

ĐẠI HỌC THÁI NGUYÊN
TRƯỜNG ĐẠI HỌC CÔNG NGHỆ THÔNG TIN VÀ TRUYỀN THÔNG

NGUYỄN QUANG HUY

NGHIÊN CỨU PHƯƠNG PHÁP NHẬN DẠNG CHỮ VIẾT
TAY HẠN CHẾ BẰNG MÔ HÌNH SVM
(SUPPORT VECTOR MACHINES)

LUẬN VĂN THẠC SĨ KHOA HỌC MÁY TÍNH

Thái Nguyên - 2014

ĐẠI HỌC THÁI NGUYÊN
TRƯỜNG ĐẠI HỌC CÔNG NGHỆ THÔNG TIN VÀ TRUYỀN THÔNG

NGUYỄN QUANG HUY

NGHIÊN CỨU PHƯƠNG PHÁP NHẬN DẠNG CHỮ VIẾT
TAY HẠN CHẾ BẰNG MÔ HÌNH SVM
(SUPPORT VECTOR MACHINES)

Chuyên ngành: KHOA HỌC MÁY TÍNH

Mã số: 60 48 01

LUẬN VĂN THẠC SĨ KHOA HỌC MÁY TÍNH

NGƯỜI HƯỚNG DẪN KHOA HỌC: PGS-TS. NGÔ QUỐC TẠO

Thái Nguyên - 2014

LỜI CẢM ƠN

Để đạt được những kết quả trong quá trình nghiên cứu luận văn, học viên xin chân thành cảm ơn thầy PGS. TS Ngô Quốc Tạo luôn tận tình chỉ bảo, hướng dẫn và giúp đỡ em trong suốt quá trình làm luận văn.

Học viên xin cảm ơn các thầy cô giáo trường Đại học Công nghệ thông tin và Truyền thông đã hướng dẫn và tạo điều kiện cho em trong suốt thời gian học tập tại trường.

Học viên xin chân thành cảm ơn các thầy giáo trong Hội đồng xét duyệt luận văn tốt nghiệp lớp cao học CK11A năm 2014 - Đợt 1 đã nhận xét và góp ý để bài luận văn của em được hoàn thiện hơn.

LỜI CAM ĐOAN

Tôi xin cam đoan luận văn “*Nghiên cứu phương pháp nhận dạng chữ viết tay hạn chế bằng mô hình SVM (Support Vector Machines)*” là do tôi tự nghiên cứu và hoàn thành dưới sự hướng dẫn của PGS-TS. Ngô Quốc Tạo.

Các kết quả đạt được trong quá trình nghiên cứu là hoàn toàn trung thực và khách quan.

Tôi xin chịu trách nhiệm về những lời cam đoan trên.

Thái Nguyên, ngày 05 tháng 05 năm 2014

Người cam đoan

Học viên Nguyễn Quang Huy

DANH MỤC CÁC KÝ HIỆU, CÁC CHỮ VIẾT TẮT

Thuật ngữ, chữ viết tắt	Giải thích
SVM	Support Vector Machine (Máy véc tơ hỗ trợ)
MMH	Maximum Marginal Hyperplane (Siêu phẳng có biên độ lớn nhất)
HMM	Markov Model (Mô hình Markov ẩn)
Kernel	Hàm nhân
MNIST	Bộ mẫu chữ số viết tay NIST - Viện Công nghệ và Tiêu chuẩn Quốc gia Hoa Kỳ (National Institute of Standard and Technology of the United States)
NN	Neuron Network (Mạng nơ ron)
OCR	Optical Character Recognition (nhận dạng chữ quang học)
QP	Quadratic Programing (quy hoạch toàn phương)
USPS	United States Postal service
VC	Vapnik – Chervonenkis

