HỌC VIỆN CÔNG NGHỆ BƯU CHÍNH VIỄN THÔNG

LÊ TRUNG HIẾU

DỰ ĐOÁN GIỚI TÍNH NGƯỜI DÙNG INTERNET DỰA TRÊN LỊCH SỬ TRUY CẬP

LUẬN VĂN THẠC SĨ KỸ THUẬT

(Theo định hướng ứng dụng)

HÀ NỘI - 2017

HỌC VIỆN CÔNG NGHỆ BƯU CHÍNH VIỄN THÔNG

LÊ TRUNG HIẾU

DỰ ĐOÁN GIỚI TÍNH NGƯỜI DÙNG INTERNET DỰA TRÊN LỊCH SỬ TRUY CẬP

CHUYÊN NGÀNH : HỆ THỐNG THÔNG TIN

MÃ SỐ: 60.48.01.04

LUẬN VĂN THẠC SĨ KỸ THUẬT

(Theo định hướng ứng dụng)

NGƯỜI HƯỚNG DẪN KHOA HỌC: PGS.TS. TỪ MINH PHƯƠNG

HÀ NỘI - 2017

LỜI CAM ĐOAN

Luận văn này là thành quả của quá trình học tập nghiên cứu của tôi cùng sự giúp đỡ, khuyến khích của các quý thầy cô sau 2 năm tôi theo học chương trình đào tạo Thạc sĩ, chuyên ngành Hệ thống thông tin của trường Học viện Công nghệ Bưu chính Viễn thông.

Tôi cam đoan đây là công trình nghiên cứu của riêng tôi. Nội dung của luận văn có tham khảo và sử dụng một số thông tin, tài liệu từ các nguồn sách, tạp chí được liệt kê trong danh mục các tài liệu tham khảo và được trích dẫn hợp pháp.

TÁC GIẢ

Lê Trung Hiếu

LỜI CÁM ƠN

Tôi xin gửi lời cảm ơn và tri ân tới các thầy cô giáo, cán bộ của Học viện Công nghệ Bưu chính Viễn thông đã giúp đỡ, tạo điều kiện tốt cho tôi trong quá trình học tập và nghiên cứu để hoàn thành chương trình Thạc sĩ.

Tôi xin gửi lời cảm ơn sâu sắc tới PGS.TS Từ Minh Phương đã tận tình hướng dẫn, giúp đỡ và động viên tôi để hoàn thành tốt nhất Luận văn "DỰ ĐOÁN GIỚI TÍNH NGƯỜI DÙNG INTERNET DỰA TRÊN LỊCH SỬ TRUY CẬP".

Do vốn kiến thức lý luận và kinh nghiệm thực tiễn còn ít nên luận văn không tránh khỏi những thiếu sót nhất định. Tôi xin trân trọng tiếp thu các ý kiến của các thầy, cô để luận văn được hoàn thiện.

Trân trọng cám ơn.

Tác giả.

MỤC LỤC

DANH MỤC TỪ VIẾT TẮT

STT	Từ viết tắt	Ý nghĩa tiếng Anh	Ý nghĩa tiếng Việt
1	SVM	Support vector machine	Máy vector hỗ trợ
2	NB	Naïve Bayes	Thuật toán Naïve Bayes
3	MCRW	Multi-Class Real Winnow	Phân loại đa lớp
4	JVM	Java Virtual Machine	Môi trường tạo máy ảo thực thi
5	Tweet	Tweet	Bài đăng của người dùng trên mạng xã hội Twitter
6	Weka	Waikato Environment for Knowledge Analysis	Bộ phần mềm học máy

