HỌC VIỆN CÔNG NGHỆ BƯU CHÍNH VIỄN THÔNG

NGUYỄN THỊ PHƯƠNG THÚY

PHÂN LOẠI VĂN BẢN VÀ ỨNG DỤNG VÀO PHÂN LOẠI TIN TỨC ĐIỆN TỬ

Chuyên ngành: KHOA HỌC MÁY TÍNH

Mã số: 60.48.01.01

TÓM TẮT LUẬN VĂN THẠC SĨ

HÀ NỘI - 2014

Luận văn được hoàn thành tại:

HỌC VIỆN CÔNG NGHỆ BƯU CHÍNH VIỄN THÔNG

Người hướng dẫn khoa học: PGS.TS Từ Minh Phương

Phản biện 1: TS. Nguyễn Phương Thái

Phản biện 2: PGS.TS Đỗ Trung Tuấn

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: 9 giờ 00 ngày 15 tháng 02 năm 2014

Có thể tìm hiểu luận văn tại:

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

LỜI MỞ ĐẦU

Hiện nay, song song với sự phát triển mạnh mẽ của khoa học kỹ thuật, nhu cầu cập nhật thông tin của con người ngày càng nâng cao, báo điện tử ra đời nhằm cung cấp thông tin nhanh, chính xác, đầy đủ, được cập nhật kịp thời cùng cách trình bày thu hút. Với báo điện tử, độc giả có thể truy cập được tin tức ở bất kỳ đâu không phụ thuộc vào môi trường làm việc miễn là máy tính của họ có kết nối Internet và có cài đặt một trình duyệt web tuân thủ tiêu chuẩn.

Báo tự động cập nhật tin tức là loại báo điện tử có khả năng tổng hợp các tin tức mới, cập nhật từ nhiều nguồn báo điện tử, sau đó phân loại, tổ chức, sắp xếp tin tức theo. Báo giúp người đọc và tìm kiếm tin tức theo cách hoàn toàn mới. Mỗi mẩu tin được hiển thị kèm với các nguồn tin khác nhau đưa cùng tin hoặc tin tương tự. Ngoài ra, báo giúp bạn tiếp cận các báo điện tử một cách hiệu quả nhất và báo rất tiện lợi và tiết kiệm thời gian hơn khi đọc tin tức.

Tuy nhiên, mỗi ngày mỗi báo điện tử cung cấp hàng trăm tin tức và số lượng báo điện tử cũng rất lớn, vấn đề đặt ra là làm sao các trang báo điện tử tự động có thể phân loại được tin tức với số lượng lớn và từ nhiều nguồn khác nhau đó vào các chủ đề tương ứng mà vẫn đảm bảo tính chất "nhanh, cập nhật kịp thời" của báo điện tử? Việc phân loại này không thể thực hiện bởi bàn tay con người vì số lượng tin tức lớn, dẫn đến cần nhiều nhân lực, gây tốn kém và có thể phân loại không chính xác. Do vậy, cần một giải pháp phân loại tin tức tự động, để có thể phân loại chính xác và nhanh chóng.

Xuất phát từ ý tưởng này, tôi đã chọn đề tài "Phân loại văn bản và ứng dụng vào phân loại tin tức điện tử" làm đề tài luận văn thạc sĩ của mình.

Luận văn gồm 3 chương chính với các nội dung như sau:

Chương 1: Tổng quan về phân loại văn bản và bài toán phân loại tin tức điện tử

Chương 1 nêu tổng quan về phân loại văn bản, vai trò và ứng dụng của phân loại văn bản hiện nay, từ đó nêu ra bài toán phân loại tin tức điện tử. Sau đó, giới thiệu tổng quan về các kỹ thuật trích chọn đặc trưng trong văn bản và các phương pháp hiện tại đang được áp dụng để phân loại.

Chương 2: Trích chọn đặc trưng và phân loại văn bản với Naive Bayes và SVM

Chương 2 nêu đặc điểm của tin tức điện tử và tập trung nghiên cứu 2 vấn đề chính của phân loại văn bản là trích chọn đặc trưng văn bản và phân loại văn bản mới (cụ thể

trong luận văn, văn bản đó là tin tức điện tử). Luận văn lựa chọn 2 phương pháp là Naïve Bayes và SVM để phân loại một văn bản mới, trong chương này sẽ trình bày chi tiết cơ sở lý thuyết và phương thức phân loại của 2 phương pháp đã được lựa chọn.