DANH MỤC CÁC HÌNH

Hình 1.1. Các giai đoạn trong quá trình xử lý và nhận dạng ảnh.....	7
Hình 1.2. Nhi phân hóa ảnh.....	8
Hình 1.3. Nhiều đốm và nhiễu vết.....	8
Hình 1.4. Chuẩn hóa kích thước ảnh các ký tự “A” và “P”	8
Hình 1.5. (a) Ảnh gốc, (b) Ảnh sau khi được làm trơn biên	9
Hình 1.6. Làm mảnh chữ	9
Hình 1.7. Hiệu chỉnh độ nghiêng của văn bản	10
Hình 1.8. Tách dòng chữ dựa trên histogram theo chiều ngang của khối chữ.....	10
Hình 1.9. Xác định khoảng cách giữa hai kí tự và giữa hai từ dựa trên histogram theo chiều thẳng đứng của dòng chữ.....	11
Hình 1.10. Mô hình mạng nơron nhân tạo	17
Hình 1.11. Mô hình mạng MLP 3 lớp	17
Hình 1.12. Phân lớp bằng mạng nơron.....	18
Hình 1.13. a) Các lớp phân tách tuyến tính b)Siêu phẳng tối ưu và biên lề tương ứng, các vectơ hỗ trợ.....	19
Hình 1.14. Ánh xạ các điểm dữ liệu không thể phân tách tuyến tính vào không gian số chiều lớn hơn có thể phân tách được tuyến tính.....	20
Hình 1.15. a) One-vs-One b) One-vs-All	21
Hình 2.1. Với 3 điểm không thẳng hàng trong R^2 thì luôn tách được bởi đường thẳng .	25
Hình 2.2. Phân lớp bằng siêu phẳng	26
Hình 2.3. Đường phân chia đối với tập dữ liệu gồm hai thuộc tính.....	27
Hình 2.4. Một bộ dữ liệu hai chiều được phân chia tuyến tính.	28
Hình 2.5. Hai siêu phẳng phân chia tuyến tính cùng với biên độ của nó.	29
Hình 2.6. Đường biểu diễn H1 và H2. Đường màu đỏ là khoảng cách Euclidean của hai điểm 1 và 2, đường màu xanh là khoảng cách Euclidean nhỏ nhất.	30
Hình 2.7. Các support vector trong SVM.....	31
Hình 2.8. Trường hợp trên không gian 2 chiều không thể vẽ một đường thẳng phân chia 2 lớp	35
Hình 2.9. Bước 1- Học để xây dựng mô hình phân lớp	37

Hình 2.10. Bước 2 - Kiểm tra và đánh giá	38
Hình 2.11. Mô hình nhận dạng chữ viết tay rời rạc.....	45
Hình 2.12. Trích chọn đặc trưng trọng số vùng	45
Hình 2.13. Kiến trúc của hệ nhận dạng chữ viết tay tiếng Việt	48
Hình 2.14. Chuẩn hóa ảnh: (a) Ảnh gốc, (b) Xác định các vùng liên thông và đánh thứ tự các vùng liên thông	49
Hình 2.15. Chuẩn hóa các vùng liên thông	49
Hình 2.16. Quá trình trích chọn đặc trưng.....	51
Hình 3.1. Các bước cơ bản của quá trình nhận dạng văn bản bằng mô hình SVM	55
Hình 3.2. Các mẫu chữ số viết tay trích từ tập các tập dữ liệu USPS và MNIST.....	59
Hình 3.3. Giao diện chương trình.....	61
Hình 3.4. Hộp thoại tiền xử lý.....	61
Hình 3.5. Hộp thoại trích chọn đặc trưng.....	62
Hình 3.6. Hộp thoại lưu file mô hình huấn luyện.....	62
Hình 3.7. Hộp thoại chọn file ảnh cần nhận dạng.....	63
Hình 3.8. Hộp thoại thông báo kết quả nhận dạng.....	63

MỞ ĐẦU

Biết sử dụng các phương pháp nhận dạng đóng vai trò hết sức quan trọng trong xử lý ảnh, phân tích tài liệu văn bản, đặc biệt là đối với các dạng văn bản viết tay. Hiện nay, nhu cầu cần nhận dạng nội dung văn bản từ các ảnh là rất lớn và thiết thực. Để nâng cao độ tin cậy của các phương pháp phân tích nhận dạng đã có những công trình nghiên cứu theo hướng ứng dụng lớp bài toán đánh giá lựa chọn thông tin để lựa chọn những tổ hợp thông tin chất lượng cao trước khi tiến hành phân tích nhận dạng. Cũng từ đó đề xuất những cách tiếp cận mới giải quyết bài toán nhận dạng trong xử lý số liệu văn bản và thu được kết quả tốt.

Nhận dạng chữ viết và đặc biệt nhận dạng chữ viết tay là bài toán có nhiều ứng dụng thực tế. Máy tính xử lý, nhận dạng các biểu mẫu, phiếu điều tra tự động, bằng cách này ta có thể tiết kiệm được nhiều chi phí về thời gian, công sức cũng như các chi phí khác cho việc nhập dữ liệu.

Ngày nay cùng với sự phát triển về mặt lý thuyết, công nghệ, có rất nhiều hướng đi cho việc giải quyết bài toán nhận dạng chữ viết dựa trên cấu trúc hay cách tiếp cận khác như dùng: logic mờ, giải thuật di truyền, mô hình xác suất thống kê, mô hình Markov ẩn HMM (Hidden Markov Models), mô hình mạng nơron NN (Neural Network Model), mô hình SVM (Support Vector Machine).

