NHẬN DẠNG CHỮ VIẾT TAY RỜI RẠC TRÊN CƠ SỞ PHƯƠNG PHÁP MÁY VÉC TƠ TỰA

PHẠM ANH PHƯƠNG *
LÊ THANH LONG**
VÕ VĂN LƯỜNG**

ABSTRACT

This paper proposes an isolated handwritten character recognition model based on Support Vector Machines. Our experiments on the benchmark database MNIST and samples of Vietnamese handwriting character show that our recognition model reached higher accuracy than neuron network model. We also evaluate the advantages and disadvantages of SVM and propose research solutions.

Key words: *Handwritten Character Recognition; SVM.*

1. Giới thiệu

Cho đến nay, việc nhận dạng chữ viết tay vẫn chưa có được một giải pháp tổng thể, các ứng dụng của nó cũng chỉ giới hạn trong phạm vi hẹp. Các kết quả chủ yếu về lĩnh vực này chỉ tập trung trên các tập dữ liệu chữ số viết tay chuẩn như USPS và MNIST [2,3,6,8], bên cạnh đó cũng có một số công trình nghiên cứu trên các hệ chữ cái tiếng La tinh, Hy Lạp, Trung Quốc, Việt Nam... tuy nhiên các kết quả đạt được cũng còn nhiều hạn chế [4,5,7,8].

Các giải pháp tiếp cận để giải bài toán nhận dạng chữ viết tay khá phong phú, một số phương pháp học máy thường được áp dụng như: mô hình Markov ẩn, mạng nơ ron hay phương pháp máy véc tơ tựa (SVM - Support Vector Machines). Trong đó SVM được đánh giá là phương pháp học máy tiên tiến đang được áp dụng rộng rãi trong các lĩnh khai phá dữ liệu và thị giác máy tính… SVM gốc được thiết kế để giải bài toán phân lớp nhị phân, ý tưởng chính của phương pháp này là tìm một siêu phẳng phân cách sao cho khoảng cách lề giữa hai lớp đạt cực đại. Khoảng cách này được xác định bởi các véc tơ tựa (SV - Support Vector), các SV này được lọc ra từ tập mẫu huấn luyện bằng cách giải một bài toán tối ưu lồi [3].

Trong bài báo này, chúng tôi sẽ xây dựng mô hình nhận dạng chữ viết tay rời rạc dựa trên phương pháp SVM, đồng thời tiến hành cài đặt thử nghiệm trên các tập dữ liệu chữ số viết tay chuẩn MNIST và dữ liệu chữ viết tay tiếng Việt do chúng tôi tự thu thập.

 ** TS, ** ThS, Trường Đại học Duy Tân*

Phần còn lại của bài báo này có cấu trúc như sau: Phần 2 trình bày kiến trúc của mô hình nhận dạng chữ viết tay rời rạc. Phần 3 trình bày một số kết quả thực nghiệm trên tập dữ liệu chữ số viết tay MNIST và dữ liệu chữ viết tay tiếng Việt, so sánh kết quả đạt được với mô hình nhận dạng sử dụng mạng nơ ron. Phần 4 đánh giá các ưu và nhược điểm khi áp dụng SVM vào bài toán nhận dạng chữ viết tay tiếng Việt. Cuối cùng là phần kết luận và hướng phát triển.

2. Mô hình nhận dạng chữ viết tay rời rạc

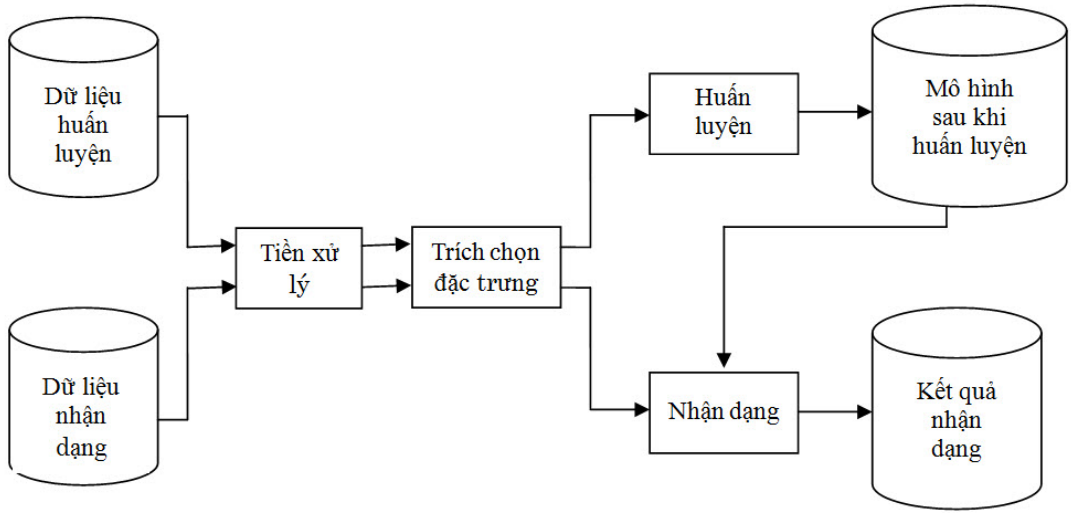
Trong phần này, chúng tôi sẽ tập trung xây dựng mô hình nhận dạng chữ viết tay rời rạc theo phương pháp phân lớp SVM. Công việc được thực hiện theo hai bước chính sau đây:

