BỘ GIÁO DỤC VÀ ĐÀO TẠO
ĐẠI HỌC ĐÀ NẴNG

HỒ THỊ NGỌC

NGHIÊN CỨU ỨNG DỤNG HỌC BÁN GIÁM SÁT

Chuyên ngành: **KHOA HỌC MÁY TÍNH**
Mã số: **60.48.01**

TÓM TẮT LUẬN VĂN THẠC SĨ KỸ THUẬT

Đà Nẵng – Năm 2012

Công trình được hoàn thành tại
ĐẠI HỌC ĐÀ NẴNG

Người hướng dẫn khoa học: **PGS.TS Võ Trung Hùng**

Phản biện 1: TS. Nguyễn Thanh Bình

Phản biện 2: PGS.TS. Đoàn Văn Ban

Luận văn đã được bảo vệ trước Hội đồng chấm Luận văn tốt nghiệp thạc sĩ kỹ thuật họp tại Đại học Đà Nẵng vào ngày 04 tháng 03 năm 2012

Có thể tìm hiểu luận văn tại:
- Trung tâm Thông tin – Học liệu, Đại học Đà Nẵng.
- Trung tâm Học liệu, Đại học Đà Nẵng.

MỞ ĐẦU

1. Lý do chọn đề tài

Công nghệ thông tin phát triển mạnh đã đem lại nhiều tiện ích cho cuộc sống, được ứng dụng rộng rãi ở nhiều lĩnh vực, 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.

Hầu hết số lượng thông tin đồ sộ là chưa được gán nhãn, một yêu cầu lớn đặt ra là làm sao tổ chức và tìm kiếm thông tin, dữ liệu có hiệu quả nhất. Để giải quyết vấn đề trên thì bài toán phân lớp là một trong những giải pháp hợp lý. Trong thực tế là số lượng thông tin quá lớn, sử dụng phương pháp phân lớp dữ liệu bằng thủ công là điều không thể. Hướng giải quyết là tìm một chương trình máy tính tự động phân lớp các thông tin dữ liệu trên.

Để xử lý các bài toán phân lớp tự động thì phải 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. Tuy nhiên giải quyết vấn đề này thường gặp nhiều khó khăn vì các dữ liệu huấn luyện này thường rất hiếm và đắt vì đòi hỏi phải tốn nhiều thời gian và công sức của con người. Để khắc phục những hạn chế trên 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 đã được gán nhãn,

phương pháp học này đang được sử dụng rất phổ biến vì khả năng tiện lợi của nó.

Vì vậy, luận văn 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 VectorMachine – SVM) vào bài toán phân lớp (loại) văn bản và trang Web.

2. Mục đích của đề tài

Đề tài tập trung nghiên cứu các kỹ thuật học máy và nghiên cứu một số giải thuật thường sử dụng trong học máy. Sau đó ứng dụng kỹ thuật học bán giám sát vào bài toán phân lớp văn bản và trang Web.

3. Mục tiêu và nhiệm vụ nghiên cứu

Mục tiêu của đề tài là: Ứng dụng thành công kỹ thuật học máy "bán giám sát" vào một bài toán thực tế.

Nhiệm vụ chính của đề tài bao gồm: Nghiên cứu cơ sở lý thuyết về học bán giám sát và áp dụng kỹ thuật học bán giám sát vào thực tế trong các bài toán xử lý ngôn ngữ tự nhiên.

4. Đối tượng và phạm vi nghiên cứu

Đối tượng nghiên cứu bao gồm: các vấn đề liên quan đến học máy, liên quan đến học bán giám sát và nghiên cứu các giải thuật học bán giám sát.

Phạm vi nghiên cứu của luận văn tập trung vào kỹ thuật học bán giám sát và ứng dụng kỹ thuật này để giải quyết bài toán phân loại văn bản và trang Web.

5. Phương pháp nghiên cứu

Bao gồm phương pháp tài liệu và phương pháp thực nghiệm. Đối với phương pháp tài liệu tập trung nghiên cứu về cơ sở lý thuyết về học máy, cơ sở lý thuyết về kỹ thuật học bán giám sát và cơ sở lý thuyết về xử lý ngôn ngữ tự nhiên. Còn đối với phương pháp thực nghiệm tập trung vào việc xây dựng kho dữ liệu huấn luyện và xây dựng chương trình thử nghiệm.