DANH MỤC CÁC BẢNG BIỂU

DANH MỤC CÁC HÌNH VẼ

MỞ ĐẦU

Ngày nay, người ta thường dành một lượng lớn thời gian trong ngày để truy cập Internet. Internet được người dùng sử dụng cho việc tìm kiếm thông tin, đọc tin tức, mua sắm, chơi trò chơi v.v. Và các nhà quảng cáo không thể bỏ lỡ cơ hội để tiếp thị trực tuyến đến với khách hàng của họ nhằm cung cấp các dịch vụ phù hợp với nhu cầu của tổ chức, cá nhân sử dụng mạng Internet. Tuy nhiên, hiện nay các nhà quảng cáo đang cung cấp toàn bộ thông tin của mình đến tất cả khách hàng họ có. Chính vì vậy người dùng thường phải đối mặt với số lượng lớn các thông tin không phù hợp ví dụ như không phù hợp về độ tuổi, về nghề nghiệp, về văn hóa và giới tính.

Tình trạng quá tải thông tin không đến đích này dẫn đến sự sụt giảm đáng kể trong việc tiếp thị trực tuyến. Từ đó việc phân loại người dùng Internet để đưa ra các số liệu thống kê, kế hoạch quảng cáo giúp hệ thống tiếp cận cung cấp thông tin phù hợp, hữu ích cho từng đối tượng tương đối quan trọng. Xuất phát từ thực trạng đang xảy ra, luận văn sẽ trình bày về phương pháp xác định giới tính để phân loại người dùng Internet được thực hiện bằng kỹ thuật học máy, sử dụng thông tin người dùng đã biết giới tính và các thông tin về lịch sử truy cập web của họ để huấn luyện máy nhận biết giới tính của những người dùng khác khi ta chỉ biết lịch sử truy cập các trang web và dữ liệu danh mục mà người đó quan tâm.

Với mục tiêu đặt ra như vậy, nội dung và kết quả của luận văn được trình bày qua 3 chương như sau:

Chương 1 giới thiệu về dữ liệu truy cập của người dùng Internet thông qua thống kê, các khái niệm và đặc trưng trong tập dữ liệu này, bao gồm các mối quan hệ giữa các trang thông tin và người dùng mạng, những hành vi của người dùng khi truy cập Internet, cách thức truy cập, tìm kiếm thông tin. Giới thiệu những phương pháp nhắm mục tiêu theo hành vi hiện nay được áp dụng cho người dùng Internet và những hạn chế của các phương pháp này.

Chương 2 trình bày tổng quan về kỹ thuật học máy, một số kỹ thuật học máy và tập trung vào kỹ thuật được sử dụng trong luận văn là kỹ thuật học máy SVM. Dựa

vào những đặc trưng việc truy cập thông tin của người dùng Internet, đưa ra phương pháp dự đoán giới tính áp dụng kỹ thuật học máy và xếp hạng tỉ lệ độ chính xác nhằm tăng hiệu quả dự đoán so với các phương pháp đang tồn tại.

Chương 3 trình bày kết quả thực nghiệm và đánh giá. Sử dụng dữ liệu có sẵn PAKDD'15 được cung cấp bởi Công ty Cổ phần FPT (http://www.fpt.com.vn), thực hiện xây dựng bộ dữ liệu từ dữ liệu thực tế chưa chuẩn hóa hiện có PAKDD'15 cho một số lượng người dùng, sử dụng kỹ thuật học máy SVM ở chương 2 và một số công cụ để đưa ra tỉ lệ, độ chính xác của phương pháp dự đoán giới tính dựa trên lịch sử truy cập. Đánh giá kết quả so với các phương pháp dự đoán khác, và so sánh với cách làm việc hiện tại trong việc dự đoán giới tính.