Bước 1: Xây dựng mô hình huấn luyện.

Tập dữ liệu huấn luyện sau khi qua các khâu tiền xử lý và trích chọn đặc trưng sẽ được đưa vào máy huấn luyện phân lớp SVM. Sau khi kết thúc quá trình huấn luyện, hệ thống sẽ lưu lại giá trị các tham số của hàm quyết định phân lớp để phục vụ cho việc nhận dạng sau này. Quá trình huấn luyện tiêu tốn khá nhiều thời gian, tốc độ huấn luyện nhanh hay chậm tùy thuộc vào từng thuật toán huấn luyện, chiến lược phân lớp SVM cũng như số lượng mẫu tham gia huấn luyện.

Bước 2: Phân lớp nhận dạng.

Dựa vào giá trị các tham số của hàm quyết định thu được ở Bước 1, một mẫu mới x sau khi đã qua các khâu tiền xử lý và trích chọn đặc trưng sẽ được đưa vào tính toán thông qua hàm quyết định để xác định lớp của mẫu x (Hình 1).



Hình 1. Mô hình nhận dạng chữ viết tay rời rạc.

2.1. Tiền xử lý

Sau khi đã khử nhiễu, ảnh được chuẩn hóa về kích thước chuẩn 16´16. Việc chuẩn hóa kích thước ảnh được thực hiện theo các bước sau:

Bước 1: Nhị phân hóa ảnh.

Bước 2: Tìm hình chữ nhật R bé nhất chứa các điểm đen trên ảnh.

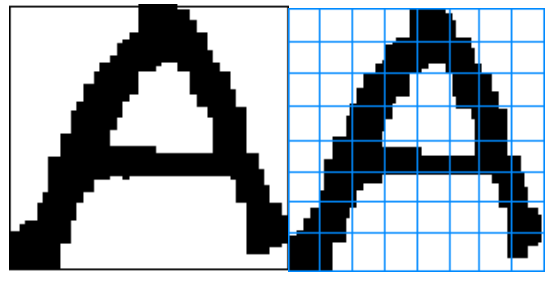
Bước 3: Lấy vùng ảnh I nằm trong hình chữ nhật R.

Bước 4: Chuẩn hóa ảnh I về kích thước chuẩn 16´16.

2.2. Trích chọn đặc trưng

Trong phần này, chúng tôi sẽ chọn phương pháp trích chọn đặc trưng đơn giản nhưng hiệu quả, có thể áp dụng cho các tập dữ liệu chữ viết tay rời rạc.

Ảnh ký tự sau khi đã chuẩn hóa về kích thước chuẩn sẽ được chia thành N´N vùng (Hình 2). Tổng số điểm đen của mỗi vùng sẽ được chọn để tạo thành các vectơ đặc trưng.



Hình 2. Trích chọn đặc trưng trọng số vùng.

Trong thực nghiệm, với ảnh kích thước 16´16, chọn N=8, như vậy có 8´8 = 64 đặc trưng.

2.3. Lựa chọn thuật toán huấn luyện phân lớp

Trong phần cài đặt thực nghiệm, chúng tôi áp dụng thuật toán SMO để huấn luyện phân lớp SVM nhị phân, sử dụng và kế thừa một số chức năng của phần mềm mã nguồn mở LibSVM [1] để phát triển ứng dụng nhận dạng chữ viết tay rời rạc.