6. Ý nghĩa khoa học và thực tiễn

Ý nghĩa khoa học: Nghiên cứu các kỹ thuật học máy và một số giải thuật thường sử dụng trong học máy. Đã ứng dụng thành công kỹ thuật học bán giám sát vào bài toán thực tế đó là "Sử dụng phương pháp SVM và bán giám sát SVM vào bài toán phân lớp văn bản và trang Web".

Ý nghĩa thực tiễn: Học bán giám sát là phương pháp học tốn ít thời gian và đảm bảo tối đa hiệu quả công việc. Nó là sự kết hợp của "học không có giám sát" và "học có giám sát", vì vậy rất thích hợp để xử lý vào các bài toán thực tế. Phương pháp học này có ứng dụng rất cao trong việc truy tìm dữ liệu, phân loại văn bản, nhận dạng ngôn ngữ văn bản, nhận dạng tiếng nói và chữ viết, dịch tự động, ….Đây là kỹ thuật chưa được nghiên cứu phổ biến ở Việt Nam điều đó mở ra hướng nghiên cứu, ứng dụng mới trong tương lai.

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

Chương 1: Nghiên cứu tổng quan.

Chương này trình bày khái quát về bài toán phân lớp dữ liệu, phân loại văn bản, học máy, các kỹ thuật học học máy.

Chương 2: Một số thuật toán học máy.

Chương này trình bày một số thuật toán học máy có giám sát, bán giám sát. Sử dụng SVM và bán giám sát SVM vào bài toán phân lớp văn bản và trang Web.

Chương 3: Thử nghiệm.

Ứng dụng phần mềm mã nguồn mở SVMLin đã được biên dịch chạy trên Windows vào thuật toán SVM và bán giám sát SVM để phân lớp văn bản và trang Web.

CHƯƠNG 1 - TỔNG QUAN VỀ PHÂN LỚP VĂN BẢN VÀ HỌC MÁY

1.1. Tổng quan về phân lớp dữ liệu

1.1.1. Khái niệm

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).

1.1.2. Mô tả bài toán phân lớp dữ liệu

Về phân lớp dữ liệu có nhiều bài toán như: phân lớp dữ liệu nhị phân, phân lớp dữ liệu đa lớp, phân lớp dữ liệu đơn trị, phân lớp dữ liệu đa trị,….

Phân lớp dữ liệu nhị phân là quá trình phân lớp dữ liệu vào một trong hai lớp cho trước khác nhau.

Phân lớp dữ liệu đa lớp là quá trình phân lớp với số lượng lớp cho trước lớn hơn hai.

Phân lớp dữ liệu đơn trị là quá trình phân lớp mà mỗi đối tượng dữ liệu trong tập dữ liệu huấn luyện được gán vào chính xác một lớp.

Phân lớp dữ liệu đa trị là quá trình phân lớp mà mỗi đối tượng dữ liệu trong tập dữ liệu huấn luyện (training data set) sau khi được phân lớp có thể thuộc vào từ hai lớp trở lên.

1.1.3. Quá trình phân lớp dữ liệu

Quá trình phân lớp dữ liệu có thể chia thành hai bước như sau:

❖ **Bước thứ nhất: Học (learning)**

Quá trình học nhằm xây dựng một mô hình phân lớp 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. Trong khi sử dụng một tập dữ liệu kiểm tra (test data set) cần phải tính độ chính xác của mô hình. Nếu độ chính xác đạt mức cao có nghĩa là chấp nhận được thì mô hình sẽ được sử dụng để xác định nhãn lớp cho các dữ liệu khác mới trong tương lai.

❖ **Bước thứ hai: Phân lớp (classification)**

Tiếp theo dùng mô hình đã xây dựng ở bước trước để phân lớp dữ liệu mới.

1.2. Phân lớp văn bản

1.2.1. Khái niệm

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.

