HỌC VIỆN CÔNG NGHỆ BƯU CHÍNH VIỄN THÔNG

----------o0o----------

NGUYỄN THỊ VÂN TRANG

NGHIÊN CỨU MỘT SỐ THUẬT TOÁN HỌC MÁY CÓ GIÁM SÁT VÀ ỨNG DỤNG TRONG LỌC THƯ RÁC

Chuyên ngành : Truyền dữ liệu và mạng máy tính

Mã số : 60.48.15

TÓM TẮT LUẬN VĂN THẠC SỸ KỸ THUẬT

HÀ NỘI – NĂM 2012

Luận văn được hoàn thành tại:

HỌC VIỆN CÔNG NGHỆ BƯU CHÍNH VIỄN THÔNG

Người hướng dẫn khoa học: **TS HOÀNG XUÂN DẬU**

Phản biện 1: ……

Phản biện 2: ……

Luận văn sẽ được bảo vệ trước Hội đồng chấm luận văn thạc sĩ tại Học viện Công nghệ Bưu chính Viễn thông

Vào lúc:.....giờ ..... ngày.....tháng.....năm .....

Có thể tìm hiểu luận văn tại:

- Thư viện của Học viện Công nghệ Bưu chính Viễn thông

LỜI MỞ ĐẦU

Hiện nay, việc trao đổi thông tin, liên lạc qua Internet đã trở nên quen thuộc, phổ biến ở hầu hết các quốc gia, các lĩnh vực trong đời sống xã hội. Thư điện tử (email) là một trong những dịch vụ truyền thông tiện ích, được ứng dụng thường xuyên, giúp con người trao đổi thông tin một cách nhanh chóng, chính xác.

Cùng với sự phát triển mạnh mẽ của mạng Internet, các dịch vụ thư điện tử đã được mở rộng với số lượng lớn các nhà cung cấp dịch vụ và lượng người dùng khổng lồ. Thư điện tử được truyền qua mạng Internet dưới dạng các tín hiệu điện nên tốc độ di chuyển gần như là tức thời.

Tuy nhiên, ngoài những lợi ích mà thư điện tử mang lại, chúng có thể gây ra những phiền phức, thiệt hại nếu không biết cách khắc phục, loại bỏ và phòng chống. Một trong những vấn đề nhức nhối luôn song hành với thư điện tử là thư rác hay còn gọi là "spam emails". Đó là những thư quảng cáo, hay các thư mang nội dung với mục đích tấn công ăn cắp thông tin hoặc phá hoại gây thiệt hại cho người dùng. Theo thống kê của MessageLabs vào tháng 10 năm 2005, số lượng thư rác đã chiếm 68% trên tổng số tất cả các thư được gửi đi.

Để ngăn chặn thư rác, nhiều tổ chức, cá nhân đã nghiên cứu và phát triển những kỹ thuật phân loại thư điện tử thành các nhóm (group); từ đó xác định, nhận biết giữa thư rác và thư có giá trị. Tuy nhiên, những người tạo nên spam emails

(spammer) luôn tìm mọi cách vượt qua các bộ phân loại này và phát tán chúng. Do vậy, cần có một giải pháp có khả năng tự học để lọc thư rác một cách hiệu quả hơn.

Xuất phát từ thực trạng đó, tôi chọn đề tài *"Nghiên cứu một số thuật toán học máy có giám sát và ứng dụng trong lọc thư rác"* với mục đích nghiên cứu một số thuật toán học máy có giám sát và thử nghiệm ứng dụng cho bài toán lọc thư rác. Nội dung của luận văn được trình bày theo 3 chương:

Chương 1: Giới thiệu tổng quát về học máy bao gồm khái niệm, ứng dụng và phần trình bày chi tiết về học máy có giám sát, các kỹ thuật của học máy có giám sát dùng cho phân loại như Naïve Bayes, SVM, cây quyết định,…Chương cũng giới thiệu khái quát về thư rác, các đặc trưng của thư rác và bài toán lọc thư rác.

Chương 2: Đi sâu nghiên cứu hai thuật toán học máy có giám sát là Naïve Bayes và phương pháp SVM (Support Vector Machine).

Chương 3: Phần đầu chương giới thiệu bộ dữ liệu thử nghiệm và cài đặt chi tiết hai thuật toán đề cập ở chương 2. Phần cuối của chương trình bày kết quả thu được và đưa ra đánh giá về hai thuật toán được sử dụng trong bài toán lọc thư rác.

CHƯƠNG 1: TỔNG QUAN VỀ HỌC MÁY

1.1. Tổng quan về học máy

1.1.1. Khái quát về học máy

Học máy (tiếng Anh: Machine Learning) là một lĩnh vực của trí tuệ nhân tạo liên quan đến việc phát triển các kĩ thuật cho phép các máy tính có thể "học". Học máy được xem là phương pháp tạo ra các chương trình máy tính sử dụng kinh nghiệm, quan sát hoặc dữ liệu trong quá khứ để cải thiện công việc của mình trong tương lai.

1.1.2. Phân loại học máy

Học máy chủ yếu được phân thành 3 loại chính:

a) Học có giám sát (supervised learning)

Với cách học này, kinh nghiệm được cho một cách tường minh dưới dạng đầu vào và đầu ra của hàm đích, ví dụ cho trước tập các mẫu cùng nhãn phân loại tương ứng.

b) Học không có giám sát (unsupervised learning)

Ngược với học có giám sát, học không giám sát là cách học mà kinh nghiệm chỉ gồm các mẫu và không có nhãn hoặc giá trị hàm đích đi kèm.

c) Học tăng cường (reinforcement)

Đối với dạng học này, kinh nghiệm không được cho trực tiếp dưới dạng đầu vào/ đầu ra. Thay vào đó, hệ thống nhận được một giá trị tăng cường là kết quả cho một chuỗi hành động nào đó.

