Phân loại văn bản tiếng Việt với bộ phân loại vectơ hỗ trợ SVM

Classification of Vietnamese Documents Using Support Vector Machine

Nguyễn Linh Giang, Nguyễn Mạnh Hiển

Abstract: *In this paper, we present studies on Vietnamese document classification problem using Support Vector Machine (SVM). SVM is a learning method with ability to automatically tune the capacity of the learning machine by maximizing the margin between positive and negative examples in order to optimize the generalization performance, SVM has a large potential for the successful applications in the field of text categorization. This paper presents the results of the experiment on Vietnamese text categorization with SVM.*

Từ khóa: *Phân loại văn bản, Support Vector Machine*

I. GIỚI THIỆU

Bài toán tự động phân loại là một trong những bài toán kinh điển trong lĩnh vực xử lý dữ liệu văn bản. Đây là vấn đề có vai trò quan trọng khi phải xử lý một số lượng lớn dữ liệu. Trên thế giới đã có nhiều công trình nghiên cứu đạt những kết quả khả quan về hướng này. Tuy vậy, các nghiên cứu và ứng dụng đối với văn bản tiếng Việt còn có nhiều hạn chế. Phần nhiều lý do là đặc thù của tiếng Việt trên phương diện từ vựng và câu.

Trong lĩnh vực khai phá dữ liệu, các phương pháp phân loại văn bản đã dựa trên những phương pháp quyết định như quyết định Bayes, cây quyết định, k-láng giềng gần nhất, mạng nơron, ... Những phương pháp này đã cho kết quả chấp nhận được và được sử dụng trong thực tế. Trong những năm gần đây, phương pháp phân loại sử dụng Bộ phân loại vector hỗ trợ (SVM) được quan tâm và sử dụng nhiều trong những lĩnh vực nhận dạng và phân loại. SVM là một họ các phương pháp dựa trên cơ sở các hàm nhân (kernel) để tối thiểu hóa rủi ro ước lượng. Phương pháp SVM ra đời từ *lý thuyết học thống kê* do Vapnik và Chervonenkis xây dựng [11, 12] và có nhiều tiềm năng phát triển về mặt lý thuyết cũng như ứng dụng trong thực tiễn. Các thử nghiệm thực tế cho thấy, phương pháp SVM có khả năng phân loại khá tốt đối với bài toán phân loại văn bản cũng như trong nhiều ứng dụng khác (như nhận dạng chữ viết tay, phát hiện mặt người trong các ảnh, ước lượng hồi quy, ...). So sánh với các phương pháp phân loại khác, khả năng phân loại của SVM là tương đương hoặc tốt hơn đáng kể [1, 2, 3, 4, 10].

Vấn đề phân loại văn bản tiếng Việt được nhiều cơ sở nghiên cứu trong cả nước quan tâm trong những năm gần đây. Một số công trình nghiên cứu cũng đạt được những kết quả khả quan. Các hướng tiếp cận bài toán phân loại văn bản đã được nghiên cứu bao gồm: hướng tiếp cận bài toán phân loại bằng lý thuyết đồ thị [14], cách tiếp cận sử dụng lý thuyết tập thô [13], cách tiếp cận thống kê [15], cách tiếp cận sử dụng phương pháp học không giám sát và đánh chỉ mục [16, 17]. Nhìn chung, những cách tiếp cận này đều cho kết quả chấp nhận được. Tuy vậy để đi đến những triển khai khả thi thì vẫn cần đẩy mạnh nghiên cứu

trên hướng này. Một trong những khó khăn trong việc áp dụng những thuật toán phân loại văn bản vào tiếng Việt là xây dựng được tập hợp từ vựng của văn bản. Vấn đề này liên quan tới việc phân tách một câu thành các từ một cách chính xác. Để giải quyết vấn đề này, chúng tôi sử dụng từ điển các thuật ngữ tiếng Việt với khoảng 11.000 từ và cụm từ. Văn bản được biểu diễn dưới dạng vector và được phân loại theo phương pháp SVM.

Trong bài báo này, trước hết chúng tôi trình bày cơ sở của phương pháp SVM và các thuật toán giải bài toán quy hoạch toàn phương phát sinh từ phương pháp này. Phần tiếp theo đề cập tới bài toán phân loại văn bản trong biểu diễn vector. Chúng tôi nhấn mạnh vào khía cạnh tiền xử lý văn bản, trích chọn đặc trưng, biểu diễn văn bản, và phân tích sự phù hợp của phương pháp SVM áp dụng vào bài toán phân loại văn bản. Phần cuối là các kết quả thí nghiệm ứng dụng SVM vào phân loại văn bản tiếng Việt. Những thí nghiệm này nhằm kiểm chứng khả năng phân loại của SVM đối với văn bản tiếng Việt. Đồng thời xác định các tham số của SVM thích hợp cho các phân lớp xác định trong bài toán phân loại văn bản.

II. BỘ PHÂN LOẠI VECTOR HỖ TRỢ (SVM)

Đặc trưng cơ bản quyết định khả năng phân loại của một bộ phân loại là hiệu suất tổng quát hóa, hay là khả năng phân loại những dữ liệu mới dựa vào những tri thức đã tích lũy được trong quá trình huấn luyện. Thuật toán huấn luyện được đánh giá là tốt nếu sau quá trình huấn luyện, hiệu suất tổng quát hóa của bộ phân loại nhận được cao. Hiệu suất tổng quát hóa phụ thuộc vào hai tham số là *sai số huấn luyện* và *năng lực* của máy học. Trong đó sai số huấn luyện là tỷ lệ lỗi phân loại trên tập dữ liệu huấn luyện. Còn năng lực của máy học được xác định bằng *kích thước Vapnik-Chervonenkis (kích thước VC)*. Kích thước VC là một khái niệm quan trọng đối với một họ hàm phân tách (hay là bộ phân loại). Đại lượng này được xác định bằng số điểm cực đại mà họ hàm có thể phân tách hoàn toàn trong không gian đối tượng. Một bộ phân loại tốt là bộ phân loại có năng lực thấp nhất (có nghĩa là đơn giản nhất) và đảm bảo sai số huấn luyện nhỏ. Phương pháp SVM được xây dựng dựa trên ý tưởng này.

