Tạp chí Khoa học Đại học Huế: Khoa học Tự nhiên; ISSN 1859–1388

Tập 126, Số 1A, 2017, Tr. 151–163; http://dx.doi.org/10.26459/hueuni-jns.v126i1A.4014

CƠ SỞ TOÁN HỌC CỦA MÁY VÉC TƠ HỖ TRỢ

Nguyễn Thế Cường*

Khoa Toán, Trường ĐH Thông tin Liên lạc.
101 Mai Xuân Thưởng, Nha Trang, Khánh Hòa.

Tóm tắt: Máy véc-tơ hỗ trợ là một kĩ thuật phân lớp rất triển vọng đã được Vapnik và các cộng sự phát triển vào những năm đầu thập niên 90 của thế kỷ hai mươi. Trong hơn hai thập kỷ qua, kỹ thuật này đã được áp dụng rất thành công vào việc xây dựng các chương trình nhận dạng mà được sử dụng cho nhiều lĩnh vực khác nhau của đời sống. Nhận thấy đây là một vấn đề thiết thực nên chúng tôi đã lựa chọn để nghiên cứu, nhằm mục tiêu tìm thêm các ứng dụng thực tiễn của thuật toán "máy véc tơ hỗ trợ". Từ khi được đề xuất đến nay, kỹ thuật này đã được mở rộng không ngừng. Đầu tiên là cho bài toán với tập dữ liệu tách được tuyến tính, sau đó đến trường hợp tập dữ liệu có chồng lấn hoặc phức tạp hơn khi dữ liệu các lớp trộn lẫn vào nhau. Trong tất cả các trường hợp, phương pháp nhân tử Lagrange vẫn tỏ ra hiệu quả để đưa bài toán về dạng tường minh nhất mà từ đó có thể đưa ra thuật toán giải. Bài báo này nhằm trình bày cơ sở toán học của kỹ thuật phân lớp dữ liệu cho các trường hợp từ đơn giản đến phức tạp.

Từ khoá: Support vector machines.

1 Giới thiệu

Trong bài báo này chúng tôi đề cập đến nguồn gốc toán học của kỹ thuật phân lớp máy véc-tơ hỗ trợ, kỹ thuật này đã được V.N. Vapnik giới thiệu năm 1995 [2]. Thực tế kỹ thuật này đã được áp dụng thành công trong nhiều bài toán như nhận dạng chữ viết tay của nhóm Vapnik, nhận dạng khuôn mặt của nhóm Girosi, hay bài toán phân loại văn bản... Thực chất giải quyết các bài toán dạng này là đưa về giải quyết bài toán tối ưu dạng:

$$\begin{cases} f(x) \to min \\ x \in M. \end{cases}$$

và sử dụng phương pháp nhân tử Lagrange để đưa bài toán về dạng tường minh mà từ đó đưa ra thuật toán giải. Chúng tôi sẽ trình bày kỹ thuật này qua từng trường hợp từ đơn giản đến phức tạp. Để hiểu tường minh hơn độc giả nên tham khảo về phương pháp nhân tử Lagrange ở [1], tiếp cận máy véc-tơ hỗ trợ ở [2] và một vài ứng dụng thực tế ở [7].

2 Bài toán phân lớp dữ liệu

Trong thực tế chúng ta thường gặp bài toán phân loại dữ liệu trên cơ sở một số mẫu thử cho trước. Cụ thể, cho q mẫu thử $x_1, x_2, \ldots, x_q \in \mathbb{R}^n$ tương ứng với các đầu ra $y_1, y_2, \ldots, y_q \in \Omega$,

* *Liên hệ: nckcbnckcb@gmail.com*

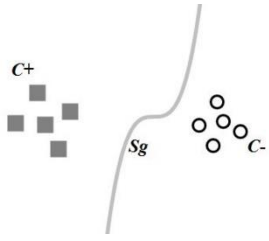
Nhận bài: 17–1–2017; Hoàn thành phản biện: 4–8–2017; Ngày nhận đăng: 7–8–2017

trong đó Ω là một tập hợp rời rạc cố định. Ở đây chúng ta chủ yếu xét bài toán phân hai lớp. Nghĩa là $\Omega = \{-1,1\}$. Lúc đó dữ liệu x_i được xếp vào lớp C^+ nếu tương ứng ta có $y_i = 1$, và được xếp vào lớp C^- nếu $y_i = -1$. Bài toán đặt ra là: Cần tìm một hàm phân lớp

$$f : \mathbb{R}^n \to \{-1,1\} thỏa mãn : f(x_i) = y_i, \forall i \in Q := \{1,2,\ldots,q\} \quad (1)$$