Chương 3: Thử nghiệm và đánh giá

Chương 3 trình bày mô hình phân loại mà luận văn đã đề xuất ở chương 2 và cách thức cài đặt mô hình này. Tiếp theo là thử nghiệm 2 bộ phân loại Naïve Bayes và SVM trên tập dữ liệu tin tức điện tử đã thu thập và cây phân lớp đã xây dựng được. Cuối cùng thực hiện đánh giá và so sánh kết quả thử nghiệm của 2 bộ phân loại.

CHƯƠNG 1 – TỔNG QUAN VỀ PHÂN LOẠI VĂN BẢN VÀ BÀI TOÁN PHÂN LOẠI TIN TỨC

1.1 Tổng quan về phân loại văn bản

1.1.1 Khái niệm phân loại văn bản

Phân loại văn bản là quá trình gán nhãn (tên lớp/nhãn lớp) các văn bản ngôn ngữ tự nhiên vào một hay nhiều lớp cho trước.

1.1.2 Phân loại bài toán phân lớp văn bản

1.2 Phân loại tin tức báo điện tử

1.2.1 Báo điện tử

1.2.2 Phân loại tin tức báo điện tử

Bài toán phân loại tin tức điện tử được phát biểu như sau:

Gọi X là tập các tin tức cần phân loại và Y là tập các chủ đề có thể được gán cho các các tin tức. Khi đó ta cần phải chỉ ra một tin tức $x \in X$ thuộc vào chủ đề $y \in Y$ nào. Trong đó, x bao gồm các từ, cụm từ, câu được dùng cho nhiệm vụ phân loại.

1.3 Tiền xử lý và trích chọn đặc trưng

1.4 Các phương pháp phân loại văn bản

1.4.1 Phương pháp K-Nearest Neighbor (kNN)

1.4.2 Phương pháp Naïve Bayes

1.4.3 Phương pháp SVM

1.4.4 Phương pháp cây quyết định

1.4.5 Phương pháp sử dụng mạng Nơron

1.4.6 So sánh các phương pháp phân loại văn bản

Phương pháp Naïve Bayes và SVM thích hợp trong việc phân loại văn bản với dữ liệu lớn một cách nhanh chóng và hiệu quả. Đây là lý do mà luận văn chọn thuật toán Naïve Bayes và SVM để nghiên cứu giải quyết bài toán phân loại tin tức điện tử.

1.5 Kết luận

Chương 1 đã trình bày tổng quan về bài toán phân loại văn bản và phát biểu ứng dụng của phân loại văn bản đó là bài toán phân loại tin tức điện tử. Sau khi tìm hiểu về các

phương pháp phân loại khác nhau, trong chương 1, luận văn đã nêu lên lý do chọn hai phương pháp Naïve Bayes và SVM để nghiên cứu.

CHƯƠNG 2 – TRÍCH CHỌN ĐẶC TRƯNG VÀ PHÂN LOẠI VĂN BẢN VỚI NAÏVE BAYES VÀ SVM

2.1 Đặc điểm của tin tức điện tử

2.2 Tiền xử lý

2.2.1 Lọc nhiễu

2.2.2 Loại bỏ stop-word

2.2.3 Cây phân lớp

2.3 Xây dựng đặc trưng

2.3.1 Lựa chọn đặc trưng

2.3.2 Đánh trọng số cho từng đặc trưng

2.4 Phương pháp phân loại Naïve Bayes

2.2.1 Lý thuyết xác suất Bayes

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:

$$y = c_{MAP} = \arg\max_{c_j \in C} P(c_j | x_1, x_2, \ldots, x_n) \qquad (2.7)$$

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

$$\begin{aligned} c_{MAP} &= \arg\max_{c_j \in C} \frac{P(x_1, x_2, \ldots, x_n | c_j) P(c_j)}{P(x_1, x_2, \ldots, x_n)} \\ &= \arg\max_{c_j \in C} P(x_1, x_2, \ldots, x_n | c_j) P(c_j) \end{aligned} \qquad (2.8)$$

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 $P(x_1, x_2,...,x_n|c_j)$ khó khăn hơn nhiều.