1.1.3. Ứng dụng của học máy

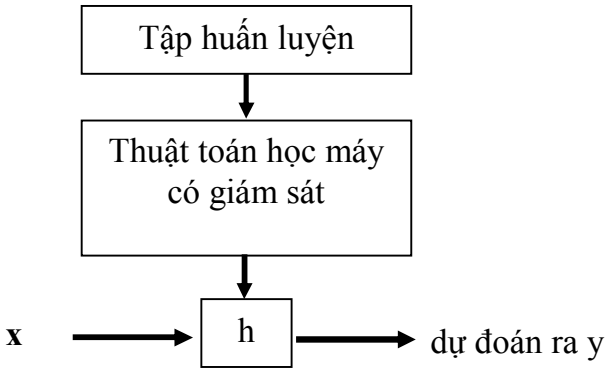
Học máy là một nhánh nghiên cứu rất quan trọng của trí tuệ nhân tạo với khá nhiều ứng dụng thành công trong thực tế. Cụ thể:

- Xử lý ngôn ngữ tự nhiên
- Phát hiện và nhận dạng mặt người
- Lọc thư rác, phân loại văn bản
- …

1.1.4. Học máy có giám sát

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).

Mục đích chính của bài toán học có giám sát là để học một ánh xạ từ x tới y. Mô hình chung của học có giám sát được khái quát như hình 1.2:



Hình 1.2: Mô hình thuật toán học có giám sát

Hiện nay đã có rất nhiều thuật toán được sử dụng để tạo những trình học có giám sát, phổ biến nhất là:

- Thuật toán K hàng xóm (KNN)
- Mô hình xác suất Naïve Bayes
- Phương pháp Support Vector Machines
- ….

1.2. Tổng quan về thư rác và các đặc trưng của thư rác

1.2.1. Khái quát về thư rác (spam – emails)

Thư rác (spam) là những bức thư điện tử không yêu cầu, không mong muốn và được gửi hàng loạt tới nhiều người nhận.

1.2.2. Các đặc trưng của thư rác

Các loại thư rác hiện này có một số đặc điểm sau:

- Thư rác được gửi đi một cách tự động
- Thư rác được gửi đến những địa chỉ ngẫu nhiên trên một diện rộng
- Nội dung của thư rác thường là những nội dung bất hợp pháp, gây phiền hà cho người dùng
- Địa chỉ của người gửi thư rác thường là những địa chỉ trá hình

1.2.3. Phân loại thư rác

Có rất nhiều cách phân loại thư rác:

- Dựa trên kiểu phát tán thư rác

- Dựa vào quan hệ với người gửi thư rác
- Dựa vào nội dung thư rác.
- Dựa trên động lực của người gửi

1.2.4. Quy trình và thủ đoạn gửi thư rác

Để phát tán thư rác, những người gửi thư rác phải có được những điều kiện sau: một là có danh sách địa chỉ email nhận thư, hai là có các server cho phép gửi thư, ba là phải soạn được nội dung thư theo yêu cầu quảng cáo và qua mặt được các bộ lọc nội dung, cuối cùng cần có những chương trình để gửi thư đi.

1.2.4.1. Thu thập địa chỉ email

Danh sách địa chỉ email cần gửi có thể thu thập được từ nhiều nguồn khác nhau, họ có thể mua từ các trang web thương mại có nhiều thành viên đăng ký hoặc sử dụng các kỹ thuật như kỹ thuật *Phishing email*,..

Người gửi thư rác còn sử dụng các máy tìm kiếm chỉ để tìm kiếm địa chỉ email trên các trang web.

Danh sách các địa chỉ cũng có thể được sinh tự động theo một cơ chế nào đó.

1.2.4.2. Tìm kiếm các máy tính trên Internet cho phép gửi thư

Muốn gửi được thư rác, người gửi thư rác cần có trong tay một danh sách các server để gửi thư đi. Các server này có thể là những server chuyên để gửi thư rác do người gửi thư rác

sở hữu hoặc thuê, hoặc là những server bị người gửi thư rác lợi dụng.

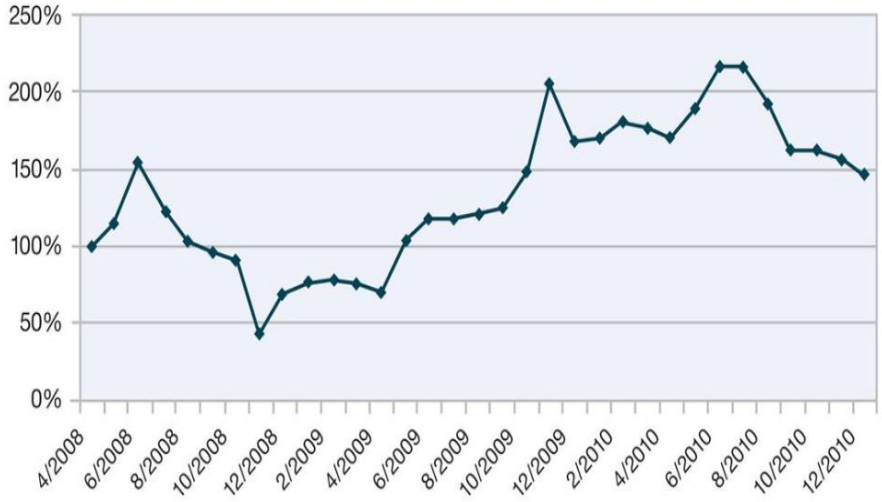
1.3. Bài toán phân loại thư rác dựa trên học máy có giám sát

1.3.1. Sự cần thiết phân loại thư rác

a) Tốc độ phát triển của thư rác

Theo số liệu thống kê của hãng bảo mật Symantec cho biết, có tổng số 70 tỷ thư rác được gửi đi mỗi ngày trên toàn cầu, những thư với nội dung mời gọi mua dược phẩm chiếm tới 64%.

