ĐẠI HỌC QUỐC GIA HÀ NỘI
TRƯỜNG ĐẠI HỌC CÔNG NGHỆ

Nguyễn Thị Hải Yến

PHÂN LỚP BÁN GIÁM SÁT VÀ ỨNG DỤNG THUẬT TOÁN SVM VÀO PHÂN LỚP TRANG WEB

KHOÁ LUẬN TỐT NGHIỆP ĐẠI HỌC HỆ CHÍNH QUY

Ngành: Công nghệ thông tin

HÀ NỘI - 2007

ĐẠI HỌC QUỐC GIA HÀ NỘI
TRƯỜNG ĐẠI HỌC CÔNG NGHỆ

Nguyễn Thị Hải Yến

PHÂN LỚP BÁN GIÁM SÁT VÀ ỨNG DỤNG THUẬT TOÁN SVM VÀO PHÂN LỚP TRANG WEB

KHOÁ LUẬN TỐT NGHIỆP ĐẠI HỌC HỆ CHÍNH QUY

Ngành: Công nghệ thông tin

Cán bộ hướng dẫn: PGS – TS Hà Quang Thụy

Cán bộ đồng hướng dẫn: ThS. Đặng Thanh Hải

HÀ NỘI – 2007

LỜI CẢM ƠN

Trước tiên, em xin bày tỏ lòng biết ơn chân thành và sâu sắc nhất tới Thầy giáo, PGS-TS Hà Quang Thụy và Thầy giáo, ThS. Đặng Thanh Hải đã tận tình hướng dẫn, động viên, giúp đỡ em trong suốt quá trình thực hiện đề tài.

Em xin gửi lời cảm ơn sâu sắc tới quý Thầy Cô trong Khoa Công nghệ thông tin đã truyền đạt kiến thức quý báu cho em trong những năm học vừa qua.

Em xin gửi lời cảm ơn các anh chị trong nhóm seminar về khai phá dữ liệu đã nhiệt tình chỉ bảo trong quá trình em làm khoá luận.

Con xin nói lên lòng biết ơn đối với Ông Bà, Cha Mẹ luôn là nguồn chăm sóc, động viên trên mỗi bước đường học vấn của con.

Xin chân thành cảm ơn các Anh Chị và Bạn bè, đặc biệt là các thành viên trong lớp K48CD đã ủng hộ, giúp đỡ và động viên tôi trong suốt thời gian học tập bốn năm trên giảng đường đại học và thực hiện đề tài.

Mặc dù đã cố gắng hoàn thành luận văn trong phạm vi và khả năng cho phép nhưng chắc chắn sẽ không tránh khỏi những thiếu sót. Em kính mong nhận được sự cảm thông và tận tình chỉ bảo của quý Thầy Cô và các Bạn.

Em xin chân thành cảm ơn!

Hà Nội, ngày 31 tháng 05 năm 2007

Sinh viên

Nguyễn Thị Hải Yến

TÓM TẮT NỘI DUNG

Hiện nay, với một lượng lớn các dữ liệu thì phân lớp dữ liệu có vai trò rất quan trọng, là một trong những bài toán luôn thời sự trong lĩnh vực xử lý dữ liệu văn bản. Một yêu cầu cơ bản được đặt ra là cần tăng tính hiệu quả của thuật toán phân lớp, nâng cao giá trị của các độ đo hồi tưởng, chính xác của thuật toán. Mặt khác, nguồn tài nguyên về ví dụ học có nhãn không phải luôn được đáp ứng vì vậy cần có các thuật toán phân lớp sử dụng các ví dụ chưa có nhãn. Phân lớp bán giám sát đáp ứng được hai yêu cầu nói trên [5, 7, 8, 16, 17]. Các thuật toán phân lớp bán giám sát tận dụng các nguồn dữ liệu chưa gán nhãn rất phong phú có trong tự nhiên kết hợp với một số dữ liệu đã được gán nhãn cho sẵn.

Trong những năm gần đây, phương pháp sử dụng bộ phân loại máy hỗ trợ vector (Support Vector Machine - SVM) được quan tâm và sử dụng nhiều trong lĩnh vực nhận dạng và phân loại. Từ các công trình khoa học [4, 7, 8, 11] được công bố 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.

Trong khoá luận này, em khảo sát thuật toán học bán giám sát SVM và trình bày các nội dung về phần mềm SVMlin do V. Sindhwani đề xuất [18]. Trong năm 2006-2007, V. Sindhwani đã dùng SVMlin tiến hành phân lớp văn bản từ nguồn 20-Newsgroups cho các kết quả tốt [14,15].

MỤC LỤC

MỞ ĐẦU 9

Chương 1 TỔNG QUAN VỀ PHÂN LỚP BÁN GIÁM SÁT 11

1.1. Phân lớp dữ liệu 11

1.1.1. Bài toán phân lớp dữ liệu 11

1.1.2. Quá trình phân lớp dữ liệu 12

1.2. Phân lớp văn bản 13

1.2.1. Đặt vấn đề 13

1.2.2. Mô hình vector biểu diễn văn bản 14

1.2.3. Phương pháp phân lớp văn bản 19

1.2.4. Ứng dụng của phân lớp văn bản 19

1.2.5. Các bước trong quá trình phân lớp văn bản 20

1.2.6. Đánh giá mô hình phân lớp 22

1.2.7. Các yếu tố quan trọng tác động đến phân lớp văn bản 23

1.3. Một số thuật toán học máy phân lớp 23

1.3.1. Học có giám sát 23

1.3.1.1. Bài toán học có giám sát 23

1.3.1.2. Giới thiệu học có giám sát 24

1.3.1.3. Thuật toán học có giám sát k-nearest neighbor (kNN) 25

1.3.1.4. Thuật toán học có giám sát Support vector machine (SVM) 26

1.3.2. Thuật toán phân lớp sử dụng quá trình học bán giám sát 27

1.3.2.1. Khái niệm 27

1.3.2.2. Lịch sử phát triển sơ lược của học bán giám sát 28

1.3.2.3. Một số phương pháp học bán giám sát điển hình 29
Chương 2 SỬ DỤNG SVM VÀ BÁN GIÁM SÁT SVM VÀO BÀI TOÁN PHÂN LỚP 32
2.1. SVM – Support Vector Machine.....32
2.1.1. Thuật toán SVM 33
2.1.2. Huấn luyện SVM.....35
2.1.3. Các ưu thế của SVM trong phân lớp văn bản 35
2.2. Bán giám sát SVM và phân lớp trang Web.....37
2.2.1. Giới thiệu về bán giám sát SVM.....37
2.2.2. Phân lớp trang Web sử dụng bán giám sát SVM 38
2.2.2.1. Giới thiệu bài toán phân lớp trang Web (Web Classification).....38
2.2.2.3. Áp dụng S3VM vào phân lớp trang Web.....39
Chương 3 THỬ NGHIỆM HỌC BÁN GIÁM SÁT PHÂN LỚP TRANG WEB..... 41
3.1. Giới thiệu phần mềm SVMlin.....41
3.2. Download SVMlin 42
3.3. Cài đặt.....42
3.4. Cách sử dụng phần mềm 42
KẾT LUẬN 45
Những công việc đã làm được của khoá luận 45
Hướng nghiên cứu trong thời gian tới.....45
TÀI LIỆU THAM KHẢO..... 46
I. Tiếng Việt.....46
II. Tiếng Anh 46

DANH SÁCH BẢNG VÀ TỪ VIẾT TẮT

Ký hiệu viết tắt	***Cụm từ***
kNN	k Nearest Neighbor
SVM	Support Vector Machine
S3VM	Semi Supervised Support Vector Machine

DANH MỤC HÌNH ẢNH

Hình 1. Bài toán phân lớp.

Hình 2. Văn bản được biểu diễn là vector đặc trưng.

Hình 3. Sơ đồ khung quá trình phân lớp văn bản.

Hình 4. Siêu phẳng h phân chia dữ liệu huấn luyện thành 2 lớp + và - với khoảng cách biên lớn nhất. Các điểm gần h nhất là các vector hỗ trợ (Support Vector - được khoanh tròn).

Hình 5. Phương pháp học bán giám sát Self-training.

Hình 6. Phương pháp học bán giám sát Co-training.

MỞ ĐẦU

Trong những năm gần đây, sự phát triển vượt bậc của công nghệ thông tin đã làm tăng số lượng giao dịch thông tin trên mạng Internet một cách đáng kể đặc biệt là thư viện điện tử, tin tức điện tử… Do đó mà số lượng văn bản xuất hiện trên mạng Internet cũng tăng với một tốc độ chóng mặt, và tốc độ thay đổi thông tin là cực kỳ nhanh chóng. Với số lượng thông tin đồ sộ như vậy, một yêu cầu lớn đặt ra là làm sao tổ chức và tìm kiếm thông tin, dữ liệu có hiệu quả nhất. Bài toán phân lớp là một trong những giải pháp hợp lý cho yêu cầu trên. Nhưng một thực tế là khối lượng thông tin quá lớn, việc phân lớp dữ liệu thủ công là điều không thể. Hướng giải quyết là một chương trình máy tính tự động phân lớp các thông tin dữ liệu trên.

Tuy nhiên, khi xử lý các bài toán phân lớp tự động thì gặp phải một số khó khăn là để xây dựng được bộ phân lớp có độ tin cậy cao đòi hỏi phải có một lượng lớn các mẫu dữ liệu huấn luyện tức là các văn bản đã được gán nhãn lớp tương ứng. Các dữ liệu huấn luyện này thường rất hiếm và đắt vì đòi hỏi thời gian và công sức của con người. Do vậy cần phải có một phương pháp học không cần nhiều dữ liệu gán nhãn và có khả năng tận dụng được các nguồn dữ liệu chưa gán nhãn rất phong phú như hiện nay, phương pháp học đó là học bán giám sát. Học bán giám sát chính là cách học sử dụng thông tin chứa trong cả dữ liệu chưa gán nhãn và tập huấn luyện, phương pháp học này được sử dụng rất phổ biến vì tính tiện lợi của nó.

Vì vậy, khoá luận tập trung vào nghiên cứu bài toán phân lớp sử dụng quá trình học bán giám sát, và việc áp dụng thuật toán bán giám sát máy hỗ trợ vector (Support Vector Machine – SVM) vào phân lớp trang Web.

Nội dung của khoá luận được trình bày bao gồm 3 chương. Tổ chức cấu trúc như sau:

- **Chương 1 Tổng quan về phân lớp bán giám sát.** Phần đầu trình bày khái quát về bài toán phân lớp dữ liệu, phân lớp văn bản, một số nét sơ bộ về học có giám sát. Phần cuối của chương giới thiệu các nội dung cơ bản về phương pháp học bán giám sát, trong đó đã giới thiệu một số thuật toán học bán giám sát điển hình.

- **Chương 2 Sử dụng SVM và bán giám sát SVM vào bài toán phân lớp.** Khóa luận trình bày những bước hoạt động cơ bản nhất của thuật toán SVM, sau đó nghiên cứu thuật toán học bán giám sát SVM, một cải tiến của SVM được trình bày trong [11]. Khoá luận trình bày một số áp dụng học bán giám sát vào bài toán phân lớp trang Web trong phần cuối cùng của chương.

- **Chương 3 Hệ thống thử nghiệm phân loại trang Web và đánh giá.** Trình bày kết quả nghiên cứu của V. Sindhwani về phần mềm nguồn mở SVMlin [14, 15, 18] mà do chính tác giả đề xuất và công bố. Các nghiên cứu này cho thấy phần mềm SVMlin phân lớp bán giám sát văn bản cho độ chính xác cao.