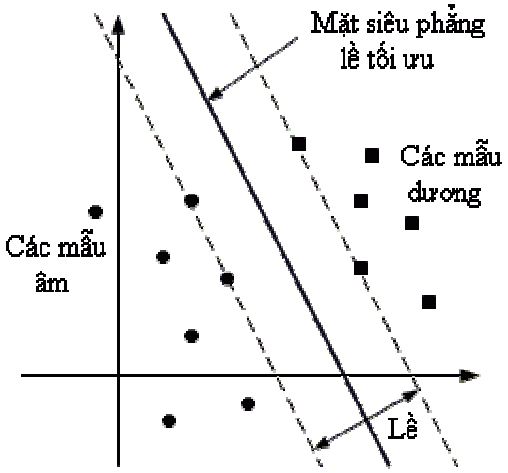
Xét bài toán phân loại đơn giản nhất - phân loại hai phân lớp với tập dữ liệu mẫu:

$$\{(\boldsymbol{x}_i, y_i) | \; i = 1, 2, ..., N, \boldsymbol{x}_i \in R^m \}$$

Trong đó mẫu là các vector đối tượng được phân loại thành các mẫu dương và mẫu âm:

– Các mẫu dương là các mẫu $\mathbf{x}_i$ thuộc lĩnh vực quan tâm và được gán nhãn $y_i = 1$;

– Các mẫu âm là các mẫu $\mathbf{x}_i$ không thuộc lĩnh vực quan tâm và được gán nhãn $y_i = -1$;



Hình 1. *Mặt siêu phẳng tách các mẫu dương khỏi các mẫu âm.*

Trong trường hợp này, bộ phân loại *SVM là mặt siêu phẳng phân tách các mẫu dương khỏi các mẫu âm với độ chênh lệch cực đại*, trong đó *độ chênh lệch – còn gọi là* ***lề*** (margin) xác định bằng khoảng cách giữa các mẫu dương và các mẫu âm gần mặt siêu phẳng nhất (hình 1). Mặt siêu phẳng này được gọi là *mặt siêu phẳng lề tối ưu*.

Các mặt siêu phẳng trong không gian đối tượng có phương trình là $\mathbf{w}^T\mathbf{x} + b = 0$, trong đó $\mathbf{w}$ là vector trọng số, b là độ dịch. Khi thay đổi $\mathbf{w}$ và b, hướng và

khoảng cách từ gốc tọa độ đến mặt siêu phẳng thay đổi. Bộ phân loại SVM được định nghĩa như sau:

$$f(\mathbf{x}) = \text{sign}(\mathbf{w}^T\mathbf{x} + b) \quad (1)$$

Trong đó

$$\text{sign}(z) = +1 \text{ nếu } z \geq 0,$$

$$\text{sign}(z) = -1 \text{ nếu } z < 0.$$

Nếu $f(\mathbf{x}) = +1$ thì $\mathbf{x}$ thuộc về lớp dương (lĩnh vực được quan tâm), và ngược lại, nếu $f(\mathbf{x}) = -1$ thì $\mathbf{x}$ thuộc về lớp âm (các lĩnh vực khác).

Máy học SVM là một họ các mặt siêu phẳng phụ thuộc vào các tham số $\mathbf{w}$ và b. Mục tiêu của phương pháp SVM là ước lượng $\mathbf{w}$ và b để cực đại hóa lề giữa các lớp dữ liệu dương và âm. Các giá trị khác nhau của lề cho ta các họ mặt siêu phẳng khác nhau, và lề càng lớn thì năng lực của máy học càng giảm. Như vậy, cực đại hóa lề thực chất là việc tìm một máy học có năng lực nhỏ nhất. Quá trình phân loại là tối ưu khi sai số phân loại là cực tiểu.

Nếu tập dữ liệu huấn luyện là *khả tách tuyến tính*, ta có các ràng buộc sau:

$$\mathbf{w}^T\mathbf{x}_i + \text{b} \geq +1 \text{ nếu } y_i = +1 \quad (2)$$

$$\mathbf{w}^T\mathbf{x}_i + \text{b} \leq -1 \text{ nếu } y_i = -1 \quad (3)$$

Hai mặt siêu phẳng có phương trình là $\mathbf{w}^T\mathbf{x} + \text{b} = \pm 1$ được gọi là các mặt siêu phẳng hỗ trợ (các đường nét đứt trên hình 1).

Để xây dựng một mặt siêu phẳng lề tối ưu, ta phải giải bài toán quy hoạch toàn phương sau:

Cực đại hóa:

$$\sum_{i=1}^{N}\alpha_i - \frac{1}{2}\sum_{i=1}^{N}\sum_{j=1}^{N}\alpha_i\alpha_j y_i y_j \mathbf{x}_i^T\mathbf{x}_j \quad (4)$$

với các ràng buộc:

$$\alpha_i \geq 0 \quad (5)$$

$$\sum_{i=1}^{N}\alpha_i y_i = 0 \quad (6)$$

trong đó các hệ số Lagrange α_i, $i = 1, 2, ..., N$, là các biến cần được tối ưu hóa.