Số lượng thư rác năm từ tháng 4/2008 đến tháng 12/2010:



Hình 1.6: Số lượng thư rác năm từ tháng 4/ 2008 đến tháng 12/2010

b) Sự bùng nổ của thư rác ở Việt Nam

Hãng bảo mật Sophos vừa công bố danh sách "dirty dozen" mới nhất, trong đó nêu danh tính top 12 quốc gia phát tán thư rác nhiều nhất thế giới tính đến thời điểm tháng 1-3/2012. Đứng đầu là Ấn Độ, tiếp theo là Mỹ và Hàn Quốc còn Việt Nam đứng thứ 10.

Việt Nam có tên trong cả danh sách của Sophos và Trend Micro được thể hiện trong bảng 1.1.

Bảng 1.1:Danh sách top 10 quốc gia phát tán spam nhất thế giới quí I/2012 của Sophos. Việt Nam đứng thứ 10/12.

STT	TÊN NƯỚC	TỶ LỆ PHẦN TRĂM PHÁT TÁN THƯ RÁC
1	India	9.3%
2	USA	8.3%
3	S Korea	5.7%
4	Indonesia	5.0%
5	Russia	5.0%
6	Italy	4.9%
7	Brazil	4.3%
8	Poland	3.9%
9	Pakistan	3.3%
10	VietNam	3.2%
11	Taiwan	2.9%
12	Peru	2.5%
13	Khác	41.7%

1.3.2. Bài toán phân loại thư rác

Bài toán phân loại thư rác thực chất là bài toán phân loại các thư nhận được thành hai nhóm chính là nhóm thư rác và nhóm thư bình thường.

Việc phân loại tiến hành như sau. Trước tiên, nội dung thư được biểu diễn dưới dạng các *đặc trưng* hay các *thuộc tính*, mỗi đặc trưng thường là một từ hoặc cụm từ xuất hiện trong thư. Tiếp theo, trong giai đoạn huấn luyện, tập thư đã được gán nhãn {rác, bình thường} - gọi là *dữ liệu huấn luyện* hay dữ liệu mẫu - được sử dụng để huấn luyện một bộ phân loại. Sau khi huấn luyện xong, bộ phân loại được sử dụng để xác định thư mới (thư chưa biết nhãn) thuộc vào loại nào trong hai loại nói trên. Trong cả giai đoạn huấn luyện và phân loại, thuật toán phân loại chỉ làm việc với nội dung thư đã được biểu diễn dưới dạng các đặc trưng.

1.3.3. Biểu diễn nội dung thư rác

Biểu diễn nội dung thư dưới dạng tập hợp từ ("túi từ")

Các phương pháp lọc thư bằng cách tự động phân loại theo nội dung đều sử dụng cách biểu diễn thư dưới dạng vector. Mặc dù có nhiều cách xây dựng vector nhưng cách đơn giản nhất là mô hình "*túi từ*" ("bag-of-words"). Nguyên tắc cơ bản của phương pháp này là không quan tâm tới vị trí xuất hiện các từ hay cụm từ trong thư mà coi thư như một tập hợp không có thứ tự các từ. Mỗi thư khi đó được biểu diễn bởi một vector. Số

phần tử của vector bằng số lượng từ khác nhau trên toàn bộ tập dữ liệu huấn luyện.

Có nhiều cách tính giá trị các phần tử của vector. Cách đơn giản nhất là sử dụng giá trị nhị phân {1,0} tùy thuộc vào từ tương ứng có xuất hiện trong thư tương ứng với vector hay không.

Dưới đây là một ví dụ đơn giản minh hoạ cho cách biểu diễn nội dung nói trên. Dữ liệu huấn luyện bao gồm bốn thư, trong đó hai thư là thư rác và hai là thư bình thường được thể hiện trong bảng 1.2 và bảng 1.3.

Bảng 1.2. Ví dụ nội dung của 4 thư.

Số TT	Nội dung	Nhãn
1	Mua và quay số	Rác
2	Mua một tặng một	Rác
3	Tôi mua rồi	Bình thường
4	mới nhận được	Bình thường

Bảng 1.3. Biểu diễn vector cho dữ liệu trong bảng 1.2

TT	mua	và	quay	số	một	tặng	tôi	rồi	mới	nhận	được
1	1	1	1	1	0	0	0	0	0	0	0
2	1	0	0	0	2	1	0	0	0	0	0
3	1	0	0	0	0	0	1	1	0	0	0
4	0	0	0	0	0	0	0	0	1	1	1

Một số phương pháp biểu diễn nội dung thư khác

Đặc điểm chung của phương pháp không dùng "túi từ" là sử dụng các đặc trưng chứa nhiều thông tin về ngữ nghĩa hơn

để biểu diễn nội dung văn bản. Tiêu biểu nhất là phương pháp sử dụng *cụm từ có ngữ nghĩa* (*phrase*) và phương pháp sử dụng *phân cụm từ* (*word clusters*).

1.4. Kết luận chương

Chương này đã giới thiệu được tổng quát về học máy bao gồm khái niệm, ứng dụng và phần trình bày chi tiết về học máy có giám sát, các kỹ thuật của học máy có giám sát dùng cho phân loại như Naïve Bayes, SVM, cây quyết định,…Chương cũng giới thiệu khái quát về thư rác, các đặc trưng của thư rác và bài toán lọc thư rác.