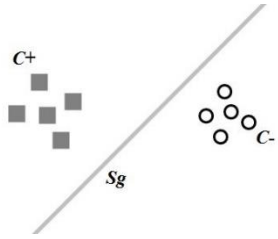
Để hiểu ý nghĩa của bài toán này ta xét ví dụ sau. Trong mùa dịch sốt xuất huyết, tại một trung tâm y tế X người ta đã tiếp nhận và đã có hồ sơ bệnh án đầy đủ của 100 bệnh nhân. Thông tin của mỗi bệnh nhân được số hóa thành một vec-tơ $x \in \mathbb{R}^9$ gồm các thành phần: giới tính, tuổi, cân nặng, nhiệt độ, huyết áp, đau đầu, đau bụng, buồn nôn (trong đó huyết áp gồm 2 số, giới tính, đau đầu, đau bụng, buồn nôn được số hóa bởi 0 hoặc 1). Chẳng hạn, $x = (1,25,57,39,90,150,1,1,0)$ biểu thị cho bệnh nhân nam, 25 tuổi, nặng 57 kg, thân nhiệt 39 $^\circ$ C, huyết áp 150/90, có triệu chứng đau đầu, đau bụng, nhưng không buồn nôn. Với mỗi bệnh nhân x_i như vậy sau một thời gian điều trị ta đã biết người ấy có bị sốt xuất huyết (và gán $y_i = 1$) hay không (gán $y_i = -1$). Trên cơ sở q = 100 mẫu thử như thế chúng ta cần thiết lập một hàm $f : \mathbb{R}^9 \to \{-1,1\}$ nhằm để chẩn đoán cho những bệnh nhân mới. Dĩ nhiên hàm f đó khi thực hiện trên các mẫu cũ phải cho chẩn đoán đúng, nghĩa là $f(x_i) = y_i$ với mọi $i \in \{1,2,\ldots,100\}$.

Trở lại bài toán (1). Nếu có một hàm $g : \mathbb{R}^n \to \mathbb{R}$ sao cho $g(x_i) > 0$ với mọi $x_i \in C^+$, và $g(x_i) < 0$ với mọi $x_i \in C^-$, thì hiển nhiên ta có thể chọn $f(x) = sgn(g(x))$ để làm hàm phân lớp. Lúc đó mặt mức $S_g = \{x \in \mathbb{R}^n \,/\, g(x) = 0\}$ được gọi là *mặt biên* hay *mặt quyết định* (Hình 1).



Hình. 1. Mặt quyết định

Mặt này chia $\mathbb{R}^n$ ra làm hai miền, tương ứng với hai lớp C^+ và C^-. Đặc biệt nếu g là hàm affine: $g(x) = w \cdot x + b$ ($w \in \mathbb{R}^n$, $b \in \mathbb{R}$; $w \cdot x$ là tích vô hướng của w và x) thì S_g là một siêu phẳng nhận w làm vec-tơ pháp mà ta gọi là *mặt quyết định tuyến tính* (Hình 2).



Hình 2. Mặt quyết định tuyến tính

Ngược lại, nếu g không phải là hàm affine, ta nói mặt quyết định là *phi tuyến*.

3 Hàm phân lớp tuyến tính

Trước hết ta xét trường hợp đơn giản nhất khi tập dữ liệu là *tách được tuyến tính*, nghĩa là tồn tại cặp $(w,b) \in \mathbb{R}^n \times \mathbb{R}$ sao cho

$$\begin{cases} w \cdot x_i + b \geq 1, & \forall x_i \in C^+, \\ w \cdot x_i + b \leq -1, & \forall x_i \in C^-. \end{cases} \quad (2)$$

Trong trường hợp này ta xét hàm phân lớp $f_{w,b}(x) = sgn(w \cdot x + b)$ và mặt quyết định là siêu phẳng

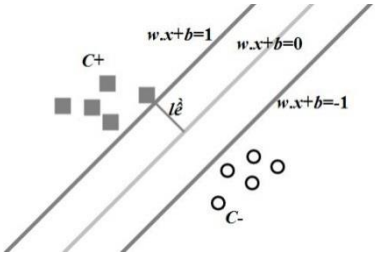
$$S_{(w,b)} = \{x \in \mathbb{R}^n \ / \ w \cdot x + b = 0\}.$$

Siêu phẳng này chia $\mathbb{R}^n$ ra làm hai nửa không gian, đặt C^+ và C^- về hai phía khác nhau. Thông thường, khi dữ liệu là tách được tuyến tính, sẽ có rất nhiều siêu phẳng tách và vì vậy cũng có nhiều hàm phân lớp. Nhiệm vụ của chúng ta là cần chọn ra một hàm phân lớp tốt nhất.