2.4. Thuật toán nhận dạng chữ viết tay rời rạc

Cả hai chiến lược phân lớp OVO và OVR đều có thể áp dụng để phân lớp dữ liệu một cách tổng quát mà không cần phải can thiệp sâu để phân tích các đặc trưng khác nhau giữa các lớp dữ liệu [6]. Vì vậy hai chiến lược phân lớp này sẽ được chúng tôi lựa chọn để cài đặt thử nghiệm thuật toán nhận dạng đối với dữ liệu chữ viết tay rời rạc.

Procedure **SVMClassify**

//Thuật toán phân lớp theo 2 chiến lược OVO và OVR

Input:

- Mẫu x;
- Số lớp N;
- Chiến lược phân lớp Strategy;
- Các mô hình đã huấn luyện {OVOModel, OVRModel}

Output:

label; // *Nhãn lớp của mẫu x*

Method

1. Case Strategy of
2. OVO:// *Chiến lược một đối một*
3. Khởi tạo Count[i] = 0; // *i=0,..,N-1*
4. LoadModel(OVOModel);
5. for (i=0; i < N-1; i++)
6. for (j=i+1; j < N; j++)
7. Count[BinarySVM(x,i,j)]++;
8. Count[label]=Max(Count[i]);
9. OVR: // *Chiến lược một đối phần còn lại*
10. LoadModel(OVRModel);
11. label=-1;
12. for (i=0; i < N; i++)
13. {
14. label=BinarySVM(x,i,Rest);
15. if(label=i) break;
16. }
17. EndCase;
18. Return label;

Trong đó:

BinarySVM(x,i,j) là hàm xếp x vào một trong hai lớp i hoặc j,

Count[ ] là mảng biến đếm để lưu số lần nhận diện của các lớp.

3. Kết quả thực nghiệm

Các kết quả thực nghiệm được cài đặt và chạy thử nghiệm trên môi trường Window XP, máy PC Pentium 4 tốc độ 2.4 Ghz với dung lượng bộ nhớ RAM 1Gb.

3.1. Chuẩn bị các bộ dữ liệu thực nghiệm

Bộ dữ liệu chuẩn MNIST

Bộ dữ liệu MNIST bao gồm 60.000 mẫu huấn luyện và 10.000 mẫu khác để nhận dạng, mỗi mẫu là một ảnh kích thước 28´28.

Bộ dữ liệu chữ viết tay tiếng Việt

Chúng tôi xây dựng bộ dữ liệu chữ viết tay tiếng Việt **(VietData)** phục vụ cho việc thực nghiệm bao gồm 89 lớp chữ cái in hoa, mỗi lớp chọn ra 200 mẫu, như vậy bộ dữ liệu VietData có tổng cộng 17800 mẫu.

3.2. Kết quả thực nghiệm trên bộ dữ liệu MNIST

Đầu tiên chúng tôi thử nghiệm hiệu quả của Thuật toán **SVMClassify** trên bộ dữ liệu MNIST với các chiến lược OVO và OVR. Mô hình SVM được sử dụng với hàm nhân Gauss và các tham số C = 10 (tham số hàm phạt), Cache = 1000 (kích thước vùng nhớ để lưu trữ các vectơ tựa).

Bảng 1: Kết quả thực nghiệm trên tập MNIST với hàm nhân RBF(s =0.08).

Chiến lược	*Số vectơ tựa*	*Thời gian huấn luyện*	*Thời gian Test*	*Độ chính xác*
OVR	8542	> 9 giờ	~ 3 phút	96,1%
OVO	31280	~ 2 giờ	~ 5 phút	97,2%

Kết quả thực nghiệm ở Bảng 1 cho thấy các chiến lược OVO và OVR đều có các ưu điểm và nhược điểm riêng.

Chúng tôi so sánh hiệu quả phân lớp của SVM so với phương pháp sử dụng mô hình mạng nơ ron 4 lớp (144 nơ ron lớp vào, 72+36 nơ ron ở các lớp ẩn, 10 nơ ron lớp ra) [8] trên cùng một bộ dữ liệu chuẩn MNIST (Bảng 2).

Bảng 2: So sánh kết quả nhận dạng của VM với mô hình mạng nơ ron.

