ĐẠI HỌC THÁI NGUYÊN

ĐẠI HỌC CÔNG NGHỆ THÔNG TIN & TRUYỀN THÔNG

NGUYỄN THỊ THOA

"PHÂN LỚP QUAN ĐIỂM KHÁCH HÀNG VÀ ỨNG DỤNG"

LUẬN VĂN THẠC SỸ

THÁI NGUYÊN – 2016

ĐẠI HỌC THÁI NGUYÊN

ĐẠI HỌC CÔNG NGHỆ THÔNG TIN & TRUYỀN THÔNG

NGUYỄN THỊ THOA

"PHÂN LỚP QUAN ĐIỂM KHÁCH HÀNG VÀ ỨNG DỤNG"

CHUYÊN NGÀNH: KHOA HỌC MÁY TÍNH

MÃ SỐ CHUYÊN NGÀNH: 60.48.01.01

NGƯỜI HƯỚNG DẪN KHOA HỌC

PGS.TS ĐOÀN VĂN BAN

THÁI NGUYÊN - 2016

MỤC LỤC

CHƯƠNG 1 – PHÂN LỚP DỮ LIỆU 3
1.1 Giới thiệu về phân lớp dữ liệu 3
1.2 Quá trình phân lớp dữ liệu 4
1.3 Các vấn đề liên quan đến phân lớp dữ liệu 8
1.3.1 Chuẩn bị dữ liệu cho việc phân lớp 8
1.3.2 So sánh các mô hình phân lớp 9
1.3.3 Các phương pháp đánh giá độ chính xác của mô hình phân lớp 10
1.4 Kết luận chương 1 11
CHƯƠNG 2 – MỘT SỐ KỸ THUẬT TRONG PHÂN LOẠI VĂN BẢN 12
2.1 Xử lý văn bản 12
2.1.1 Đặc điểm của từ trong tiếng việt 12
2.1.2 Tách từ 13
2.2 Biểu diễn văn bản 18
2.2.1 Mô hình logic 18
2.2.2 Mô hình phân tích cú pháp 19
2.2.3 Mô hình không gian vector 20
2.2.4 Mô hình Boolean 22
2.2.5 Mô hình tần suất 23
2.3 Độ tương đồng 25
2.3.1 Khái niệm độ tương đồng 25
2.3.2 Độ tương đồng 26
2.3.3 Các phương pháp tính độ tương đồng 26
2.4 Các phương pháp phân loại văn bản 29

2.4.1 Phương pháp pháp Naïve Bayes (NB) 29
2.4.2 Phương pháp Support Vector Machine (SVM) 31
2.4.3 Phương pháp K-Nearest Neighbor (K-NN) 35
2.4.4 Phương pháp Linear Least Square Fit (LLSF) 37
2.4.5 Phương pháp Centroid – based vector 38
2.4.6 Kết luận 38
2.5 Kết luận chương 2 40
CHƯƠNG 3 – CHƯƠNG TRÌNH THỬ NGHIỆM 41
3.1 Xây dựng mô hình ứng dụng khai phá ý kiến phản hồi của khách hàng trên website dựa trên SVM 41
3.1.1 Phát biểu bài toán 41
3.1.2 Mô hình ứng dụng khai phá ý kiến phản hồi của khách hàng trên website dựa trên SVM 41
3.2 Yêu cầu phần cứng và phần mềm 44
3.2.1 Cấu hình máy thực nghiệm 44
3.2.2 Công cụ và phần mềm sử dụng 44
3.3 Một số kết quả và đánh giá 45
3.3.1 Kết quả thử nghiệm 45
3.3.2 Đánh giá kết quả 56
3.4 Kết luận chương 3 57
KẾT LUẬN VÀ ĐỀ NGHỊ 58

DANH MỤC HÌNH ẢNH

Hình 1.1 Quy trình phân loại văn bản [3] 4
Hình 1.2 Bước xây dựng mô hình phân lớp - Training..... 5
Hình 1.3 Ước lượng độ chính xác của mô hình 6
Hình 1.4 Phân lớp dữ liệu mới 7
Hình 1.5 Ước lượng độ chính xác của mô hình phân lớp bằng phương pháp holdout..... 10
Hình 2.1 Biểu diễn vector văn bản trong không gian 2 chiều 21
Hình 2.2 Mô hình SVM [18]..... 32
Hình 2.3 Margin - khoảng cách của các điểm tới biên 32
Hình 2.4 Mô hình SVM trong không gian 33
Hình 2.5 Mô hình thuật toán K-NN 35
Hình 3.1 Sơ đồ xử lý dữ liệu 41
Hình 3.2 Giao diện Weka..... 45
Hình 3.3 Chuyển đổi dữ liệu sang .arff..... 50
Hình 3.4 vector hóa dữ liệu..... 51
Hình 3.5 Giao diện huấn luyện 55
Hình 3.6 Kết quả huấn luyện..... 55