Vector $\mathbf{w}$ sẽ được tính từ các nghiệm của bài toán toàn phương nói trên như sau:

$$\mathbf{w} = \sum_{i=1}^{N}\alpha_i y_i \mathbf{x}_i \quad (7)$$

Để xác định độ dịch b, ta chọn một mẫu $\mathbf{x}_i$ sao cho với $\alpha_i > 0$, sau đó sử dụng điều kiện Karush–Kuhn–Tucker (KKT) như sau:

$$\alpha_i [\, y_i(\mathbf{w}^T\mathbf{x}_i + \text{b}) - 1] = 0 \quad (8)$$

Các mẫu $\mathbf{x}_i$ tương ứng với $\alpha_i > 0$ là những mẫu nằm gần mặt siêu phẳng quyết định nhất (thỏa mãn dấu đẳng thức trong (2), (3)) và được gọi là các *vector hỗ trợ*. Những vector hỗ trợ là những thành phần quan trọng nhất của tập dữ liệu huấn luyện. Bởi vì nếu chỉ có các vector hỗ trợ, ta vẫn có thể xây dựng mặt siêu phẳng lề tối ưu như khi có một tập dữ liệu huấn luyện đầy đủ.

Nếu tập dữ liệu huấn luyện không khả tách tuyến tính thì ta có thể giải quyết theo hai cách.

Cách thứ nhất sử dụng một *mặt siêu phẳng lề mềm*, nghĩa là cho phép một số mẫu huấn luyện nằm về phía sai của mặt siêu phẳng phân tách hoặc vẫn ở vị trí đúng nhưng rơi vào vùng giữa mặt siêu phẳng phân tách và mặt siêu phẳng hỗ trợ tương ứng. Trong trường hợp này, các hệ số Lagrange của bài toán quy hoạch toàn phương có thêm một cận trên C dương - tham số do người sử dụng lựa chọn. Tham số này tương ứng với giá trị phạt đối với các mẫu bị phân loại sai.

Cách thứ hai sử dụng một ánh xạ phi tuyến Φ để ánh xạ các điểm dữ liệu đầu vào sang một không gian mới có số chiều cao hơn. Trong không gian này, các điểm dữ liệu trở thành khả tách tuyến tính, hoặc có thể phân tách với ít lỗi hơn so với trường hợp sử dụng không gian ban đầu. Một mặt quyết định tuyến tính trong không gian mới sẽ tương ứng với một mặt quyết định phi tuyến trong không gian ban đầu. Khi đó, bài toán quy hoạch toàn phương ban đầu sẽ trở thành:

Cực đại hóa:

$$\sum_{i=1}^{N}\alpha_i - \frac{1}{2}\sum_{i=1}^{N}\sum_{j=1}^{N}\alpha_i\alpha_j y_i y_j k(\mathbf{x}_i, \mathbf{x}_j) \quad (9)$$

với các ràng buộc:

$$0 \le \alpha_i \le C \quad (10)$$

$$\sum_{i=1}^{N} \alpha_i y_i = 0 \quad (11)$$

trong đó k là một *hàm nhân* thỏa mãn:

$$k(x_i, x_j) = \Phi(x_i)^T \cdot \Phi(x_j) \quad (12)$$

Với việc dùng một hàm nhân, ta không cần biết rõ về ánh xạ Φ. Hơn nữa, bằng cách chọn một nhân phù hợp, ta có thể xây dựng được nhiều bộ phân loại khác nhau. Chẳng hạn, nhân đa thức $k(\mathbf{x}_i, \mathbf{x}_j) = (\mathbf{x}_i^T \mathbf{x}_j + 1)^p$ dẫn đến bộ phân loại đa thức, nhân Gaussian $k(\mathbf{x}_i, \mathbf{x}_j) = \exp(-\gamma\|\mathbf{x}_i - \mathbf{x}_j\|^2)$ dẫn đến bộ phân loại RBF (Radial Basis Functions), và nhân sigmoid $k(\mathbf{x}_i, \mathbf{x}_j) = \tanh(\kappa \mathbf{x}_i^T \mathbf{x}_j + \delta)$, trong đó tanh là hàm tang hyperbol, dẫn tới mạng nơron sigmoid hai lớp (một lớp nơron ẩn và một nơron đầu ra). Tuy nhiên, một ưu điểm của cách huấn luyện SVM so với các cách huấn luyện khác là hầu hết các tham số của máy học được xác định một cách tự động trong quá trình huấn luyện.

Huấn luyện SVM

Huấn luyện SVM là việc giải bài toán quy hoạch toàn phương SVM. Các phương pháp số giải bài toán quy hoạch này yêu cầu phải lưu trữ một ma trận có kích thước bằng bình phương của số lượng mẫu huấn luyện. Trong những bài toán thực tế, điều này là không khả thi vì thông thường kích thước của tập dữ liệu huấn luyện thường rất lớn (có thể lên tới hàng chục nghìn mẫu). Nhiều thuật toán khác nhau được phát triển để giải quyết vấn đề nêu trên. Những thuật toán này dựa trên việc phân rã tập dữ liệu huấn luyện thành những nhóm dữ liệu. Điều đó có nghĩa là bài toán quy hoạch toàn phương lớn được phân rã thành các bài toán quy hoạch toàn phương với kích thước nhỏ hơn. Sau đó, những thuật toán này kiểm tra các điều kiện KKT để xác định phương án tối ưu.

