

ỨNG DỤNG MÔ HÌNH MÁY HỌC VÉC-TƠ TỰA (SVM) TRONG PHÂN TÍCH DỮ LIỆU ĐIỂM SINH VIÊN

APPLICATION OF THE SUPPORT VECTOR MACHINE FOR ANALYSES OF STUDENT SCORE DATA

Nguyễn Đức Hiền

Trường Cao đẳng Công nghệ Thông tin, Đại học Đà Nẵng

Email: hiencit@gmail.com

TÓM TẮT

Bài báo này đề xuất ứng dụng mô hình kết hợp máy học véc-tơ tựa và hệ thống mờ trong việc trích xuất luật mờ từ dữ liệu điểm sinh viên. Máy học Véc-tơ tựa (SVMs) và hệ thống luật mờ có sự tương đương nhau với một số điều kiện nhất định. Trên cơ sở phân tích sự tương đương giữa mô hình máy học Véc-tơ tựa với mô hình mờ (Fuzzy model), chúng tôi đề xuất một mô hình tích hợp SVMs và Fuzzy model để trích xuất luật mờ từ kết quả huấn luyện SVMs. Thuật toán f-SVM cho phép sản xuất được các luật mờ từ dữ liệu huấn luyện. Dữ liệu điểm thực tế của sinh viên được sử dụng để kiểm tra khả năng thực hiện của mô hình đề xuất. Tập luật mờ trích xuất được từ tập dữ liệu huấn luyện bằng thuật toán f-SVM, sẽ được sử dụng để suy luận trên tập dữ liệu thử nghiệm.

Từ khóa: Máy học véc-tơ tựa; mô hình mờ; khai phá luật kết hợp; khai phá dữ liệu; luật mờ

ABSTRACT

This paper proposed an application of a combining model based on the support vector machine and fuzzy system to extract fuzzy rules from the student score data. The support vector machines and fuzzy rule systems are functionally equivalent under some conditions. Based on the discussions about the equivalence between SVMs and the Fuzzy model, we proposed a combining model for extracting fuzzy rules from the trained SVMs. The f-SVM algorithm allows us to extract the fuzzy rules from the training data set. The real data sets of student scores were used to examine the performance of the proposed model. The fuzzy rules obtained from the training data using f-SVM algorithm is tested on the testing data.

Key words: support vector machine; Fuzzy model; Association Rule Discovery / Association Rule mining; data mining; fuzzy rules

1. Đặt vấn đề

Khai phá tri thức từ dữ liệu (Data mining) là một lĩnh vực ứng dụng của Trí tuệ nhân tạo đã và đang được rất nhiều nhà khoa học, tổ chức trên thế giới nghiên cứu và phát triển ứng dụng. Nhiều kỹ thuật mới đã được nghiên cứu và đề xuất như Neural Network, Mô hình mờ (Fuzzy modeling), Support Vector Machine (SVM), Self Organizing Map (SOM), Lý thuyết tập thô (Rough Sets), các thuật toán Phân cụm, Phân lớp, Hồi quy,... [2], [3], [5], [9], [10]. Đặc biệt gần đây xu hướng nghiên cứu cải tiến và tích hợp nhiều công cụ khác nhau nhằm nâng cao hiệu năng của các giải pháp đang được nhiều nhà khoa học trên thế giới quan tâm; ví dụ như việc kết hợp giữa SVM và Fuzzy modeling, SOM và SVM, Rough Sets và SVM, hoặc kết hợp Giải thuật di truyền (GA) với các thuật toán Phân

cụm, Phân lớp,... [6], [11]. Mô hình trích xuất luật mờ từ SVMs được giới thiệu như là một trong những hướng nghiên cứu mới của mô hình mờ - mô hình mờ hướng dữ liệu (data-driven model) [2], [3], [5].

Đối với mô hình đào tạo theo học chế tín chỉ, công tác cố vấn học tập, tư vấn cho sinh viên chiến lược lựa chọn môn học hợp lý, số tín chỉ đăng ký phù hợp,... là những yêu cầu khá bức thiết hiện nay. Những kinh nghiệm, qui tắc theo suy nghĩ chủ quan của Cố vấn học tập có thể không phù hợp, hoặc là chưa khách quan và đầy đủ; vì vậy việc tìm ra những qui tắc dựa vào những dữ liệu lịch sử là một bổ sung rất cần thiết cho việc đưa ra những qui luật tư vấn cho sinh viên. Trong nghiên cứu này, chúng tôi đề xuất một mô hình kết hợp SVMs với Fuzzy model để sản xuất các luật mờ từ dữ liệu điểm sinh viên.