1.2.2. Cách biểu diễn văn bản

Cách biểu diễn thông thường nhất là bằng mô hình vector:

❖ **Mô tả:**

Mỗi văn bản được biểu diễn bằng một vector trọng số. Độ dài của vector là số các từ khóa (keyword) xuất hiện trong ít nhất trong một mẫu dữ liệu huấn luyện. Biểu diễn trọng số có thể là nhị phân (từ khóa đó có hay không xuất hiện trong văn bản tương ứng) hoặc không nhị phân (từ khóa đó xuất hiện bao nhiêu lần trong văn bản đó).

❖ **Biểu diễn trang Web**

Biểu diễn trang web theo mô hình vector như sau:

- **Cách 1:** Cách này sẽ liệt kê tần số xuất hiện của mỗi từ khóa trong một trang web.
- **Cách 2:** Sử dụng đến chức năng liên kết của trang web
- **Cách 3:** Dùng một vector cấu trúc
- **Cách 4:** Xây dựng một vector có cấu trúc.

1.2.3. Phương pháp phân lớp văn bản

Dùng các thuật toán học máy (machine learning).

1.2.4. Ứng dụng của phân lớp văn bản

- Tìm kiếm văn bản.
- Lọc các văn bản hoặc một phần các văn bản chứa dữ liệu cần tìm.
- Trích lọc thông tin trên.

1.2.5. Các bước trong quá trình phân lớp văn bản

Gồm 4 bước:

Đánh chỉ số (indexing

Xác định độ phân lớp

So sánh

Phản hồi (thích nghi.

1.3. Học máy (Machine Learning)

1.3.1. Định nghĩa về học máy

❖ *Với:*

Một tập dữ liệu vũ trụ X

- *Một tập mẫu S, cho S là tập hợp con của X*
- *Một số hàm đích (quá trình ghi nhãn) f: X → {đúng, sai}*
- *Một tập huấn luyện D được gán, D = {(x, y) | x thuộc S và y = f(x)}*
- *Tính toán một hàm f': X → {đúng, sai} bằng cách sử dụng D như là:*

$$f'(x) \cong f(x) \quad (1.4)$$

cho tất cả các x thuộc X.

1.3.2. Các kỹ thuật học máy

1.3.2.1. Học không có giám sát (Unsupervised learning)

Học với tập dữ liệu huấn luyện ban đầu hoàn toàn chưa được gán nhãn.

1.3.2.2. Học có giám sát (Supervised learning)

Học với tập dữ liệu huấn luyện ban đầu hoàn toàn được gán nhãn.

1.3.2.3. Học bán giám sát (Semi-supervised learning)

❖ **Khái niệm**

Học cả dữ liệu gán nhãn và chưa gán nhãn.

❖ **Lịch sử phát triển**

1.3.3. Một số ứng dụng hiện có bằng phương pháp thống kê

1.3.3.1. Nhận dạng ngôn ngữ (Language identification)

1.3.3.2. Dịch tự động (Machine translation)

1.3.3.3. Phân loại văn bản (Text categorization)

CHƯƠNG 2 - MỘT SỐ THUẬT TOÁN HỌC MÁY

2.1. Thuật toán học bán giám sát Self-training

2.1.1. Giới thiệu

Nội dung chính là thuật toán học - sử dụng lặp nhiều lần một phương pháp học giám sát. Self-training là một trong những kỹ thuật học bán giám sát được sử dụng rất phổ biến. Với một bộ phân lớp (classifier) ban đầu được huấn luyện bằng một số lượng nhỏ các dữ liệu gán nhãn. Tiếp theo sử dụng bộ phân lớp này để gán nhãn các dữ liệu chưa gán nhãn. Các dữ liệu được gán nhãn có độ tin cậy cao (vượt trên một ngưỡng nào đó) và nhãn tương ứng của chúng được đưa vào tập huấn luyện (training set). Sau đó, bộ phân lớp được học lại trên tập huấn luyện mới ấy và thủ tục lặp tiếp tục. Ở mỗi vòng lặp, bộ học sẽ chuyển một vài các mẫu có độ tin cậy cao nhất sang tập dữ liệu huấn luyện cùng với các dự đoán phân lớp của chúng. Tên gọi self-training xuất phát từ việc nó sử dụng dự đoán của chính nó để dạy chính nó.