Các thông số	**Mạng nơ ron**	**SVM**
Số mẫu học	60.000	60.000
Thời gian học	~ 24 giờ	~ 2 giờ
Số mẫu test	10.000	10.000
Thời gian test	~ 2 phút	~ 5 phút
Tỷ lệ test lỗi (%)	4.6	**2.8**

Kết quả ở Bảng 2 cho thấy kết quả nhận dạng theo mô hình SVM có độ chính xác cao hơn so với mô hình mạng nơ ron, tuy nhiên tốc độ nhận dạng của SVM thì chậm hơn.

3.3. Kết quả thực nghiệm trên dữ liệu chữ viết tay tiếng Việt

Việc thực nghiệm trên dữ liệu chữ viết tay tiếng Việt được tiến hành theo phương thức thẩm định chéo (Cross-Validation). Bộ dữ liệu **VietData** được chia thành k phần (ở đây k được chọn =10), sau đó sử dụng k-1 phần để huấn luyện và 1 phần còn lại để nhận dạng, quá trình được này được lặp đi lặp lại k lần. Các kết quả thực nghiệm được thể hiện trên Bảng 3.

Kết quả thực nghiệm ở Bảng 3 cho thấy tốc độ phân lớp của SVM đối với bài toán phân đa lớp là quá chậm, không thể đáp ứng được đối với một hệ thống nhận dạng thời gian thực. Vì vậy, cần phải có những giải pháp phù hợp để tăng tốc độ cũng như độ chính xác phân lớp đối với dữ liệu chữ viết tay tiếng Việt.

Bảng 3: Thực nghiệm trên tập dữ liệu chữ viết tay tiếng Việt.

Chiến lược	*Thời gian huấn luyện*	*Thời gian Test*	*Độ chính xác*
OVR	~ 49 phút	~ 2 phút	82.7%
OVO	~ 16 phút	~ 6 phút	83.6%

4. Đânhs giá hiệu quả phân lớp của SVM

Áp dụng phương pháp phân lớp SVM vào bài toán nhận dạng chữ viết tay rời rạc, chúng tôi có một số nhận xét sau đây:

- SVM là một phương pháp học máy tiên tiến có cơ sở toán học chặt chẽ và đạt độ chính xác phân lớp cao. Tuy nhiên, hạn chế lớn nhất của SVM là tốc độ phân lớp chậm, tùy thuộc vào số lượng vectơ tựa thu được sau khi huấn luyện. Một hạn chế khác của SVM là pha huấn luyện đòi hỏi không gian nhớ lớn, vì vậy việc huấn luyện đối với các bài toán có số lượng mẫu lớn sẽ gặp trở ngại trong vấn đề lưu trữ.

- Bản chất nhị phân cũng là một hạn chế của SVM, việc mở rộng khả năng của SVM để giải quyết các bài toán phân loại nhiều lớp là vấn đề không tầm thường. Có nhiều chiến lược được đề xuất để mở rộng SVM cho bài toán phân loại nhiều lớp với những điểm mạnh, yếu khác nhau tùy thuộc vào từng loại dữ liệu cụ thể. Cho đến nay, việc lựa chọn các chiến lược phân lớp vẫn thường được tiến hành trên cơ sở thực nghiệm.

- Bài toán huấn luyện SVM thực chất là bài toán qui hoạch toàn phương (QP) trên một tập lồi, do đó luôn luôn tồn tại nghiệm toàn cục và duy nhất, đây là điểm khác biệt rõ nhất giữa SVM so với mạng nơ ron, vì mạng nơ ron vốn tồn tại nhiều cực trị địa phương. Bản chất của SVM là việc phân lớp được thực hiện gián tiếp trong không gian đặc trưng với số chiều cao hơn số chiều của không gian đầu vào thông qua hàm nhân. Do đó, hiệu quả phân lớp của SVM phụ thuộc vào hai yếu tố: giải bài toán QP và lựa chọn hàm nhân. Việc giải bài toán QP luôn luôn đạt được giải pháp tối ưu nên mọi cố gắng trong nghiên cứu lý thuyết SVM tập trung vào việc lựa chọn hàm nhân. Lựa chọn hàm nhân và các tham số của nó như thế nào để SVM phân lớp tốt nhất vẫn là một bài toán mở.

- Tốc độ phân lớp của SVM bị đánh giá là chậm so với các phương pháp phân lớp khác, tùy thuộc vào số lượng vectơ tựa thu được sau khi huấn luyện. Vì vậy, có nhiều công trình tập trung nghiên cứu để giảm tối đa số lượng vectơ tựa nhằm tăng tốc độ phân lớp của SVM, một số kết quả nghiên cứu có giá trị về SVM đã được công bố trong các công trình [1,3,5].