Chương 1 TỔNG QUAN VỀ PHÂN LỚP BÁN GIÁM SÁT

1.1. Phân lớp dữ liệu

1.1.1. Bài toán phân lớp dữ liệu

Là quá trình phân lớp một đối tượng dữ liệu vào một hay nhiều lớp cho trước nhờ một mô hình phân lớp mà mô hình này được xây dựng dựa trên một tập hợp các đối tượng dữ liệu đã được gán nhãn từ trước gọi là tập dữ liệu học (tập huấn luyện) [1-3]. Quá trình phân lớp còn được gọi là quá trình gán nhãn cho các đối tượng dữ liệu.

Như vậy, nhiệm vụ của bài toán phân lớp dữ liệu là cần xây dựng mô hình (bộ) phân lớp để khi có một dữ liệu mới vào thì mô hình phân lớp sẽ cho biết dữ liệu đó thuộc lớp nào.

Có nhiều bài toán phân lớp dữ liệu, như phân lớp nhị phân, phân lớp đa lớp, phân lớp đa trị,….

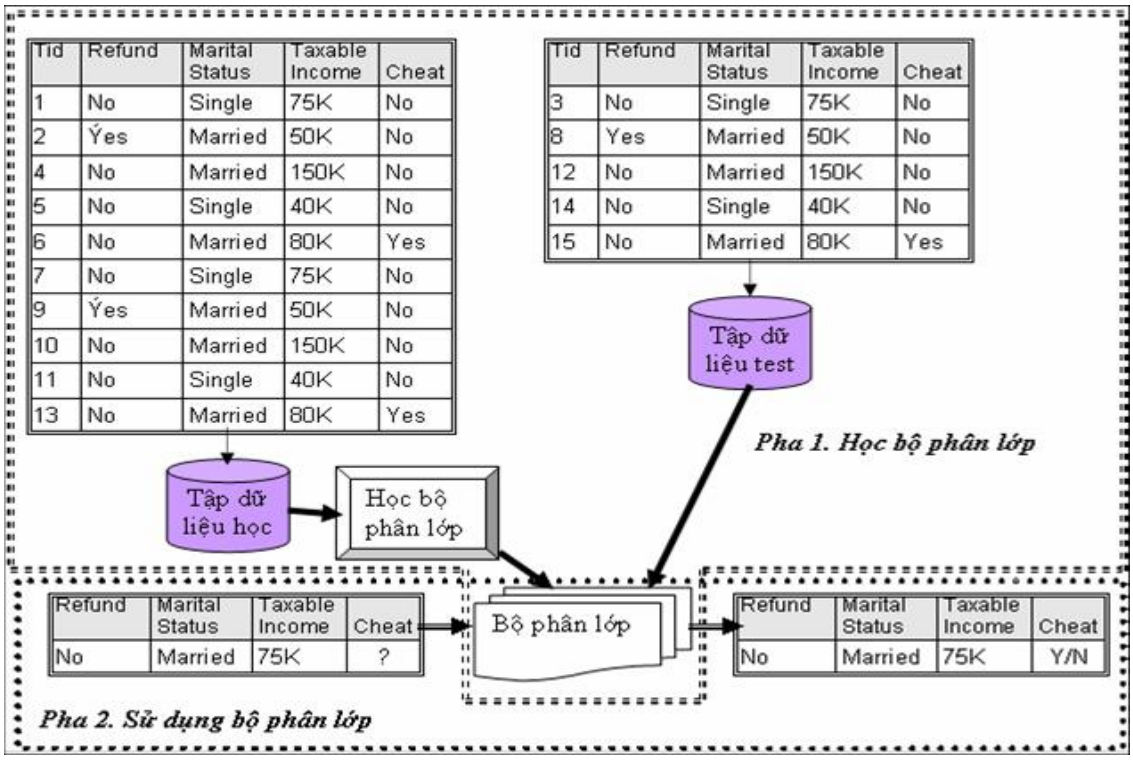
Phân lớp nhị phân là quá trình tiến hành việc phân lớp dữ liệu vào một trong hai lớp khác nhau dựa vào việc dữ liệu đó có hay không một số đặc tính theo quy định của bộ phân lớp.

Phân lớp đa lớp là quá trình phân lớp với số lượng lớp lớn hơn hai. Như vậy, tập hợp dữ liệu trong miền xem xét được phân chia thành nhiều lớp chứ không đơn thuần chỉ là hai lớp như trong bài toán phân lớp nhị phân. Về bản chất, bài toán phân lớp nhị phân là trường hợp riêng của bài toán phân lớp đa lớp.

Trong phân lớp đa trị, mỗi đối tượng dữ liệu trong tập huấn luyện cũng như các đối tượng mới sau khi được phân lớp có thể thuộc vào từ hai lớp trở lên. Ví dụ như trang web về việc bùng phát bệnh cúm gia cầm, thủy cầm tại một số tính phía Bắc vừa thuộc về lĩnh vực y tế liên quan đến lây bệnh sang người nhưng cũng thuộc về lĩnh vực kinh tế liên quan đến ngành chăn nuôi… Trong những trường hợp như vậy, việc sắp xếp một tài liệu vào nhiều hơn một lớp là phù hợp với yêu cầu thực tế.

Sau đây chúng ta sẽ tìm hiểu khái quát về quá trình phân lớp dữ liệu và sơ bộ về phương pháp phân lớp dữ liệu.

1.1.2. Quá trình phân lớp dữ liệu



Hình 1. Bài toán phân lớp

Quá trình phân lớp dữ liệu thường gồm hai bước: xây dựng mô hình (tạo bộ phân lớp) và sử dụng mô hình đó để phân lớp dữ liệu.

• Bước 1: một mô hình sẽ được xây dựng dựa trên việc phân tích các đối tượng dữ liệu đã được gán nhãn từ trước. Tập các mẫu dữ liệu này còn được gọi là ***tập dữ liệu huấn luyện*** (***training data set***). Các nhãn lớp của tập dữ liệu huấn luyện được xác định bởi con người trước khi xây dựng mô hình, vì vậy phương pháp này còn được gọi là *học có giám sát (supervised learning)*. Trong bước này, chúng ta còn phải tính độ chính xác của mô hình, mà cần phải sử dụng ***một tập dữ liệu kiểm tra*** (***test data set***). Nếu độ chính xác là chấp nhận được (tức là cao), mô hình sẽ được sử dụng để xác định nhãn lớp cho các dữ liệu khác mới trong tương lai. Trong việc test mô hình, sử dụng các độ đo để đánh

giá chất lượng của tập phân lớp, đó là độ hồi tưởng, độ chính xác, độ đo F_1 ... Nội dung chi tiết về các độ đo này được trình bày trong mục (1.2.6).

Tồn tại nhiều phương pháp phân lớp dữ liệu để giải quyết bài toán phân lớp tùy thuộc vào cách thức xây dựng mô hình phân lớp như phương pháp Bayes, phương pháp cây quyết định, phương pháp k-người láng giềng gần nhất, phương pháp máy hỗ trợ vector.... Các phương pháp phân lớp khác nhau chủ yếu về mô hình phân lớp. Mô hình phân lớp còn được gọi là thuật toán phân lớp.

- Bước 2: sử dụng mô hình đã được xây dựng ở bước 1 để phân lớp dữ liệu mới.

Như vậy, thuật toán phân lớp là một ánh xạ từ miền dữ liệu đã có sang một miền giá trị cụ thể của thuộc tính lớp, dựa vào giá trị các thuộc tính của dữ liệu.

1.2. Phân lớp văn bản

1.2.1. Đặt vấn đề

Ngày nay phương thức sử dụng giấy tờ trong giao dịch đã dần được số hoá chuyển sang các dạng văn bản lưu trữ trên máy tính hoặc truyền tải trên mạng. Bởi nhiều tính năng ưu việt của tài liệu số như cách lưu trữ gọn nhẹ, thời gian lưu trữ lâu dài, tiện dụng trong trao đổi đặc biệt là qua Internet, dễ dàng sửa đổi… nên càng ngày, số lượng văn bản số tăng lên một cách nhanh chóng đặc biệt là trên World Wide Web. Cùng với sự gia tăng về số lượng văn bản, nhu cầu tìm kiếm văn bản cũng tăng theo. Trong đời thường, phân lớp các văn bản được tiến hành một cách thủ công, nghĩa là chúng ta thực hiện công việc đọc từng văn bản một, xem xét và sau đó là gán nó vào một lớp cụ thể nào đó. Cách này sẽ tốn rất nhiều thời gian và công sức của con người vì các văn bản là vô vàn, để gán mỗi văn bản vào một lớp đã cho là một vấn đề không thể và do đó không khả thi. Với số lượng văn bản đồ sộ thì việc ***phân lớp văn bản tự động*** là một nhu cầu bức thiết.

Vậy phân lớp văn bản là gì? *Phân lớp văn bản* (Text Categorization) là việc phân lớp áp dụng đối với dữ liệu văn bản, tức là phân lớp một văn bản vào một hay nhiều lớp văn bản nhờ một mô hình phân lớp; mô hình này được xây dựng dựa trên một tập hợp các văn bản đã được gán nhãn từ trước.

Phân lớp văn bản là một lĩnh vực được chú ý nhất và đã được nghiên cứu trong những năm gần đây.

1.2.2. Mô hình vector biểu diễn văn bản

Như đã trình bày ở phần trên, bước đầu tiên trong qui trình phân lớp văn bản là thao tác chuyển văn bản đang được mô tả dưới dạng chuỗi các từ thành một mô hình khác, sao cho phù hợp với các thuật toán phân lớp.

Thông thường người ta thường biểu diễn văn bản bằng mô hình vector, mỗi văn bản được biểu diễn bằng một vector trọng số. Ý tưởng của mô hình này là xem mỗi một văn bản D_i được biểu diễn theo dạng $D_i = (\vec{d}_i, i)$, trong đó i là chỉ số dùng để nhận diện văn bản này và $\vec{d}_i$ là vector đặc trưng của văn bản D_i này, trong đó : $\vec{d}_i = (w_{i1}, w_{i2}, \ldots w_{in})$, và n là số lượng đặc trưng của vector văn bản, w_{ij} là trọng số của đặc trưng thứ j , $j \in \{1,2,...,n\}$.

Trong quá trình chuyển thể văn bản sang thành dạng vector, vấn đề mà chúng ta cần quan tâm là việc lựa chọn đặc trưng và số chiều cho không gian vector, chọn bao nhiêu từ, là các từ nào, phương pháp chọn ra sao?

Việc lựa chọn phương pháp biểu diễn văn bản để áp dụng vào bài toán phân lớp tuỳ thuộc vào độ thích hợp, phù hợp, độ đo đánh giá mô hình phân lớp của phương pháp đó sử dụng so với bài toán mà chúng ta đang xem xét giải quyết. Ví dụ nếu văn bản là một trang Web thì sẽ có phương pháp để lựa chọn đặc trưng khác so với các loại văn bản khác.