2.1.2. Thuật toán

❖ **Mục đích:** Mở rộng tập các mẫu gán nhãn ban đầu bằng cách chỉ cần một bộ phân lớp với một khung nhìn của dữ liệu.

❖ **Dữ liệu vào:**

- L: là tập các dữ liệu gán nhãn.
- U: là tập các dữ liệu chưa gán nhãn.

❖ **Dữ liệu ra:**

- Gán nhãn cho tập con U’ của U có độ tin cậy cao nhất.

❖ **Giải thuật:**

Loop

- Huấn luyện bộ phân lớp h trên tập dữ liệu huấn luyện L.
- Sử dụng h để phân lớp dữ liệu trong tập U.
- Tìm tập con U' của U có độ tin cậy cao nhất.
- L + U' -> L
- U – U'-> U

Until (U = ∅)

2.2. Thuật toán học bán giám sát Co-training

2.2.1. Giới thiệu

Thuật toán co-training dựa trên giả thiết rằng các đặc trưng (features) có thể được phân chia thành 2 tập con. Mỗi tập con phù hợp để huấn luyện một bộ phân lớp tốt. Hai tập con đó phải thoả mãn tính chất độc lập điều kiện (conditional independent) khi cho trước lớp (class). Thủ tục học được tiến hành như sau:

- Học 2 bộ phân lớp riêng rẽ bằng dữ liệu đã được gán nhãn trên hai tập thuộc tính con tương ứng.
- Mỗi bộ phân lớp sau đó lại phân lớp các dữ liệu chưa gán nhãn (unlabel data). Sau đó, chúng lựa chọn ra các dữ liệu chưa gán nhãn + nhãn dự đoán của chúng (các mẫu (examples) có độ tin cậy cao) để dạy cho bộ phân lớp kia.
- Sau đó, mỗi bộ phân lớp được học lại (re-train) với các mẫu huấn luyện được cho bởi bộ phân lớp kia và tiến trình lặp bắt đầu.

Cái khó của co-training là ở chỗ: hai bộ phân lớp phải dự đoán trùng khớp trên dữ liệu chưa gán nhãn rộng lớn cũng như dữ liệu gán nhãn.

2.2.2. Thuật toán

❖ **Mục đích:** Mở rộng tập các mẫu gán nhãn ban đầu bằng cách sử dụng hai bộ phân lớp với hai khung nhìn của dữ liệu.

❖ **Dữ liệu vào:**

- L: là tập các mẫu huấn luyện đã gán nhãn.
- U: là tập các mẫu chưa gán nhãn.

❖ **Dữ liệu ra:**

- Tạo một tập dữ liệu gán nhãn U'gồm u mẫu được chọn ngẫu nhiên từ U.

❖ **Giải thuật [2]:**

For i=1 **to** k **do**

- Sử dụng L huấn luyện bộ phân lớp h_1 trên phần x_1 của x .
- Sử dụng L huấn luyện bộ phân lớp h_2 trên phần x_2 của x .
- Cho h_1 gán nhãn p mẫu dương và n mẫu âm từ tập U' .
- Cho h_2 gán nhãn p mẫu dương và n mẫu âm từ tập U' .
- Thêm các mẫu tự gán nhãn này vào tập L .
- Chọn ngẫu nhiên 2 p + 2n mẫu từ tập U bổ sung vào tập U'

2.3. Thuật toán học có giám sát SVM và bán giám sát SVM

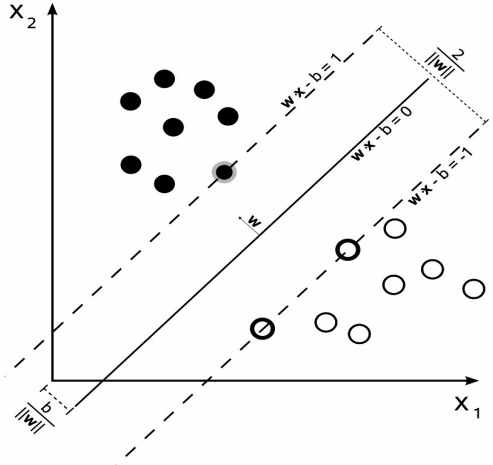
2.3.1. Giới thiệu

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

2.3.2. Thuật toán SVM

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 siêu mặt 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 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 họa cho thuật toán này



Hình 2.4. Siêu mặt tối ưu và biên

2.3.3. Huấn luyện SVM

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

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…).