Một số thuật toán huấn luyện dựa vào tính chất [6]: nếu trong tập dữ liệu huấn luyện của bài toán quy hoạch toàn phương con cần giải ở mỗi bước có ít nhất một mẫu vi phạm các điều kiện KKT, thì sau khi giải bài toán này, hàm mục tiêu sẽ tăng. Như vậy, một chuỗi các bài toán quy hoạch toàn phương con với ít nhất một mẫu vi phạm các điều kiện KKT được đảm bảo hội tụ đến một phương án tối ưu. Do đó, ta có thể duy trì một tập dữ liệu làm việc đủ lớn có kích thước cố định và tại mỗi bước huấn luyện, ta loại bỏ và thêm vào cùng một số lượng mẫu.

Chúng tôi tập trung vào nghiên cứu thuật toán huấn luyện SVM tối ưu hóa tuần tự cực tiểu (*Sequential Minimal Optimization* - SMO) [7]. Thuật toán này sử dụng tập dữ liệu huấn luyện (còn gọi là *tập làm việc*) có kích thước nhỏ nhất bao gồm hai hệ số Lagrange. Bài toán quy hoạch toàn phương nhỏ nhất phải gồm hai hệ số Lagrange vì các hệ số Lagrange phải thỏa mãn ràng buộc đẳng thức (11). Phương pháp SMO cũng có một số heuristic cho việc chọn hai hệ số Lagrange để tối ưu hóa ở mỗi bước. Mặc dù có nhiều bài toán quy hoạch toàn phương con hơn so với các phương pháp khác, mỗi bài toán con này được giải rất nhanh dẫn đến bài toán quy hoạch toàn phương tổng thể cũng được giải một cách nhanh chóng.

III. PHÂN LOẠI VĂN BẢN VÀ SVM

Phân loại văn bản là một tiến trình đưa các văn bản chưa biết chủ đề vào các lớp văn bản đã biết (tương ứng với các chủ đề hay lĩnh vực khác nhau). Mỗi lĩnh vực được xác định bởi một số tài liệu mẫu của lĩnh vực đó. Để thực hiện quá trình phân loại, các phương pháp huấn luyện được sử dụng để xây dựng bộ phân loại từ các tài liệu mẫu, sau đó dùng bộ phân loại này để dự đoán lớp của những tài liệu mới (chưa biết chủ đề).

Trong quá trình phân loại, các văn bản được biểu diễn dưới dạng vector với các thành phần (chiều) của vector này là các trọng số của các từ. Ở đây, chúng ta bỏ qua thứ tự giữa các từ cũng như các vấn đề ngữ pháp khác. Dưới đây là một số phương pháp *định trọng số từ* thông dụng:

1. **Tần suất từ** (term frequency - *TF*): Trọng số từ là tần suất xuất hiện của từ đó trong tài liệu. Cách định trọng số này nói rằng một từ là quan trọng cho

một tài liệu nếu nó xuất hiện nhiều lần trong tài liệu đó.

2. **TFIDF**: Trọng số từ là tích của tần suất từ *TF* và tần suất tài liệu nghịch đảo của từ đó và được xác định bằng công thức

$$\text{IDF} = \log(N / DF) + 1 \quad (13)$$

trong đó:

N là kích thước của tập tài liệu huấn luyện;

DF là tần suất tài liệu: là số tài liệu mà một từ xuất hiện trong đó.

Trọng số *TFIDF* kết hợp thêm giá trị tần suất tài liệu *DF* vào trọng số *TF*. Khi một từ xuất hiện trong càng ít tài liệu (tương ứng với giá trị *DF* nhỏ) thì khả năng phân biệt các tài liệu dựa trên từ đó càng cao.

Các từ được dùng để biểu diễn các tài liệu cũng thường được gọi là các *đặc trưng*. Để nâng cao tốc độ và độ chính xác phân loại, tại bước tiền xử lý văn bản, ta loại bỏ các từ không có ý nghĩa cho phân loại văn bản. Thông thường những từ này là những từ có số lần xuất hiện quá ít hoặc quá nhiều. Tuy vậy việc loại bỏ những từ này có thể không làm giảm đáng kể số lượng các đặc trưng. Với số lượng các đặc trưng lớn bộ phân loại sẽ học chính xác tập tài liệu huấn luyện, tuy vậy nhiều trường hợp cho kết quả dự đoán kém chính xác đối với các tài liệu mới. Để tránh hiện tượng này, ta phải có một tập tài liệu mẫu đủ lớn để huấn luyện bộ phân loại. Tuy vậy, thu thập được tập mẫu đủ lớn tương ứng với số lượng đặc trưng thường khó thực hiện được trong thực tế. Do đó để cho bài toán phân loại có hiệu quả thực tiễn, cần thiết phải làm giảm số lượng đặc trưng.

Có nhiều phương pháp chọn đặc trưng hiệu quả. Ở đây, chúng tôi sử dụng phương pháp *lượng tin tương hỗ* . Phương pháp này sử dụng độ đo lượng tin tương hỗ giữa mỗi từ và mỗi lớp tài liệu để chọn các từ tốt nhất. Lượng tin tương hỗ giữa từ *t* và lớp *c* được tính như sau:

$$MI(t,c) = \sum_{t\in\{0,1\}} \sum_{c\in\{0,1\}} P(t,c) \log \frac{P(t,c)}{P(t)P(c)} \quad (14)$$

trong đó:

P(t, c) là xác suất xuất hiện đồng thời của từ *t* trong lớp *c*;

P(t) là xác suất xuất hiện của từ *t* và

P(c) là xác suất xuất hiện của lớp *c*.