CHƯƠNG 2: MỘT SỐ THUẬT TOÁN HỌC MÁY CÓ GIÁM SÁT VÀ ỨNG DỤNG TRONG BÀI TOÁN LỌC THƯ RÁC

2.1. Thuật toán Naïve Bayes

2.1.1. Định lý

Theo lý thuyết học Bayes, nhãn phân loại được xác định bằng cách tính xác suất điều kiện của nhãn khi quan sát thấy tổ hợp giá trị thuộc tính $<x_1, x_2, \ldots, x_n>$. Thuộc tính được chọn, ký hiệu c_{MAP} là thuộc tính có xác suất điều kiện cao nhất tức là:

$$y = c_{MAP} = \arg\max_{c_j \in C} P(c_j | x_1, x_2, \ldots, x_n) \quad (2.1)$$

Sử dụng quy tắc Bayes, biểu thức trên được viết lại như sau:

$$\begin{aligned} c_{MAP} &= \arg\max_{c_j \in C} \frac{P(x_1, x_2, \ldots, x_n | c_j) P(c_j)}{P(x_1, x_2, \ldots, x_n)} \\ &= \arg\max_{c_j \in C} P(x_1, x_2, \ldots, x_n | c_j) P(c_j) \end{aligned} \quad (2.2)$$

Giá trị $P(c_j)$ được tính bằng tần suất quan sát thấy nhãn c_j trên tập huấn luyện, tức là bằng số mẫu có nhãn là c_j chia cho tổng số mẫu. Việc tính $P(x_1, x_2, \ldots, x_n \mid c_j)$ khó khăn hơn nhiều. Để tính xác suất này được chính xác, mỗi tổ hợp giá trị thuộc tính phải xuất hiện cùng nhãn phân loại đủ nhiều trong khi số mẫu huấn luyện thường không đủ lớn.

Để giải quyết vấn đề này, ta giả sử các thuộc tính là độc lập về xác suất với nhau khi biết nhãn phân loại c_j.

Với giả thiết về tính độc lập xác suất có điều kiện được viết lại như sau:

$$P(x_1, x_2,...,x_n | c_j) = P(x_1 | c_j) P(x_2 | c_j) \dots P(x_n | c_j) \quad (2.3)$$

Thay vào biểu thức (2.2) ta được bộ phân loại Bayes đơn giản (có đầu ra ký hiệu là c_{NB}) như sau:

$$c_{NB} = \arg\max_{c_j \in C} P(c_j) \prod_i P(x_i | c_j) \qquad (2.4)$$

Trong đó $P(x_i | c_j)$ được tính từ dữ liệu huấn luyện bằng số lần x_i xuất hiện cùng với c_j chia cho số lần x_i xuất hiện. Việc tính xác suất này đòi hỏi ít dữ liệu hơn nhiều so với tính $P(x_1, x_2,...,x_n | c_j)$.

2.1.2. Thuật toán

Các bước thực hiện thuật toán Naïve Bayes:

Bước 1: Huấn luyện Naïve Bayes(dựa vào tập dữ liệu)

- Tính xác suất $P(C_i)$
- Tính xác suất $P(x_k|C_i)$

Bước 2: X_{new} được gán vào lớp có giá trị lớn nhất theo công thức

$$\max\left(P(C_i) \prod_{k=1}^{n} P(x_k | C_i) \right) \qquad (2.5)$$

Để minh họa thuật toán Bayes một cách đơn giản, ta sử dụng bài toán phân chia ngày thành phù hợp hay không phù

hợp với việc chơi tennis theo điều kiện thời tiết được đưa ra trong bảng 2.1:

Bảng 2.1: Bộ dữ liệu huấn luyện cho bài toán phân loại "Chơi Tennis"

Ngày	Trời	Nhiệt độ	Độ ẩm	Gió	Chơi Tennis
D1	Nắng	Nóng	Cao	Yếu	Không
D2	Nắng	Nóng	Cao	Mạnh	Không
D3	Nhiều mây	Nóng	Cao	Yếu	Có
D4	Mưa	Trung bình	Cao	Yếu	Có
D5	Mưa	Ấm áp	Bình thường	Yếu	Có
D6	Mưa	Lạnh	Bình thường	Mạnh	Không
D7	Nhiều mây	Lạnh	Bình thường	Mạnh	Có
D8	Nắng	Ấm áp	Cao	Yếu	Không
D9	Nắng	Lạnh	Bình thường	Yếu	Có
D10	Mưa	Ấm áp	Bình thường	Yếu	Có
D11	Nắng	Ấm áp	Bình thường	Mạnh	Có
D12	Nhiều mây	Ấm áp	Cao	Mạnh	Có
D13	Nhiều mây	Nóng	Bình thường	Yếu	Có
D14	Mưa	Ấm áp	Cao	Mạnh	Không

Trong đó: có 9 mẫu tích cực (có chơi Tennis) và 5 mẫu tiêu cực (Không chơi Tennis):

- Độ ẩm = Cao có 3 tích cực và 4 tiêu cực.
- Độ ẩm = Bình thường có 6 tích cực và 1 tiêu cực
- Gió = Yếu có 6 tích cực và 2 tiêu cực
- Gió = Mạnh có 3 tích cực và 3 tiêu cực.

Vậy từ các dữ liệu trên bạn hãy xác định xem với các điều kiện <Trời = nắng, Nhiệt độ = trung bình, Độ ẩm = cao, Gió = mạnh> thì người chơi có chơi Tennis không ?