Lưu ý rằng với hệ số $\lambda > 0$ tùy ý ta có $S_{(w,b)} \equiv S_{(\lambda w, \lambda b)}$, và các hàm phân lớp $f_{w,b}$, $f_{\lambda w, \lambda b}$ là bằng nhau. Vì vậy để bảo đảm tính duy nhất của cặp (w,b) đối với một mặt quyết định ta đưa thêm ràng buộc:

$$\min_{i \in Q} | w \cdot x_i + b | = 1. \quad (3)$$

Siêu phẳng $S_{(w,b)}$, với (w,b) thỏa mãn (2) và (3), được gọi là *mặt quyết định chính tắc* (Hình 3).



Hình 3. Mặt quyết định chính tắc

Ta gọi lề của mặt quyết định chính tắc $S_{(w,b)}$ là khoảng cách ngắn nhất từ nó đến tập các điểm dữ liệu

$$\rho(w,b) = min\{d(x_i; S_{(w,b)}) : i \in Q\}.$$

Rõ ràng, hàm $f_{w,b}$ sẽ phân lớp dữ liệu tốt nhất khi lề $\rho(w,b)$ là lớn nhất. Ta đã biết công thức tính khoảng cách từ một điểm $x_i \in \mathbb{R}^n$ đến siêu phẳng $S_{(w,b)}$ là

$$d(x_i; S_{(w,b)}) = \frac{/\ w \cdot x_i + b\ /}{\| w \|}.$$

Do đó,

$$\rho(w,b) = min\left\{\frac{/\ w \cdot x_i + b\ /}{\| w \|} : i \in Q\right\} = \frac{min\{/\ w \cdot x_i + b\ /: i \in Q\}}{\| w \|} = \frac{1}{\| w \|}.$$

Đẳng thức cuối cùng có được là do (3). Vậy để mặt quyết định có lề cực đại thì $\| w \|$ phải cực tiểu. Chú ý rằng điều kiện (2) có thể viết lại là

$$y_i(w \cdot x_i + b) \geq 1;\ \forall i \in Q. \tag{4}$$

Tóm lại, để tìm mặt quyết định chính tắc có lề cực đại chúng ta cần giải bài toán tối ưu sau:

$$\begin{cases} \Phi(w,b) := \dfrac{1}{2} \| w \|^2 \to min, \\ 1 - y_i(w \cdot x_i + b) \leq 0, i \in Q. \end{cases} \tag{5}$$

Bài toán này giải được dễ dàng bởi các thuật toán quy hoạch toàn phương cơ bản.

Tuy nhiên, để có thể mở rộng cho trường hợp dữ liệu không tách được tuyến tính, chúng ta sử dụng phương pháp nhân tử Lagrange ở ([1] Định lí 6.27–6.30). Thực chất, phương pháp này cũng chính là cơ sở lí thuyết của phần tiếp theo. Cụ thể, ta có hàm Lagrange của bài toán là:

$$L(w,b,\lambda) = \frac{1}{2} \| w \|^2 + \sum_{i=1}^{q} \lambda_i (1 - y_i (w \cdot x_i + b)), (w,b,\lambda) \in \mathbb{R}^n \times \mathbb{R} \times \mathbb{R}_+^q.$$

Với $\lambda = (\lambda_1, \lambda_2, ..., \lambda_q), \lambda_i \in \mathbb{R}_+, \forall i \in Q$. Lúc đó, để $(\overline{w}, \overline{b})$ là nghiệm của (5) điều kiện cần và đủ là tồn tại $\overline{\lambda} \geq 0$ sao cho $(\overline{w}, \overline{b}, \overline{\lambda})$ là điểm yên ngựa của hàm L :

$$L(\overline{w}, \overline{b}, \overline{\lambda}) = \inf_{(w,b) \in \mathbb{R}^{n+1}} \sup_{\lambda \in \mathbb{R}_+^q} L(w,b,\lambda) = \sup_{\lambda \in \mathbb{R}_+^q} \inf_{(w,b) \in \mathbb{R}^{n+1}} L(w,b,\lambda).$$

Đó cũng là điểm dừng của hàm Lagrange, tức là nghiệm của hệ sau đây:

$$3\frac{\partial L}{\partial w} = w - \sum_{i=1}^{q} \lambda_i y_i x_i = 0, \tag{6}$$

$$\frac{\partial L}{\partial b} = -\sum_{i=1}^{q} \lambda_i y_i = 0, \tag{7}$$

$$\frac{\partial L}{\partial \lambda_i} = 1 - y_i (w \cdot x_i + b) \leq 0, \forall i \in Q, \tag{8}$$

$$\lambda_i \frac{\partial L}{\partial \lambda_i} = \lambda_i (1 - y_i (w \cdot x_i + b)) = 0, \forall i \in Q. \tag{9}$$