2. Trích xuất luật mờ từ máy học véc-tơ tựa – Thuật toán F-SVM

2.1. Máy học véc-tơ tựa – SVM

Máy học véc-tơ tựa SVM được Vapnik giới thiệu năm 1995, đây là mô hình học dựa trên lý thuyết học thống kê (Statistical Learning Theory) [1] và là một kỹ thuật được đề nghị để giải quyết cho các bài toán phân lớp. Một số nghiên cứu gần đây [2], [4], [6] đã đề xuất sử dụng SVM giải quyết bài toán tối ưu hóa hồi quy; đồng thời SVM cũng được sử dụng để khai phá luật mờ từ dữ liệu số [2], [3], [5]. Với vai trò giải quyết vấn đề tối ưu hóa hồi quy, lý thuyết cơ bản của SVM có thể được vắn tắt như sau:

Cho một tập dữ liệu huấn luyện $\{(x_1, y_1), \dots, (x_l, y_l)\} \subset \mathcal{X} \times \mathbb{R}$, trong đó $\mathcal{X}$ xác định miền dữ liệu đầu vào. Mục tiêu của ε -SV hồi quy (ε -Support Vector Regression) là tìm một hàm quyết định phẳng nhất có thể, đồng thời độ sai lệch trên các y_i của cả tập dữ liệu huấn luyện là không lớn hơn ε . Trong trường hợp hồi quy phi tuyến, hàm quyết định $f(x)$ có thể xác định như sau:

$$f(x) = \sum_{i=1}^l (\alpha_i - \alpha_i^*) K(x_i, x) + b \quad (1)$$

Sao cho:

$$\sum_{i=1}^l (\alpha_i - \alpha_i^*) = 0, \text{ and } C \geq \alpha_i, \alpha_i^* \geq 0, \forall i, \quad (2)$$

Trong đó, C là hằng số xác định sự cân bằng của biên độ lỗi giữa độ phẳng của f và lượng sai lệch vượt quá ε được chấp nhận, α_i, α_i^* là những nhân tử Lagrange; và $K(x_i, x)$ là hàm kernel được định nghĩa như sau:

$$K(x_i, x_j) = \langle \Phi(x_i), \Phi(x_j) \rangle \quad (3)$$

với Φ là ánh xạ thuộc tính cho kernel K .

Những điểm đầu vào x_i với $(\alpha_i - \alpha_i^*) \neq 0$ gọi là những véc-tơ tựa (SVs).

2.2. Mô hình mờ - Fuzzy model

Ta biết rằng các luật mờ (fuzzy rules) được biểu diễn ở dạng IF – THEN, là cơ sở của phép suy luận mờ [2]. Giả sử có m luật mờ được biểu diễn như sau:

$$R_j: \text{IF } x_1 \text{ is } A_1^j \text{ and } x_2 \text{ is } A_2^j \text{ and } \dots \text{ and } x_n \text{ is } A_n^j \\ \text{THEN } y \text{ is } B^j, \text{ for } j = 1, 2, \dots, M \quad (4)$$

Trong đó $x_i (i = 1, 2, \dots, n)$ là các biến điều kiện; y là các biến quyết định của hệ thống mờ; A_i^j và B^j là những thuật ngữ ngữ nghĩa xác định bởi các hàm thành viên (membership functions) tương ứng $\mu_{A_i^j}(x_i)$ và $\mu_{B^j}(y)$. Lưu ý, $\mu_{A_i^j}(x_i)$ được định nghĩa như sau:

$$\mu_{A^j}(x_i) = \prod_{i=1}^n \mu_{A_i^j}(x_i) \quad (5)$$

Quá trình suy luận được thực hiện như sau: 1) Kích hoạt các giá trị thành viên.

$$\prod_{i=1}^n \mu_{A_i^j}(x_i) \quad (6)$$

2) Kết quả đầu ra của suy luận được tính như sau:

$$f(x) = \frac{\sum_{j=1}^M \bar{z}^j \left(\prod_{i=1}^n \mu_{A_i^j}(x_i) \right)}{\sum_{j=1}^M \prod_{i=1}^n \mu_{A_i^j}(x_i)} \quad (7)$$

Trong đó, $\bar{z}^j$ là giá trị đầu ra khi hàm thành viên $\mu_{B^j}(y)$ đạt giá trị cực đại.