Trả lời:

Bước1:

$$P(\text{Chơi Tennis} = \text{Có}) = \frac{9}{14} = 0.64$$

$$P(\text{Chơi Tennis} = \text{Không}) = \frac{5}{14} = 0.36$$

$$P(\text{Gió} = \text{Mạnh} | \text{Chơi Tennis} = \text{Có}) = \frac{3}{9} = 0.33$$

$$P(\text{Gió} = \text{Mạnh} | \text{Chơi Tennis} = \text{Không}) = \frac{3}{5} = 0.6$$

Bước 2: Tính xác suất

$R_{Có}$ = P(Có) × P(Trời = Nắng |Có) × P(Nhiệt độ = Lạnh |Có) × P(Độ ẩm = Cao | Có) × P(Gió = Mạnh | Có) = 0.005

$R_{Không}$ = P(Không) × P(Trời = Nắng | Không) × P(Nhiệt độ = Lạnh | Không) × P(Độ ẩm = Cao | Không) × P(Gió = Mạnh | Không) = 0.021

→ Vì 0.021 > 0.005 nên kết luận lại là người chơi KHÔNG chơi Tennis khi có điều kiện thời tiết như trên.

2.1.3. Áp dụng trong phân loại thư rác

Với phương pháp phân loại Bayes đơn giản, mỗi thư (phần nội dung) được biểu diễn bởi một vector $\vec{x} = (x_1, x_2, \ldots, x_n)$, trong đó $x_1, x_2, \ldots, x_n$ là giá trị của đặc trưng $X_1, X_2, \ldots, X_n$.

Mỗi đặc trưng có thể là một từ hoặc một cụm từ. Ở đây, n là số lượng đặc trưng được xác định từ toàn bộ tập dữ liệu huấn luyện, tức là số lượng từ/cụm từ khác nhau trong tập dữ liệu huấn luyện. Mỗi thư được gán một nhãn phân loại Y có thể nhận một trong hai giá trị: $Y = 1$ cho trường hợp thư rác và $Y = 0$ cho trường hợp thư bình thường.

Để xác định nhãn phân loại cho thư, bộ phân loại Bayes tính xác suất điều kiện

$$P(Y = y \mid X_1 = x_1, \ldots, X_n = x_n) \qquad (2.9)$$

tức là xác suất một thư với nội dung $(x_1, x_2, \ldots, x_n)$ nhận nhãn phân loại y, $y \in \{1,0\}$. Sử dụng công thức Bayes, xác suất trên được tính như sau:

$$P(Y = y \mid X_1 = x_1, \ldots, X_n = x_n) = \frac{P(X_1 = x_1, \ldots, X_n = x_n \mid Y = y) \cdot P(Y = y)}{P(X_1 = x_1, \ldots, X_n = x_n)} \qquad (2.10)$$

Trong công thức (2.10), giá trị mẫu số không phụ thuộc vào nhãn phân loại và do vậy có thể bỏ qua. Nhãn phân loại Y là nhãn tương ứng với giá trị lớn nhất của tử số. Cụ thể, trong trường hợp phân loại thư rác, nhãn của thư được xác định bằng cách tính giá trị biểu thức:

$$\frac{P(Y = 1 \mid X_1 = x_1, \ldots, X_n = x_n)}{P(Y = 0 \mid X_1 = x_1, \ldots, X_n = x_n)} = \frac{P(X_1 = x_1, \ldots, X_n = x_n \mid Y = 1) \cdot P(Y = 1)}{P(X_1 = x_1, \ldots, X_n = x_n \mid Y = 0) \cdot P(Y = 0)} \qquad (2.11)$$

Giá trị biểu thức (2.11) lớn hơn 1 có nghĩa xác suất thư là thư rác lớn hơn xác suất thư bình thường và thư sẽ được gán nhãn thư rác. Giá trị biểu thức (2.11) nhỏ hơn 1 cho kết quả ngược lại.

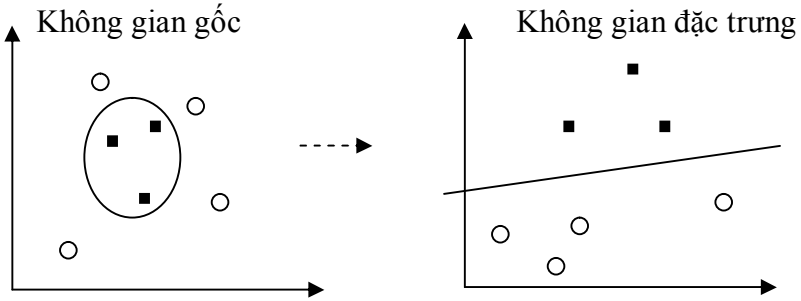
2.2. Thuật toán SVM

2.2.1. Mô tả thuật toán

Xét bài toán phân loại đơn giản nhất - phân loại hai phân lớp với tập dữ liệu huấn luyện bao gồm *n* mẫu được cho dưới dạng $< \vec{x}_i, y_i >$, *i=1,....n*. Trong đó, $\vec{x}_i \in \Re^m$ là vector bao gồm *m* phần tử chứa giá trị của *m* thuộc tính hay đặc trưng và y_i là nhãn phân loại có thể nhận giá trị +1 (tương ứng với các mẫu x_i thuộc lĩnh vực quan tâm) hoặc -1 (tương ứng các mẫu x_i không thuộc lĩnh vực quan tâm).