Trong thực hành, ta giải bài toán $max\,min$: $\sup_{\lambda \in \mathbb{R}_+^q} \inf_{(w,b) \in \mathbb{R}^{n+1}} L(w,b,\lambda)$. Với mỗi $\lambda \in \mathbb{R}_+^q$ thỏa mãn (7), giải (6) ta được $\overline{w}$ ($\overline{b}$ tùy ý) thỏa mãn

$$F(\lambda) = \inf_{(w,b) \in \mathbb{R}^{n+1}} L(w,b,\lambda) = L(\overline{w}, \overline{b}, \lambda)$$

Chú ý rằng $\overline{w}$ thỏa (6), ngoài ra để ý (7), ta được

$$F(\lambda) = \frac{1}{2} \| \overline{w} \|^2 + \sum_{i=1}^{q} \lambda_i - \sum_{i=1}^{q} \lambda_i y_i x_i \cdot \overline{w} - \overline{b} \sum_{i=1}^{q} \lambda_i y_i = -\frac{1}{2} \| \overline{w} \|^2 + \sum_{i=1}^{q} \lambda_i. \tag{10}$$

Vì

$$\overline{w} = \sum_{i=1}^{q} \lambda_i y_i x_i$$

nên

$$\| \overline{w} \|^2 = \overline{w} \cdot \overline{w} = (\sum_{i=1}^{q} \lambda_i y_i x_i) \cdot (\sum_{i=1}^{q} \lambda_i y_i x_i) = \sum_{i=1}^{q} \sum_{j=1}^{q} (y_i y_j x_i \cdot x_j) \lambda_i \lambda_j = \lambda \cdot D\lambda,$$

trong đó D là ma trận vuông, đối xứng có phần tử ở hàng i cột j là $D_{ij} = y_i y_j x_i \cdot x_j$. Nếu ta kí hiệu 1 là vec-tơ gồm các thành phần đều bằng 1, nghĩa là $1 = (1,1,\ldots,1)$, thì $F(\lambda)$ trong (10) có thể viết lại dưới dạng hàm toàn phương:

$$F(\lambda) = -\frac{1}{2} \lambda \cdot D\lambda + 1 \cdot \lambda.$$

Nếu đặt $y = (y_1, y_2, \ldots, y_q)$ thì điều kiện (7) có thể viết lại dưới dạng

$$\lambda \cdot y = 0.$$

Nên cuối cùng ta giải bài toán quy hoạch toàn phương:

$$\begin{cases} F(\lambda) = -\frac{1}{2} \lambda \cdot D\lambda + 1 \cdot \lambda \to max, \\ \lambda \geq 0, \\ \lambda \cdot y = 0. \end{cases} \tag{11}$$

Giải (11) ta được nghiệm $\overline{\lambda}$. Từ (6) ta tính được

$$\overline{w} = \sum_{i=1}^{q} \overline{\lambda}_i y_i x_i. \tag{12}$$

Chọn $i \in Q$ sao cho $\overline{\lambda}_i > 0$. Từ (9) ta suy ra

$$\overline{b} = \frac{1}{y_i} - \overline{w} \cdot x_i = y_i - \overline{w} \cdot x_i.$$

Đến đây ta thu được $(\overline{w}, \overline{b})$ và cả $\overline{\lambda}$.

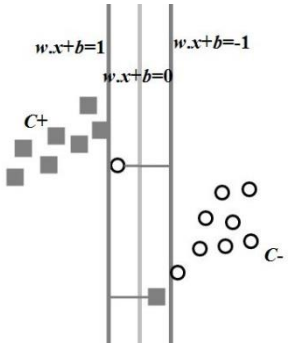
Nhận xét 3.1 *Với giả thiết dữ liệu tách được tuyến tính, bài toán (5) thỏa mãn điều kiện Slater. Vì vậy, (6–9) là hệ điều kiện cần và đủ tối ưu. Không những thế, nghiệm* $(\overline{w}, \overline{b})$ *thỏa mãn* $\Phi(\overline{w}, \overline{b}) = \frac{1}{2} \| \overline{w} \|^2 > 0$. *Từ đó suy ra* $\overline{\lambda} \neq 0$. *Nên chắc chắn tồn tại* $i \in Q$ *sao cho* $\overline{\lambda}_i > 0$.

Nhận xét 3.2 *Từ (12) ta thấy chỉ những vec-tơ* x_i *ứng với các* $\overline{\lambda}_i > 0$ *mới tham gia vào việc cấu tạo nên vec-tơ pháp* $\overline{w}$ *của mặt quyết định. Lại do điều kiện (9),* $y_i(\overline{w} \cdot x_i + \overline{b}) = 1$, *nên những vec-tơ như thế có khoảng cách đến* $S_{(\overline{w},\overline{b})}$ *đúng bằng lề* $\rho(\overline{w}, \overline{b})$. *Những vec-tơ như thế được gọi là vec-tơ hỗ trợ. Vì*

$y_i \in \{-1,1\}$, từ (7) ta thấy tồn tại cả những $y_i = 1$ và những $y_i = -1$ ở đó $\bar{\lambda}_i > 0$. Nói cách khác, trong C^+ và C^- đều có những vec-tơ hỗ trợ. Vậy mặt quyết định $S_{(\bar{w},\bar{b})}$ nằm cách đều hai miền C^+ và C^-.