2.4. Bán giám sát SVM và phân lớp trang Web

2.4.1. Giới thiệu về bán giám sát SVM

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

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

2.4.2.1. Giới thiệu bài toán phân lớp trang Web

Phân lớp trang web là một trường hợp đặc biệt của phân lớp văn bản. Trong trang web có 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….

2.4.2.2. Áp dụng S3VM vào phân lớp trang Web

Khi áp dụng thuật toán S3VM vào quá trình phân lớp nó sẽ tìm ra được nhãn lớp của các trang web chưa gán nhãn bằng cách thay thế vector trọng số biểu diễn trang web đó vào phương trình siêu phẳng của S3VM. Từ đó suy ra thực chất của quá trình phân lớp bán giám sát các trang web là: tập dữ liệu huấn luyện (training set) 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.

2.5. Học ghép đôi của mô hình khai thác văn bản

2.5.1. Giới thiệu

Ý tưởng trong phần này là chúng ta có thể đạt được độ chính xác cao hơn với phương pháp học bán giám sát cho các bộ khai thác thông tin bằng cách ghép cặp đôi một cách đồng thời việc đào tạo của nhiều bộ khai thác thông tin. Chúng ta có thể hiểu rằng các nhiệm vụ học bán giám sát dưới mức hạn chế có thể được thực hiện dễ dàng hơn bằng cách thêm nhiều hạn chế mới phát sinh từ việc ghép cặp đôi việc đào tạo của nhiều bộ khai thác thông tin. Chúng ta có thể xác định tổng quát có ba loại ghép cặp đôi giữa các chức năng mục tiêu mà có thể được kết hợp để tạo thành một mạng lưới dày đặc của các vấn đề học tập ghép cặp đôi.

2.5.2. Các mẫu

Chúng tôi sử dụng các mẫu văn bản để biểu diễn cho việc trích thông tin từ văn bản tự do.

2.5.3. Học theo mẫu bẫy khởi động (Bootstrapped)

“Bootstrap learning” là phương pháp khởi động học tập để học bán giám sát. Phương pháp này tự khởi động từ một số lượng nhỏ dữ liệu có nhãn.

2.5.3.1. Độ suy giảm về ngữ nghĩa

Sau bước lặp của phương pháp Bootstrapping thì độ chính xác suy giảm dần vì có sai sót trong quá trình ghi nhãn tích lũy, vấn đề này được gọi với tên là độ suy giảm về ngữ nghĩa (seSmantic drift).

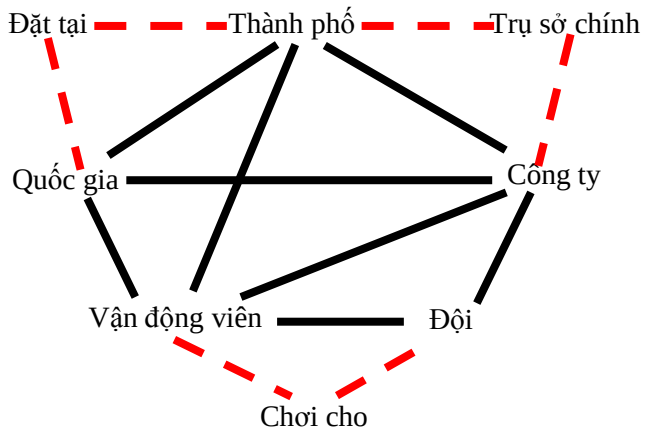
2.5.3.2. Đào tạo theo phương pháp ghép đôi bẫy khởi động (Coupled bootstrapped)

❖ **Có ba loại ghép đôi thông dụng sau đây**

1. Các ràng buộc đầu ra

2. Các ràng buộc thành phần

3. Các ràng buộc phù hợp đa chiều



Hình 2.5. Các ràng buộc loại trừ lẫn nhau

Lưu ý: Các ràng buộc loại trừ lẫn nhau (đường liền nét), các ràng buộc kiểu kiểm tra (đường nét đứt).