Để (1) và (5) bằng nhau, trước tiên chúng ta phải đồng nhất giữa hàm kernel trong (1) và hàm thành viên trong (7). Ở đây, để thỏa mãn điều kiện Mercer [7] hàm thành viên Gaussian được chọn làm hàm kernel; đồng thời giá trị của ab trong (1) phải bằng 0.

2.3. Thuật toán trích xuất luật mờ từ SVMs

Khi hàm Gaussian được chọn làm hàm thành viên và hàm kernel, đồng thời số luật mờ bằng với số SVs (Support vectors) thì (1) và (7) trở thành

$$f(x) = \sum_{i=1}^l (\alpha_i - \alpha_i^*) \exp \left(-\frac{1}{2} \left(\frac{x_i - x}{\sigma_i} \right)^2 \right) \quad (8)$$

và

$$f(x) = \frac{\sum_{j=1}^l \bar{z}^j \exp \left(-\frac{1}{2} \left(\frac{x_j - x}{\sigma_j} \right)^2 \right)}{\sum_{j=1}^l \exp \left(-\frac{1}{2} \left(\frac{x_j - x}{\sigma_j} \right)^2 \right)} \quad (9)$$

Như cách biến đổi trong [2], hàm suy luận mờ có thể viết lại như sau:

$$f(x) = \sum_{j=1}^l \bar{z}^j \exp \left(-\frac{1}{2} \left(\frac{x_j - x}{\sigma_j} \right)^2 \right) \quad (10)$$

Và trung tâm của hàm thành viên Gaussian được chọn là

$$\bar{z}^j = (\alpha_i - \alpha_i^*) \quad (11)$$

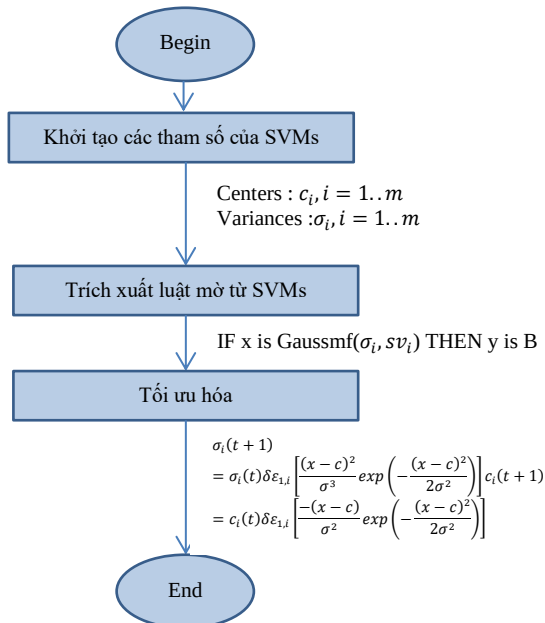
Một cách tiếp cận khác là thiết lập hàm kernel của SVMs:

$$K(x_i, x) = \frac{\exp\left(-\frac{1}{2}\left(\frac{x_i - x}{\sigma_i}\right)^2\right)}{\sum_{i=1}^l \exp\left(-\frac{1}{2}\left(\frac{x_i - x}{\sigma_i}\right)^2\right)} \quad (12)$$

Như vậy đầu ra của SVMs trở thành

$$f(x) = \frac{\sum_{i=1}^l (\alpha_i - \alpha_i^*) \exp\left(-\frac{1}{2}\left(\frac{x_i - x}{\sigma_i}\right)^2\right)}{\sum_{i=1}^l \exp\left(-\frac{1}{2}\left(\frac{x_i - x}{\sigma_i}\right)^2\right)} \quad (13)$$

Khi đó chúng ta chỉ cần thiết lập trung tâm của hàm thành viên bằng với $(\alpha_i - \alpha_i^*)$, thì chắc chắn rằng đầu ra của SVMs (13) và đầu ra của hệ thống mờ (9) là bằng nhau. Lưu ý rằng biểu thức (12) chỉ có thể có nếu số lượng Support-vector l được biết trước.



Hình 1. Sơ đồ khối thuật toán f-SVM

Trên cơ sở sự tương tự của SVMs và hệ thống mờ, thuật toán f-SVM đề xuất ở Hình 1 cho phép trích xuất các luật mờ từ SVMs.