4 Siêu phẳng lề mềm

Trong các bài toán thực tế, dữ liệu thu được từ thế giới thực thường không tách được tuyến tính. Điều đó là do thông tin bị nhiễu, dẫn đến sự chồng lấp một phần nhỏ dữ liệu giữa hai lớp C^+ và C^-.



Hình 4. Siêu phẳng lề mềm

Nói cách khác, không tồn tại $(w,b) \in \mathbb{R}^{n+1}$ thỏa mãn thực sự (4). Trong trường hợp như vậy ta cần nới lỏng ràng buộc và cho phép giải bài toán với đòi hỏi tối thiểu hóa sự vi phạm các ràng buộc. Cụ thể, ta đưa vào bộ biến phụ $\xi = (\xi_1, \xi_2, \ldots, \xi_q)$ thể hiện cho mức độ vi phạm điều kiện (4):

$$y_i(w \cdot x_i + b) \geq 1 - \xi_i, \forall i \in Q. \tag{13}$$

Song song với việc nới lỏng ràng buộc như vậy, ta đưa thêm vào hàm mục tiêu một hệ số phạt $C(\xi_1 + \xi_2 + \cdots + \xi_q)$, với C là hằng số dương, để giảm thiểu sự vi phạm. Như vậy, thay vì bài toán (5) ta giải bài toán sau:

$$\begin{cases} \Phi(w,b,\xi) = \dfrac{1}{2} \| w \|^2 + C \sum\limits_{i=1}^{q} \xi_i \to min, \\ 1 - y_i(w \cdot x_i + b) - \xi_i \leq 0, \forall i \in Q. \end{cases} \tag{14}$$

Với bài toán này ta có Hàm Lagrange:

$$L(w,b,\xi,\lambda,\mu) = \frac{1}{2} \| w \|^2 + C \sum_{i=1}^{q} \xi_i + \sum_{i=1}^{q} \lambda_i (1 - y_i(w \cdot x_i + b) - \xi_i) - \sum_{j=1}^{q} \mu_j \xi_j,$$

với $(w,b,\xi,\lambda,\mu) \in \mathbb{R}^n \times \mathbb{R} \times \mathbb{R}^q \times \mathbb{R}_+^q \times \mathbb{R}_+^q$.

Thay cho hệ (6–9), điều kiện cần và đủ tối ưu của bài toán (14) là hệ sau đây:

$$3\frac{\partial L}{\partial w} = w - \sum_{i=1}^{q} \lambda_i y_i x_i = 0, \tag{15}$$

$$\frac{\partial L}{\partial b} = -\sum_{i=1}^{q} \lambda_i y_i = 0, \tag{16}$$

$$\frac{\partial L}{\partial \xi_i} = C - \lambda_i - \mu_i = 0, \forall i \in Q, \tag{17}$$

$$\frac{\partial L}{\partial \lambda_i} = 1 - y_i(w \cdot x_i + b) - \xi_i \leq 0, \forall i \in Q, \tag{18}$$

$$\lambda_i \frac{\partial L}{\partial \lambda_i} = \lambda_i(1 - y_i(w \cdot x_i + b) - \xi_i) = 0, \forall i \in Q, \tag{19}$$

$$\frac{\partial L}{\partial \mu_i} = -\xi_i \leq 0, \forall i \in Q, \tag{20}$$

$$\mu_i \frac{\partial L}{\partial \mu_i} = -\mu_i \xi_i = 0, \forall i \in Q. \tag{21}$$

Nghĩa là, $(\bar{w},\bar{b},\bar{\xi})$ là nghiệm của bài toán (14) khi và chỉ khi tồn tại $\bar{\lambda} \geq 0$ và $\bar{\mu} \geq 0$ sao cho $(\bar{w},\bar{b},\bar{\xi},\bar{\lambda},\bar{\mu})$ là nghiệm của hệ trên. Hơn nữa, ta có

$$\Phi(\bar{w},\bar{b},\bar{\xi}) = \sup_{(\lambda,\mu) \in \mathbb{R}_+^q \times \mathbb{R}_+^q} \inf_{(w,b,\xi) \in \mathbb{R}^{n+1} \times \mathbb{R}^q} L(w,b,\xi,\lambda,\mu).$$