Độ đo MI toàn cục (tính trên toàn bộ tập tài liệu huấn luyện) cho từ *t* được tính như sau:

$$MI_{avg}(t) = \sum_i P(c_i) MI(t, c_i) \quad (15)$$

Khi sử dụng các phương pháp chọn đặc trưng, ta có thể loại bỏ đi nhiều từ quan trọng, dẫn đến mất mát nhiều thông tin, điều đó làm cho độ chính xác phân loại sẽ giảm đi đáng kể. Trong thực tế, theo thí nghiệm của Joachims [4], rất ít đặc trưng không có liên quan, và hầu hết đều mang một thông tin nào đó, vì vậy một bộ phân loại tốt nên được huấn luyện với nhiều đặc trưng nhất nếu có thể. Điều này làm cho SVM trở thành một phương pháp thích hợp cho phân loại văn bản, bởi vì giải thuật SVM có khả năng điều chỉnh năng lực phân loại tự động đảm bảo hiệu suất tổng quát hóa tốt, thậm chí cả trong không gian dữ liệu có số chiều cao (số đặc trưng rất lớn) và lượng tài liệu mẫu là có hạn.

Trong các thực nghiệm đối với bài toán phân loại văn bản tiếng Anh, phương pháp SVM cho kết quả phân loại tương đối khả quan [4]. Một trong những lý do là dữ liệu văn bản thường khả tách tuyến tính, và SVM thực hiện việc xác định mặt siêu phẳng phân tách dữ liệu tối ưu. Trong những thí nghiệm phân loại văn bản tiếng Việt được thực hiện, chúng tôi cũng nhận thấy dữ liệu văn bản tiếng Việt nói chung là khả tách. Khi dữ liệu là khả tách thì giải thuật SVM chỉ cần tập trung vào cực đại hóa lề, do đó có thể dẫn tới một hiệu suất tổng quát hóa tốt.

Một điểm đáng chú ý nữa khi huấn luyện SVM cho phân loại văn bản là ta có thể xây dựng được nhiều bộ phân loại khác nhau bằng cách chọn những hàm nhân phù hợp như đã nói trong phần II. Nhưng không như các phương pháp khác, mô hình của máy học (các

tham số **w**, *b* tối ưu) được học một cách tự động trong quá trình huấn luyện SVM.

Những phân tích trên đây cho thấy SVM có nhiều điểm phù hợp cho việc ứng dụng trong phân loại văn bản. Và trên thực tế, các thí nghiệm phân loại văn bản tiếng Anh chỉ ra rằng SVM đạt được độ chính xác phân loại cao và tỏ ra xuất sắc hơn so với các phương pháp phân loại văn bản khác. Trong phần IV của bài báo này, chúng tôi đưa ra các kết quả thí nghiệm ứng dụng SVM vào phân loại văn bản tiếng Việt.

IV. KẾT QUẢ THỰC NGHIỆM

Chúng tôi đã thực hiện một thí nghiệm ứng dụng SVM vào phân loại văn bản tiếng Việt. Tập tài liệu mẫu được sử dụng gồm 4162 tài liệu được lấy từ trang http://vnexpress.net (bảng 1). Tập tài liệu này được chia thành hai phần: 50% được dùng làm tập tài liệu huấn luyện, 50% được dùng làm tập tài liệu kiểm thử.

Việc lựa chọn các văn bản để kiểm thử thuật toán dựa vào những giả thiết sau:

– Các tài liệu được phân lớp thành những phân nhóm tách biệt. Trên thực tế, các tài liệu trên Vnexpress.net được phân loại không chính xác. Các phân lớp tài liệu có sự giao thoa và do đó một tài liệu thuộc một phân lớp có thể có những đặc trưng thuộc một phân lớp khác.

– Sự phân bố tài liệu trong một phân nhóm không ảnh hưởng tới sự phân bố tài liệu trong phân nhóm khác. Giả thiết này được đặt ra để có thể chuyển bài toán phân loại nhiều phân lớp thành các bài toán phân loại hai phân lớp.

Bộ phân loại SVM sẽ được huấn luyện trên tập tài liệu huấn luyện và hiệu suất tổng quát hóa (độ chính xác) được đánh giá trên tập tài liệu kiểm thử (tập tài liệu kiểm thử không tham gia vào quá trình huấn luyện, do đó cho phép đánh giá khách quan hiệu suất tổng quát hóa).

Bảng 1. *Tập tài liệu mẫu được dùng trong thí nghiệm phân loại văn bản tiếng Việt.*

Loại tài liệu	Huấn luyện	Kiểm thử
Âm nhạc	119	119
Ẩm thực	109	109
Bất động sản	119	119
Gia đình	85	86
Giáo dục	165	166
Hội họa	111	112
Khảo cổ	45	45
Khoa học	119	118
Kinh doanh	193	194
Pháp luật	155	154
Phim ảnh	117	117
Sức khỏe	109	108
Tâm lý	47	46
Thế giới	85	85
Thể thao	257	256
Thời trang	107	106
Vi tính	140	140