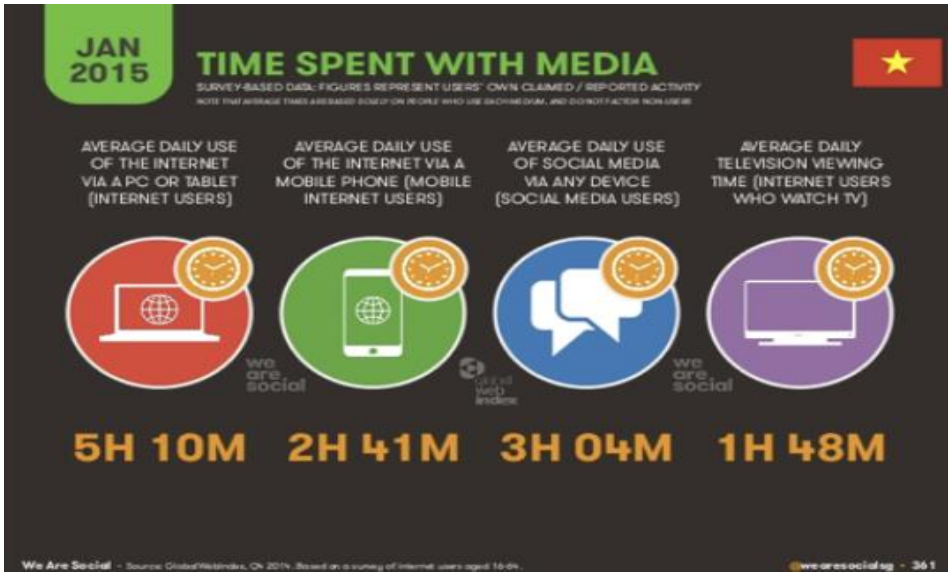
CHƯƠNG 1: TỔNG QUAN VỀ DỰ ĐOÁN GIỚI TÍNH NGƯỜI DÙNG INTERNET

1.1. Bài toán xác định giới tính và ứng dụng của bài toán vào thực tiễn

1.1.1. Mở đầu

Ngày nay, với sự phát triển không ngừng của khoa học công nghệ trên thế giới nói chung và ở Việt Nam nói riêng có những bước tiến vượt bậc. Cơ sở hạ tầng và các trang thiết bị tương đối hiện đại và không ngừng phát triển. Theo báo cáo tổng kết của Bộ TT&TT năm 2016, tỷ lệ người sử dụng Internet ở Việt Nam đạt 62,76% dân số, trong đó tỷ lệ hộ gia đình có truy cập Internet đạt 24,38%, tức là cứ 5 gia đình thì có một hộ sử dụng băng thông rộng cố định. Trong đó, theo thống kê của Cục Viễn thông (Bộ TT&TT) tháng 11/2016, tổng số thuê bao Internet băng rộng cố định đạt hơn 9 triệu thuê bao và số thuê bao băng rộng di động đạt hơn 12,6 triệu thuê bao.

Bên cạnh đó, theo thống kê của "wearesocial.net", tháng 1-2015, người Việt Nam đang đứng thứ 4 trên thế giới về thời gian sử dụng Internet với 5,2 giờ mỗi ngày, chỉ sau Philippines đứng đầu là 6 giờ, tiếp đó là Thái Lan với 5,5 giờ, và Brazin là 5,4 giờ/ngày.