Những tham số của hàm thành viên có thể được tối ưu hóa dùng những thuật toán gradient descent hoặc thuật toán di truyền (GAs). Trong trường hợp này, để nhận được tập mờ tối ưu, chúng tôi cập nhật giá trị các tham số của hàm thành viên theo các hàm thích nghi sau đây:

$$\sigma_i(t+1) = \sigma_i(t) \delta \varepsilon_{1,i} \left[\frac{(x-c)^2}{\sigma^3} \exp\left(-\frac{(x-c)^2}{2\sigma^2}\right) \right] \quad (14)$$

$$c_i(t+1) = c_i(t) \delta \varepsilon_{1,i} \left[\frac{(x-c)}{\sigma^2} \exp\left(-\frac{(x-c)^2}{2\sigma^2}\right) \right] \quad (15)$$

3. Ứng dụng khai phá luật mờ từ dữ liệu điểm sinh viên dựa trên f-SVM

3.1. Bài toán tìm kiếm các luật thể hiện sự liên quan giữa các môn học

Một trong những vấn đề được quan tâm của các nhà quản lý đào tạo là việc xác định sự liên quan thật sự khách quan giữa các môn học trong chương trình đào tạo. Bên cạnh đó, sinh viên cũng cần sự tư vấn để lựa chọn các môn học nhằm đạt được kết quả học tập cao nhất. Ở đây, chúng tôi đề xuất sử dụng thuật toán f-SVM để khai phá các luật thể hiện sự liên quan giữa các môn học từ dữ liệu điểm học tập của sinh viên.

Giả sử cần tìm luật thể hiện sự liên hệ của môn Lập trình C++ với các môn Lập trình C, Toán rời rạc và Cấu trúc dữ liệu & giải thuật. Trong trường hợp này ta chọn các thuộc tính đầu vào và đầu ra như Bảng 1.

Bảng 1. Các thuộc tính lựa chọn

ký hiệu	Thuộc tính
x_1	Điểm môn học Lập trình C
x_2	Điểm môn học Toán rời rạc
x_3	Điểm môn học CTDL & GT
y	Điểm môn học Lập trình C++

3.2. Kết quả thực nghiệm

Dữ liệu kết quả điểm các môn học của gần 1000 sinh viên ngành Công nghệ thông tin – Trường CĐ Công nghệ Thông tin, được xử lý bằng cách:

- Trích lọc các thuộc tính tương ứng như trong Bảng 1.
- Loại bỏ các dòng dữ liệu bị thiếu (không có điểm).
- Scale toàn bộ dữ liệu trong $[-1,1]$.

Để trích xuất các luật mờ từ dữ liệu huấn luyện, chúng tôi xây dựng một hệ thống thử nghiệm dựa trên bộ công cụ Matlab. Thuật toán học SVM của thư viện LIBSVM được phát triển bởi nhóm của Chih-Chung Chang [8], được sử dụng để sản xuất ra các Support vector (SVs), làm cơ sở để xây dựng thuật toán trích xuất các luật mờ, f-SVM.

Bảng 2. Tập các luật sản xuất được

Luật	Chi tiết
R1	IF $x_1 = \text{Gaussmf}(0.35, -0.97)$ and $x_2 = \text{Gaussmf}(0.35, -0.29)$ and $x_3 = \text{Gaussmf}(0.16, -0.33)$ THEN $y = -0.05$
R2	IF $x_1 = \text{Gaussmf}(0.45, 0.88)$ and $x_2 = \text{Gaussmf}(0.32, -0.36)$ and $x_3 = \text{Gaussmf}(0.19, 0.34)$ THEN $y = 0.96$
R3	IF $x_1 = \text{Gaussmf}(0.32, 0.28)$ and $x_2 = \text{Gaussmf}(0.21, -0.36)$ and $x_3 = \text{Gaussmf}(0.29, -0.32)$ THEN $y = 0.07$
R4	IF $x_1 = \text{Gaussmf}(0.30, -0.29)$ and $x_2 = \text{Gaussmf}(0.27, -1.01)$ and $x_3 = \text{Gaussmf}(0.31, -1.00)$ THEN $y = -0.05$
R5	IF $x_1 = \text{Gaussmf}(0.46, 0.31)$ and $x_2 = \text{Gaussmf}(0.29, -0.34)$ and $x_3 = \text{Gaussmf}(0.14, 0.30)$ THEN $y = -0.34$

Kết quả một phần tập luật mờ sản xuất được từ 600 mẫu dữ liệu huấn luyện, được thể hiện trong Bảng 2.