Để tính giá trị 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 $P(x_1, x_2,...,x_n|c_j)$ được viết lại như sau:

$$P(x_1, x_2,...,x_n | c_j) = P(x_1 | c_j)\, P(x_2 | c_j) \ldots P(x_n | c_j) \qquad (2.9)$$

Tức là xác suất đồng thời quan sát thấy các thuộc tính bằng tích xác suất điều kiện của từng thuộc tính riêng lẻ. Thay vào biểu thức (2.8) ta được bộ phân loại Bayes đơn giản (có đầu ra ký hiệu là c_{NB}) như sau:

$$c_{NB} = \arg\max_{c_j \in C} P(c_j) \prod_i P(x_i \mid c_j) \qquad (2.10)$$

2.2.4 *Phân loại văn bản dựa trên Naïve Bayes*

Để sử dụng phân loại Bayes đơn giản, mỗi nội dung tin tức được biểu diễn bởi một vectơ $\vec{x} = (x_1, x_2, \ldots, x_n)$, 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 tin tức được gán một nhãn phân loại $Y = \{y_1, y_2, \ldots, y_m\}$.

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

$$P(Y = y \mid X_1 = x_1, \ldots, X_n = x_n) \qquad (2.11)$$

tức là xác suất một tin tức với nội dung $(x_1, x_2, \ldots, x_n)$ nhận nhãn phân loại y, $y \in \{y_1, y_2, \ldots, y_m\}$. Sử dụng công thức Bayes, xác suất trên được tính như sau:

$$P(Y = y \mid X_1 = x_1, \ldots, X_n = x_n) = \frac{P(X_1 = x_1, \ldots, X_n = x_n \mid Y = y) \cdot P(Y = y)}{P(X_1 = x_1, \ldots, X_n = x_n)} \qquad (2.12)$$

Trong công thức (2.12), 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 tin tức điện tử, nhãn của tin tức được xác định bằng cách tính giá trị biểu thức:

$$Y_{MAP} = \arg\max_{y_j \in Y} \frac{P(x_1, x_2, \ldots, x_n \mid y_j) P(y_j)}{P(x_1, x_2, \ldots, x_n)} = \arg\max_{y_j \in Y} P(x_1, x_2, \ldots, x_n \mid y_j) P(y_j)$$

(2.13)

Xác suất $P(Y = y)$ trên tập dữ liệu huấn luyện có thể tính dễ dàng bằng cách đếm tần suất xuất hiện của tin tức có nhãn y. Việc xác định $P(\vec{X} = \vec{x} \mid Y = y)$ phức tạp hơn nhiều do phải tính tất cả các tổ hợp giá trị của vectơ $\vec{X}$ và đòi hỏi lượng dữ liệu huấn luyện lớn tương ứng. Có một số cách tính giá trị $P(\vec{X} = \vec{x} \mid Y = y)$ khác nhau tương ứng với các phiên bản khác nhau của phương pháp phân loại văn bản sử dụng Bayes đơn giản. Trong nghiên cứu này, luận văn sẽ tìm hiểu hai phiên bản thông dụng nhất: *Bayes đơn giản với mô hình Bécnuli đa trị* (multivariate Bernoulli naïve Bayes) và *Bayes đơn giản với mô hình đa thức* (multinomial naïve Bayes).

Phân loại Bayes đơn giản với mô hình Bécnuli đa trị

$$P(X_1 = x_1,..., X_n = x_n \mid Y = y) = \prod_{i=1}^{n} P(f_i \mid Y = y)^{x_i}.(1 - P(f_i \mid Y = y))^{(1-x_i)} \quad (2.14)$$

Trong đó, xác suất $P(f_i| Y = y)$ là tỷ lệ tin tức với nhãn y đồng thời có chứa f_i trong số tin tức có nhãn y. Tỷ lệ này được tính trên tập dữ liệu huấn luyện.