DANH MỤC BẢNG BIỂU

Bảng 2.1 Biểu diễn văn bản trong mô hình Logic 18
Bảng 2.2 Biểu diễn văn bản mô hình Vector 21
Bảng 2.3 Biểu diễn văn bản mô hình Boolean..... 22
Bảng 3.1 kết quả huấn luyện và kiểm thử..... 56

MỞ ĐẦU

I. ĐẶT VẤN ĐỀ

Hầu hết các doanh nghiệp đều luôn muốn quan tâm đến ý kiến, phản hồi của khách hàng về sản phẩm, dịch vụ của họ như thế nào. Các đánh giá của khách hàng một mặt giúp cho những người dùng khác định hướng trong việc chọn lựa sản phẩm, mặt khác giúp cho các doanh nghiệp định hướng cải tiến chất lượng. Số lượng đánh giá về một sản phẩm mà chúng ta nhận được ngày càng tăng và có thể đến từ nhiều nguồn khác nhau (web bán hàng, diễn đàn, blog, mạng xã hội ...). Vì vậy, để có thể tổng hợp ý kiến phản hồi của khách hàng về chất lượng, thì phải tự động hóa được công việc thu thập và phân tích đánh giá.

Công nghệ phân lớp dữ liệu đã, đang và sẽ phát triển mạnh mẽ trước những khao khát tri thức của con người. Trong những năm qua, phân lớp dữ liệu đã thu hút sự quan tâm các nhà nghiên cứu trong nhiều lĩnh vực khác nhau như học máy (machine learning), hệ chuyên gia (expert system), thống kê (statistics) ... Công nghệ này cũng ứng dụng trong nhiều lĩnh vực thực tế như: thương mại, nhà băng, maketing, nghiên cứu thị trường, bảo hiểm, y tế, giáo dục ...

Phân lớp văn bản là bài toán cơ bản trong khai phá quan điểm. Các hệ thống phân lớp văn bản là các hệ thống phải có khả năng xác định, khai phá ra nội dung thông tin. Có thể coi phân lớp quan điểm là bài toán phân lớp văn bản theo hai lớp tích cực và tiêu cực.

Do đó tôi chọn đề tài "**Đánh giá sản phẩm trên các website thương mại điện tử dựa trên nhận xét của người dùng trên internet**" đề tài nghiên cứu một số kỹ thuật phân lớp văn bản như K-means, Naïve Bayes, Maximum entropy và SVM để sử dụng trong phương pháp học máy phân lớp quan điểm khách hàng.

II. ĐỐI TƯỢNG VÀ PHẠM VI NGHIÊN CỨU

- **Đối tượng nghiên cứu:**

  Các kỹ thuật phân lớp văn bản và ứng dụng để phân lớp quan điểm khách hàng đưa vào các website TMĐT bán hàng trực tuyến với số lượng truy cập và giao dịch lớn.

- **Phạm vi nghiên cứu**

  Nghiên cứu các tài liệu bài viết trong và ngoài nước về kỹ thuật phân lớp dữ liệu để xây dựng phát triển bài toán "Phân lớp quan điểm khách hàng và ứng dụng" hiệu quả trong công việc phân tích, khai thác các nguồn ý kiến khách hàng.

III. HƯỚNG NGHIÊN CỨU CỦA ĐỀ TÀI

- Đề tài sẽ kết hợp phương pháp nghiên cứu lý thuyết với kết quả thực nghiệm.
- Phân tích các tài liệu và thông tin liên quan.
- Mô phỏng và thử nghiệm.

IV. PHƯƠNG PHÁP NGHIÊN CỨU

Nghiên cứu lý thuyết dựa trên các tài liệu về phân lớp dữ liệu, các thuật toán, phương pháp phân lớp … của các tác giả trong và ngoài nước. Thực nghiệm dựa trên các website TMĐT để xây dựng, đánh giá phương pháp.

CHƯƠNG 1 – PHÂN LỚP DỮ LIỆU

1.1 Giới thiệu về phân lớp dữ liệu

Bài toán phân lớp quan điểm

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). Quá trình phân lớp còn được gọi là quá trình gán nhãn cho các đối tượng dữ liệu [5][3].

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,….