❖ Các đặc trưng của văn bản khi biểu diễn dưới dạng vector

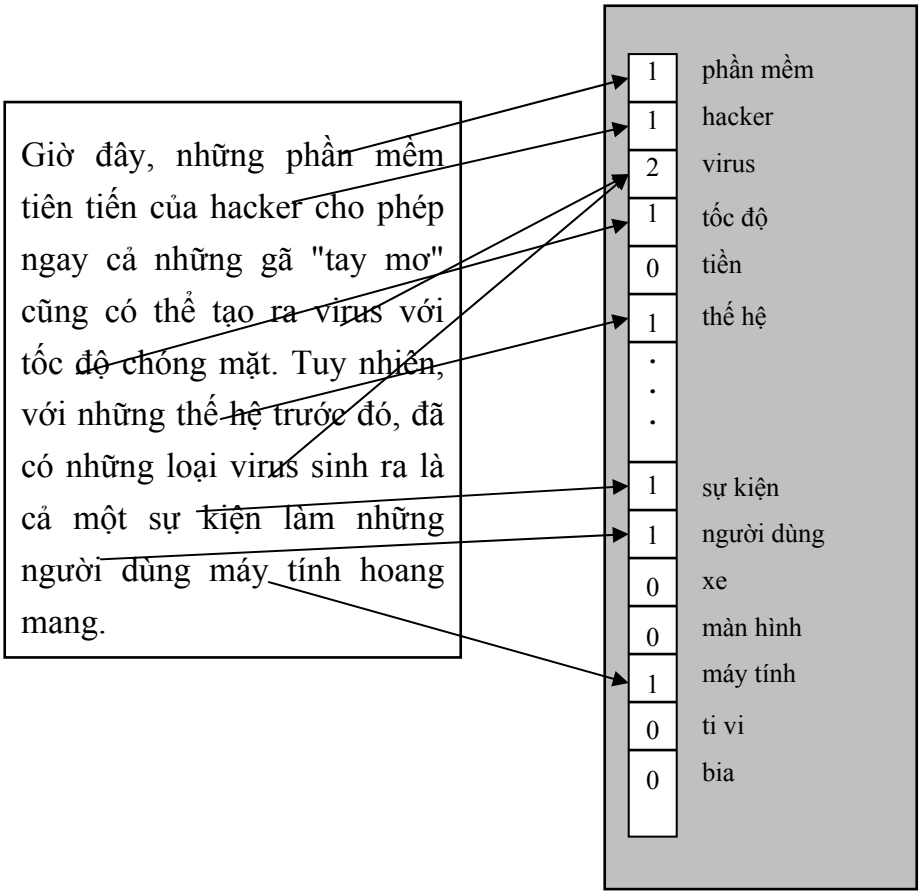
- Số nhiều không gian đặc trưng thường lớn. Các văn bản càng dài, lượng thông tin trong nó đề cập đến nhiều vấn đề thì không gian đặc trưng càng lớn.
- Các đặc trưng độc lập nhau, sự kết hợp các đặc trưng này thường không có ý nghĩa trong phân lớp.
- Các đặc trưng rời rạc: vector đặc trưng d_i có thể có nhiều thành phần mang giá trị 0 do có nhiều đặc trưng không xuất hiện trong văn bản d_i (nếu chúng ta tiếp cận theo cách sử dụng giá trị nhị phân 1, 0 để biểu diễn cho việc có xuất hiện hay không một đặc trưng nào đó trong văn bản đang được biểu diễn thành vector), tuy nhiên nếu đơn thuần cách tiếp cận sử dụng giá trị nhị phân 0, 1 này thì kết quả

phân lớp phần nào hạn chế là do có thể đặc trưng đó không có trong văn bản đang xét nhưng trong văn bản đang xét lại có từ khóa khác với từ đặc trưng nhưng có ngữ nghĩa giống với từ đặc trưng này, do đó một cách tiếp cận khác là không sử dụng số nhị phân 0, 1 mà sử dụng giá trị số thực để phần nào giảm bớt sự rời rạc trong vector văn bản.

- Hầu hết các văn bản có thể được phân chia một cách tuyến tính bằng các hàm tuyến tính.

Như vậy, độ dài của vector là số các từ khoá xuất hiện trong ít nhất một mẫu dữ liệu huấn luyện. Trước khi đánh trọng số cho các từ khoá cần tiến hành loại bỏ các ***từ dừng***. Từ dừng là những từ thường xuất hiện nhưng không có ích trong việc đánh chỉ mục, nó không có ý nghĩa gì trong việc phân lớp văn bản. Có thể nêu một số từ dừng trong tiếng Việt như "và", "là", "thì", "như vậy",…, trong tiếng Anh như "and", "or", "the",…**.** Thông thường từ dừng là các trạng từ, liên từ, giới từ.

Có thể lấy một ví dụ về việc biểu diễn văn bản dưới dạng vector trọng số như sau:



Hình 2. Văn bản được biểu diễn là vector đặc trưng

❖ Biểu diễn trang Web

Các trang Web về bản chất là siêu văn bản. Ngoài các văn bản và các thành phần đa phương tiện, các trang Web còn bao gồm những đặc trưng như là các siêu liên kết (Hyperlink), các thẻ HTML và các dữ liệu biến đổi (meta data). Hầu hết các nghiên cứu cho thấy rằng các thành phần văn bản của các trang Web cung cấp thông tin chính cho công việc phân lớp Web trong khi những thành phần không phải văn bản có thể được sử dụng để hoàn thiện hiệu suất phân lớp [6, 9].

Hiện nay tồn tại rất nhiều cách biểu diễn trang Web, với mỗi mục đích khác nhau thì sẽ có cách biểu diễn trang Web riêng. Trong các máy tìm kiếm như Yahoo, Altavista, Google... không sử dụng mô hình vector mà sử dụng hệ thống từ khoá móc nối song

không biểu diễn nội dung văn bản. Hiện nay cách tiếp cận biểu diễn Website là một cách tiếp cận nhận được nhiều sự quan tâm của nhiều người trên thế giới, đối tượng quan tâm không phải là Webpage mà là Website, nghĩa là đối tượng tìm kiếm không phải là các trang Web đơn nữa mà là cả một Website [2, 9].

Trong lĩnh vực văn bản truyền thống từ trước đến nay thì thông thường vẫn thực hiện các công việc như biểu diễn, tìm kiếm, phân lớp... trên cơ sở xem trang Web như là các trang văn bản thông thường và sử dụng mô hình không gian vector để biểu diễn văn bản. Việc sử dụng siêu liên kết giữa các trang Web có thể lấy được thông tin về mối liên hệ giữa nội dung các trang, và dựa vào đó để nâng cao hiệu quả phân lớp và tìm kiếm, đây chính là việc khai thác thế mạnh của siêu liên kết trong văn bản. Một số nhà nghiên cứu đã đưa ra cách cải tiến định hướng bằng cách liệt kê thêm các từ khoá xuất hiện từ các trang Web láng giềng bằng cách bổ sung thêm các từ khoá xuất hiện trong đoạn văn bản lân cận với siêu liên kết.

Trong khoá luận này, chúng ta sẽ nghiên cứu cách biểu diễn trang Web theo mô hình vector vì nó là một phương pháp rất phổ biến hiện nay. Với việc sử dụng các thông tin liên kết nhằm tăng độ chính xác tìm kiếm cũng như phân lớp các trang Web nên cần thiết phải đưa thêm các thông tin về các trang Web láng giềng vào vector biểu diễn của trang đang xét.

Tồn tại bốn cách biểu diễn trang Web theo mô hình vector như sau [2]:

- ***Cách thứ nhất***

Mỗi từ khóa trong một trang Web được lưu trữ cùng tần số xuất hiện nó ở trong trang Web. Cách này bỏ qua tất cả các thông tin về vị trí của từ khoá trong trang, thứ tự của các từ trong trang cũng như các thông tin về siêu liên kết.

Trong nhiều trường hợp khi mà các tài liệu đã liên kết độc lập với các nhãn của các lớp thì cách biểu diễn này là lựa chọn tốt nhất. Tuy nhiên trong một số trường hợp thì cách này không khai thác được tính cân đối trong tài liệu siêu liên kết.

- ***Cách thứ hai***

Sử dụng các thông tin về liên kết của trang Web, móc nối nó tới các trang láng giềng để tạo ra một siêu trang (super document). Vector biểu diễn bao gồm các từ xuất

hiện trong một trang cùng với tất cả các từ xuất hiện trong các trang láng giềng của nó cùng với tần số xuất hiện của các từ. Cách này bỏ qua thông tin về vị trí của các từ trong trang và thứ tự của chúng.

Nhược điểm của cách này là làm loãng đi nội dung của trang mà chúng ta đang quan tâm. Tuy nhiên đây là cách lựa chọn tốt trong trường hợp cần biểu diễn một tập các trang Web có nội dung về cùng một chủ đề, nhưng hiện nay số lượng các trang Web liên kết tới nhau có cùng một chủ đề tương đối ít, vì vậy cách biểu diễn này hiếm khi được sử dụng.

- ***Cách thứ ba***

Dùng một vector cấu trúc để biểu diễn trang Web. Một vector có cấu trúc được chia một cách logic thành hai phần hoặc nhiều hơn. Mỗi phần được sử dụng để biểu diễn một tập các trang láng giềng. Độ dài của một vector cố định nhưng mỗi phần của vector thì chỉ dùng để biểu diễn các từ xuất hiện trong một tập nào đó.

Cách này tránh được khả năng các trang láng giềng của một trang Web có thể làm loãng nội dung của nó. Nếu thông tin của các trang láng giềng này hữu ích cho quá trình phân lớp một trang nào đó thì máy học vẫn có thể truy cập đến toàn bộ nội dung của chúng để học.

- ***Cách thứ tư***

Xây dựng một vector có cấu trúc:

1. Xác định một số d được xem là bậc cao nhất của các trang trong tập
2. Xây dựng một vector cấu trúc với d + 1 phần như sau
    a. Phần đầu tiên biểu diễn chính tài liệu của một trang Web.
    b. Các phần tiếp theo đến d+1 biểu diễn các tài liệu láng giềng của nó, mỗi tài liệu được biểu diễn trong một phần.

Như vậy qua bốn cách biểu diễn vector trên thì ta thấy rằng hầu hết các phương pháp biểu diễn vector có kết hợp các thông tin về trang láng giềng cho kết quả phân lớp tốt hơn so với phương pháp biểu diễn vector với thông tin về tần số xuất hiện của các từ.

1.2.3. Phương pháp phân lớp văn bản

Như đã giới thiệu, tồn tại nhiều phương pháp phân lớp văn bản như phương pháp Bayes, phương pháp cây quyết định, phương pháp k-người láng giềng gần nhất, phương pháp máy hỗ trợ vector.... [1-3].

Để xây dựng công cụ phân lớp văn bản tự động người ta thường dùng các thuật toán ***học máy*** (machine learning). Tuy nhiên còn có các thuật toán đặc biệt hơn dùng cho phân lớp trong các lĩnh vực đặc thù của văn bản một cách tương đối máy móc, như là khi hệ thống thấy trong văn bản có một cụm từ cụ thể thì hệ thống sẽ phân văn bản đó vào một lớp nào đó. Tuy nhiên khi phải làm việc với các văn bản ít đặc trưng hơn thì cần phải xây dựng các thuật toán phân lớp dựa trên nội dung của văn bản và so sánh độ phù hợp của chúng với các văn bản đã được phân lớp bởi con người. Đây là tư tưởng chính của thuật toán học máy. Trong mô hình này, các văn bản đã được phân lớp sẵn và hệ thống của chúng ta phải tìm cách để tách ra đặc trưng của các văn bản thuộc mỗi nhóm riêng biệt. Tập văn bản mẫu dùng để huấn luyện gọi là ***tập huấn luyện (train set),*** hay ***tập mẫu (pattern set)***, còn quá trình máy tự tìm đặc trưng của các nhóm gọi là quá trình ***học (learning)***. Sau khi máy đã ***học*** xong, người dùng sẽ đưa các văn bản mới vào và nhiệm vụ của máy là tìm ra xem văn bản đó phù hợp nhất với nhóm nào mà con người đã huấn luyện nó.

1.2.4. Ứng dụng của phân lớp văn bản

Một trong những ứng dụng quan trọng nhất của phân lớp văn bản là trong tìm kiếm văn bản. Từ một tập dữ liệu đã phân lớp các văn bản sẽ được đánh số đối với từng lớp tương ứng. Người dùng có thể xác định chủ để phân lớp văn bản mà mình mong muốn tìm kiếm thông qua các câu hỏi [2, 3].

Một ứng dụng khác của phân lớp văn bản là có thể được sử dụng để lọc các văn bản hoặc một phần các văn bản chứa dữ liệu cần tìm mà không làm mất đi tính phức tạp của ngôn ngữ tự nhiên.