Đối với việc tiền xử lý các tài liệu, chúng tôi sử dụng một bộ từ tiếng Việt gồm 11.210 từ. Sở dĩ chúng tôi phải sử dụng từ điển từ là do đặc điểm khác biệt của tiếng Việt so với tiếng Anh trên phương diện từ vựng. Các từ tiếng Anh được ngăn cách bằng những dẫu cách, dấu câu. Do đó việc xác định ranh giới từ trong câu văn tiếng Anh có thể dựa hoàn toàn vào các dấu ngắt từ. Trong khi đó, việc xác định ranh giới từ trong câu tiếng Việt là khá khó khăn nếu không hiểu ngữ nghĩa của từ trong từng nghữ cảnh và ngữ nghĩa của câu. Ví dụ, từ "phản" và từ "động" là những từ độc lập và đều có ý nghĩa khi đứng riêng lẻ. Tuy vậy khi chúng đứng cạnh nhau tạo thành từ ghép "phản động" thì đây cũng là một từ độc lập và có ý nghĩa khác tùy theo ngữ cảnh. Như vậy để tìm ranh giới từ trong câu tiếng Việt, không thể chỉ dựa vào các dấu ngắt từ nhưn dấu cách thông thường. Để làm đơn giản hóa vấn đề nay, chúng tôi sử dụng một bộ từ tiếng Việt để hỗ trợ quá trình phân tách từ.

Bước đầu tiên của tiền xử lý là đếm số lần xuất hiện của mỗi từ trong mỗi tài liệu. Vì các từ tiếng Việt có thể bao nhau (như "áo" và "áo sơ mi"), các từ dài hơn

(theo số âm tiết) sẽ được tách ra trước. Những từ không xuất hiện lần nào (trong tập tài liệu huấn luyện) bị loại bỏ, kết quả là còn lại 7721 từ. Để thử nghiệm với những số đặc trưng khác nhau, 100 từ có tần suất cao nhất và các từ xuất hiện ít hơn 3 lần bị loại bỏ, thu được 5709 từ; sau đó, phương pháp lượng tin tương hỗ được sử dụng để chọn ra lần lượt 5000, 4000, 3000, 2000 và 1000 từ. Với mỗi số đặc trưng được chọn, các tài liệu được biểu diễn dưới dạng các vector thưa dùng cách định trọng số từ TFIDF. Mỗi vector thưa gồm hai mảng: một mảng số nguyên lưu chỉ số của các giá trị khác 0, và một mảng số thực lưu các giá trị khác 0 tương ứng. Sở dĩ dùng các vector thưa là do số từ xuất hiện trong mỗi tài liệu là rất nhỏ so với tổng số từ được sử dụng; điều này một mặt tiết kiệm bộ nhớ, mặt khác làm tăng tốc độ tính toán lên đáng kể. Các vector cũng được tỷ lệ sao cho các thành phần của nó nằm trong khoảng [0, 1], qua đó giúp tránh việc các thành phần có giá trị lớn lấn át các thành phần có giá trị nhỏ, và tránh được các khó khăn khi tính toán với các giá trị lớn.

Để thực hiện phân loại văn bản bằng phương pháp SVM, chúng tôi đã sử dụng phần mềm LIBSVM 2.71 với công cụ *grid.py* cho phép chọn tham số tối ưu cho giải thuật SVM với nhân Gaussian. Điều này được thực hiện bằng cách chia tập tài liệu huấn luyện thành v phần bằng nhau, và lần lượt mỗi phần được kiểm thử bằng bộ phân loại được huấn luyện trên $v - 1$ phần còn lại. Độ chính xác ứng với mỗi bộ giá trị của các tham số (C và γ) được tính bằng tỷ lệ tài liệu trong tập tài liệu huấn luyện được dự đoán đúng. Chú ý rằng ở đây hoàn toàn không có sự tham gia của các tài liệu trong tập tài liệu kiểm thử.

Sau khi đã chọn được các tham số C và γ tối ưu, bộ phân loại SVM sẽ được huấn luyện trên toàn bộ tập tài liệu huấn luyện, và độ chính xác của nó được đánh giá bằng cách thực hiện phân loại trên tập tài liệu kiểm thử. LIBSVM thực hiện phân loại đa lớp (trong trường hợp của bài báo này là 17 lớp) theo kiểu "một-đấu-một" (one-against-one), nghĩa là cứ với hai lớp thì sẽ huấn luyện một bộ phân loại, kết quả là sẽ có tổng cộng $k(k-1)/2$ bộ phân loại, với k là số lớp. Đối với hai lớp thứ i và thứ j, một tài liệu chưa biết **x** sẽ được phân loại bằng bộ phân loại được huấn luyện trên hai lớp này. Nếu **x** được xác định là thuộc lớp i thì điểm số cho lớp i được tăng lên 1, ngược lại điểm số cho lớp j được tăng lên 1. Ta sẽ dự đoán **x** nằm trong lớp có điểm số cao nhất. Trong trường hợp có hai lớp bằng nhau về điểm số này, ta chỉ đơn giản chọn lớp có số thứ tự nhỏ hơn.

Trở lại thí nghiệm, các tham số tối ưu được tìm trong số 110 bộ giá trị (C, γ) thử nghiệm (với $C = 2^{-5}$, 2^{-3}, ..., 2^{15}, $\gamma = 2^{3}$, 2^{1}, ..., 2^{-15}). Kết quả chọn tham số được đưa ra trong bảng 2.

Từ bảng 2, ta thấy các tham số tốt nhất là 7721 đặc trưng, $C = 2^{15}$ và $\gamma = 2^{-13}$. Như vậy, trong trường hợp thí nghiệm này, các phương pháp chọn đặc trưng đã không đem lại kết quả như mong muốn – chúng làm giảm độ chính xác. Với các tham số trên, bộ phân loại SVM được huấn luyện trên toàn bộ tập tài liệu huấn luyện, sau đó độ chính xác của nó được đánh giá trên tập tài liệu kiểm thử, cho kết quả như trong bảng 3.