Thuật toán phân lớp là yếu tố có vai trò quyết định đến chất lượng của một hệ thống nhận dạng. Các phương pháp nhận dạng truyền thống như đối sánh mẫu, nhận dạng cấu trúc đã được ứng dụng khá phổ biến trong các hệ thống nhận dạng và cũng đã thu được những thành công nhất định. Tuy vậy, với những trường hợp văn bản đầu vào có chất lượng không tốt (nhiều, đứt nét, dính nét...) thì các thuật toán này tỏ ra không hiệu quả. Để khắc phục điều này, trong những năm gần đây nhiều nhóm nghiên cứu đã sử dụng các thuật toán phân lớp dựa trên mô hình SVM cho các bài toán nhận dạng nói chung và nhận dạng chữ viết

tay nói riêng. Trong luận văn này, học viên xin trình bày thuật toán SVM đối với việc nhận dạng chữ viết tay hạn chế.

MỤC LỤC

LỜI CẢM ƠN.....	i
LỜI CAM ĐOAN.....	ii
DANH MỤC CÁC KÝ HIỆU, CÁC CHỮ VIẾT TẮT.....	iii
DANH MỤC CÁC HÌNH.....	iv
MỞ ĐẦU.....	1
MỤC LỤC.....	2
Chương 1. GIỚI THIỆU VỀ CHỮ VIẾT VÀ NHẬN DẠNG CHỮ VIẾT.....	5
1.1. Trình bày về lịch sử của nhận dạng chữ viết tay.....	5
1.2. Giới thiệu các hướng tiếp cận trong việc nhận dạng chữ viết tay.....	5
1.2.1. Nhận dạng chữ in.....	5
1.2.2. Nhận dạng chữ viết tay.....	6
1.3. Tiền xử lý.....	7
1.3.1. Nhị phân hóa ảnh.....	7
1.3.2. Lọc nhiễu.....	8
1.3.3. Chuẩn hóa kích thước ảnh.....	8
1.3.4. Làm trơn biên chữ.....	9
1.3.5. Làm đầy chữ.....	9
1.3.6. Làm mảnh chữ.....	9
1.3.7. Điều chỉnh độ nghiêng của văn bản.....	9
1.4. Khôi tách chữ.....	10
1.4.1. Tách chữ theo chiều nằm ngang và thẳng đứng.....	10
1.4.2. Tách chữ dùng lược đồ sáng.....	11
1.5. Trích chọn đặc trưng.....	11
1.5.1. Biến đổi toàn cục và khai triển chuỗi.....	12
1.5.2. Đặc trưng thống kê.....	13
1.5.3. Đặc trưng hình học và hình thái.....	14

1.6. Huấn luyện và nhận dạng	15
1.7. Hậu xử lý	15
1.8. Một số thuật toán phân lớp nhận dạng chữ viết tay	16
1.8.1. Giới thiệu	16
1.8.2. Các mô hình nhận dạng chữ viết tay	16
1.8.3. Đánh giá, so sánh các phương pháp nhận dạng chữ.....	22
Chương 2. MÔ HÌNH SVM VÀ ỨNG DỤNG TRONG NHẬN DẠNG CHỮ	25
2.1. Giới thiệu chung	25
2.2. Lý thuyết chiều VC (Vapnik Chervonenkis dimension)	26
2.3. Hàm phân lớp	27
2.4. Siêu phẳng phân cách	28
2.5. Support vector	30
2.6. SVM với dữ liệu không nhiễu	32
2.7. SVM với dữ liệu có nhiễu	34
2.8. Biên độ (Margin)	34
2.9. Phân lớp dữ liệu tuyến tính và không tuyến tính.....	35
2.10. Sự cần thiết của SVM nhận dạng chữ viết tay hạn chế	37
2.10.1. Học máy có giám sát	37
2.10.2. Phân lớp dữ liệu.....	37
2.10.3. Nhận xét.....	40
2.10.4. Bài toán cho mô hình SVM.....	40
2.10.5. Xây dựng mô hình học cho SVM.....	43
2.11. Mô hình nhận dạng chữ viết tay rời rạc.....	46
2.11.1. Tiền xử lý.....	46
2.11.2. Trích chọn đặc trưng.....	47
2.11.3. Lựa chọn thuật toán huấn luyện phân lớp	47
2.11.4. Thuật toán nhận dạng chữ viết tay rời rạc	47
2.12. Áp dụng SVM vào nhận dạng chữ Việt viết tay rời rạc	49
2.12.1. Tiền xử lý.....	49