Hình 1.1 Thời gian sử dụng Internet trung bình một ngày của người Việt Nam

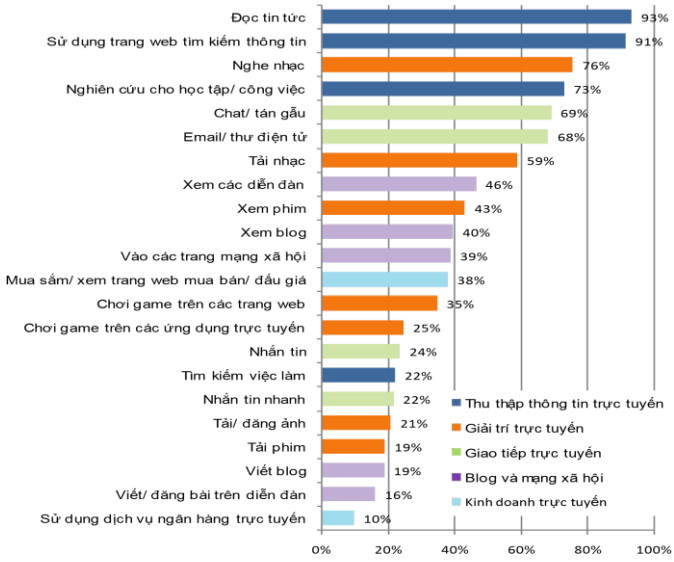
Chính vì sự phát triển không ngừng của công nghệ thông tin và mức độ phổ biến của Internet ngày nay mà thông tin đến với người dùng vô cùng phong phú và liên tục. Người sử dụng Internet hiện nay thường có thói quen truy cập và tìm kiếm đến những các vấn đề mình quan tâm. Hầu hết các thông tin được lưu vào như một phiên làm việc trên mạng. Các thông tin đó có thể là các bài báo, các tài liệu kinh doanh, sản phẩm, các thông tin kinh tế, thương mại điện tử, các thông tin cá nhân khác, ... Từ thực tế đó đã xuất hiện các nhu cầu phân tích thông tin để phân loại các thông tin đó cho các mục đích khác nhau như học tập, nghiên cứu, kinh doanh, tiếp thị thương mại.

Với thực tế đó, ta phải phân loại những thông tin hữu ích từ các nguồn dữ liệu phong phú và các phiên làm việc sử dụng Internet của người dùng sao cho phù hợp với đối tượng cụ thể. Ngoài ra cần áp dụng các công cụ tự động hoá trợ giúp trong việc phát hiện tri thức và khai thác thông tin.

1.1.2. Bài toán xác định giới tính

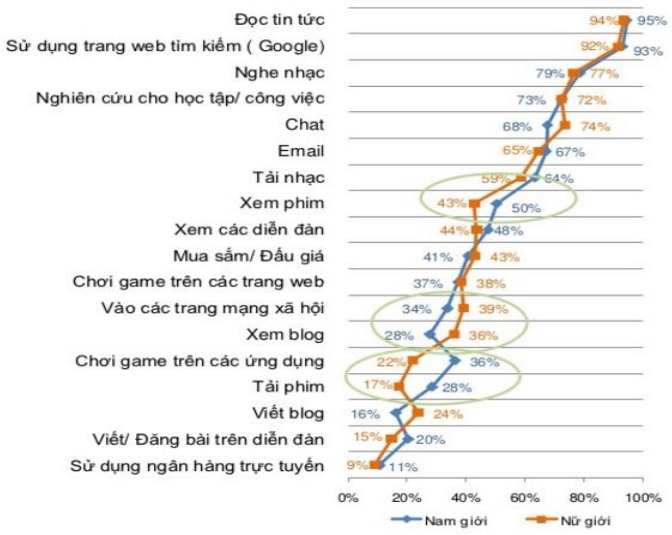
Nhìn chung, hoạt động thực hiện thường xuyên nhất trên Internet của người dùng là thu thập thông tin, như đọc tin tức hay sử dụng các trang web tìm kiếm. Hơn 90% số lượng người sử dụng Internet đã sử dụng những trang web tìm kiếm, khoảng một nửa trong số họ thậm chí sử dụng hàng ngày. Internet cũng được sử dụng để nghiên hoặc cho công việc bởi một nửa số người sử dụng Internet 1 lần 1 tuần hay thường xuyên hơn. Với các trang web và ứng dụng tương tác trực tuyến mới, người sử dụng không chỉ có cơ hội tìm được thông tin mà cũng đóng góp phần nội dung của riêng họ.

Thương mại điện tử hiện nay, số lượng truy cập đạt mức tăng trưởng đáng kể. Hầu hết các trang phổ biến là các trang web đấu giá và mua bán, nơi có 40% người sử dụng đã từng viếng thăm. Ngân hàng trực tuyến vẫn đang ở giai đoạn sơ khai tuy nhiên cũng đã được rất nhiều người trên thế giới quan tâm. Mức độ sử dụng các trang web mua hàng trực tuyến và ngân hàng trực tuyến đã phát triển rất mạnh trong vòng vài năm trở lại đây.