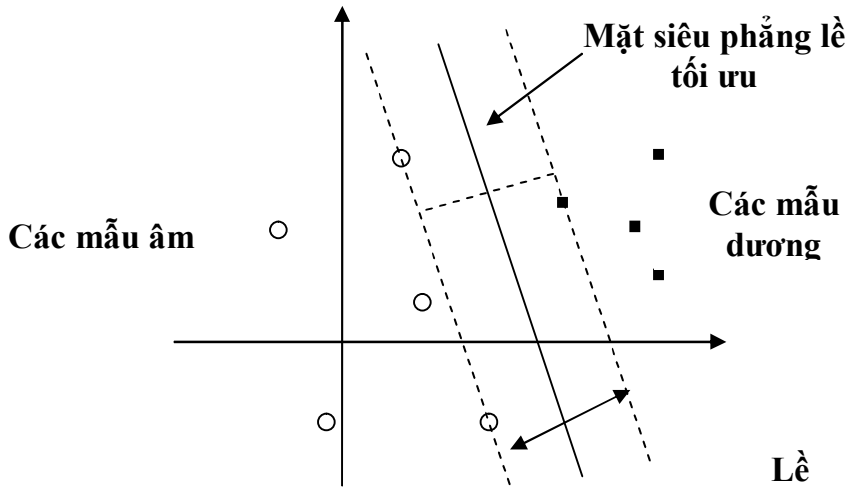
Có thể hình dung dữ liệu như các điểm trong không gian ơclit *m* chiều và được gán nhãn. SVM được xây dựng trên cơ sở hai ý tưởng chính.

Ý tưởng thứ nhất là ánh xạ dữ liệu gốc sang một không gian mới gọi là *không gian đặc trưng* với số chiều lớn hơn sao cho trong không gian mới có thể xây dựng một siêu phẳng cho phép phân chia dữ liệu thành hai phần riêng biệt, mỗi phần bao gồm các điểm có cùng nhãn phân loại. Ý tưởng ánh xạ sang không gian đặc trưng được minh hoạ trên hình 2.1.



Hình 2.1: Ánh xạ dữ liệu từ không gian gốc sang không gian đặc trưng cho phép phân chia dữ liệu bởi siêu phẳng

Ý tưởng thứ hai là trong số những siêu phẳng như vậy cần lựa chọn siêu phẳng có lề lớn nhất. Lề ở đây là khoảng cách từ siêu phẳng tới các điểm gần nhất nằm ở hai phía của siêu phẳng (mỗi phía tương ứng với một nhãn phân loại). Lưu ý rằng siêu phẳng nằm cách đều các điểm gần nhất với nhãn khác nhau. Trên hình 2.2. là minh hoạ siêu phẳng (đường liền nét) với lề cực đại tới các điểm dữ liệu biểu diễn bởi các hình tròn và hình vuông.



Hình 2.2: Siêu phẳng với lề cực đại cho phép phân chia các hình vuông khỏi các hình tròn trong không gian đặc trưng

Để tránh việc tính toán trực tiếp với dữ liệu trong không gian mới, ta sử dụng một phương pháp gọi là *thủ thuật nhân* bằng cách tìm một *hàm nhân* (kernel function) *K* sao cho:

$$K(\vec{a},\vec{b}) = <\vec{a},\vec{b}> \qquad (2.19)$$

Sử dụng phương pháp nhân tử Lagrăng và thay thế tích vô hướng của hai vector bằng giá trị hàm nhân theo công thức (2.19), bài toán tìm lề cực đại của SVM được đưa về bài toán quy hoạch toán học bậc hai như sau:

Tìm vector hệ số $\vec{\alpha} = (\alpha_1, \alpha_2, ..., \alpha_n)$ cho phép cực tiểu hoá hàm mục tiêu

$$W(\vec{\alpha}) = -\frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n} y_i y_j \alpha_i \alpha_j K(\vec{x}_i, \vec{x}_j) + \sum_{i=1}^{n}\alpha_i \qquad (2.20)$$

đồng thời thoả mãn các điều kiện

$$\sum_{i=1}^{n} y_i \alpha_i = 0 \qquad (2.21)$$

Và $0 \le \alpha_i \le C$

Trong (2.20), (2.21), (2.22), $\vec{x}_i$ và y_i tương ứng là dữ liệu và nhãn phân loại của ví dụ huấn luyện thứ i, α_i là hệ số cần xác định. Trong ràng buộc (2.22), C là số lượng tối đa các điểm dữ liệu có phân loại sai, tức là các điểm nằm ở phía này của siêu phẳng nhưng lại có nhãn của các điểm nằm ở bên kia. Việc sử dụng C cho phép khắc phục tình trạng dữ liệu huấn luyện có các ví dụ bị gán nhãn không chính xác.

Sau khi huấn luyện xong, giá trị nhãn phân loại cho một ví dụ mới $\vec{x}$ sẽ được tính bởi

$$f(\vec{x}) = sign(\sum_{i=1}^{n} y_i \alpha_i K(\vec{x}_i, \vec{x}) + b) \qquad (2.23)$$

Ở đây, b được tính trong giai đoạn huấn luyện theo công thức sau

$$b = y_i - \sum_{j=1}^{n} y_j \alpha_j K(\vec{x}_i, \vec{x}_j) \qquad (2.24)$$

trong đó i là một hệ số thoả mãn điều kiện $0<\alpha<C$.

2.2.2. 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.

2.2.3. Áp dụng SVM trong phân loại thư rác

Đối với bài toán phân loại rác, giống như phần phân loại Bayes (mục 2.1.3), thuật toán SVM xem mỗi vector $\vec{x}_i$ là một vector đặc trưng biểu diễn cho nội dung thư và y_i là nhãn phân loại đối với dữ liệu huấn luyện. Tương tự như phần phân loại Bayes, giá trị x_i có thể là 0 hoặc 1.

