhou (2011), “Improving semi-supervised support vector machines through unlabeled instances selection,” in *Proc. 25th AAAI Conf. Artif. Intell*, pp. 500-505.

[14] Y. Wang, S. Chen, and Z.-H. Zhou (2012), “New semi-supervised classification method based on modified cluster assumption,” *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 23, no. 5, pp. 689-702.

[15] Z.-H. Zhou and M. Li (2010), “Semi-supervised learning by disagreement”, *Knowl. Inf. Syst.*, vol. 24, no. 3, pp. 415-439.

[16] Z. Xu, R. Jin, I. King, M. R. Lyu, and Z. Yang (2009), “Adaptive regularization for transductive support vector machine,” in *Proc. Adv. Neural Inf.Process. Syst*, pp. 2125-2133.

A SOLUTION MANAGING CLASSIFICATION PARTICIPATION DATA IN SEMI-SUPERVISED LEARNING MODELS

Pham Anh Phuong, Quach Hai Tho

ABSTRACT: *Advanced machine learning model with semi-supervised classification has attracted the attention of many researchers. Some studies have shown that in some cases semi-supervised classification methods have performed not as effectively as those of supervised classification methods when using unlabeled data, which reduces reliability in practical applications. With the idea of developing a method that does not reduce its performance when applying the semi-supervised classification method, this paper proposes a solution managing classification participation data for semi-supervised learning models by controlling the balance having adaptability between semi-supervised and supervised subclassification involving unlabeled data in classification. The empirical results show that the overall performance of our proposed solution is competitive to apply to the semi-supervised learning model.*

Keywords: *Semi-supervised classification; Semi-supervised improvement; Manifold; Clustering; least-square support vector machine (LS-SVM).*