Xác suất $P(f_i \mid Y = y)$ được tính như sau:

$$P(f_i \mid Y = y) = \frac{N_{y,f_i} + 1}{N_y + 2} \quad (2.15)$$

Phân loại Bayes đơn giản với mô hình đa thức

$$P(X_1 = x_1,..., X_n = x_n \mid Y = y) = P(|d|).|d|!.\prod_{i=1}^{n} \frac{P(f_i \mid Y = y)^{x_i}}{x_i!} \quad (2.16)$$

Xác suất $P(f_i \mid Y = y)$ được tính từ dữ liệu huấn luyện theo công thức

$$P(f_i \mid Y = y) = \frac{N_{y,f_i} + 1}{N_y + n} \quad (2.17)$$

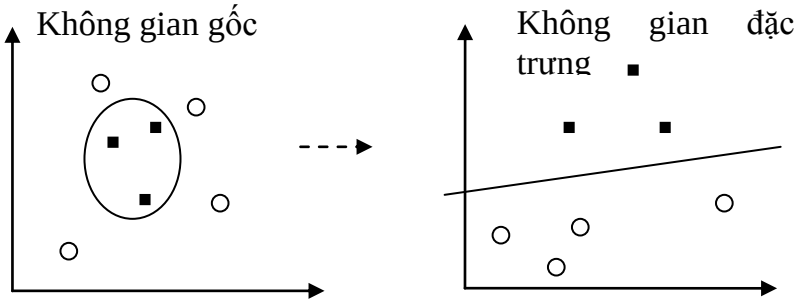
2.5 Phương pháp phân loại SVM

2.5.1 Ý tưởng của SVM

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 véctơ 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).

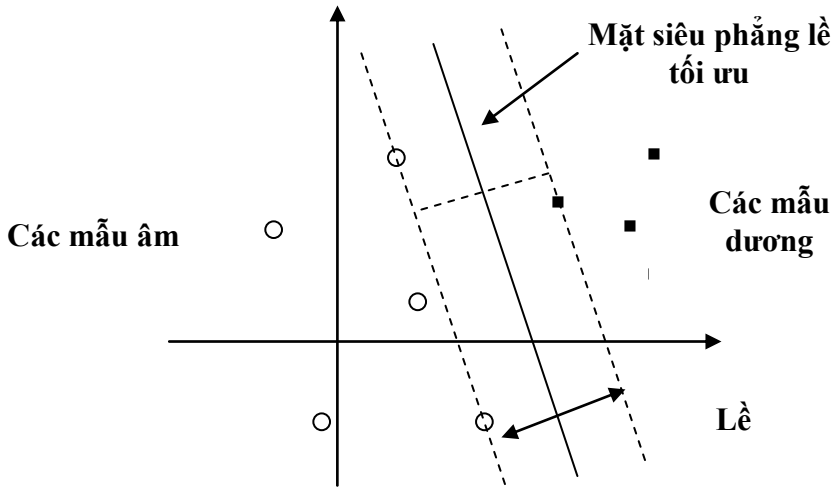
Có thể hình dung dữ liệu như các điểm trong không gian ơclit m chiều và được gán nhãn. SVM được xây dựng trên cơ sở hai ý tưởng chính.

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 2.2.



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 2.3 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:

$$K(\vec{a},\vec{b}) = <\vec{a},\vec{b}> \tag{2.18}$$

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

$$W(\vec{\alpha}) = -\frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n} y_i y_j \alpha_i \alpha_j K(\vec{x}_i, \vec{x}_j) + \sum_{i=1}^{n} \alpha_i \qquad (2.19)$$

(2.19)

đồng thời thoả mãn các điều kiện:

$$\begin{cases} \sum_{i=1}^{n} y_i \alpha_i = 0 \\ 0 \le \alpha_i \le C \end{cases} \qquad (2.20)$$

Trong (2.18, (2.19), (2.20), $\vec{x}_i$ 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 (2.20), *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.

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.

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

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

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

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

Trong đó, *i* là một hệ số thoả mãn điều kiện $0 < \alpha_i < C$.