Thư mới $\vec{x}$ được phân loại theo công thức (2.23):

$$f(\vec{x}) = sign(\sum_{i=1}^{n} y_i \alpha_i K(\vec{x}_i, \vec{x}) + b)$$

Nếu $f(\vec{x}) \geq 0$ thì thư đó thuộc lớp +1 là thư rác; ngược lại $f(\vec{x}) < 0$ thì thư đó thuộc lớp -1 tương ứng với thư bình thường.

2.3. Xây dựng mô hình lọc thư rác dựa trên học máy có giám sát

2.3.1. Lựa chọn mô hình và thuật toán

Theo nhiều báo cáo, thuật toán Support Vector Machine cho kết quả phân loại tốt và ổn định. Trong khi đó, thuật toán Naïve Bayes đơn giản và cũng cho kết quả phân loại tương đối tốt với chi phí thấp.

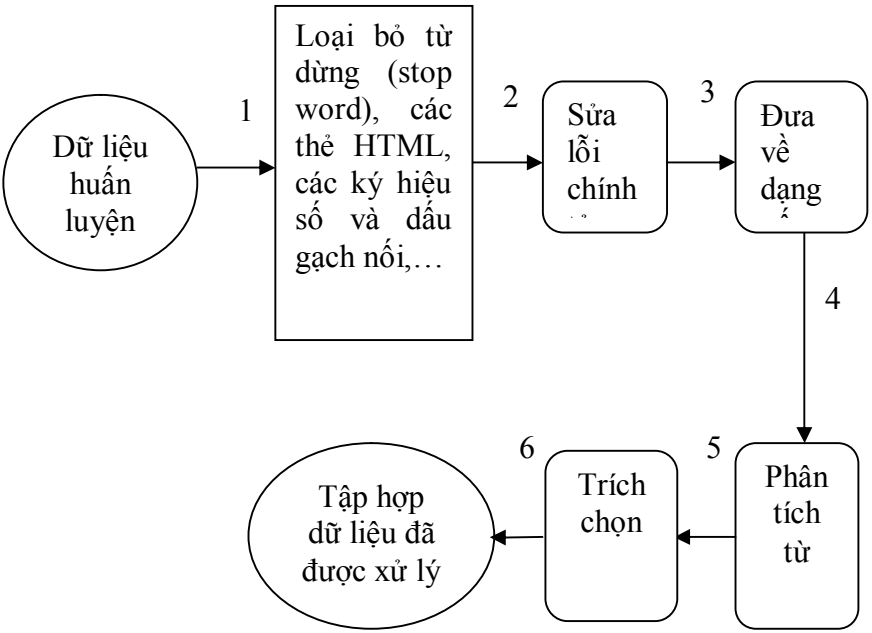
Mô hình sẽ bao gồm 3 bước chính:

- Tiền xử lý dữ liệu
- Huấn luyện dữ liệu
- Thử nghiệm đánh giá độ chính xác của mô hình học máy

2.3.2. Xây dựng hệ thống

2.3.2.1 Tiền xử lí dữ liệu

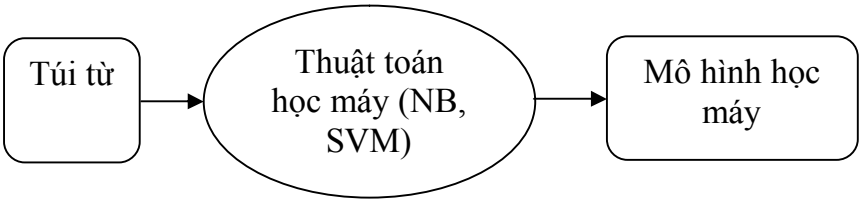
Phần tiền xử lí dữ liệu được coi là một trong những phần quan trọng nhất trong phân loại thư rác. Tuy nhiên, cho đến này vẫn chưa đưa ra được phương pháp tiếp cận có hiệu quả nhất vì nhiều lý do. Nhưng có lẽ lý do quan trọng nhất là độ phức tạp, tính linh hoạt của ngôn ngữ. Ví dụ: các từ động âm, các cụm động từ, các thành ngữ ... phong thái ngôn ngữ khác nhau của từng vùng miền.



Hình 2.3: Tiền xử lý dữ liệu

2.3.2.2. Huấn luyện dữ liệu

Đầu vào của bước này là các túi từ được đưa ra từ bước tiền xử lí. Kết quả của bước này là đưa ra mô hình học máy phù hợp với tập dữ liệu đầu vào.



Hình 2.4: Huấn luyện dữ liệu

Hai phương pháp phân loại được thử nghiệm bao gồm hai phiên bản phân loại Bayes đơn giản – phiên bản sử dụng mô hình đa thức (Bayes đa thức) – và SVM.

CHƯƠNG 3: CÀI ĐẶT, THỬ NGHIỆM VÀ ĐÁNH GIÁ KẾT QUẢ

3.1. Bộ dữ liệu thử nghiệm

Dữ liệu thử nghiệm trong luận văn gồm có hai tập dữ liệu: LingSpam, PU1 được trình bày trong bảng 3.1.

Bảng 3.1: Bộ dữ liệu thử nghiệm

Tập dữ liệu	Thư rác	Thư bình thường	Tổng số thư
PU1	481	618	1099
LingSpam	481	2412	2893

3.2. Cài đặt thử nghiệm và kết quả

Như đã đề cập ở chương 2, luận văn tập trung vào cài đặt thử nghiệm hai phương pháp phân loại gồm phân loại Naïve Bayes đơn giản và phương pháp Support Vector Machine (SVM). Để thử nghiệm các phương pháp này, luận văn sử dụng bộ công cụ WEKA có tại địa chỉ http://www.cs.waikato.ac.nz/ml/weka/. Đối với SVM, hàm nhân sử dụng là hàm tuyến tính tức là việc phân loại được tiến hành trong không gian gốc của dữ liệu. Tham số C trong công thức (2.22) được đặt bằng 1.