Để giải bài toán $\max\min$ trên, với mỗi $(\lambda,\mu) \in \mathbb{R}_+^q \times \mathbb{R}_+^q$, thỏa mãn các điều kiện (16–17), ta tìm được $\bar{w} \in \mathbb{R}^n$ ($\bar{\xi}$, $\bar{b}$ tùy ý) thỏa mãn (15). Thay vào hàm Lagrange:

$$F(\lambda,\mu) = \inf_{(w,b,\xi)} L(w,b,\xi,\lambda,\mu) = L(\bar{w},\bar{b},\bar{\xi},\lambda,\mu) = -\frac{1}{2}\| \bar{w} \|^2 + \sum_{i=1}^{q} \lambda_i = -\frac{1}{2}\lambda \cdot D\lambda + \lambda \cdot 1,$$

với D được xác định như ở Mục 3. Vì hàm F không phụ thuộc vào μ nên cực đại hàm F ta dẫn đến bài toán

$$\begin{cases} -\frac{1}{2}\lambda \cdot D\lambda + \lambda \cdot 1 \to max, \\ \lambda \geq 0, \\ \lambda \cdot y = 0, \\ \lambda \leq C \cdot 1. \end{cases} \quad (22)$$

Giải (22) ta được nghiệm $\bar{\lambda}$. Từ (17) và (15) ta tính được $\bar{\mu}$ và

$$\bar{w} = \sum_{i=1}^{q} \bar{\lambda}_i y_i x_i. \quad (23)$$

Chọn $i \in Q$ sao cho $\bar{\lambda}_i > 0$ và $\bar{\mu}_i > 0$. Từ (19), (21) ta suy ra $\bar{\xi}_i = 0$ và

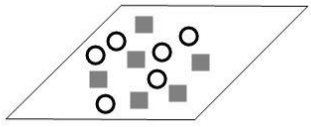
$$\bar{b} = \frac{1}{y_i} - \bar{w} \cdot x_i = y_i - \bar{w} \cdot x_i. \quad (24)$$

Bây giờ sử dụng (19) và (21) ta thu được $\bar{\xi}$.

Nhận xét 4.1 *Bằng cách chọn $C > 0$ khá lớn, nghiệm $\bar{\lambda}$ của bài toán (22) sẽ có ít nhất một tọa độ $0 < \bar{\lambda}_i < C$, nên cũng sẽ có $\bar{\mu}_i > 0$. Nhờ chỉ số i này ta tính được $\bar{b}$ theo (24). Do (17) nên với mọi $i \in Q$ ta phải có $\bar{\lambda}_i > 0$ hoặc $\bar{\mu}_i > 0$. Vì vậy, dùng (19), (21) ta sẽ tính được mọi $\bar{\xi}_i$. Những $\bar{\xi}_i > 0$ chính là những ràng buộc bị vi phạm.*

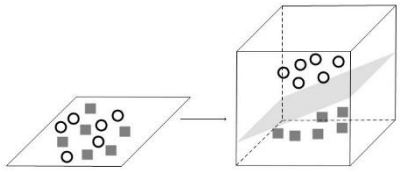
5 Hàm phân lớp phi tuyến

Trong Mục 4 ta đã sử dụng kĩ thuật lề mềm để xét bài toán phân lớp khi dữ liệu bị chồng lấp một phần nhỏ. Trong trường hợp tồi hơn khi dữ liệu hoàn toàn trộn lẫn vào nhau (Hình 5) thì kĩ thuật lề mềm không còn hiệu quả.



Hình 5. Dữ liệu hoàn toàn trộn lẫn vào nhau

Lúc đó ta sử dụng biện pháp nâng cao số chiều. Cụ thể, ta thiết lập một ánh xạ từ $\mathbb{R}^n$ vào một không gian mới $\mathbb{R}^m$ nhiều chiều hơn sao cho ảnh của các điểm dữ liệu trong không gian đó là tách được tuyến tính. Một cách trực quan ta tưởng tượng có hai nhóm vỏ ốc "lớn" và "nhỏ" nằm xen kẽ nhau trên mặt bàn mà không thể phân lớp tuyến tính chúng bằng một đường thẳng được. Bằng cách hất mạnh lên không thì những con ốc nhỏ bay cao hơn những con ốc lớn, khi ấy ta có thể phân lớp tuyến tính chúng bằng một mặt phẳng (Hình 6).