2.6 Kết luận chương

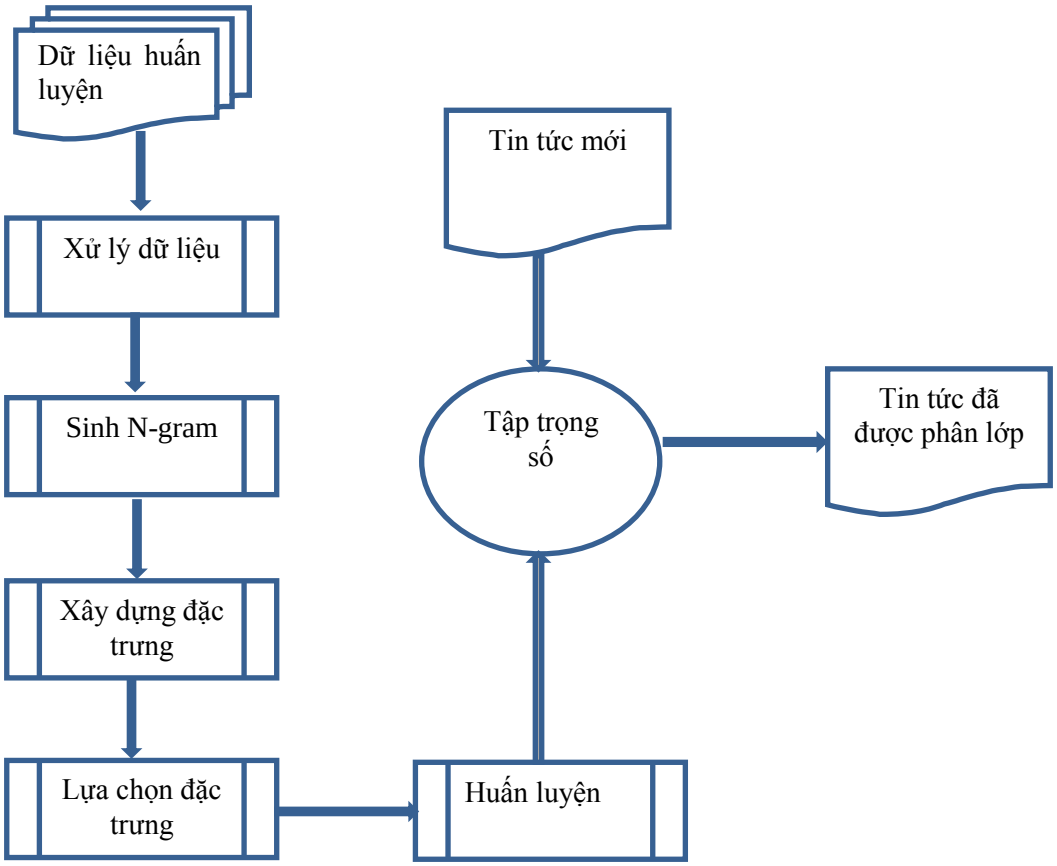
Chương 2 đã trình bày cụ thể các bước trong tiến trình phân loại tin tức điện tử. Trong đó tập trung nghiên cứu về 2 thuật toán đó là Naïve Bayes và SVM để hiểu rõ việc thực hiện huấn luyện và phân loại văn bản mới của các bộ phân loại.

CHƯƠNG 3: THỬ NGHIỆM VÀ ĐÁNH GIÁ

3.1 Mở đầu

Chương 3 sẽ trình bày mô hình phân loại để giải quyết bài toán phân loại tin tức điện tử tiếng Việt sử dụng 2 bộ phân loại Naïve Bayes và SVM đã đề xuất trong chương 2. Tiếp theo là thử nghiệm 2 bộ phân loại Naïve Bayes và SVM trên tập dữ liệu tin tức điện tử đã thu thập được từ trang báo http://vnexpress.net/. Trong phần cuối của chương, luận văn thực hiện áp dụng phương pháp phân loại Naïve Bayes đa thức để phân lớp dữ liệu mới đưa vào.

3.2 Mô hình phân loại tin tức điện tử



3.3 Đánh giá bộ phân lớp

3.2.1 Các độ đo

Các độ đo sẽ được sử dụng để đánh giá đó là độ chính xác, độ nhậy, fmeasure.

3.3.2 Phương pháp ước lượng chéo trên k tập con

3.4 Thử nghiệm và đánh giá kết quả phân loại

3.4.1 Dữ liệu thử nghiệm

Dữ liệu được sử dụng trong huấn luyện và kiểm thử là những bài báo được lọc ra từ trang web http://www.vnexpress.net/ bao gồm 7 chủ đề: kinh doanh, pháp luật, thể thao, văn hóa, khoa hoc, công nghệ và xã hội. Mỗi chủ đề tương ứng với một thư mục với tên: kinh-doanh, phap-luat, the-thao, van-hoa, cong-nghe, khoa-hoc và xa-hoi. Dữ liệu được chia làm 2 phần: một phần gồm 3789 file và phần còn lại gồm 1932 file.