2.5.4. Thuật toán

❖ **Giới thiệu:**

❖ **Mục đích:**

Mục đích nghiên cứu thuật toán Người học mẫu có ghép đôi (**C**oupled **P**attern **L**earner – CPL) là để trích xuất thể loại và các ví

dụ quan hệ từ văn bản không có cấu trúc. CPL tìm các mẫu ngữ cảnh có bộ trích xuất với độ tin cậy cao cho mỗi vị từ (ví dụ: “X và các công ty phần mềm khác” và “X thắng 1 điểm cho Y”) và dùng chúng để tạo ra một tập trường hợp vị từ có độ tin cậy, các cụm danh từ điền vào các mẫu bỏ trống của “X” và “Y” tại các câu trong một tập ngữ liệu là xảy ra đồng thời với các mẫu kia. Tại thời điểm bắt đầu quá trình xử lý, CPL khởi tạo các tập trường hợp và các mẫu đã có cùng với các trường hợp tiềm năng và các mẫu đã dùng như là các dữ kiện đầu vào. Trong mỗi lần lặp, CPL mở rộng các tập trường hợp và mẫu đã có cho mỗi vị từ, đồng thời tuân thủ nguyên tắc loại trừ lẫn nhau và những hạn chế trong việc kiểm tra chủng loại. Việc này đạt được là nhờ vào bước lọc ra các trường hợp triển vọng xảy ra đồng thời với các trường hợp hoặc các mẫu ra khỏi nhóm loại trừ lẫn nhau và nhờ vào việc đòi hỏi các đối số của các mối quan hệ có triển vọng để trở thành các trường hợp có triển vọng.

❖ **Đầu vào**: một bản thể O và một tập ngữ liệu lớn C

❖ **Đầu ra:** đề xuất các trường hợp/ các mẫu ngữ cảnh cho mỗi vị từ

❖ **Giải Thuật:** Người học mẫu có ghép đôi (**C**oupled **P**attern **L**earner – CPL)

For i=1, 2, …, vô cùng do

Foreach vị từ p thuộc O **do**

Rút trích (Extract) các trường hợp có triển vọng mới / các mẫu ngữ cảnh đang sử dụng các mẫu / các trường hợp đã đề cập gần đây;

Lọc (Filter) các trường hợp có triển vọng nhưng bị lỗi trong việc ghép đôi

Sắp xếp (Rank) các trường hợp/ các mẫu theo ngữ cảnh có triển vọng;

Đề xuất (Promote) các trường hợp/ các mẫu theo ngữ cảnh có triển vọng.

CHƯƠNG 3 - THỬ NGHIỆM

3.1. Mô tả ứng dụng

Ứng dụng thuật toán học bán giám sát SVM được cài đặt trên phần mềm mã nguồn mở SVMlin (đã được biên dịch lại chạy trên Windows) để phân lớp bán giám sát văn bản và các tài liệu web.

3.2. Yêu cầu của ứng dụng

- **Đầu vào:**

Để sử dụng ứng dụng trên ta phải chuẩn bị khâu dữ liệu đầu vào gồm có 3 tập tin sau:

Tập tin **traindt.dat**: là tập tin chứa dữ liệu huấn luyện.

Tập tin **trainlb.dat**: là tập tin chứa nhãn của dữ liệu huấn luyện.

Tập tin **test.dt**: là tập tin chứa tập dữ liệu được đưa vào kiểm tra (dữ liệu này có thể đã được gán nhãn hoặc chưa được gán nhãn).

- **Đầu ra:**

Sau khi đưa tập dữ liệu **test.dt** vào kiểm tra thì dữ liệu kiểm tra sẽ gán nhãn và được xuất ra tập tin **test.dt.outputs**.

3.3. Lựa chọn công cụ

Dùng phần mềm mã nguồn mở SVMlin viết bằng ngôn ngữ C hoặc C++.