Ngoài ra phân lớp văn bản có rất nhiều ứng dụng trong thực tế, điển hình là các ứng dụng trích lọc thông tin trên Internet. Hiện nay, có rất nhiều trang Web thương mại quảng cáo hoặc các trang web phản động, có văn hoá không lành mạnh, vì mục đích làm tăng lượng người truy cập, chúng trà trộn vào kết quả trả về của máy tìm kiếm, chúng vào

hòm thư của chúng ta theo chu kỳ và gây nhiều phiền toái, các ứng dụng cụ thể là lọc thư rác (spam mail), lọc trang web phản động, các trang web không lành mạnh…

Như vậy phân lớp văn bản là công cụ không thể thiếu trong thời đại Công nghệ thông tin phát triển lớn mạnh như hiện nay, vì thế phân lớp văn bản là vấn đề đáng được quan tâm để xây dựng và phát triển được những công cụ hữu ích làm cho hệ thống công nghệ thông tin hiện nay ngày càng phát triển và lớn mạnh.

1.2.5. Các bước trong quá trình phân lớp văn bản

Quá trình phân lớp văn bản trải qua 4 bước [1] cơ bản sau:

Đánh chỉ số (indexing): Các văn bản ở dạng thô cần được chuyển sang một dạng biểu diễn nào đó để xử lý, quá trình này được gọi là quá trình biểu diễn văn bản, dạng biểu diễn phải có cấu trúc và dễ dàng trong khi xử lý, ở đây văn bản được biểu diễn dưới dạng phổ biến nhất là vector trọng số. Tốc độ đánh chỉ số có vai trò quan trọng trong quá trình phân lớp văn bản.

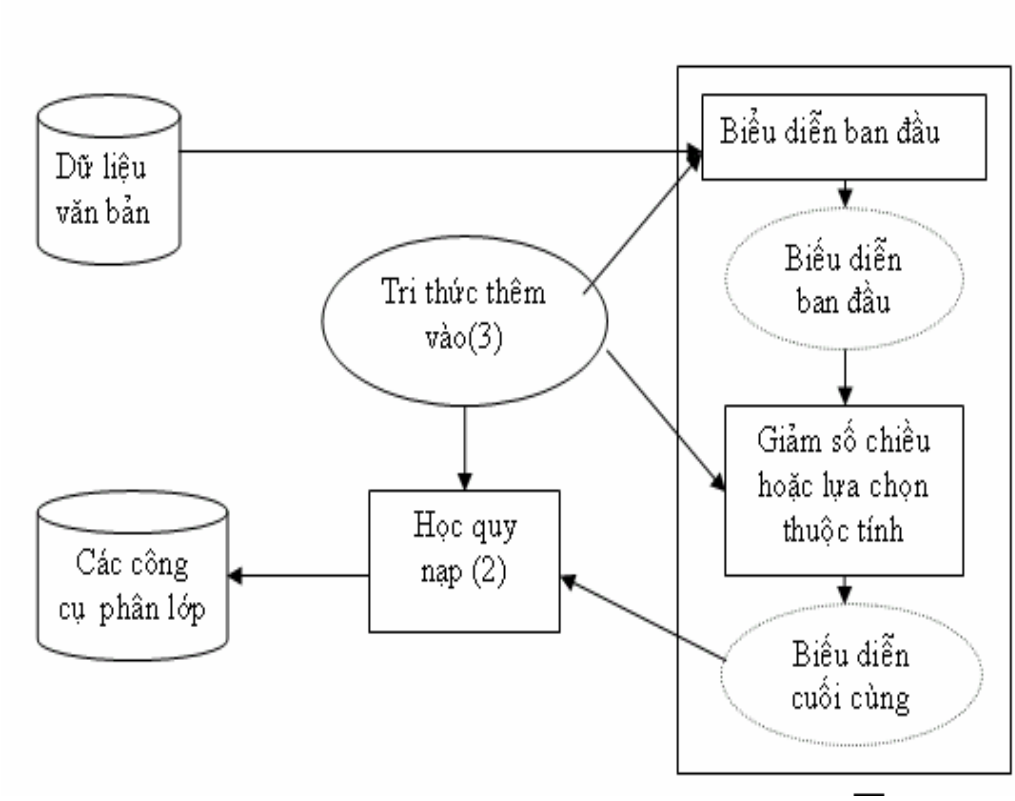
Xác định độ phân lớp: Cần nêu lên cách thức xác định lớp cho mỗi văn bản như thế nào, dựa trên cấu trúc biểu diễn của văn bản đó. Nhưng trong khi những câu hỏi mang tính nhất thời thì tập phân lớp được sử dụng một cách ổn định và lâu dài cho quá trình phân lớp.

So sánh: Trong hầu hết các tập phân lớp, mỗi văn bản đều được yêu cầu gán đúng sai vào một lớp nào đó.

Phản hồi (thích nghi): Quá trình phản hồi đóng hai vai trò trong hệ phân lớp văn bản. Thứ nhất là, khi phân lớp thì phải có một số lượng lớn các văn bản đã được xếp loại bằng tay trước đó, các văn bản này được sử dụng làm mẫu huấn luyện để hỗ trợ xây dựng tập phân lớp. Thứ hai là, đối với việc phân lớp văn bản này, không dễ dàng thay đổi các yêu cầu bởi vì người dùng có thể thông tin cho người bảo trì hệ thống về việc xoá bỏ, thêm vào hoặc thay đổi các lớp văn bản nào đó mà mình yêu cầu.

Hình sau là một sơ đồ khung cho việc phân lớp văn bản, trong đó bao gồm ba công đoạn chính:

- Công đoạn đầu: Biểu diễn văn bản, tức là chuyển các dữ liệu văn bản thành một dạng có cấu trúc nào đó, tập hợp các mẫu cho trước thành một tập huấn luyện.
- Công đoạn thứ hai: Việc sử dụng các kỹ thuật học máy để học trên các mẫu huấn luyện vừa biểu diễn. Như vậy là việc biểu diễn ở công đoạn một sẽ là đầu vào cho công đoạn thứ hai.
- Công đoạn thứ ba: Việc bổ sung các kiến thức thêm vào do người dùng cung cấp để làm tăng độ chính xác trong biểu diễn văn bản hay trong quá trình học máy.



Hình 3. Sơ đồ khung quá trình phân lớp văn bản

1.2.6. Đánh giá mô hình phân lớp

Chúng ta không thể khẳng định một phương pháp phân lớp văn bản cụ thể nào là chính xác hoàn toàn. Bất kỳ phương pháp nào cũng có độ sai lệch không nhiều thì ít. Vì vậy việc đưa ra độ đo để đánh giá hiệu quả của thuật toán phân lớp giúp chúng ta có thể xác định được mô hình nào là tốt nhất, kém nhất, từ đó áp dụng thuật toán đó vào việc phân lớp. Sau đây chúng ta sẽ đưa ra công thức chung để đánh giá độ chính xác của các thuật toán.

Độ hồi tưởng (Recall) và độ chính xác (Precision), độ và độ đo F_1 được dùng để đánh giá chất lượng của thuật toán phân lớp.

- $$recall = \frac{true\_positive}{(true\_positive) + (false\_positive)} \times 100\,\% \qquad (1.1)$$
- $$precision = \frac{true\_positive}{(true\_positive) + (true\_negative)} \times 100\,\% \qquad (1.2)$$
- $$F_1(recall, precision) = \frac{2 \times recall \times precision}{recall \times precision} \qquad (1.3)$$

Để dễ hiểu hơn, chúng ta có công thức:

$$\textit{Độ hồi tưởng} = \frac{\textit{Số văn bản được phân vào lớp dương và đúng}}{\textit{Tổng số văn bản phân vào lớp dương}}$$

$$\textit{Độ chính xác} = \frac{\textit{Số văn bản phân vào lớp dương và đúng}}{\textit{Tổng số văn bản được phân lớp và đúng}}$$

$$\textit{Tiêu chuẩn đánh giá} = \frac{\textit{2 * độ hồi tưởng * độ chính xác}}{\textit{Độ hồi tưởng + độ chính xác}}$$

1.2.7. Các yếu tố quan trọng tác động đến phân lớp văn bản

Ngày nay phân lớp văn bản có vai trò rất quan trọng trong sự phát triển của Công nghệ thông tin, tuy nhiên độ phức tạp của từng loại văn bản khác nhau, vì thế khả năng mà từng tập phân lớp có thể thực thi được là khác nhau dẫn đến kết quả phân lớp khác nhau. Chúng ta có thể liệt kê 3 yếu tố quan trọng tác động đến kết quả phân lớp như sau:

- Cần một tập dữ liệu huấn luyện chuẩn và đủ lớn để cho thuật toán học phân lớp. Nếu chúng ta có được một tập dữ liệu chuẩn và đủ lớn thì quá trình huấn luyện sẽ tốt và khi đó chúng ta sẽ có kết quả phân lớp tốt sau khi đã được học.
- Các phương pháp trên hầu hết đều sử dụng mô hình vector để biểu diễn văn bản, do đó phương pháp tách từ trong văn bản đóng vai trò quan trọng trong quá trình biểu diễn văn bản bằng vector. Yếu tố này rất quan trọng, vì có thể đối với một số ngôn ngữ như tiếng Anh chẳng hạn thì thao tác tách từ trong văn bản đơn giản chỉ là dựa vào các khoảng trắng, tuy nhiên trong các ngôn ngữ đa âm tiết như tiếng Việt và một số ngôn ngữ khác thì sử dụng khoảng trắng khi tách từ là không chính xác, do đó phương pháp tách từ là một yếu tố quan trọng.
- Thuật toán sử dụng để phân lớp phải có thời gian xử lý hợp lý, thời gian này bao gồm: thời gian học, thời gian phân lớp văn bản, ngoài ra thuật toán này phải có tính tăng cường (incremental function) nghĩa là không phân lớp lại toàn tập tập văn bản khi thêm một số văn bản mới vào tập dữ liệu mà chỉ phân lớp các văn bản mới mà thôi, khi đó thuật toán phải có khả năng giảm độ nhiễu (noise) khi phân lớp văn bản.

1.3. Một số thuật toán học máy phân lớp

1.3.1. Học có giám sát

1.3.1.1. Bài toán học có giám sát

Mục đích là để học một ánh xạ từ x tới y. Khi cho trước một tập huấn luyện gồm các cặp (x_i, y_i), trong đó $y_i \in \mathrm{Y}$ gọi là các nhãn của các mẫu x_i. Nếu nhãn là các số,

$y = (y_i)_{i\in[n]}^T$ biểu diễn vector cột của các nhãn. Hơn nữa, một thủ tục chuẩn là các cặp (x_i, y_i) được thử theo giả thiết i.i.d (independent and identically distributed random variables) trên khắp X × Y [15].

1.3.1.2. Giới thiệu học có giám sát

Học có giám sát là một kỹ thuật của ngành học máy để xây dựng một hàm từ dữ liệu huấn luyện. Dữ liệu huấn luyện bao gồm các cặp đối tượng đầu vào (thường dạng vector) và đầu ra thực sự. Đầu ra của một hàm có thể là một giá trị liên tục (gọi là hồi quy), hay có thể là dự đoán một nhãn phân lớp cho một đối tượng đầu vào (gọi là phân lớp). Nhiệm vụ của chương trình học có giám sát là dự đoán giá trị của hàm cho một đối tượng bất kỳ là đầu vào hợp lệ, sau khi đã xem xét một số ví dụ huấn luyện (nghĩa là, các cặp đầu vào và đầu ra tương ứng). Để đạt được điều này, chương trình học phải tổng quát hoá từ các dữ liệu sẵn có để dự đoán những tình huống chưa gặp phải theo một cách hợp lý.