3.4.2 Các công cụ hỗ trợ

3.4.3 Tiền xử lý dữ liệu

3.4.4 Huấn luyện

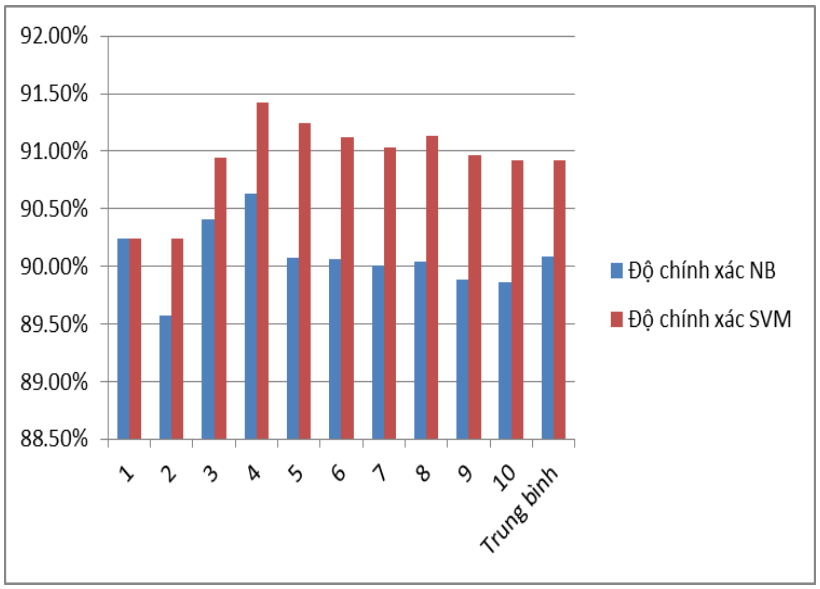
Hai phương pháp phân loại được thử nghiệm bao gồm hai phiên bản phân loại Bayes đơn giản – phiên bản sử dụng mô hình đa thức (Bayes đa thức) – và SVM.

Đối với SVM, hàm nhân được lựa chọn là hàm RBF.

3.4.5 Kết quả thử nghiệm

3.4.5.1 Đánh giá theo cross-validation

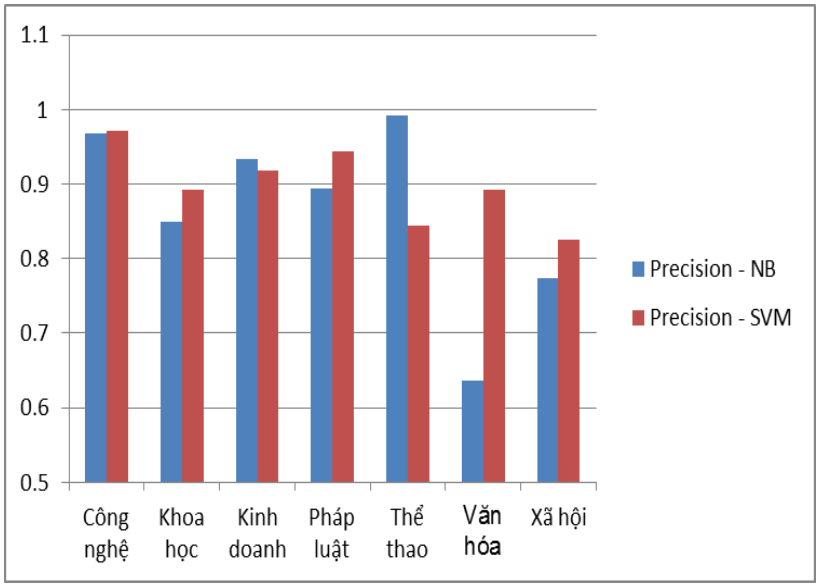
Chia dữ liệu gồm 3789 file thành 10 phần, lấy một phần để kiểm thử và 9 phần còn lại là dữ liệu huấn luyện, sau đó thực hiện đánh giá 2 bộ phân lớp NB và SVM. Thực hiện 10 lần với lần lượt các tập dữ liệu kiểm thử và huấn luyện khác nhau, cuối cùng lấy độ chính xác trung bình sau 10 lần thực hiện đánh giá. Kết quả: Độ chính xác của NB là 90.08% và độ chính xác của SVM là 90.92%