Bảng 2. *Các tham số tối ưu tương ứng với mỗi số lượng đặc trưng.*

Số đặc trưng	(C, γ) tốt nhất	Độ chính xác (%)
7721	$(2^{15}, 2^{-13})$	82,90
5709	$(2^{13}, 2^{-11})$	82,04
5000	$(2^{11}, 2^{-11})$	80,40
4000	$(2^{5}, 2^{-5})$	78,58
3000	$(2^{5}, 2^{-5})$	78,34
2000	$(2^{7}, 2^{-5})$	73,87
1000	$(2^{3}, 2^{-3})$	71,57

Bảng 3. *Độ chính xác phân loại trên mỗi lớp và trên toàn bộ tập tài liệu kiểm thử.*

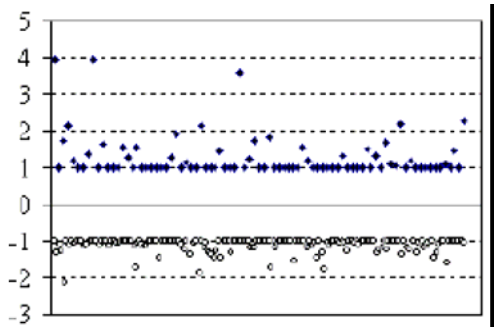
Loại tài liệu	Độ chính xác (%)
Âm nhạc	72,27
Ẩm thực	93,58
Bất động sản	94,12
Gia đình	72,09

Giáo dục	79,52
Hội họa	82,14
Khảo cổ	51,11
Khoa học	65,25
Kinh doanh	83,51
Pháp luật	94,81
Phim ảnh	66,67
Sức khỏe	78,70
Tâm lý	39,13
Thế giới	71,76
Thể thao	98,05
Thời trang	76,42
Vi tính	79,29
Tất cả	80,72

Trong bảng 3, độ chính xác trên tất cả các lớp tài liệu là 80,72% được tính bằng tỷ số giữa số tài liệu được dự đoán đúng trên tổng số tài liệu của tập tài liệu kiểm thử.

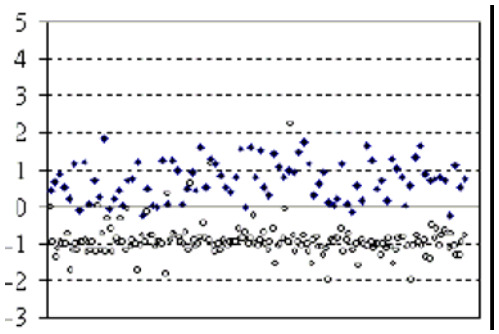
Hình 2 là đồ thị minh họa cho trường hợp bộ phân loại SVM được huấn luyện trên hai lớp tài liệu gia đình và giáo dục. Hình 2a cho thấy sự phân bố của các điểm dữ liệu huấn luyện, còn hình 2b cho thấy sự phân bố của các điểm dữ liệu kiểm thử. Ta nhận thấy rằng không có một lỗi vị trí nào trên hình 2a, nhưng lại có một vài lỗi vị trí trên hình 2b. Trong trường hợp này, máy học SVM đã học chính xác tập tài liệu huấn luyện (khả tách tuyến tính) nhưng mắc phải một vài sai sót khi dự đoán các tài liệu chưa biết (các tài liệu kiểm thử).

Những kết quả thực nghiệm trong thí nghiệm phân loại các văn bản tiếng Việt bằng bộ phân loại SVM có độ chính xác chưa được cao (khoảng 80,72%). Điều này có thể do quá trình tiền xử lý văn bản và những dữ liệu huấn luyện cùng với dữ liệu thử nghiệm được phân loại chưa chính xác. Thật vậy đây là những dữ liệu thu thập trên Vnexpress.net và không được phân loại chuẩn. Một văn bản, ví dụ thuộc lĩnh vực "Bất động sản" hoàn toàn có thể thuộc cả lĩnh vực "Kinh doanh". Như vậy các phân lớp văn bản mẫu trên thực tế không hoàn toàn phân tách tuyến tính mà có vùng không gian mập mờ. Điều này ảnh hưởng khá mạnh đến quá trình huấn luyện bộ phân loại.



(a) huấn luyện chỉ gồm hai lớp gia đình và giáo dục

♦ Các tài liệu gia đình ◊ Các tài liệu giáo dục



(b) kiểm thử chỉ gồm hai lớp gia đình và giáo dục

Hình 2. *Đồ thị giá trị của* $f(\mathbf{x}) = \mathbf{w}^T \mathbf{x} + b$ *tại các tài liệu* $\mathbf{x}$ *trong tập tài liệu*

Tuy vậy trong những ứng dụng thực tế như phân loại trang Web, hoặc xử lý phân loại khối lớn văn bản thì kết quả này có thể chấp nhận được. Vấn đề đặt ra cho những nghiên cứu tiếp theo là:

- Xây dựng được hệ thống dữ liệu thử nghiệm tiêu chuẩn. Đây là vấn đề lớn và cần tập trung nhiều công sức;
- Thử nghiệm bộ phân loại với những hàm nhân khác nhau để chọn được nhân tối ưu đối với một tập hợp dữ liệu kiểm thử.