Hình 1.2 Các hoạt động trực tuyến được người dùng mạng sử dụng (Nguồn: Cimigo NetCitizens)

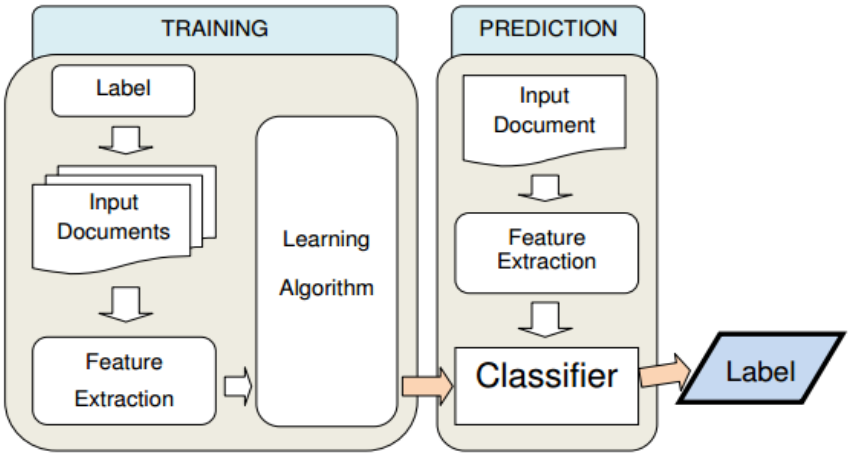
Việc sử dụng các hoạt động và truy cập Internet có sự khác nhau giữa nam giới và nữ giới. Trung bình một ngày nam giới dành thời gian nhiều hơn cho Internet. Nam giới cũng có một số hoạt động trực tuyến giống với nữ giới. Tuy nhiên có những khác nhau cụ thể ví dự như nam giới có khuynh hướng truy cập những đặc trưng như tin tức thời sự, bóng đá, hay trò chơi và các mặt hàng dành cho nam giới. Trái lại nữ giới thường thích thú với các mục mua sắm, thương mại điện tử, chat và tham gia các trang mạng xã hội và blog.



Hình 1.3 Các hoạt động trực tuyến được người dùng mạng theo giới tính (Nguồn: Cimigo NetCitizens)

Dự đoán giới tính (hay Determination Gender hoặc Gender Prediction) là phương pháp phân loại và xác định các hoạt động được truy cập bởi giới tính Nam hoặc giới tính Nữ từ những hoạt động khác đã biết trước nhãn. Ví dụ một bài báo trong một trang web có thể được truy cập bởi giới tính nam hoặc giới tính nữ (như thể thao, giáo dục, pháp luật, công nghệ thông tin, mỹ phẩm, quần áo …). Việc phân loại có thể được tiến hành một cách thủ công: đọc nội dung của từng hoạt động và gán nó vào một nhãn nào đó. Tuy nhiên, đối với hệ thống gồm rất bản ghi thì phương pháp này sẽ tốn rất nhiều thời gian và công sức. Do vậy cần phải có phương pháp tự động để phân loại giới tính. Phương pháp này giúp cho việc xác định giới tính đạt độ chính xác cao và sử dụng cho các mục đích như học tập, nghiên cứu, kinh doanh, tiếp thị thương mại.

Dưới đây là hình vẽ mô tả quy trình của bài toán xác định giới tính:



Hình 1.4 Quy trình phân loại xác định giới tính

Để tiến hành phân loại xác định giới tính nói chung, chúng ta sẽ thực hiện các bước sau đây:

- Bước 1: Xây dựng bộ dữ liệu huấn luyện dựa trên tập dữ liệu thu thập của người dùng đã được phân loại sẵn. Tiến hành học cho bộ dữ liệu, xử lý và thu thập được dữ liệu của quá trình học là các đặc trưng riêng biệt cho từng nội dung.