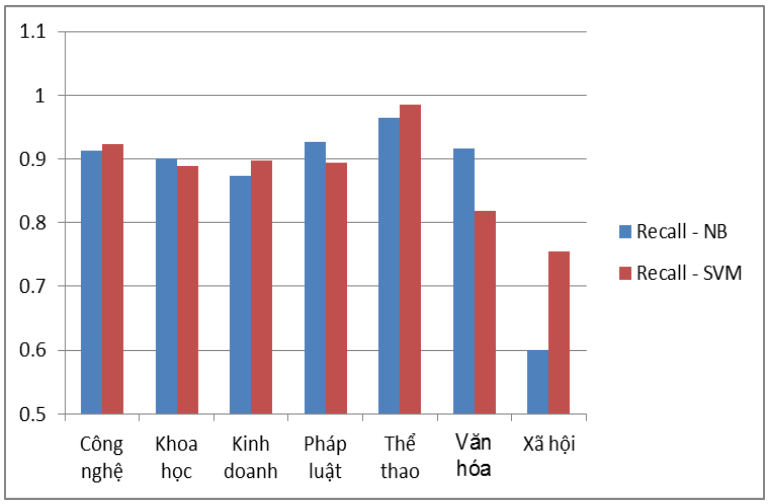
Hình 3.1: Độ chính xác phân loại của Naive Bayes và SVM

3.4.5.2 Đánh giá trên tập dữ liệu kiểm thử mới

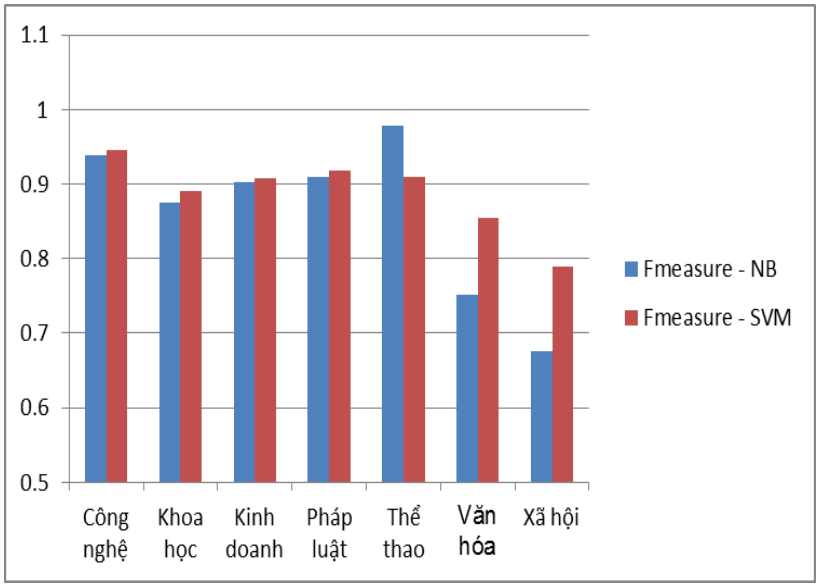
Ở mục trên, luận văn đã đánh giá 2 bộ phân loại trên tập dữ liệu thứ nhất gồm 4642 file, trong phần này, luận văn sẽ sử dụng tập dữ liệu thứ nhất làm tập huấn luyện và sử dung tập dữ liệu thứ hai gồm 2267 file làm tập kiểm thử. Kết quả huấn luyện sẽ so sánh dựa trên 3 tiêu chí: độ chính xác (precision), độ nhậy (recall) và fmeasure. Kết quả được thể hiện trong Bảng 8.