3.4. Các bước triển khai

3.4.1. Xây dựng kho dữ liệu cho các tập tin đầu vào

Thành lập bảng dữ liệu huấn luyện như sau:

Bảng 3.1. Bảng tính năng cho một số lượng nhỏ các đối tượng huấn luyện

	Chân	Cánh	Lông thú	Lông chim	Động vật có vú
Mèo	4	không	có	không	đúng
Quạ	2	có	không	có	sai
Ếch	4	không	không	không	sai
Dơi	4	có	có	không	đúng
Ghế đẩu	3	không	không	không	sai

Từ bảng 3.1, xây dựng ma trận dữ liệu với 5 dữ liệu và 4 đặc trưng như sau:

4 0 1 0
2 1 0 1
4 0 0 0
4 1 1 0
3 0 0 0

Tập tin đầu vào **traindt.dat** được mô tả như sau:

1:4 3:1
1:2 2:14:1
1:4
1:4 2:1 3:1
1:3

Tập tin đầu vào **trainlb.dat** (đưa vào cột cuối cùng trong bảng 3.1 để gán nhãn) được mô tả như sau:

+1
-1
-1
+1
-1

Tập tin đầu vào **test.dt** được mô tả như sau:

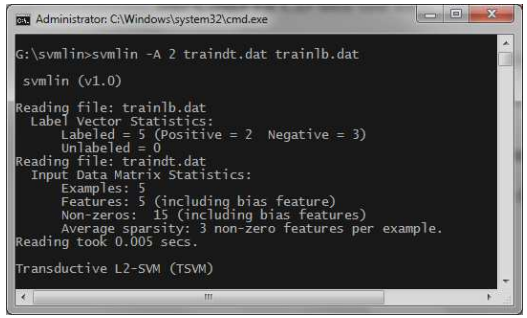
3:1
1:2 2:1 4:1
1:4
1:4 2:1 3:1

Trong file đính kèm có một tập tin tên svmlin.zip.

❖ **Chạy chương trình và giao diện:**

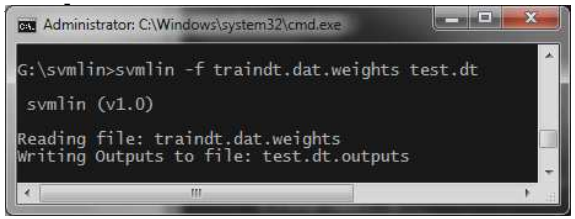
Mở CMD và CD đến thư mục vừa giải nén.

- Để học gõ lệnh sau: ***svmlin -A 2 traindt.dat trainlb.dat***. Sau khi quá trình hoàn tất, sẽ tạo ra 2 file: *traindt.dat.weights* và *traindt.dat.outputs*.



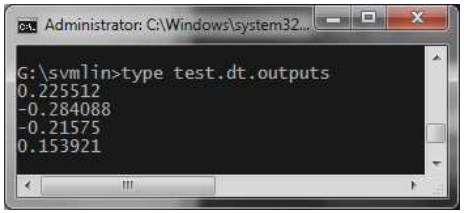
Hình 3.1. Giao diện chương trình – Học dữ liệu

- Để kiểm tra dữ liệu, gõ lệnh **svmlin -f traindt.dat.weights**



Hình 3.2. Giao diện chương trình – Kiểm tra dữ liệu

- Sau khi quá trình hoàn tất, tập kết quả dữ liệu kiểm tra sẽ được tạo ra ở tập tin *test.dt.outputs*. Xem kết quả tập tin này như sau:



Hình 3.3. Giao diện chương trình – Xem kết quả

3.5. Đánh giá:

Phần mềm SVMlin chỉ thực hiện phân lớp nhị phân.

KẾT LUẬN

Đề tài 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.

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 văn bản và 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. Tôi đã tự biên dịch lại chạy trên môi trường Windows cho thuận tiện và đã tự xây dựng được kho dữ liệu huấn luyện để đưa vào chạy chương trình. Đề tài đã trình bày khá chi tiết cách sử dụng phần mềm và chạy cho ra kết quả.

Phần mềm mã nguồn mở SVMlin chỉ dùng để phân lớp văn bản theo phương pháp nhị phân nên còn nhiều hạn chế. Để có được phần mềm phân lớp văn bản hoàn chỉnh thì có thể tiếp tục tự cài đặt thuật toán phân lớp hoặc dựa trên nền tảng phần mềm SVMlin để phát triển bài toán phân lớp theo phương pháp đa lớp.