- Bước 2: Dữ liệu cần phân loại được xử lý, rút ra các đặc trưng kết hợp với đặc trưng được học trước đó để phân loại và đưa ra kết quả.

Đặc điểm nổi bật của bài toán này là sự đa dạng của hoạt động và đặc trưng của nam giới và nữ giới. Các đặc trưng làm cho sự phân loại chỉ mang tính tương đối và có phần chủ quan, nếu do con người thực hiện có thể dễ bị nhập nhằng. Ví dụ có hoạt động truy cập về xem thông tin mua sắm quần áo tại một trang web thương mại điện tử, hoạt động truy cập này vẫn có thể được truy cập bởi nam giới hoặc nữ giới.

1.1.3. Ứng dụng của bài toán vào thực tiễn

Trên thế giới đã có một số công trình nghiên cứu với các hướng tiếp cận khác nhau cho bài toán xác định giới tính, bao gồm các tập dữ liệu có đặc trưng thể hiện giới tính. Theo các kết quả trình bày trong các công trình đó thì những cách tiếp cận đều cho kết quả khả quan. Tuy nhiên khó có thể so sánh các kết quả ở trên với nhau vì tập dữ liệu thực nghiệm của mỗi phương pháp là khác nhau.

Hiện nay, công nghệ ngày càng phát triển, đặc biệt với sự ra đời của các trang mạng xã hội, thương mại điện tử nên lượng thông tin lớn, phi cấu trúc, phức tạp, thậm chí là các thông tin rác cũng rất nhiều. Cần thiết phải có những nghiên cứu để xác định được thông tin gì là cần thiết và thông tin nào là dư thừa. Các nhà nghiên cứu xử lý ngôn ngữ tự nhiên và trích chọn thông tin đều đi tìm câu trả lời cho câu hỏi đó. Hầu hết các thông tin đều là các hoạt động trực tuyến như tìm kiếm thông tin, chat, email, mua sắm trực tuyến ... Từ thực tế đó đã xuất hiện các nhu cầu phân tích thông tin của người dùng Internet để phân loại các thông tin đó sao cho phù hợp với giới tính nhằm đưa ra các số liệu thống kê, kế hoạch quảng cáo giúp hệ thống tiếp cận cung cấp thông tin phù hợp, hữu ích cho từng đối tượng.

Trong những năm gần đây, phương pháp phân loại sử dụng Máy vector hỗ trợ (SVM) được quan tâm và sử dụng nhiều trong những lĩnh vực nhận dạng và phân loại. 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 loại

khá tốt đối với bài toán phân loại 2 lớp và đa lớp cũng như trong nhiều ứng dụng khác (như phân loại văn bản theo chủ đề, phát hiện mặt người trong các ảnh, ước lượng hồi quy, dự đoán lỗi phần mềm...). So sánh với các phương pháp phân loại khác, khả năng phân loại của SVM là tương đương hoặc tốt hơn đáng kể. Vì những lý do đó mà tôi đã chọn phương pháp này cho việc dự đoán giới tính của người dùng Internet, cụ thể thuật toán và ứng dụng sẽ được trình bày trong các chương sau.

1.2. Các dạng dữ liệu lịch sử có thể dự đoán

Có nhiều loại dữ liệu lịch sử có thể được sử dụng để dự đoán. Ở giai đoạn đầu phân loại giới tính, hầu hết các nghiên cứu về lĩnh vực này tập trung vào việc nghiên cứu tác giả, đó là những nhiệm vụ xác định hoặc dự đoán các đặc điểm tác giả bằng cách phân tích các câu chuyện, tác phẩm, tiểu thuyết được tạo ra bởi tác giả nam hay tác giả nữ. Các phương pháp mà các nhà nghiên cứu sử dụng trong các nghiên cứu này chủ yếu dựa trên việc phân tích các phong cách viết, văn phong sử dụng các đặc trưng về ngữ pháp chẳng hạn như từ vựng, cú pháp, hoặc các đặc trưng dựa trên nội dung. Nghiên cứu đầu tiên trong lĩnh vực này bắt đầu vào thế kỷ 19 khi Mendenhall (1887) [16] đã nghiên cứu các tác phẩm của Shakespeare.