Hình 3.2: So sánh độ chính xác của hai bộ phân loại theo precision trên từng lớp



Hình 3. 3: So sánh hai bô phân loại theo recall trên từng lớp



Hình 3. 4: So sánh hai bộ phân loại theo Fmeasure trên từng lớp

3.4.5.2 Đánh giá kết quả thử nghiệm

Kết quả sau 2 lần thực nghiệm cho thấy phương pháp Naïve Bayes đa thức cho kết quả kém hơn so với phương pháp SVM, nhưng sự chênh lệch không đáng kể (theo mục 3.4.5.1, độ chính xác khi phân loại bằng Naïve Bayes đa thức là 90.8%, trong khi độ chính xác của SVM là 90.9%). Ngoài ra, 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 Naïve Bayes từ 10 tới 50 lần (trong lần đánh giá với tập dữ liệu mới, tổng thời gian huấn luyện và phân loại của Naïve Bayes là khoảng 5 giây, trong khi, SVM thực hiện hết 258 giây).

Do tính chất của tin tức điện tử là nhanh, chính xác và dựa trên kết quả thực nghiệm như trên, luận văn sẽ chọn bộ phân loại Naïve Bayes đa thức để tạo một ứng dụng phân loại tin tức điện tử.

3.5 Phân lớp tin tức điện tử mới

Tin tức điện tử mới sẽ được lấy từ các nguồn khác nhau như http://vietnamnet.vn/, http://dantri.com.vn/..., sau khi qua bộ phân lớp mà luận văn xây dựng sẽ được gán một nhãn tương ứng với nội dung của tin tức điện tử.

Ứng dụng phân loại tin tức điện tử sẽ gồm phần:

- Phần 1: **Huấn luyện dữ liệu**: dữ liệu huấn luyện sẽ được thực hiện tiền xử lý và huấn luyện qua bộ phân loại Naïve Bayes
- Phân 2: **Gán nhãn**: một file tin tức bất kì sẽ được gán một trong các nhãn: Kinh Doanh, Pháp Luật, Thể Thao, Khoa Học, Văn Hóa, Công Nghệ, Xã hội.

3.5.1 Giao diện huấn luyện dữ liệu

3.5.2 Giao diện gán nhãn

3.6 Kết luận chương

Chương 3 đã tiến hành thử nghiệm hai bộ phân loại Naïve Bayes và SVM. Kết quả thực nghiệm đã thể hiện rằng hai phương pháp Naïve Bayes và SVM đều đưa ra kết quả phân loại tương đối cao. Tuy nhiên phân loại tin tức điện tử bằng Naïve Bayes đa thức có độ phức tạp và thời gian tính toán thấp hơn so với SVM. Từ đó, luận văn đã lựa chọn Naïve Bayes để tiến hành cài đặt ứng dụng gán nhãn tin tức điện tử mới.

KẾT LUẬN

Với mục tiêu nghiên cứu, xây dựng mô hình tin tức điện tử có hiệu quả, luận văn đã đi sâu nghiên cứu hai thuật toán phân loại văn bản, bao gồm Naïve Bayes và SVM và áp dụng thử nghiệm trong bài toán phân loại tin tức điện tử. Những kết quả chính đã đạt được trong luận văn như sau:

1) Nghiên cứu tổng quan về phân loại văn bản và bài toán phân loại tin tức điện tử.
2) Nghiên cứu hai thuật toán phân loại là Naïve Bayes và SVM; từ đó đưa ra bài toán áp dụng vào phân loại tin tức điện tử.
3) Xây dựng mô hình, cài đặt thử nghiệm và đánh giá kết quả phân loại tin tức điện tử tiếng Việt dựa trên hai thuật toán đã nghiên cứu. 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. Do đó, luận văn lựa chọn Naïve Bayes làm bộ phân loại cho ứng dụng phân loại tin tức điện tử.

Các kết quả nghiên cứu trên có thể sử dụng làm cơ sở cho việc xây dựng những hệ thống phân loại tin tức điện tử tự động ở trên các website của 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 phân loại tin tức điện tử nhiều nhãn. Trong tương lai, luận văn có thể sẽ được nghiên cứu tiếp theo hướng sau:

Khi thực hiện phân loại tin tức điện tử, cây phân lớp văn bản không chỉ là bẩy lớp như trong luận văn trình bày. Tập các lớp có thể rất nhiều, điều này dẫn đến một tin tức có thể thuộc nhiều lớp khác nhau. Luận văn có thể phát triển theo hướng nghiên cứu mở rộng tập các lớp và nghiên cứu để phân loại tin tức vào nhiều lớp khác nhau.