Để giải quyết một bài toán nào đó của học có giám sát, người ta phải xem xét nhiều bước khác nhau:

- Xác định loại của các ví dụ huấn luyện. Trước khi làm bất cứ điều gì, người làm nhiệm vụ phân lớp nên quyết định loại dữ liệu nào sẽ được sử dụng làm ví dụ. Chẳng hạn đó có thể là một kí tự viết tay đơn lẻ, toàn tập một từ viết tay, hay toàn tập một dòng chữ viết tay.
- Thu thập tập huấn luyện. Tập huấn luyện cần đặc trưng cho thực tế sử dụng của hàm chức năng. Vì thế, một tập các đối tượng đầu vào được thu thập và đầu ra tương ứng được thu thập, hoặc từ các chuyên gia hoặc từ việc đo dạc tính toán.
- Xác định việc biểu diễn các đặc trưng đầu vào cho hàm chức năng cần tìm. Sự chính xác của hàm chức năng phụ thuộc lớn vào cách các đối tượng đầu vào được biểu diễn. Thông thường, đối tượng đầu vào được chuyển đổi thành một vector đặc trưng, chứa một số các đặc trưng nhằm mô tả cho đối tượng đó. Số lượng các đặc trưng không nên quá lớn, do sự bùng nổ tổ hợp (curse of dimensionality), nhưng phải đủ lớn để dự đoán chính xác đầu ra.

- Xác đinh cấu trúc của hàm chức năng cần tìm và giải thuật học tương ứng. Ví dụ người thực hiện quá trình phân lớp có thể lựa chọn việc sử dụng mạng nơ-ron nhân tạo hay cây quyết định….
- Hoàn thiện thiết kế. Người thiết kế sẽ chạy giải thuật học từ một tập huấn luyện thu thập được. Các tham số của giải thuật học có thể được điều chỉnh bằng cách tối ưu hoá hiệu năng trên một tập con (gọi là tập kiểm chứng – validation set) của tập huấn luyện, hay thông qua kiểm chứng chéo (cross-validation). Sau khi học và điều chỉnh tham số, hiệu năng của giải thuật có thể được đo dạc trên một tập kiểm tra độc lập với tập huấn luyện.

1.3.1.3. Thuật toán học có giám sát k-nearest neighbor (kNN)

Có rất nhiều thuật toán học có giám sát, ở đây em sẽ giới thiệu một thuật toán học có giám sát điển hình, đó là *k-nearest neighbor* (kNN hay k-láng giềng gần nhất)

kNN là phương pháp truyền thống khá nổi tiếng theo hướng tiếp cận thống kê đã được nghiên cứu trong nhiều năm qua. kNN được đánh giá là một trong những phương pháp tốt nhất được sử dụng từ những thời kỳ đầu trong nghiên cứu về phân loại văn bản

Ý tưởng của phương pháp này đó là khi cần phân loại một văn bản mới, thuật toán sẽ xác định khoảng cách (có thể áp dụng các công thức về khoảng cách như Euclide, Cosine, Manhattan, …) của tất cả các văn bản trong tập huấn luyện đến văn bản này để tìm ra k văn bản gần nhất, gọi là k nearest neighbor – k láng giềng gần nhất, sau đó dùng các khoảng cách này đánh trọng số cho tất cả các chủ đề. Khi đó, trọng số của một chủ đề chính là tổng tất cả các khoảng cách ở trên của các văn bản trong k láng giềng có cùng chủ đề, chủ đề nào không xuất hiện trong k láng giềng sẽ có trọng số bằng 0. Sau đó các chủ đề sẽ được sắp xếp theo giá trị trọng số giảm dần và các chủ đề có trọng số cao sẽ được chọn làm chủ đề của văn bản cần phân loại.

Trọng số của chủ đề c_j đối với văn bản x được tính như sau :

$$W\left(\vec{x}, c_j\right) = \sum_{\vec{d_i} \in \{kNN\}} sim\left(\vec{x}, \vec{d_i}\right).y\left(\vec{d_i}, c_j\right) - b_j \qquad (1.4)$$

Trong đó :

y (d_i, c) thuộc {0,1}, với:

y = 0: văn bản d_i không thuộc về chủ đề c_j

y = 1: văn bản d_i thuộc về chủ đề c_j

sim (x, d): độ giống nhau giữa văn bản cần phân loại x và văn bản d. Chúng ta có thể sử dụng độ đo cosine để tính khoảng cách:

$$sim\left(\vec{x}, \vec{d_i}\right) = \cos\left(\vec{x}, \vec{d_i}\right) = \frac{\vec{x} \cdot \vec{d_i}}{\left\|\vec{x}\right\| \left\|\vec{d_i}\right\|} \qquad (1.5)$$

b_j là ngưỡng phân loại của chủ đề c_j được tự động học sử dụng một tập văn bản hợp lệ được chọn ra từ tập huấn luyện.

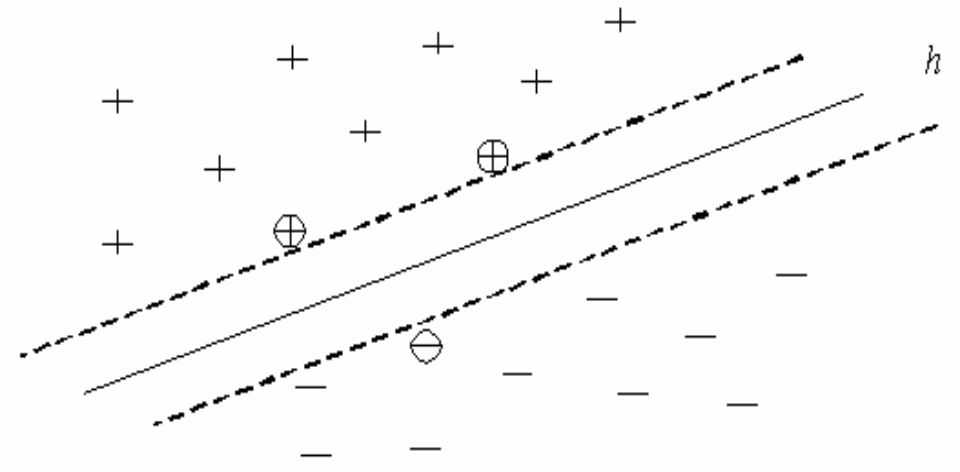
Để chọn được tham số k tốt nhất cho thao tác phân loại, thuật toán cần được chạy thử nghiệm trên nhiều giá trị k khác nhau, giá trị k càng lớn thì thuật toán càng ổn định và sai sót càng thấp.

1.3.1.4. Thuật toán học có giám sát Support vector machine (SVM)

Theo [4, 7], SVM là phương pháp phân lớp rất hiệu quả được Vapnik giới thiệu vào năm 1995 để giải quyết nhận dạng mẫu hai lớp sử dụng nguyên lý *Cực tiểu hoá Rủi ro Cấu trúc* (Structural Risk Minimization).

Ý tưởng chính của thuật toán này là cho trước một tập huấn luyện được biểu diễn trong không gian vector trong đó mỗi tài liệu là một điểm, phương pháp này tìm ra một mặt phẳng *h* quyết định tốt nhất có thể chia các điểm trên không gian này thành hai lớp riêng biệt tương ứng lớp + và lớp -. Chất lượng của siêu mặt phẳng này được quyết định bởi khoảng cách (gọi là biên) của điểm dữ liệu gần nhất của mỗi lớp đến mặt phẳng này. Khoảng cách biên càng lớn thì mặt phẳng quyết định càng tốt đồng thời việc phân loại càng chính xác. Mục đích thuật toán SVM tìm ra được khoảng cách biên lớn nhất để tạo kết quả phân lớp tốt.

Hình sau minh hoạ cho thuật toán này:



Hình 4. Siêu phẳng h phân chia dữ liệu huấn luyện thành 2 lớp + và - với khoảng cách biên lớn nhất. Các điểm gần h nhất là các vector hỗ trợ (Support Vector - được khoanh tròn)

Trong chương 2 sẽ trình bày chi tiết về thuật toán học SVM và bán giám sát SVM.

1.3.2. Thuật toán phân lớp sử dụng quá trình học bán giám sát

1.3.2.1. Khái niệm

Theo Xiaojin Zhu [16], khái niệm học bán giám sát được đưa ra năm 1970 khi bài toán đánh giá quy tắc Linear Discrimination Fisher cùng với dữ liệu chưa gán nhãn được nhiều sự quan tâm của các nhà khoa học trên thế giới.

Trong khoa học máy tính, *học bán giám sát* là một phương thức của ngành học máy sử dụng cả dữ liệu gán nhãn và chưa gán nhãn, nhiều nghiên cứu của ngành học máy có thể tìm ra được dữ liệu chưa gán nhãn khi sử dụng với một số lượng nhỏ dữ liệu gán nhãn [15]. Công việc thu được kết quả của dữ liệu gán nhãn thường đòi hỏi ở trình độ tư duy và khả năng của con người, công việc này tốn nhiều thời gian và chi phí, do vậy dữ liệu gán nhãn thường rất hiếm và đắt, trong khi dữ liệu chưa gán nhãn thì lại rất phong phú. Trong trường hợp đó, chúng ta có thể sử dụng học bán giám sát để thi hành các công việc ở quy mô lớn.

Học bán giám sát bao gồm dữ liệu gán nhãn và chưa gán nhãn. Học bán giám sát có thể được áp dụng vào việc phân lớp và phân cụm. Mục tiêu của học bán giám sát là huấn luyện tập phân lớp tốt hơn học có giám sát từ dữ liệu gán nhãn và chưa gán nhãn.

Như vậy, *có thể nói học bán giám sát là phương pháp học có giám sát kết hợp với việc tận dụng các dữ liệu chưa gán nhãn.* Trong phần bổ sung thêm vào cho dữ liệu gán nhãn, thuật toán cung cấp một vài thông tin giám sát, việc này không cần thiết cho tất cả các mẫu huấn luyện. Thông thường thông tin này sẽ được kết hợp với một vài mẫu cho trước.

Học bán giám sát là một nhánh của ngành ***học máy (machine learning)***. Các dữ liệu gán nhãn thường hiếm, đắt và rất mất thời gian, đòi hỏi sự nỗ lực của con người, trong khi đó dữ liệu chưa gán nhãn thì vô vàn nhưng để sử dụng vào mục đích cụ thể của chúng ta thì rất khó, vì vậy ý tưởng kết hợp giữa dữ liệu chưa gán nhãn và dữ liệu đã gán nhãn để xây dựng một tập phân lớp tốt hơn là nội dung chính của học bán giám sát. Bởi vậy học bán giám sát là một ý tưởng tốt để giảm bớt công việc của con người và cải thiện độ chính xác lên mức cao hơn.

1.3.2.2. Lịch sử phát triển sơ lược của học bán giám sát

Theo [16, 17], quá trình học bán giám sát đã được nghiên cứu phát triển trong một thập kỷ gần đây, nhất là từ khi xuất hiện các trang Web với số lượng thông tin ngày càng lớn, chủ đề ngày càng phong phú. Chúng ta có thể nêu lên quá trình phát triển của học bán giám sát trải qua các thuật toán được nghiên cứu như sau.

Cùng với số liệu lớn của dữ liệu chưa gán nhãn, các thành phần hỗn hợp có thể được nhận ra cùng với thuật toán Cực đại kỳ vọng EM (expectation-maximization). Chỉ cần một mẫu đơn đã gán nhãn cho mỗi thành phần để xác định hoàn toàn được mô hình hỗn hợp. Mô hình này được áp dụng thành công vào việc phân lớp văn bản. Một biến thể khác của mô hình này chính là self-training. Cả 2 phương pháp này được sử dụng cách đây một thời gian khá dài. Chúng được sử dụng phổ biến vì dựa trên khái niệm đơn giản của chúng và sự dễ hiểu của thuật toán.

Co-training là thuật toán học bán giám sát điển hình tiếp theo mà các nhà khoa học đầu tư nghiên cứu. Trong khi self-training là thuật toán mà khi có một sự phân lớp lỗi thì

có thể tăng cường thêm cho chính nó, thì co-training giảm bớt được lỗi tăng cường có thể xảy ra khi có một quá trình phân lớp bị lỗi.