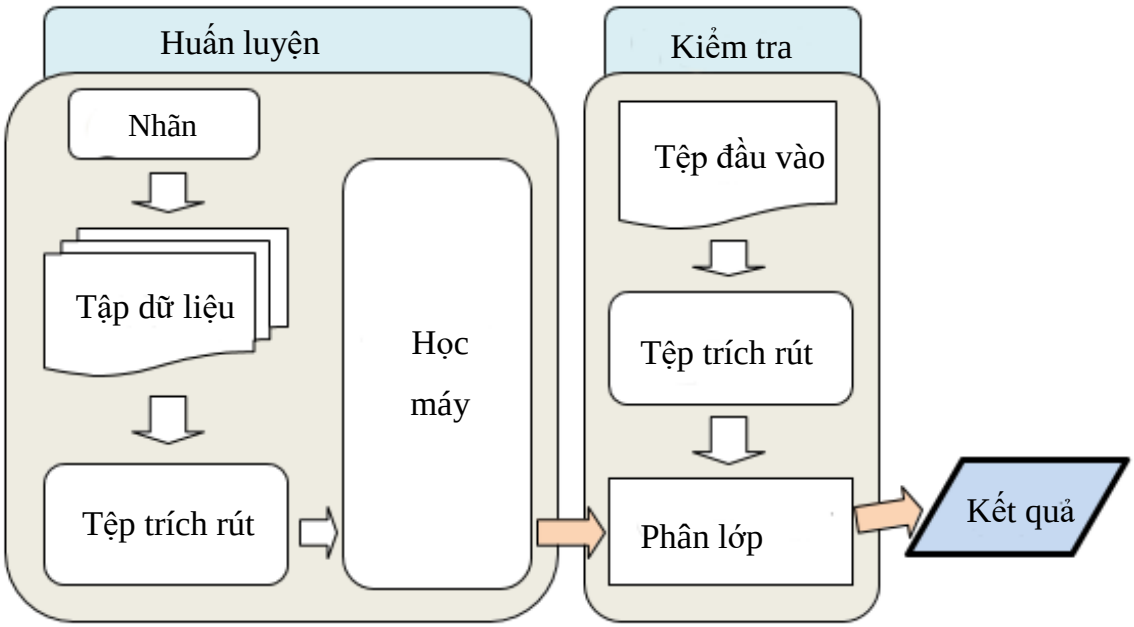
Phân lớp nhị phân là quá trình tiến hành việc phân lớp dữ liệu vào một trong hai lớp khác nhau dựa vào việc dữ liệu đó có hay không một số đặc tính theo quy định của bộ phân lớp.

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.

1.2 Quá trình phân lớp dữ liệu



Hình 1.1 Quy trình phân loại văn bản [3]

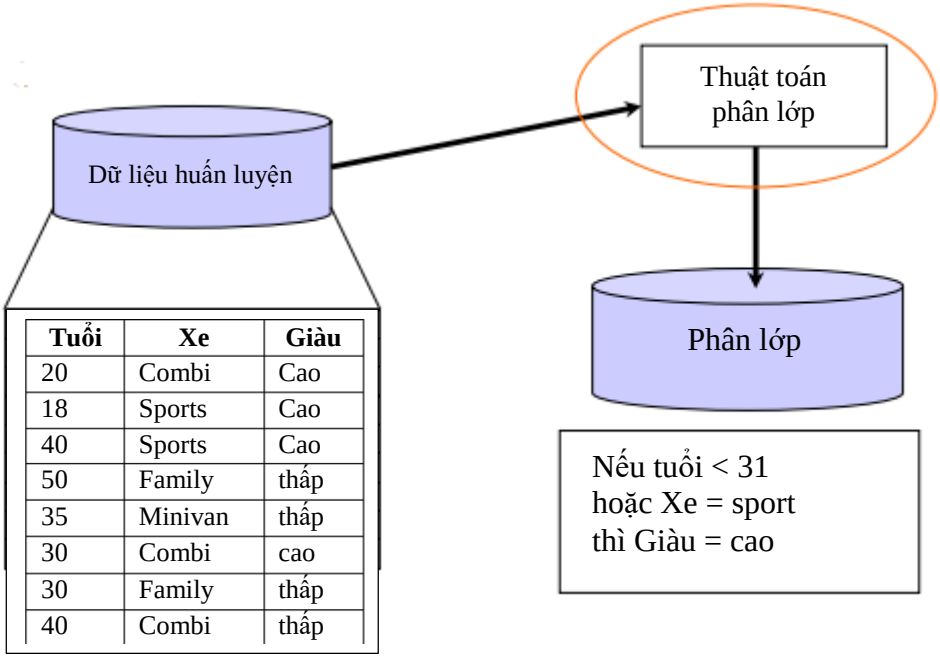
Quá trình phân lớp dữ liệu thường gồm hai bước: xây dựng mô hình (tạo bộ phân lớp) và sử dụng mô hình đó để phân lớp dữ liệu.

Bước 1 Bước xây dựng mô hình phân lớp (Training)

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.

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.



Tuổi	Xe	Giàu
20	Combi	Cao
18	Sports	Cao
40	Sports	Cao
50	Family	thấp
35	Minivan	thấp
30	Combi	cao
30	Family	thấp
40	Combi	thấp

Hình 1.2 Bước xây dựng mô hình phân lớp - Training

Bước 2 Phân lớp (classification)

Bước thứ hai dùng mô hình đã xây dựng ở bước trước để phân lớp dữ liệu mới. Trước tiên độ chính xác mang tính chất dự đoán của mô hình phân lớp vừa tạo ra được ước lượng. Holdout là một kỹ thuật đơn giản để ước lượng độ chính xác đó. Kỹ thuật này sử dụng một tập dữ liệu kiểm tra với các mẫu đã được gán nhãn lớp. Các mẫu này được chọn ngẫu nhiên và độc lập với các mẫu trong tập dữ liệu đào tạo. Độ chính xác của mô hình trên tập dữ liệu kiểm tra