Hiệu quả lọc thư được đánh giá theo nhiều tiêu chí như *độ nhậy* (recall), *độ chính xác* (precision), và *độ chính xác phân loại chung* tức là phần trăm thư được phân loại đúng

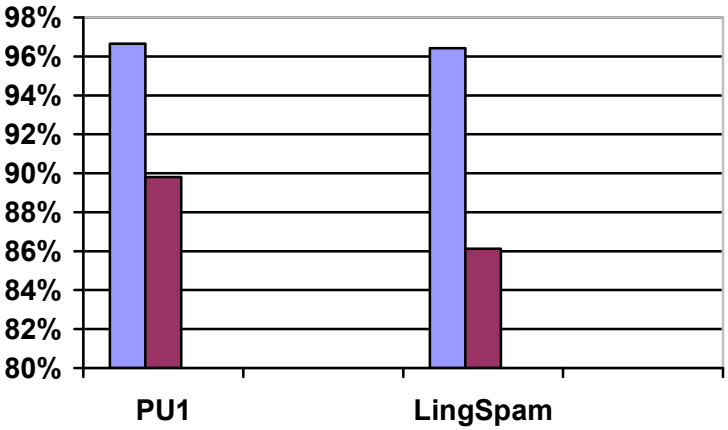
không phụ thuộc vào đó là thư rác hay thư bình thường. Trong luận văn, tôi chủ yếu tập trung đánh giá hiệu quả lọc thư qua tiêu chí về *độ chính xác* (precision) được định nghĩa như sau:

$$\textit{độ chính xác} = \frac{\textit{số thư rác phát hiện chính xác}}{\textit{Tổng số thư được phân loại là thư rác}}$$

Kết quả thực nghiệm của hai phương pháp Naïve Bayes và SVM với tập dữ liệu mẫu được thể hiện trong hình 3.1 và chi tiết ở bảng 3.2.

Bảng 3.2: Độ chính xác phân loại với hai phương pháp phân loại khác nhau

Tập dữ liệu	NB	SVM
PU1	89.81 %	96.65 %
LingSpam	86.12 %	96.42 %



Hình 3.1: Độ chính xác phân loại của NB và SVM

3.3. Đánh giá kết quả

Theo kết quả thực nghiệm cho thấy phương pháp Naïve Bayes cho kết quả kém hơn so với phương pháp SVM. Tuy nhiên, phương pháp Bayes có ưu thế rõ rệt về tốc độ phân loại do có độ phức tạp tính toán thấp hơn trong khi SVM đòi hỏi khối lượng và thời gian tính toán lớn hơn nhiều. Trong các thử nghiệm, tổng thời gian huấn luyện và phân loại bằng SVM lớn hơn Bayes đơn giản từ 10 tới 50 lần.

KẾT LUẬN

Với mục tiêu nghiên cứu, xây dựng mô hình lọc thư rác có hiệu quả, luận văn đã đi sâu nghiên cứu hai thuật toán học máy có giám sát, bao gồm Naïve Bayes và SVM và áp dụng thử nghiệm trong bài toán lọc thư rác. Những kết quả chính đã đạt được trong luận văn:

1) Khái quát được một số vấn đề về học máy, học máy có giám sát bao gồm ứng dụng và một số thuật toán học máy áp dụng vào bài toán phân loại, trong đó chú trọng các phương pháp học máy có giám sát. Ngoài ra, luận văn cũng giới thiệu được tổng quan về thư rác, đặc trưng của thư rác, từ đó xây dựng bài toán lọc thư rác.

2) Nghiên cứu hai thuật toán phân loại học máy có giám sát là Naïve Bayes và SVM; từ đó đưa ra bài toán áp dụng vào phân loại thư rác.

3) Xây dựng mô hình, cài đặt thực nghiệm và đánh giá kết quả lọc thư rác dựa trên các thuật toán học máy có giám sát. Kết quả thực nghiệm khẳng định thuật toán Naïve Bayes cho kết quả phân loại tương đối tốt, đơn giản, dễ cài đặt và đặc biệt là chi phí tính toán không cao; thuật toán SVM cho kết quả phân loại tốt hơn nhưng đòi hỏi chi phí tính toán cho huấn luyện và phân loại cao hơn nhiều so với Naïve Bayes.

Các kết quả nghiên cứu trên có thể sử dụng làm cơ sở cho việc xây dựng những hệ thống lọc thư rác thương mại sử dụng cho các mail server tại Việt Nam.

Tuy nhiên, do còn hạn chế về mặt thời gian và kiến thức nên luận văn chưa đi sâu vào nghiên cứu bài toán lọc thư rác tiếng Việt. Trong tương lai, luận văn có thể sẽ được nghiên cứu tiếp theo hướng sau:

Khi áp dụng những thuật toán phân loại một khó khăn gặp phải là xây dựng được tập hợp từ vựng và các mẫu huấn luyện đủ lớn. Vấn đề này liên quan tới việc phân tách một câu thành các từ và cụm từ một cách chính xác. Luận văn có thể được tiếp tục phát triển theo hướng nghiên cứu mở rộng ứng dụng các bộ từ điển sẵn có và xây dựng các mẫu huấn luyện tiêu chuẩn về thư tiếng Việt bao gồm có dấu và không có dấu cũng như điều chỉnh các tham số của giải thuật phân loại để nâng cao độ chính xác.