Tạp chí KTKT Mỏ - Địa chất, số 43/7-2013, tr. 100-103

THUẬT TOÁN LỰA CHỌN PHƯƠNG PHÁP TỈ LỆ DỮ LIỆU

ĐẶNG HỮU NGHỊ, HOÀNG KIM BẢNG, BÙI THỊ VÂN ANH
Trường Đại học Mỏ - Địa chất

Tóm tắt: *Máy tựa vector (Support Vector Machine – SVM) là một kỹ thuật hữu ích cho việc phân loại dữ liệu. Việc tỉ lệ giá trị của các thuộc tính trong tập dữ liệu huấn luyện cũng như tập dữ liệu kiểm thử về cùng một phạm vi (gọi tắt là tỉ lệ dữ liệu) trước khi áp dụng SVM là một bước rất quan trọng. Khi thiếu thông tin người ta thường tỉ lệ giá trị của các thuộc tính về cùng một phạm vi với cùng một phương pháp. Có 3 phương pháp tỉ lệ dữ liệu thường được sử dụng là: trung bình 0 và độ lệch chuẩn 1, tầm trung 0 và phạm vi 2, hoặc khi ý nghĩa về độ lớn là phi tuyến giá trị của các thuộc tính có thể được tỉ lệ bằng cách lấy logarit (hoặc lấy căn bậc 3) sau đó tiếp tục tỉ lệ kết quả nhận được bằng phương pháp tầm trung 0 và phạm vi 2. Trong bài báo này chúng tôi đề xuất phương pháp sử dụng giải thuật di truyền (Genetic Algorithm - GA) để lựa chọn phương pháp tỉ lệ cho từng thuộc tính. Kết quả thực nghiệm cho thấy trong nhiều trường hợp phương pháp mà chúng tôi đề xuất tốt hơn phương pháp vẫn thường được sử dụng đó là tỉ lệ giá trị của tất cả các thuộc tính theo cùng một phương pháp.*

1. Mở đầu

SVM là một kỹ thuật mới được sử dụng cho việc phân tích hồi quy và phân loại dữ liệu. Nhằm giảm độ phức tạp tính toán (vì các giá trị kernel được tính bởi tính vô hướng của các vector đặc trưng) cũng như tăng độ chính xác, khi áp dụng SVM dữ liệu cần phải được tỉ lệ về khoảng [-1,1] hoặc [0,1]. Trong [4] các tác giả giải thích tại sao chúng ta phải tỉ lệ dữ liệu khi sử dụng mạng Noron, điều này cũng tương tự như khi chúng ta sử dụng SVM.

Một phương pháp tiêu chuẩn để điều chỉnh giá trị của các thuộc tính là lấy giá trị của mỗi thuộc tính trừ đi giá trị trung bình của nó sau đó tiếp tục chia giá trị của các thuộc tính cho giá trị độ lệch chuẩn của thuộc tính đó. Kết quả của phương pháp này là hầu hết các giá trị của các thuộc tính sau khi điều chỉnh sẽ nằm trong khoảng [-1, 1]. Phương pháp trên chỉ áp dụng khi các giá trị của các thuộc tính được phân bố theo phân phối chuẩn. Khi không biết được chính xác sự phân bố của các giá trị trong các thuộc tính thì một phương pháp thường được sử dụng là phương pháp trung bình 0 và phạm vi 2 (min = -1 và max = 1) [4, 2]. Với các phương pháp trên thì tất cả các thuộc tính trong tập dữ liệu sẽ được tỉ lệ theo cùng một phương pháp. Trong bài báo này chúng tôi đề xuất phương pháp sử dụng giải thuật di truyền để lựa chọn phương pháp tỉ lệ riêng rẽ cho từng thuộc tính.