Cùng với quá trình phát triển và việc áp dụng phổ biến và sự tăng lên về chất lượng của thuật toán SVM (Máy hỗ trợ vector - Support Vector Machine), SVM truyền dẫn (Transductive Support Vector Machine – TSVM) nổi bật lên như một SVM chuẩn mở rộng cho phương pháp học bán giám sát.

Gần đây các phương pháp học bán giám sát dựa trên đồ thị (graph-based) thu hút nhiều sự quan tâm của các nhà khoa học cũng như những người quan tâm đến lĩnh vực khai phá dữ liệu. Các phương pháp Graph-based bắt đầu với một đồ thị mà các nút là các điểm dữ liệu gán nhãn và chưa gán nhãn, và các điểm nối phản ánh được sự giống nhau giữa các nút này.

Có thể thấy học bán giám sát là một quá trình hoàn thiện dần các thuật toán để áp dụng vào các vấn đề của đời sống con người. Sau đây chúng ta sẽ giới thiệu sơ qua một số thuật toán học bán giám sát điển hình có thể xem là được áp dụng nhiều nhất.

1.3.2.3. Một số phương pháp học bán giám sát điển hình

Có rất nhiều phương pháp học bán giám sát. Có thể nêu tên các phương pháp thường được sử dụng như: *Naïve Bayes, EM với các mô hình hỗn hợp sinh, self-training, co-training, transductive support vector machine (TSVM),* và các phương pháp *graph-based.* Chúng ta không có câu trả lời chính xác cho câu hỏi phương pháp nào là tốt nhất ở đây. Có thể thấy phương pháp học bán giám sát sử dụng dữ liệu chưa gán nhãn để thay đổi hoặc giảm bớt các kết quả từ những giả thuyết đã thu được của dữ liệu đã gán nhãn.

Sau đây, chúng tôi xin trình bày sơ bộ nội dung của một số thuật toán học bán giám sát điển hình.

Self-training

Self-training là một phương pháp được sử dụng phổ biến trong học bán giám sát. Trong self-training một tập phân lớp ban đầu được huấn luyện cùng với số lượng nhỏ dữ liệu gán nhãn. Tập phân lớp sau đó sẽ được dùng để gán nhãn cho dữ liệu chưa gán nhãn. Điển hình là hầu hết các điểm chưa gán nhãn có tin cậy cao, cũng như cùng với các nhãn dự đoán trước của chúng, được chèn thêm vào tập huấn luyện. Sau đó tập phân lớp sẽ

được huấn luyện lại và lặp lại các quy trình. Chú ý rằng tập phân lớp sử dụng các dự đoán của nó để dạy chính nó. Quy trình này được gọi là self-teaching hay là bootstrapping.

Self-training được áp dụng để xử lý các bài toán của một số ngôn ngữ tự nhiên. Ngoài ra self-training còn được áp dụng để phân tách và dịch máy. Theo Xiaojin Zhu [16, 17], nhiều tác giả đã áp dụng self-training để phát hiện các đối tượng hệ thống từ các hình ảnh.

Thuật toán: Self-training

1. Lựa chọn một phương pháp phân lớp. Huấn luyện một bộ phân lớp f từ (X_l, Y_l).
2. Sử dụng f để phân lớp tất cả các đối tượng chưa gán nhãn $x \in X_u$.
3. Lựa chọn x^* với độ tin cậy cao nhất, chèn thêm $(x^*, f(x^*))$ tới dữ liệu đã gán nhãn.
4. Lặp lại các quá trình trên.

Hình 5. Phương pháp học bán giám sát Self-training

Co-training

Theo [16,17], Co-training dựa trên giả thiết rằng các đặc trưng (features) có thể được phân chia thành hai tập. Mỗi một tập đặc trưng con có khả năng huấn luyện một tập phân lớp tốt. Hai tập con này độc lập điều kiện (conditionally independent) đã cho của lớp (class).

Đầu tiên hai tập phân lớp phân tách thành dữ liệu huấn luyện và dữ liệu gán nhãn trên hai tập đặc trưng con được tách biệt ra. Sau đó mỗi tập phân lớp lại phân lớp các dữ liệu chưa gán nhãn và "dạy" tập phân lớp khác cùng với một vài mẫu chưa gán nhãn (và các nhãn dự đoán) mà chúng cảm giác có độ tin cậy cao. Cuối cùng, mỗi tập phân lớp sẽ

được huấn luyện lại cùng với các mẫu huấn luyện chèn thêm được cho bởi tập phân lớp khác và bắt đầu tiến trình lặp.

Thuật toán: Co-training

1. Huấn luyện hai bộ phân lớp: $f^{(1)}$ từ $(X_l^{(1)}, Y_l)$, $f^{(2)}$ từ $(X_l^{(2)}, Y_l)$.
2. Phân lớp X_u với $f^{(1)}$ và $f^{(2)}$ tách biệt nhau.
3. Chèn thêm vào $f^{(1)}$ k-most-confident $(x, f^{(1)}(x))$ tới các dữ liệu đã gán nhãn của $f^{(2)}$.
4. Chèn thêm vào $f^{(2)}$ k-most-confident $(x, f^{(2)}(x))$ tới các dữ liệu đã gán nhãn của $f^{(1)}$.
5. Lặp lại các quá trình trên.

Hình 6. Phương pháp học bán giám sát Co-training

Chương 2 SỬ DỤNG SVM VÀ BÁN GIÁM SÁT SVM VÀO BÀI TOÁN PHÂN LỚP

Trong lĩnh vực khai phá dữ liệu, các phương pháp phân lớp văn bản đã dựa trên những phương pháp quyết định như quyết định Bayes, cây quyết định, k-người láng giềng gần nhất, …. Những phương pháp này đã cho kết quả chấp nhận được và được sử dụng nhiều trong thực tế. Trong những năm gần đây, phương pháp phân lớp sử dụng tập phân lớp vector hỗ trợ (máy vector hỗ trợ - Support Vector Machine – SVM) được quan tâm và sử dụng nhiều trong lĩnh vực nhận dạng và phân lớp. SVM là một họ các phương pháp dựa trên cơ sở các hàm nhân (kernel) để tối thiểu hoá rủi ro ước lượng. Phương pháp SVM ra đời từ lý thuyết học thống kê do Vapnik và Chervonenkis xây dựng 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 lớp khá tốt đối với bài toán phân lớp văn bản cũng như trong nhiều ứng dụng khác (như nhận dạng chữ viết tay, phát hiên mặt người trong các ảnh, ước lượng hồi quy,…). Xét với các phương pháp phân lớp khác, khả năng phân lớp của SVM là tương đối tốt và hiệu quả.

2.1. SVM – Support Vector Machine

SVM sử dụng thuật toán học nhằm xây dựng một siêu phẳng làm cực tiểu hoá độ phân lớp sai của một đối tượng dữ liệu mới. Độ phân lớp sai của một siêu phẳng được đặc trưng bởi khoảng cách bé nhất tới siêu phẳng đấy. SVM có khả năng rất lớn cho các ứng dụng được thành công trong bài toán phân lớp văn bản.

Như đã biết, phân lớp văn bản là một cách tiếp cận mới để tạo ra tập phân lớp văn bản từ các mẫu cho trước. Cách tiếp cận này phối hợp với sự thực thi ở mức độ cao và hiệu suất cùng với những am hiểu về mặt lý thuyết, tính chất thô ngày càng được hoàn thiện. Thông thường, hiệu quả ở mức độ cao không có các thành phần suy nghiệm. Phương pháp SVM có khả năng tính toán sẵn sàng và phân lớp, nó trở thành lý thuyết học mà có thể chỉ dẫn những ứng dụng thực tế trên toàn cầu.

Đặc trưng cơ bản quyết định khả năng phân lớp là khả năng phân lớp những dữ liệu mới dựa vào những tri thức đã tích luỹ được trong quá trình huấn luyện. Sau quá trình huấn luyện nếu hiệu suất tổng quát hoá của bộ phân lớp cao thì thuật toán huấn

luyện được đánh giá là tốt. Hiệu suất tổng quát hoá phụ thuộc vào hai tham số là *sai số huấn luyện* hay và *năng lực* của máy học. Trong đó sai số huấn luyện là tỷ lệ lỗi phân lớp 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 tập phân lớp). 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 tập phân lớp tốt là tập phân lớp 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 trên ý tưởng này.

2.1.1. Thuật toán SVM

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

$$\{(x_i, y_i)\ i = 1, 2,\ldots, N, x_i \in R^m\}$$

Trong đó mẫu là các vector đối tượng được phân lớp thành các mẫu dương và mẫu âm như trong hình 4:

- 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 $y_i = -1$.

Thực chất phương pháp này là một bài toán tối ưu, mục tiêu là tìm ra một không gian ***H*** và siêu mặt phẳng quyết định h trên H sao cho sai số phân lớp là thấp nhất.

Trong trường hợp này, tập phân lớp 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:

$$C + w_1 x_1 + w_2 x_2 + \ldots + w_n x_n = 0 \qquad (2.1)$$

Tương đương với công thức

$$C + \sum_{i=1,\ldots,n} w_i x_i = 0 \qquad (2.2)$$

Với $w = w_1 + w_2 + \ldots + w_n$ là bộ hệ số siêu phẳng hay là vector trọng số, C là độ dịch, khi thay đổi w và C thì hướng và khoảng cách từ gốc toạ độ đến mặt siêu phẳng thay đổi.

Tập phân lớp SVM được định nghĩa như sau:

$$f(\mathbf{x}) = \text{sign}(C + \sum w_i x_i) \qquad (2.3)$$

Trong đó

sign(z) = +1 nếu $z \geq 0$,
sign(z) = -1 nếu $z < 0$.

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

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

Ta phải giải phương trình sau:

$$\begin{aligned} \min_{\mathbf{w},b,\eta} \quad & C\sum_{i=1}^{\ell}\eta_i + \frac{1}{2}\|\mathbf{w}\|^2 \\ s.t. \quad & y_i[\mathbf{w}\cdot x_i - b] + \eta_i \geq 1 \\ & \eta_i \geq 0, \quad i = 1,\ldots,\ell \end{aligned} \qquad (2.4)$$

tìm ra được vector trọng số w và sai số của mỗi điểm trong tập huấn luyện là η_i từ đó ta có phương trình tổng quát của siêu phẳng tìm ra được bởi thuật toán SVM là:

$$f(x_1, x_2, \ldots, x_n) = C + \sum w_i x_i \qquad (2.5)$$

Với i = 1,…, n. Trong đó n là số dữ liệu huấn luyện.

Sau khi đã tìm được phương trình của siêu phẳng bằng thuật toán SVM, áp dụng công thức này để tìm ra nhãn lớp cho các dữ liệu mới.

2.1.2. 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 với kích thước nhỏ hơn. Sau đó, những thuật toán này kiểm tra các điều kiện KKT (Karush-Kuhn-Tucker) để xác định phương án tối ưu.

Một số thuật toán huấn luyện dựa vào tính chất: 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 náy, 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.

2.1.3. Các ưu thế của SVM trong phân lớp văn bản

Như đã biết, phân lớp 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 lớp, các phương pháp huấn luyện được sử dụng để xây dựng tập phân lớp từ các tài liệu mẫu, sau đó dùng tập phân lớp này để dự đoán lớp của những tài liệu mới (chưa biết chủ đề).

Chúng ta có thể thấy từ các thuật toán phân lớp hai lớp như SVM đến các thuật toán phân lớp đa lớp đều có đặc điểm chung là yêu cầu văn bản phải được biểu diễn dưới dạng vector đặc trưng, tuy nhiên các thuật toán khác đều phải sử dụng các uớc lượng tham số và ngưỡng tối ưu trong khi đó thuật toán SVM có thể tự tìm ra các tham số tối ưu này. Trong các phương pháp thì SVM là phương pháp sử dụng không gian vector đặc

trưng lớn nhất (hơn 10.000 chiều) trong khi đó các phương pháp khác có số chiều bé hơn nhiều (như Naïve Bayes là 2000, k-Nearest Neighbors là 2415…).