Hình 6. Ánh xạ lên không gian mới nhiều chiều hơn

Cho ánh xạ $\Phi : \mathbb{R}^n \to \mathbb{R}^m$ sao cho với mỗi $x \in \mathbb{R}^n$ tương ứng với một dãy

$$\Phi(x) = (a_1\varphi_1(x), a_2\varphi_2(x), \ldots, a_m\varphi_m(x)) \in \mathbb{R}^m.$$

Ta cần tìm ánh xạ Φ sao cho hai tập $\Phi(C^+)$ và $\Phi(C^-)$ (trong $\mathbb{R}^m$) là tách được tuyến tính. Nghĩa là tồn tại cặp $(w,b) \in \mathbb{R}^m \times \mathbb{R}$ sao cho

$$y_i(w \cdot \Phi(x_i) + b) \geq 1, \forall i \in Q.$$

Đến đây ta áp dụng kĩ thuật phân lớp tuyến tính trong không gian $\mathbb{R}^m$ bằng phương pháp hàm Lagrange và đưa về bài toán quy hoạch toàn phương

$$\begin{cases} F(\lambda) = \lambda \cdot 1 - \dfrac{1}{2}\lambda \cdot D\lambda \to max, \\ \lambda \geq 0, \\ \lambda \cdot y = 0, \end{cases} \tag{25}$$

trong đó D là ma trận vuông với $D_{ij} = y_i y_j \Phi(x_i) \cdot \Phi(x_j)$. Giải bài toán này ta được nghiệm $\bar{\lambda}$. Suy ra

$$\bar{w} = \sum_{i=1}^{q} \bar{\lambda}_i y_i \Phi(x_i). \tag{26}$$

Chọn $i \in Q$ sao cho $\bar{\lambda}_i > 0$ ta tính được:

$$\overline{b} = \frac{1}{y_i} - \overline{w} \cdot \Phi(x_i) = y_i - \overline{w} \cdot \Phi(x_i).$$

Cuối cùng, ta có hàm phân lớp trên $\mathbb{R}^n$ là

$$f(x) = sgn(\overline{w} \cdot \Phi(x) + \overline{b}), x \in \mathbb{R}^n.$$

Một vấn đề quan trọng đặt ra ở đây là ta cần xây dựng hàm Φ sao cho việc tính toán ma trận D là dễ dàng. Muốn vậy chúng ta phải thiết kế sao cho công thức tính $\Phi(x) \cdot \Phi(y)$ phải đơn giản, thuận tiện. Dưới đây là một vài ví dụ về những hàm Φ như thế [7, 10–12].

Ví dụ 5.1 $\Phi : \mathbb{R}^2 \to \mathbb{R}^6$ *cho bởi*

$$\Phi(x_1, x_2) = (1, \sqrt{2}x_1, \sqrt{2}x_2, x_1^2, x_2^2, \sqrt{2}x_1x_2).$$

Lúc đó, với mọi $x = (x_1, x_2)$, $y = (y_1.y_2)$ ta có $\Phi(x) \cdot \Phi(y) = (1 + x \cdot y)^2$.

Ví dụ 5.2 $\Phi : \mathbb{R}^3 \to \mathbb{R}^{10}$ *được cho bởi*

$$\Phi(x_1, x_2, x_3) = (1, \sqrt{2}x_1, \sqrt{2}x_2, \sqrt{2}x_3, x_1^2, x_2^2, x_3^2, \sqrt{2}x_1x_2, \sqrt{2}x_2x_3, \sqrt{2}x_3x_1).$$

Lúc đó, với mọi $x, y \in \mathbb{R}^3$ ta có $\Phi(x) \cdot \Phi(y) = (1 + x \cdot y)^2$.

Ví dụ 5.3 $\Phi : \mathbb{R} \to \mathbb{R}^{m+1}$ *được cho bởi*

$$\Phi(x) = (1, \sqrt{C_m^1}\, x, \sqrt{C_m^2}\, x^2, ..., \sqrt{C_m^m}\, x^m).$$

Lúc đó, với mọi $x, y \in \mathbb{R}$ ta có

$$\Phi(x) \cdot \Phi(y) = 1 + C_m^1 xy + C_m^2 x^2 y^2 + \cdots + C_m^m x^m y^m = (1 + xy)^m.$$

Qua các ví dụ trên ta thấy việc chọn các hàm φ_i và các hệ số a_i hợp lý sẽ cho ta một công thức tính toán $\Phi(x) \cdot \Phi(y)$ đơn giản.

6 Hàm phân lớp có trọng số

Việc xây dựng một hàm phân lớp dựa vào một tập huấn luyện (tức là các mẫu thử) cho trước dĩ nhiên vẫn có những sai lầm trong thực hiện. Như ví dụ về hàm chẩn đoán bệnh sốt xuất huyết nói đến ở Mục 2. Có thể bệnh nhân không bị sốt xuất huyết mà hàm phân loại cho ra giá trị 1, hoặc ngược lại, bệnh nhân bị sốt xuất huyết mà hàm phân loại cho giá trị –1. Tổng quát, sẽ có hai loại sai lầm chính là $f(x) = 1$ với $x \in C^-$ và $f(x) = -1$ với $x \in C^+$. Trong những bài toán thực tế, một trong hai loại sai lầm là nguy hiểm hơn sai lầm kia. Nói cách khác, chúng ta có thái độ thiên vị đối với các loại sai lầm, có sai lầm chúng ta sẽ rất khắt khe trong khi lại dễ dàng tha thứ cho sai lầm loại khác. Chẳng hạn người bệnh bị đau ruột thừa mà chuẩn đoán