2. Phương pháp máy tựa Vector

Việc sử dụng phương pháp máy tựa vector SVM trong việc phân loại dữ liệu hiện đang được áp dụng trong rất nhiều lĩnh vực. Trong lĩnh vực về khoa học trái đất thì phương pháp máy tựa vector được áp dụng cho các bài toán như phân loại ảnh viễn thám [6], nhận dạng, phân loại đất v.v…

Phương pháp tựa vector ánh xạ các vector đầu vào x sang không gian đặc trưng có số chiều cao hoặc vô hạn chiều ($z = \phi(x)$) sau đó xây dựng một siêu phẳng tối ưu $w.z + b = 0$ để phân loại dữ liệu thành hai lớp. Phương pháp máy tựa vector giải quyết bài toán tối ưu sau:

$$\min \quad \frac{1}{2}\| w\|^2 + \quad \frac{1}{2}C\sum_{i=1}^{N}\xi_i \quad , \qquad (1)$$

với các ràng buộc:

$$y_i\left(\langle w.x_i\rangle - b\right) \geq 1-\xi_i , \quad \xi_i \geq 0, \quad \forall i = 1....N , \qquad (2)$$

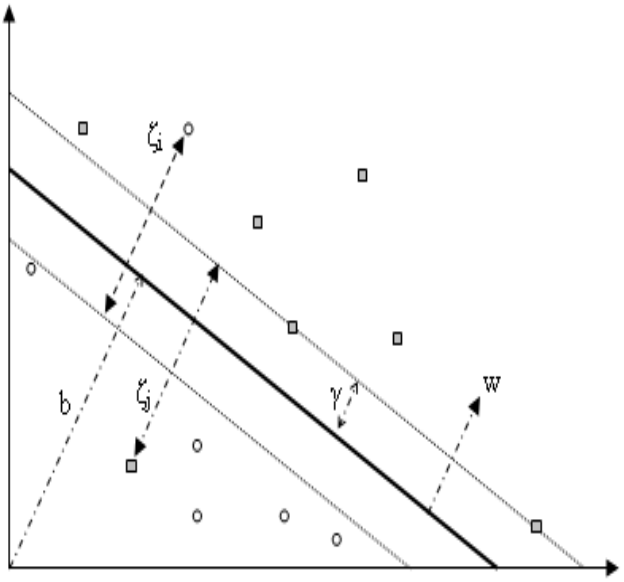
trong đó:

Mỗi x_i là một vectơ thực m chiều.

Ta cần tìm siêu phẳng có lề lớn nhất chia tách các điểm có $y_i = 1$ và các điểm có $y_i = -1$

w là một vectơ pháp tuyến của siêu phẳng.

Các biến bù ξ_i dùng để đo độ sai lệch của x_i



Bằng cách thêm các nhân tử Lagrange α, bài toán trên trở thành

$$\max \quad \sum_{i=1}^{N}\alpha_i - \frac{1}{2}\sum_{i,j=1}^{N}\alpha_i\alpha_j y_i y_j k(x_i, x_j), \quad (3)$$

với các ràng buộc

$$0 \le \alpha_i \le C, \forall i, \quad \sum_{i=1}^{N} y_i\alpha_i = 0 \ , \qquad (4)$$

trong đó $k(x_i, x_j) = \Phi(x_i).\ \Phi(x_j)$ là hàm hạt nhân (kernel function) thực hiện ánh xạ phi tuyến. Một số hàm hạt nhân thường được sử dụng là:

Gaussiankernel:

$$k(x_i, x_j) = \exp\left(-\frac{\| x_i - x_j \|^2}{2\sigma^2}\right)$$

Polynomial kernel:

$$k(x_i, x_j) = (1 + x_i.x_j)^d$$

RBF kernel:

$$k(x_i, x_j) = \exp(-\gamma \| x_i - x_j \|^2)$$

3. Lựa chọn phương pháp tỉ lệ dữ liệu sử dụng giải thuật di truyền

3.1. Giải thuật di truyền

Giải thuật di truyền là một kỹ thuật của khoa học máy tính nhằm tìm kiếm giải pháp thích hợp cho các bài toán tối ưu tổ hợp (*combinatorial optimization*). Giải thuật di truyền là một phân ngành của giải thuật tiến hóa vận dụng các nguyên lý của tiến hóa như di truyền, đột biến, chọn lọc tự nhiên, và trao đổi chéo. Giải thuật di truyền thực hiện tiến trình tìm kiếm lời giải tối ưu theo nhiều hướng bằng cách duy trì một quần thể các lời giải và thúc đẩy sự hình thành và trao đổi thông tin giữa các hướng này. Quần thể trải qua quá trình tiến hóa, ở mỗi thế hệ phát sinh lời giải tương đối "tốt", trong khi các lời giải tương đối "xấu" thì bị loại đi. Để phân biệt các lời giải khác nhau người ta sử dụng hàm mục tiêu. Mỗi cá thể trong một quần thể gọi là một nhiễm sắc thể (chromosome). Nhiễm thể là một chuỗi nhị phân gồm n bit, trong bài toán lựa chọn phương pháp tỉ lệ dữ liệu n là số thuộc tính trong tập dữ liệu. Mỗi bit trong một nhiễm sắc thể biểu diễn cho một thuộc tính. Nếu bit bằng 1 thì lấy căn bậc 3 của từng giá trị trong thuộc tính tương ứng sau đó chuyển đổi về khoảng [-1, 1] theo phương pháp trung bình 0 và phạm vi 2. Nếu bit bằng 0 thì các giá trị trong thuộc tính tương ứng sẽ được chuyển đổi về khoảng [-1, 1] theo phương pháp trung bình 0 và phạm vi 2.

Ta kí hiệu X_i là giá trị thứ i của một thuộc tính và S_i là giá trị sau khi tỉ lệ của X_i. Việc tỉ lệ X_i về trung bình 0 và phạm vi 2 được thực hiện như sau:

$$m = \frac{\max X_i + \min X_i}{2}$$

$$r = \max X_i - \min X_i$$

$$S_i = \frac{X_i - m}{r/2}$$

trong đó m là giá trị trung bình, r là giá trị phạm vi.

Chúng tôi sử dụng kết quả phân loại khi dụng SVM với hàm hạt nhân RBF làm giá trị của hàm mục tiêu. Khi sử dụng SVM với hàm hạt nhân RBF có hai tham số cần được thiết lập trước đó là tham số C và tham số γ, ở đây chúng tôi sử dụng phương pháp tìm kiếm lưới để xác định các tham số C và γ tối ưu cho từng tập dữ liệu cụ thể [1].

3.2. Thuật toán

Thuật toán sử dụng giải thuật di truyền để lựa chọn phương pháp tỉ lệ dữ liệu được mô tả như sau:

```
1. t = 0;
2. Khởi tạo một quần thể C(t), gồm m cá thể xi(t) (i = 1,m);
3. while độ chính xác chưa thỏa mãn do
      3.1 For each xi(t)
            if bit thứ i của cá thể xi(t) có giá trị = 1
                Tỉ lệ thuộc tính thứ i của tập dữ liệu huấn luyện bằng cách lấy
                căn bậc 3 của các giá trị sau đó chuyển đổi về trung bình 0,
                phạm vi 2;
            Else
                Tỉ lệ thuộc tính thứ i của tập dữ liệu huấn luyện bằng cách
                chuyển đổi về trung bình 0, phạm vi 2;
            end if
      End for
      3.2 Sử dụng SVM để phân loại tệp dữ liệu đã được tỉ lệ và đánh giá độ
chính xác đạt được (Tính giá trị hàm mục tiêu trên tệp dữ liệu đã được tỉ lệ)
      3.3 Thực hiện các thao tác di truyền;
      3.4 Lựa chọn quần thể mới C(t + 1);
      3.5 t = t + 1;
end
```


3.3. Thực hiện

Để đánh giá hiệu quả của phương pháp đề xuất, chúng tôi thực hiện một loạt các thí nghiệm trên 5 tập dữ liệu được download từ website http://www.csie.ntu.edu.tw/ ~ cjlin /libsvmtools/datasets. Đây là các tập dữ liệu chuẩn thường được sử dụng để đánh giá kết quả phân loại của các phương pháp khác nhau.