Trong công trình của mình năm 1999 [12], Joachims đã so sánh SVM với Naïve Bayesian, k-Nearest Neighbour, Rocchio, và C4.5 và đến năm 2003 [13], Joachims đã chứng minh rằng SVM làm việc rất tốt cùng với các đặc tính được đề cập trước đây của văn bản. Các kết quả cho thấy rằng SVM đưa ra độ chính xác phân lớp tốt nhất khi so sánh với các phương pháp khác.

Theo Xiaojin Zhu [15] thì trong các công trình nghiên cứu của nhiều tác giả (chẳng hạn như Kiritchenko và Matwin vào năm 2001, Hwanjo Yu và Han vào năm 2003, Lewis vào năm 2004) đã chỉ ra rằng thuật toán SVM đem lại kết quả tốt nhất phân lớp văn bản.

Kiritchenko và Matwin đã nghiên cứu và so sánh phương pháp SVM với kỹ thuật Naïve Bayesian, sau đó đã chứng minh được rằng SVM là phương pháp tốt nhất cho phân lớp thư điện tử cũng như phân lớp văn bản.

Hwanjo Yu và Han cho thấy rằng SVM hoàn toàn được tiến hành tốt nhất so với các phương pháp phân lớp văn bản khác. Tất cả các tài liệu nghiên cứu hiện nay cho thấy rằng SVM đưa ra kết quả chính xác nhất trong khía cạnh phân lớp văn bản.

Lewis đã nghiên cứu phân lớp văn bản và đã khám phá ra rằng kết quả của SVM là tốt nhất. Lewis đã đưa ra tập hợp nhỏ các tài liệu của phân lớp văn bản. Tác giả đã cố gắng cải tiến phương pháp RCV1 cho phân lớp văn bản và sử dụng phương pháp mới được ứng dụng cho một số kỹ thuật phân lớp văn bản khác nhau. SVM đã đưa ra kết quả tốt nhất khi đặt dựa vào k-người láng giềng gần nhất và kỹ thuật tập phân lớp Rocchio-Style Prototype.

Những phân tích của các tác giả trên đây cho thấy SVM có nhiều điểm phù hợp cho việc ứng dụng phân lớp văn bản. Và trên thực tế, các thí nghiệm phân lớp văn bản tiếng Anh chỉ ra rằng SVM đạt độ chính xác phân lớp cao và tỏ ra xuất sắc hơn so với các phương pháp phân lớp văn bản khác.

Vấn đề căn bản của học bán giám sát là chúng ta có thể tận dụng dữ liệu chưa gán nhãn để cải tiến hiệu quả của độ chính xác trong khi phân lớp, điều này được đưa ra để so sánh với một tập phân lớp được thiết kề mà không tính đến dữ liệu chưa gán nhãn.

Trong phần sau của chương này, khóa luận sẽ giới thiệu một phương thức cải tiến của SVM là bán giám sát SVM (semi-supervised support vector machine – S^3VM) [16, 17]. Bán giám sát SVM được đưa ra nhằm nâng SVM lên một mức cao hơn, trong khi SVM là một thuật toán học có giám sát, sử dụng dữ liệu đã gán nhãn thì bán giám sát SVM sử dụng cả dữ liệu gán nhãn (tập huấn luyện – training set) kết hợp với dữ liệu chưa gán nhãn (working set).

2.2. Bán giám sát SVM và phân lớp trang Web

2.2.1. Giới thiệu về bán giám sát SVM

Chúng ta sẽ giới thiệu phương thức cải tiến của SVM là Bán giám sát SVM (Semi Supervised Support Vector Machine - S3VM). Cho một tập huấn luyện (training set) của dữ liệu gán nhãn và có sự tham gia của một tập các dữ liệu chưa gán nhãn (working set), S3VM xây dựng một máy hỗ trợ vector sử dụng cả training set và working set. Bài toán truyền dẫn sẽ dự đoán giá trị của một hàm phân lớp tới các điểm đã cho trong working set.

Trong khi SVM là một thuật toán có giám sát sử dụng dữ liệu đã gán nhãn, thì S3VM được xây dựng sử dụng hỗn hợp dữ liệu gán nhãn (training set) và dữ liệu chưa gán nhãn (working set). Mục đích là để gán các lớp nhãn tới working set một cách tốt nhất, sau đó sử dụng hỗn hợp dữ liệu huấn luyện đã gán nhãn và dữ liệu working set sau khi đã gán nhãn để phân lớp những dữ liệu mới. Nếu working set rỗng thì phương pháp này trở thành phương pháp chuẩn SVM để phân lớp. Nếu training set rỗng, sau đó phương pháp này sẽ trở thành hình thể học không giám sát. Học bán giám sát xảy ra khi cả training set và working set không rỗng.

Để hiểu một cách rõ ràng cụ thể về S3VM, thì chúng ta cần hiểu về SVM đã được trình bày ở trên. Với thời gian và điều kiện không cho phép, trong khoá luận này em chỉ có thể tìm hiểu về thuật toán S3VM là bài toán phân lớp nhị phân.

Cho trước một tập huấn luyện gồm những dữ liệu đã gán nhãn cùng với tập dữ liệu chưa gán nhãn working set bao gồm n dữ liệu. Mục đích là gán nhãn cho những dữ liệu chưa gán nhãn này.

Với hai lớp đã cho trước gồm lớp dương (lớp +1) và lớp âm (lớp –1). Mỗi dữ liệu được xem như một điểm trong không gian vector. Mỗi điểm i thuộc tập dữ liệu huấn luyện có một sai số là η_i và mỗi điểm j thuộc working set sẽ có hai sai số ξ_j (sai số phân lớp với giả sử rằng j thuộc lớp +1) và z_i (sai số phân lớp với giả sử rằng j thuộc lớp –1). Thuật toán S3VM sẽ giải bài toán tối ưu sau (2.6) thay cho bài toán tối ưu 2.4 ở thuật toán SVM.

$$\begin{aligned} \min_{\mathbf{w},b,\eta,\xi,z} \quad & C\left[\sum_{i=1}^{\ell}\eta_i + \sum_{j=\ell+1}^{\ell+k} min(\xi_j, z_j)\right] + \| \mathbf{w} \| \\ subject\ to \quad & y_i(\mathbf{w}\cdot x_i + b) + \eta_i \geq 1 \quad \eta_i \geq 0 \quad i = 1,\ldots,\ell \\ & \mathbf{w}\cdot x_j - b + \xi_j \geq 1 \quad \xi_j \geq 0 \quad j = \ell+1,\ldots,\ell+k \\ & -(\mathbf{w}\cdot x_j - b) + z_j \geq 1 \quad z_j \geq 0 \end{aligned} \tag{2.6}$$

Sau khi đã tìm được ξ_i và z_j, chúng ta sẽ có được sai số nhỏ nhất của mỗi điểm j, Nếu $\xi_i < z_j$ thì điểm j thuộc lớp dương, ngược lại nếu $\xi_i > z_j$ thì điểm j thuộc lớp âm. Quá trình này diễn ra trên tất cả các điểm thuộc working set, sau khi quá trình này đã hoàn thành, tất cả các điểm chưa gán nhãn sẽ được gán nhãn.

Tập dữ liệu chưa gán nhãn working set sau khi đã gán nhãn sẽ được đưa vào tập dữ liệu huấn luyện, tiếp theo đó sẽ sử dung thuật toán SVM để học tạo ra SVM mới, SVM này chính là S3VM có một siêu phẳng mới. Sau đó áp dụng siêu phẳng này để phân lớp các mẫu dữ liệu mới được đưa vào.

2.2.2. Phân lớp trang Web sử dụng bán giám sát SVM

2.2.2.1. Giới thiệu bài toán phân lớp trang Web (Web Classification)

Phân lớp trang Web là một trường hợp đặc biệt của phân lớp văn bản bởi sự hiện diện của các siêu liên kết trong trang Web, cấu trúc trang Web chặt chẽ, đầy đủ hơn, dẫn đến các tính năng hỗn hợp như là plain texts, các thẻ hypertext, hyperlinks….

Internet với hơn 10 tỷ trang Web là một tập huấn luyện rất phong phú về mọi chủ đề trong cuộc sống, hơn nữa với số lượng chủ đề trên các Website là không nhiều thì việc

sử dụng Internet như cơ sở huấn luyện rất phù hợp. Trong các trang Web, tuy độ chính xác không phải là tuyệt đối, nhưng ta có thể thấy mỗi chủ đề gồm có nhiều từ chuyên môn với tần suất xuất hiện rất cao, việc tận dụng tần số phụ thuộc của các từ này vào chủ đề có thể đem lại kết quả khả quan cho phân lớp.

2.2.2.3. Áp dụng S3VM vào phân lớp trang Web

Có thể thấy trang Web là siêu văn bản (hypertext) rất phổ dụng hiện nay. Nội dung của các trang Web thường được mô tả ngắn gọn, súc tích, có các siêu liên kết chỉ đến các Web có nội dung liên quan và cho phép các trang khác liên kết đến nó.

Như đã nói trên, vì được xem như là các văn bản thông thường nên trong quá trình phân lớp trang Web việc biểu diễn văn bản sử dụng mô hình không gian vector. Việc biểu diễn và xử lý tài liệu Web cũng giống như biểu diễn và xử lý văn bản bằng mô hình này. Tuy nhiên trong phân lớp Web thì việc khai thác thế mạnh của siêu liên kết trong văn bản là một vấn đề đáng quan tâm. Với việc sử dụng các siêu liên kết giữa các trang Web từ đó có thể lấy được các thông tin về mối liên hệ giữa nội dung các trang, và dựa vào đó để nâng cao hiệu quả phân lớp và tìm kiếm.

Để áp dụng vào phân lớp trang Web, thuật toán S3VM xem mỗi trang Web là một vector $f(d_1, d_2, \ldots, d_n)$ được biểu diễn giống như văn bản. Áp dụng công thức (2.5) trong phương trình của siêu phẳng:

$$f(x_1, x_2, \ldots, x_n) = C + \sum w_i x_i$$

thay thế mỗi văn bản tương ứng với mỗi trang Web vào phương trình siêu phẳng này:

$$f(d_1, d_2, \ldots, d_n) = C + \sum w_i d_i \qquad (2.6)$$

Với i=1,…,n.

Nếu $f(d) \geq 0$ thì trang Web thuộc lớp +1.

Ngược lại nếu $f(d) < 0$ trang Web thuộc lớp –1.

Có thể thấy rằng quá trình áp dụng thuật toán S3VM vào bài toán phân lớp trang Web chính là việc thay thế vector trọng số biểu diễn trang Web đó vào phương trình siêu phẳng của S3VM, từ đó tìm ra được nhãn lớp của các trang Web chưa gán nhãn.

Như vậy, thực chất của quá trình phân lớp bán giám sát áp dụng đối với dữ liệu là các trang Web là tập dữ liệu huấn luyện là các trang Web còn tập working set (dữ liệu chưa gán nhãn) là những trang Web được các trang Web đã có nhãn trong tập huấn luyện trỏ tới.

Chương 3 THỬ NGHIỆM HỌC BÁN GIÁM SÁT PHÂN LỚP TRANG WEB

Khóa luận định hướng khai thác phần mềm nguồn mở để tiến hành thử nghiệm phân lớp bán giám sát các tài liệu web. Phần đầu của chương giới thiệu phần mềm nguồn mở SVMlin có tiêu đề là "*"Fast Linear SVM Solvers for Supervised and Semi-supervised Learning"* do Vikas Sindhwani công bố. Các phần tiếp theo khóa luận giới thiệu quá trình khai thác phần mềm nhằm thực hiện bài toán phân lớp và đánh giá. Nội dung của chương này tổng hợp từ các nội dung được trình bày trong [14,15,18].

Phần mềm SVMlin thuộc diện phần mềm nguồn mở, được công bố theo các tiêu chuẩn của giấy phép sử dụng phần mềm GNU.