V. KẾT LUẬN

Trong bài báo này, chúng tôi đã khảo sát hiệu quả phương pháp phân loại SVM. Đây là bộ phân loại có khả năng tự động điều chỉnh các tham số để tối ưu hóa hiệu suất phân loại thậm chí trong những không gian đặc trưng có số chiều cao. Bộ phân loại SVM tỏ ra phù hợp cho phân loại văn bản. Trong thử nghiệm với

bài toán phân loại văn bản tiếng Việt, độ chính xác phân loại là 80,72% có thể chấp nhận được trong những điều kiện thực tế. Hiện tại, chúng tôi đang tiếp tục nghiên cứu cải tiến khâu tiền xử lý văn bản, xây dựng các mẫu huấn luyện tiêu chuẩn cũng như điều chỉnh giải thuật SVM để có thể nâng cao độ chính xác phân loại hơn nữa.

TÀI LIỆU THAM KHẢO

[1] B. BOSER, I. GUYON, V. VAPNIK, "*A training algorithm for optimal margin classifiers",* Proceedings of the Fifth Annual Workshop on Computational Learning Theory (ACM), pp 144-152, 1992.

[2] C. BURGES, *"A tutorial on Support Vector Machines for pattern recognition"*, Proceedings of Int Conference on Data Mining and Knowledge Discovery, Vol 2, No 2, pp 121-167, 1998.

[3] S. DUMAIS, J. PLATT, D. HECKERMAN, M. SAHAMI, *"Inductive learning algorithms and representations for text categorization"*, Proceedings of Conference on Information and Knowledge Management (CIKM), pp 148-155, 1998.

[4] T. JOACHIMS, *"Text categorization with Support Vector Machines: Learning with many relevant features"*, Technical Report 23, LS VIII, University of Dortmund, 1997.

[5] S. HAYKIN, *Neural networks: A comprehensive foundation,* Prentice Hall, 1998.

[6] E. OSUNA, R. FREUND, F. GIROSI, An improved training algorithm for Support Vector Machines, Neural Networks for Signal Processing VII –Proceedings of the 1997 IEEE Workshop, pp 276-285, New York, IEEE, 1997.

[7] J. PLATT, *Sequential minimal optimization: A fast algorithm for training Support Vector Machines*, Technical Report MSR-TR-98-14, Microsoft Research, 1998.

[8] C.J. VAN RIJSBERGEN, *Information Retrieval*, Butterworths, London, 1979.

[9] Y. YANG, X. LIU, "*A re-examination of text categorization methods",* Proceedings of the 22th Ann Int ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR'99), pp 42-49, 1999.

[10]Y. YANG, J. PEDERSEN, *"A comparative study on feature selection in text categorization"*, Proceedings of the 14th International Conference on Machine Learning (ICML), pp 412-420, Morgan & Kaufmann 1997.

[11]V. VAPNIK, *"Nature of statistical learning theory"*, Springer-Verlag, 2000

[12]V. N. VAPNIK, A. YA. CHERVONENKIS, *Teoria Raspoznavaniya Obrazov*, Nauka, 1974

[13]NGUYỄN NGỌC BÌNH, *"Dùng lý thuyết tập thô và các kỹ thuật khác để phân loại, phân cụm văn bản tiếng Việt"*, Kỷ yếu hội thảo ICT.rda'04. Hà nội 2004

[14]ĐỖ BÍCH DIỆP, *"Phân loại văn bản dựa trên mô hình đồ thị",* Luận văn cao học. Trường Đại học Tổng hợp New South Wales - Australia. 2004.

[15]NGUYỄN LINH GIANG, NGUYỄN DUY HẢI, *"Mô hình thống kê hình vị tiếng Việt và ứng dụng"*, Chuyên san "Các công trình nghiên cứu, triển khai Công nghệ Thông tin và Viễn thông, Tạp chí Bưu chính Viễn thông, số 1, tháng 7-1999, trang 61-67. 1999

[16]HUỲNH QUYẾT THẮNG, ĐINH THỊ PHƯƠNG THU, *"Tiếp cận phương pháp học không giám sát trong học có giám sát với bài toán phân lớp văn bản tiếng Việt và đề xuất cải tiến công thức tính độ liên quan giữa hai văn bản trong mô hình vector*", Kỷ yếu Hội thảo ICT.rda'04, trang 251-261, Hà Nội 2005.

[17]ĐINH THỊ PHƯƠNG THU, HOÀNG VĨNH SƠN, HUỲNH QUYẾT THẮNG, *"Phương án xây dựng tập mẫu cho bài toán phân lớp văn bản tiếng Việt: nguyên lý, giải thuật, thử nghiệm và đánh giá kết quả"*, Bài báo đã gửi đăng tại Tạp chí khoa học và công nghệ, 2005.

Ngày nhận bài: 8/6/2005

SƠ LƯỢC TÁC GIẢ

NGUYỄN LINH GIANG

Sinh năm 1968 tại Hà Nội

Tốt nghiệp Đại học năm 1991 và nhận học vị Tiến sỹ tại Liên Xô cũ chuyên ngành Đảm bảo Toán học cho máy tính năm 1995.

Hiện đang Khoa Công nghệ Thông tin, Đại học Bách khoa Hà Nội.

Lĩnh vực nghiên cứu: Điều khiển tối ưu, xử lý văn bản tiếng Việt, an toàn mạng, multimedia

Email: giangnl@it-hut.edu.vn

NGUYỄN MẠNH HIỂN

Sinh năm 1981

Tốt nghiệp Đại học chuyên ngành Truyền thông và Mạng, Đại học Bách khoa Hà Nội năm 2004.

Hiện đang công tác tại Khoa Công nghệ Thông tin, Đại học Thủy Lợi .

Lĩnh vực nghiên cứu: Học máy, khai phá dữ liệu tiếng Việt

Email: nmhien@gmail.com