Bằng cách sử dụng hàm AVALFIS trong thư viện công cụ Matlab Fuzzy Logic, chúng tôi đã thử nghiệm suy luận dựa trên tập luật sản xuất được đối với 200 mẫu dữ liệu test. Kết quả so sánh giá trị suy luận được với giá trị thực, được thể hiện qua hai thông số NMSE=1.0746 và MAE=0.5827. Trong đó, NMSE (normalized mean squared error) và MAE (mean absolute error) là hai thông số đo lường độ lệch giữa giá trị thực tế và giá trị suy luận được.

4. Kết luận

Trong nghiên cứu này, dựa trên cơ sở phân tích sự tương đương giữa mô SVM và Fuzzy model, chúng tôi đề xuất một thuật toán, f-SVM, cho phép sản xuất luật mờ từ SVMs. Một mô hình ứng dụng thuật toán f-SVM để sản xuất các luật mờ thể hiện sự liên quan giữa các môn học trong chương trình đào tạo cũng được đề xuất trong phần 3. Kết quả thực nghiệm trên dữ liệu thử nghiệm cho thấy khả năng ứng dụng thuật toán đề xuất cho sản xuất các luật mờ từ dữ liệu điểm sinh viên.

Ngoài ra, một vấn đề tồn tại đáng lưu ý là đối với mô hình máy học SVM, nếu tăng tính chính xác của mô hình thì số lượng SVs cũng tăng lên, đồng nghĩa với số lượng luật mờ cũng tăng lên. Điều này làm cho tính phức tạp của hệ thống tăng lên và đặc biệt là “tính sáng sủa” của tập luật mờ giảm đi, gây nên sự khó khăn cho chuyên gia con người để có thể hiểu và phân tích các luật này.

Trong định hướng nghiên cứu tiếp theo, chúng tôi sẽ tập trung tìm giải pháp cải thiện “tính sáng sủa” của tập luật mờ trích xuất được từ SVMs. Giải quyết được vấn đề này sẽ là cơ sở để có thể ngữ nghĩa hóa và phân tích tập luật thu được, qua đó có thể tối ưu hóa tập luật nhằm nâng cao hiệu quả sử dụng chúng.

TÀI LIỆU THAM KHẢO

- [1] Corinna Cortes and Vladimir Vapnik, Support-Vector Networks. *Machine Learning*, 20, 273-297 (1995).
- [2] J.-H Chiang and P.-Y Hao, Support vector learning mechanism for fuzzy rule-based modeling: a new approach, *IEEE Trans. On Fuzzy Systems*, vol. 12, pp. 1-12 (2004).
- [3] J.L. Castro, L.D. Flores-Hidalgo, C.J. Mantas and J.M. Puche, Extraction of fuzzy rules from support vector machines. Elsevier, *Fuzzy Sets and Systems*, 158, 2057 – 2077 (2007).
- [4] L.J.Cao and Francis E.H.Tay, Support vector machine with adaptive parameters in Financial time series forecasting, *IEEE trans. on neural network*, vol. 14, no. 6 (2003).
- [5] S. Chen, J. Wang and D. Wang, Extraction of fuzzy rules by using support vector machines, *IEEE, Computer society*, pp. 438-441 (2008).
- [6] Sheng-Hsun Hsu, JJ Po-An Hsieh, Ting-Chih CHih, Kuei-Chu Hsu, A two-stage architecture for stock price forecasting by integrating self-organizing map and support vector regression, *Expert system with applications* 36, 7947-7951 (2009).
- [7] R. Courant, D. Hilbert, *Methods of Mathematical Physics*. Wiley, New York (1953).

- [8] Chih-Wei Hsu, Chih-Chung Chang, and Chih-Jen lin, A practical Guide to Support Vector Classification. <http://www.csie.ntu.edu.tw/~cjlin/libsvm/> (2010).
- [9] Bùi Công Cường, Phạm Văn Chiến, An experiment result based on adaptive neuro-fuzzy inference system for stock price predict on, *Journal of Computer Science and Cybernetics*, V.27, N.1, 2011, 51-60.
- [10] Nguyễn Công Điều, Một số thuật toán mới cho mô hình chuỗi thời gian mờ heuristic trong dự báo chỉ số chứng khoán, *Báo cáo tại Đại hội toán học toàn quốc*, Quy Nhơn, 2008.
- [11] Vạn Duy Thanh Long, Lê Minh Duy, Nguyễn Hoàng Tú Anh, *Phương pháp dự đoán xu hướng cổ phiếu dựa trên việc kết hợp K-mean và SVM với ước lượng xác suất lớp*, Đại học Quốc gia – Tp HCM, 2011.

(BBT nhận bài: 25/09/2013, phản biện xong: 24/10/2013)