Gần đây, do sự phát triển của Internet và các kênh truyền thông trực tuyến, các dạng dữ liệu được thu thập chủ yếu dựa trên nội dung truyền thông ví dụ như:

- Email: Một dạng dữ liệu lịch sử, một phương tiện thông tin rất nhanh chứa đựng các văn bản đơn thuần và thường được dùng trong việc trao đổi thông tin. Chúng ta có thể dự đoán giới tính dựa trên địa chỉ email và văn bản có trong email.
- Blog: Là một tập san dữ liệu cá nhân trực tuyến. Nội dung và chủ đề của “blog” thì rất đa dạng, nhưng thông thường là những bài viết câu chuyện cá nhân, bản tin, danh sách các liên kết web, những bài tường thuật, phê bình một bộ phim hay tác phẩm văn học mới xuất bản và cuối cùng là những sự kiện xảy ra trong một nhóm người nào đó.

- Twitter: Là một mạng xã hội và các thông điệp trên twitter của người dùng được sử dụng như một văn phong, hành vi để xác định xem thông điệp này được viết bởi giới tính nào.

1.3. Các phương pháp xác định giới tính đã có

Trên thế giới, một số công trình đi trước đã nghiên cứu các phương pháp dựa trên phân tích văn bản như De Vel et al. [17] đã sử dụng 221 đặc trưng để xác định tác giả của email. Argamon và Koppel et al. [18] đã nghiên cứu sự khác biệt trong phong cách viết của nam và nữ trong 604 tài liệu của National Corpus của Anh. Schler et al. [19] khám phá việc sử dụng các đặc trưng và dựa trên nội dung để dự đoán giới tính và độ tuổi của các blogger trên bộ dữ liệu với hơn 71,000 bài viết blog từ blogger.com. Mô hình này đã đạt được kết quả 80% cho dự đoán giới tính và 76% đối với các dự đoán tuổi. Nguyen et al. [14] đã tiến hành một nghiên cứu để dự đoán giới tính và độ tuổi của các thông điệp twitter và diễn đàn bài viết bằng cách sử dụng phương pháp hồi quy với độ chính xác khoảng 80%.

1.3.1. Phương pháp xác định giới tính sử dụng bài viết từ blog

Trong những năm trở về trước, Blog là một loại nhật ký, website cá nhân phổ biến chia sẻ những kinh nghiệm sống hoặc một thông tin gì đó trong cuộc sống hằng ngày của con người. Đây là một loại dữ liệu rất rất lớn chứa các bài viết, văn bản do hàng trăm nghìn tác giả người dùng tạo ra. Những thông tin này chứa đựng rất nhiều các đặc trưng có thể khai thác cho bài toán phân loại, cụ thể ở đây là việc xác định giới tính các blogger. Bài báo nghiên cứu cụ thể về xác định nhân khẩu học và giới tính được Schler et al [19] xây dựng năm 2007 với tập dữ liệu là tất cả blog được truy cập trong một ngày tháng 8 năm 2004.

Nội dung nghiên cứu trú trọng sự khác biệt trong việc viết blog và sự khác biệt giữa nam giới và nữ giới giữa các blogger ở các độ tuổi khác nhau. Các đặc trưng về phong cách và nội dung được đưa ra làm hạt nhân để giải quyết bài toán.