không phải rõ ràng là nguy hại hơn người không bị đau ruột thừa mà chuẩn đoán có. Thái độ thiên vị với các sai lầm vẫn được thể hiện trong hàm mục tiêu bằng cách cho hệ số phạt trên các loại sai lầm là khác nhau. Điều này được thực hiện tương tự kĩ thuật lề mềm nhưng khác ở chỗ trong kĩ thuật lề mềm hệ số phạt là như nhau cho các vi phạm. Cụ thể, ta sẽ đưa thêm các biến phụ ξ_i và giải bài toán tối ưu sau

$$\begin{cases} \Phi(w,b) = \dfrac{1}{2} \| w \|^2 + \sum\limits_{x_i \in C^+} \delta^+ \xi_i + \sum\limits_{x_i \in C^-} \delta^- \xi_i \to min, \\ y_i (w \cdot x_i + b) \geq 1 - \xi_i, i \in Q, \end{cases} \tag{27}$$

Tương tự bài toán lề mềm, bằng phương pháp nhân tử Lagrange, ta đưa bài toán (27) về bài toán $max\,min$, trong đó bài toán max là

$$\begin{cases} -\dfrac{1}{2} \lambda \cdot D\lambda + \lambda \cdot 1 \to max, \\ \lambda \cdot y = 0, \\ 0 \leq \lambda_i \leq \delta^+, với\ y_i = 1, \\ 0 \leq \lambda_i \leq \delta^-, với\ y_i = -1. \end{cases} \tag{28}$$

Giải bài toán này ta được nghiệm $\overline{\lambda}$. Sau đó tiếp tục giải bài toán min ta được $\overline{w}$ và $\overline{b}$.

7 Kết luận

Như vậy, tư tưởng toán học của máy vec-tơ hỗ trợ thực chất là tìm cách tách các lớp dữ liệu bởi một siêu phẳng có khoảng cách đến tập các dữ liệu là lớn nhất. Bằng một phương pháp nhất quán là sử dụng quy tắc nhân tử Lagrange, chúng tôi đã trình bày cơ sở toán học của máy vec-tơ hỗ trợ trong kỹ thuật phân lớp dữ liệu, cho các trường hợp khác nhau, từ đơn giản đến phức tạp. Trường hợp đơn giản nhất là hàm phân lớp tuyến tính, tiếp theo là kỹ thuật siêu phẳng lề mềm cho bài toán không tách được tuyến tính, trường hợp hàm phân lớp phi tuyến và cuối cùng là phân lớp có trọng số.

References

1. Huỳnh Thế Phùng (2012), *Cơ sở giải tích lồi*, Nxb. Giáo dục Việt Nam.
2. Vapnik V. (1995), *The nature of statistical learning theory*, Springer-Verlag, New York.
3. Bose B. E., Guyon I. M., anh Vapnik V. N. (1992), A training algorithm for optimal magrin classifier. In *Proc. 5th ACM Workshop on Computational Learning Theory*, papes 144–152, Pittsburgh, PA, July 1992.
4. Burges C. J. C. (1996), Simplified support vector decision rules, *In Proceedings 13th Int, Conference on Machine Learning*, pp.71-77.

5. Daehyon Kim (2004), Prediction performance of support vector machine on input vector normalization methods, *International Journal of Computer Mathematics*, 81:5, pp. 547–554.
6. Cortes C. and Vapnik V. (1995), Support vector networks. *Machine Learning*, 20:1–25.
7. Osuna E., Freund R. and Girosi F. (1997), *Support Vector Machines: Training and Applications*, Massachusetts Institute of Technology.

MATHEMATICAL FOUNDATION OF SUPPORT VECTOR MACHINE

Nguyen The Cuong*

Department of Mathematics, Communications University
101 Mai Xuan Thuong, Nha Trang, Khanh Hoa.

Abstract: The support vector machine is a very promising classification technique developed by V.N. Vapnik and his colleagues in the early 90s of the twentieth century. For over two decades, this technique has been successfully applied to the construction of the identification program that is used for many different areas of the life. Since it was proposed to date, this technique was generalized continuously. The first is the problem with data sets are linearly inseparable, then to the case with overlapping data sets or more complicated when data layers are mixed together. In all cases, Lagrange multiplier method has proved effective to put the problem in the most explicit form from which some algorithms can be provided. This paper aims to present the mathematical basis of the classification technique for cases ranging from simple to complex.

Keywords: Support vector machines.