Muốn áp dụng kỹ thuật phân lớp SVM vào bài toán nhận dạng chữ viết tay tiếng Việt, cần phải có những giải pháp để tránh bùng nổ số phân lớp cũng như giảm tối đa số vectơ tựa để tăng tốc độ nhận dạng.

5. Kết luận

Bài báo này đã đề xuất mô hình nhận dạng chữ viết tay rời rạc trên cơ sở phương pháp máy véc tơ tựa. Các kết quả thực nghiệm cho thấy mô hình này có kết quả nhận dạng chính xác hơn so với mô hình mạng nơ ron. Tuy nhiên, khi áp dụng SVM vào bài toán nhận dạng cũng gặp phải một số hạn chế nhất định: bùng nổ số phân lớp và số lượng véc tơ tựa thu được sau khi huấn luyện sẽ dẫn đến việc phân lớp chậm.

Chúng tôi sẽ tiếp tục nghiên cứu để đề xuất mô hình hiệu quả cho bài toán nhận dạng chữ viết tay tiếng Việt. Giảm thiểu số véc tơ tựa để cải thiện tốc độ phân lớp và lựa chọn các tham số của SVM cũng là vấn đề cần quan tâm. Mỗi phương pháp học máy đều có những ưu và nhược điểm riêng, vì vậy việc kết hợp, lai ghép giữa các phương pháp nhằm nâng cao hiệu suất nhận dạng cũng là hướng mà các nhà nghiên cứu đang quan tâm.

TÀI LIỆU THAM KHẢO

[1]. Chih-Chung Chang and Chil-Jen Lin, "LIBSVM: a Library for Support Vector Machines", *National Taiwan University*, 2004.

[2]. Gorgevik D., Cakmakov D., "An Efficient Three-Stage Classifier for Handwritten Digit Recognition", *Proceedings of 17th Int. Conference on Pattern Recognition, ICPR2004*, Vol. 4, pp. 507-510, IEEE Computer Society, Cambridge, UK, 23-26 August 2004.

[3]. Phạm Anh Phương, Ngô Quốc Tạo, Lương Chi Mai, "Ứng dụng SVM cho bài toán phân lớp nhận dạng", *Kỷ yếu Hội thảo khoa học Quốc gia lần thứ ba về nghiên cứu, phát triển và ứng dụng Công nghệ thông tin và Truyền thông (ICT.rda'06),* nhà xuất bản Khoa học và Kỹ thuật, Hà nội, trang 393-400, 20-21/05/2006.

[4]. G. Vamvakas, B. Gatos, I. Pratikakis, N. Stamatopoulos, A. Roniotis and S.J. Perantonis, "Hybrid Off-Line OCR for Isolated Handwritten Greek Characters", *The Fourth IASTED International Conference on Signal Processing, Pattern Recognition, and Applications* (SPPRA 2007), pp. 197-202, ISBN: 978-0-88986-646-1, Innsbruck, Austria, February 2007.

[5]. Phạm Anh Phương, Ngô Quốc Tạo, Lương Chi Mai, "Trích chọn đặc trưng wavelet Haar kết hợp với SVM cho việc nhận dạng chữ viết tay tiếng Việt", *Tạp chí Công nghệ Thông tin và Truyền thông*, ISSN 0866-7039, kỳ 3, số 20, 10-2008, tr 36-42.

[6]. Phạm Anh Phương, "Áp dụng một số chiến lược SVM đa lớp cho bài toán nhận dạng chữ viết tay hạn chế", *Tạp chí khoa học Đại học Huế*, ISSN 1859-1388, số 45, 2008, tr. 109-118.

[7]. Pham Anh Phuong, Ngo Quoc Tao, Luong Chi Mai, "An Efficient Model for Isolated Vietnamese Handwritten Recognition", *The Fourth International Conference on Intelligent Information Hiding and Multimedia Signal Processing (IIH-MSP 2008)*, Harbin, China, August 15 - 17, 2008, pp. 358-361.

[8]. Nguyễn Thị Thanh Tân, Lương Chi Mai, "Phương pháp nhận dạng từ viết tay dựa trên mô hình mạng nơ ron kết hợp với thống kê từ vựng", *Tạp chí Tin học và Điều khiển học*, Tập 22, số 2, 2006, tr. 141-154.