3.1. Giới thiệu phần mềm SVMlin

SVMlin là gói phần mềm dành cho SVMs tuyến tính, nó thoả mãn bài toán phân lớp một số lớn các mẫu dữ liệu và các đặc trưng. Là chương trình phần mềm được viết trên ngôn ngữ C++ (hầu hết được viết trên C).

Ngoài tập dữ liệu đã được gán nhãn, SVMlin còn có thể tận dụng tập dữ liệu chưa được gán nhãn trong quá trình học. Tập dữ liệu chưa được gán nhãn này thực sử hữu ích trong việc nâng cao độ chính xác của quá trình phân lớp khi mà số lượng dữ liệu được gán nhãn từ trước là rất ít.

Hiện tại SVMlin đã thực hiện cài đặt các thuật toán [14, 15]sau:

- Thuật toán học có giám sát (chỉ sử dụng các dữ liệu đã gán nhãn)
  - Thuật toán phân lớp bình phương tối thiểu đã được chuẩn hóa tuyến tính (Linear Regularized Least Squares Classification).
- Bán giám sát (có thể sử dụng các dữ liệu chưa gán nhãn tương đối tốt)
  - Thuật toán học tuyến tính SVM truyền dẫn sử dụng nhiều lần chuyển đổi (Multi-switch linear Transductive L2-SVMs)

Theo Vikas Sindhwani, khi dùng SVMlin phân loại văn bản (tập dữ liệu RCV1-v2/LYRL2004) với 804414 dữ liệu gán nhãn và 47326 đặc trưng, SVMlin mất ít hơn hai phút để huấn luyện SVM tuyến tính trong một máy Intel với tốc độ xử lý 3GHz và 2GB

RAM. Nếu chỉ cho 1000 nhãn, nó có thể sử dụng hàng trăm ngàn dữ liệu chưa gán nhãn để huấn luyện một SVM tuyến tính bán giám sát trong vòng khoảng 20 phút. Dữ liệu chưa gán nhãn rất hữu ích trong việc cải thiện quá trình phân lớp khi số lượng nhãn lớp không quá lớn.

3.2. Download SVMlin

Người dùng có thể tải phiên bản mới nhất của SVMlin tại trang Web:

http://www.cs.uchicago.edu/people/vikass

3.3. Cài đặt

Trước tiên, cần giải nén file cài đặt bằng các lệnh sau:

unzip svmlin.zip

tar –xvzf svmlin.tar.gz

Sau đó nó sẽ tạo ra một thư mục có tên là *svmlin-v1.0* chứa *Makefile* và 3 file nguồn là *ssl.h, ssl.cpp* và *svmlin.cpp.*

Gõ lệnh:

make

Sẽ tạo ra file thực thi

svmlin

Quá trình thực thi này được sử dụng để huấn luyện, kiểm tra và đánh giá quá trình thực hiện.

3.4. Sử dụng phần mềm và kết quả đánh giá

- **Các file dữ liệu**

Định dạng dữ liệu đầu vào cho SVMlin tương tự như định dạng của bộ công cụ SVM-Light/LIBSVM (Điểm khác biệt duy nhất là không có cột đầu tiên mô tả nhãn của các dữ liệu)

Mỗi một dòng mô tả một mẫu dữ liệu và là danh sách các cặp gồm **chỉ số đặc trưng : giá trị đặc trưng** cho các đặc trưng có giá trị khác không, được phân cách nhau bởi một ký tự trống. Mỗi hàng được kết thúc bằng một ký tự '\n'.

<feature>:<value> <feature>:<value> ... <feature>:<value>

Cho ví dụ, ma trận dữ liệu với 4 dữ liệu và 5 đặc trưng như sau:

0	3	0	0	1
4	1	0	0	0
6	5	9	2	0
6	0	0	5	3.

Được mô tả trong file đầu vào là:

2:3 5:1
1:4 2:1
2:5 3:9 4:2
1:6 4:5 5:3

Nhãn của các dữ liệu huấn luyện được chứa trong một file riêng biệt, gọi là *file mô tả nhãn dữ liệu*. Mỗi dòng của file chứa nhãn cho dữ liệu ở dòng tương ứng trong file mô tả dữ liệu ở trên. Nhãn của dữ liệu có thể nhận các giá trị sau:

+1 (dữ liệu gán nhãn thuộc lớp dương)

-1 (dữ liệu gán nhãn thuộc lớp âm)

0 (các dữ liệu chưa được gán nhãn)

Phiên bản hiện tại của bộ công cụ SVMlin chỉ có thể áp dụng cho bài toán phân lớp nhị phân.

❖ **Quá trình huấn luyện**

Gõ lệnh:

svmlin [options] training_examples training_labels

Trong đó*:*

*training_examples.weights.*File chứa dữ liệu huấn luyện

training_examples.outputs. File chứa kết quả mô hình phân lớp

❖ **Kiểm tra (testing)**

Gõ lệnh:

svmlin -f training_examples.weights test_examples_filename

Trong đó:

training_examples.weights: File chứa kết quả mô hình phân lớp

test_examples_filename: File chứa dữ liệu kiểm tra

❖ **Đánh giá**

Nếu nhãn của dữ liệu kiểm thử đã được biết trước, chúng ta sử dụng lệnh sau để tính ma trận thực thi của quá trình phân lớp:

svmlin -f weights_filename test_examples_filename test_labels_filename

❖ **Dữ liệu huấn luyện**

Dữ liệu huấn luyện được sử dụng bao gồm 1460 tài liệu (trong đó chỉ có 50 tài liệu được gán nhãn) được lấy từ bộ dữ liệu chuẩn *20-newsgroups*.

❖ **Kết quả phân lớp**

Với dữ liệu huấn luyện trên đây, SVMlin đạt độ chính xác là 92.8% khi lựa chọn chức năng multi-switch TSVM và đạt độ chính xác là 95.5% khi lựa chọn chức năng semi-supervised SVM. Điều này khẳng định tính hiệu quả của học bán giám sát SVM.

.

KẾT LUẬN

Những công việc đã làm được của khoá luận

Khoá luận đã khái quát được một số vấn đề về bài toán phân lớp bao gồm phương pháp phân lớp dữ liệu, phân lớp văn bản và các thuật toán học máy áp dụng vào bài toán phân lớp, trong đó chú trọng nghiên cứu tới phương pháp học bán giám sát được sử dụng rất phổ biến hiện nay.

Về phân lớp dữ liệu, khoá luận đã đưa ra bài toán tổng quan, cho cái gì và cần cái gì, đồng thời trình bày về phương pháp phân lớp dữ liệu tổng quát từ đó có thể giúp người đọc hiểu sơ qua về bài toán phân lớp.

Trình bày cơ bản về bài toán phân lớp văn bản, cách biểu diễn một văn bản trong bài toán phân lớp như thế nào, qua đó nêu lên các phương pháp phân lớp văn bản cơ bản hiện nay.

Tìm hiểu về các thuật toán học máy áp dụng vào bài toán phân lớp văn bản bao gồm thuật toán phân lớp sử dụng quá trình học có giám sát và học bán giám sát. Ở đây chúng ta tập trung chủ yếu nghiên cứu về quá trình học bán giám sát, nêu lên một số phương pháp học bán giám sát điển hình, trên cơ sở đó sẽ đi sâu tìm hiểu thuật toán học bán giám sát SVM.

Bài toán phân lớp trang Web áp dụng thuật toán bán giám sát SVM được nêu lên rất cụ thể. Trong phần thực nghiệm đã giới thiệu một phần mềm mã nguồn mở có tên là SVMlin, cách sử dụng phần mềm và kết quả chạy phần mềm do V. Sindhwani tiến hành trong năm 2007. Em đã tải phần mềm về nghiên cứu khảo sát song do hạn chế về thời gian và trình độ nên chưa làm chủ thực hiện phần mềm.

Hướng nghiên cứu trong thời gian tới

Như đã trình bày ở trên, do còn hạn chế về thời gian và kiến thức nên trong khoá luận chưa thể tìm hiểu sâu, đặc biệt là tiến hành thực hiện phần mềm SVMlin đã khảo sát. Vì thế trong thời gian tới em sẽ tìm hiểu kỹ hơn về phần mềm để có thể chủ động nắm vững việc thực hiện phần mềm, đặc biệt là các thuật toán học bán giám sát nền tảng lý thuyết của phần mềm [14,15].

TÀI LIỆU THAM KHẢO

I. Tiếng Việt

1. Nguyễn Việt Cường (2006). Sử dụng các khái niệm tập mờ trong biểu diễn văn bản và ứng dụng vào bài toán phân lớp văn bản. *Khóa luận tốt nghiệp đại học*, Trường Đại học Công nghệ - Đại học Quốc gia Hà Nội.
2. Phạm Thị Thanh Nam (2003). Một số giải pháp cho bài toán tìm kiếm trong CSDL Hypertext. *Luận văn tốt nghiệp cao học*, Khoa Công nghệ, ĐHQGHN, 2003.
3. Trần Thị Oanh (2006). Thuật toán self-training và co-training ứng dụng trong phân lớp văn bản. *Khóa luận tốt nghiệp đại học*, Trường Đại học Công nghệ - Đại học Quốc gia Hà Nội.

II. Tiếng Anh

4. Aixin Sun, Ee-Peng Lim, Wee-Keong Ng. Sun (2002). Web classification using support vector machine. *Proceedings of the 4th International Workshop on Web Information and Data Management*, McLean, Virginia, USA, 2002 (ACM Press).
5. Balaij Krishnapuuram, David Williams, Ya Xue,k Alex Hartemink, Lawrence Carin, Masrio A.T.Figueiredo (2005). On Semi-Supervised Classification. *NIPS*: 721-728, 2005.
6. H-J.Oh, S.H.Myaeng, and M-H.Lee (2000). A practical hypertext categorization method using links and incrementally available class information. *Proc of the 28rd ACM SIGIR2000*: 264-271, Athens, GR, 2000.
7. Kristin P. Bennett, Ayhan Demiriz (1998). Semi-Supervised Support Vector Machines. *NIPS 1998*: 368-374.
8. Linli Xu, Dale Schuurmans (2005). Unsupervised and Semi-Supervised Multi-Class Support Vector Machines. *AAAI 2005*: 904-910.
9. M. Craven and S.Slattery (2001). Relational learning with statistical predicate invention: Better models for hypertext. *Machine Learning,* **43**(1-2):97-119, 2001.

10. Panu Erastox (2001). Support Vector Machines: Background and Practice. *Academic Dissertation for the Degree of Licentiate of Philosophy*. University of Helsinki, 2001.

11. Paul Pavlidis, llan Wapinski, and William Stafford Noble (2004). Support vector machine classification on the web. *BIOINFORMATICS APPLICATION NOTE. 20(4), 586-587.*

12. T. Joachims (1999). Transductive Inference for Text Classification using Support Vector Machines. *International Conference on Machine Learning (ICML), 1999.*

13. T. Joachims (2003). Transductive learning via spectral graph partitioning. *Proceeding of The Twentieth International Conference on Machine Learning (ICML2003)*: 290-297.

14. V. Sindhwani, S. S. Keerthi (2006). Large Scale Semi-supervised Linear SVMs. *SIGIR* 2006.

15. V. Sindhwani, S.S. Keerthi (2007). Newton Methods for Fast Solution of Semi-supervised Linear SVMs**.** *Large Scale Kernel Machines*, MIT Press, 2005

16. Xiaojin Zhu (2005). Semi-Supervised Learning with Graphs. *PhD thesis, Carnegie Mellon University*, CMU-LTI-05-192, May 2005.

17. Xiaojin Zhu (2006). Semi-Supervised Learning Literature Survey. *Computer Sciences TR 1530*, University of Wisconsin – Madison, February 22, 2006.

18. http://people.cs.uchicago.edu/~vikass/svmlin.html