Bảng 1. Các tập dữ liệu

Tập dữ liệu	Huấn luyện	Kiểm thử	Số thuộc tính	Số lớp
Astroparticle	3089	4000	4	2
Vehicle	1243	41	21	2
Satimage	3104	2000	36	6
Adult	3185	29376	123	2
Svmguide4	300	312	10	6

Chúng tôi sử dụng phương pháp SVM để phân loại trên tập dữ liệu được tỉ lệ theo phương pháp thường được sử dụng (trung bình 0 và phạm vi 2) và phân loại trên tập dữ liệu được tỉ lệ theo phương pháp mà chúng tôi đề xuất. Từ đó so sánh độ chính xác của kết quả phân loại đạt được

Bảng 2. So sánh độ chính xác giữa phương pháp thông thường và phương pháp mà chúng tôi đề xuất

Tập dữ liệu	Phương pháp trung bình 0 và phạm vi 2		Phương pháp mà chúng tôi đề xuất	
	Huấn luyện	Kiểm thử	Huấn luyện	Kiểm thử
Astroparticle	96.8922%	96.875%	97.2807%	97%
Vehicle	84.0708%	87.8049%	85.35%	87.8049%
Satimage	92.1070%	90.35%	92.0631%	91.30%
Adult	83.9873%	84.4533%	85.2402%	84.2525%
Svmguide4	81%	89.4231%	86.33%	91.98%

4. Kết luận

Để nâng cao độ chính xác của việc phân loại dữ liệu bằng phương pháp SVM thì trước khi đưa vào huấn luyện các tập dữ liệu thường được tỉ lệ sao cho giá trị của các thuộc tính trong tập dữ liệu nằm trong khoảng [-1,1] hoặc [0,1]. Các phương pháp hiện đang được sử dụng thường áp dụng một phương pháp tỉ lệ chung cho tất cả các thuộc tính trong tập dữ liệu. Trong bài báo này chúng tôi đề xuất một thuật toán sử dụng giải thuật di truyền trong việc lựa chọn các phương pháp tỉ lệ thích hợp cho từng tính của tập dữ liệu huấn luyện. Thông qua kết quả thực nghiệm cho thấy trong nhiều trường hợp phương pháp mà chúng tôi đề xuất có kết quả tốt hơn so với phương pháp tỉ lệ vẫn thường được sử dụng.

TÀI LIỆU THAM KHẢO

[1]. Chih-Wei Hsu, Chih-Chung Chang, and Chih-Jen Lin, A practical guide to Support Vector Classification, Libsvm: a library for support vector machines, 2005.

[2]. David Skillicorn, Understanding Complex Datasets, Chapman & Hall/CRC, 2007

[3]. Nello Cristianini, John Shawe-Taylor, An Introduction to Support Vector Machines and Other Kernel-based Learning Methods, Cambridge University Press, 2000.

[4]. Sarle, W. S. Neural Network FAQ. Periodic posting to the Usenet newsgroup comp.ai.neural-nets, 1997.

[5]. S. N. Sivanandam, S. N. Deepa, Introduction to Genetic Algorithms, Springer, 2008.

[6]. Nghi Dang Huu, Mai Luong Chi, "An object-oriented classification techniques for high resolution satellite imagery" GeoInformatics for Spatial-Infrastructure Development in Earth and Allied Sciences (GIS-IDEAS), pp. 230-240, 2008

SUMMARY

A method to select a data normalize method

Dang Huu Nghi, Hoang Kim Bang, Bui Thi Van Anh

University of Mining and Geology

SVM (Support Vector Machine) is a useful technique for data classification. Normalize data before applying SVM is very important. For lack of better prior information, it is common to normalize attributes to the same range with the same method. Three of the most useful method to normalize attributes are: mean 0 and standard deviation 1, midrange 0 and range 2 or when the significance of magnitudes is non-linear the attribute values can be scaled by taking logarithms (or by taking cube roots) then transforming to midrange 0 and range 2. In this page we propose a method to use GA (Genetic Algorithm) to select a normalize method for each attribute. Our experimental results show that the mothod we proposed better than the method is often used that normalize attributes by same method.