Nghiên cứu sử dụng mô hình MCRW (Multi-Class Real Winnow). Đối với mỗi lớp, c_i, i = 1, ..., m, w^i là một vector trọng lượng $<w^i_1, ..., w^i_n>$, trong đó n là kích thước của tập hợp tính năng. Mỗi w^i_j, được khởi tạo bắt đầu là 1. Các tập huấn luyện được sắp xếp ngẫu nhiên và được xử lý một lần. Thuật toán chạy vòng lặp huấn luyện liên tục, ngẫu nhiên đặt lại các ví dụ sau mỗi chu kỳ. Sau mỗi mười chu kỳ, Thuật toán kiểm tra số lượng các ví dụ đào tạo được phân loại chính xác Nếu con số này đã giảm, thuật toán sẽ quay trở lại. Nếu không có cải tiến nào được tìm thấy sau năm vòng của 10 chu kỳ, thuật toán sẽ được chấm dứt. Nghiên cứu cho thấy mô hình MCRW hiệu quả hơn so với SVM về việc phân loại một số lượng lớn văn bản.

Các kết quả kiểm thử cho thấy được việc phân loại được các blogger theo giới tính theo các nhóm tuổi, kiểu viết và nội dung. Trong các trường hợp được đưa ra, thì sự kết hợp của các đặc trưng phong cách và nội dung cung cấp độ chính xác phân loại tốt nhất.

1.3.2. Phương pháp xác định giới tính sử dụng dữ liệu thông tin di động liên lạc hàng ngày

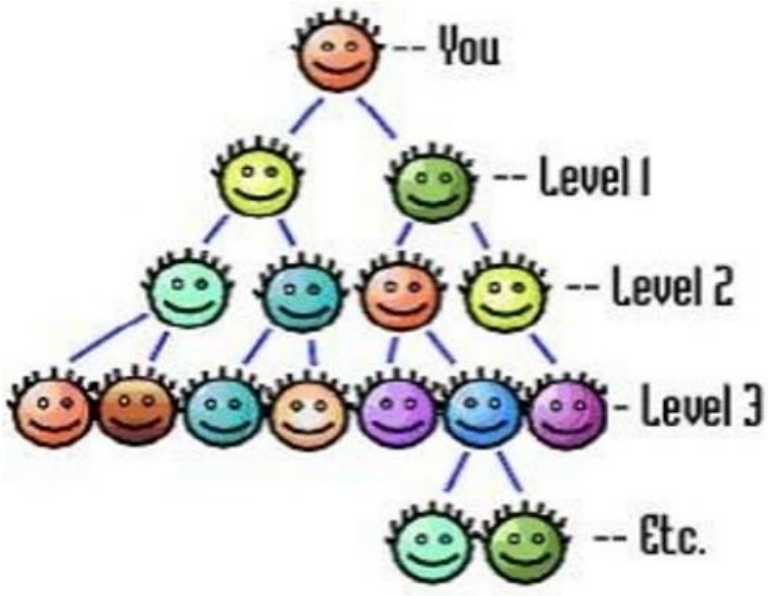
a. Giới thiệu

Phương pháp xác định giới tính thông qua dữ liệu từ các thông tin di động liên lạc hàng ngày được nghiên cứu theo bài báo Demographic Prediction Based on User's Mobile Behaviors [9] trong cuộc thi MDC Data Set. Trong bài báo này, nhóm nghiên cứu đề xuất một mô hình mới cụ thể là Multi-Level Classification Model (Mô hình phân loại Đa cấp) để giải quyết vấn đề các lớp không cân bằng hiện có trong dữ liệu. Dựa trên mô hình này, sẽ đưa ra kết quả việc dự đoán giới tính của người dùng bằng cách kết hợp nhiều mô hình phân loại vào một cấu trúc đa cấp.

b. Ý tưởng

Như đã đề cập, tài nguyên dữ liệu hiện có là dữ liệu nhật ký điện thoại di động của người dùng các vị trí khác nhau và thời gian khác nhau. Do đó, nghiên cứu trú trọng các đặc trưng hành vi người dùng và tìm kiếm các đặc trưng độc đáo của các vị trí được ghi lại trong nhật ký di động của tập dữ liệu MDC. Tập dữ liệu được trích

xuất phân loại huấn luyện và phân chia theo các tầng, từ tầng 1 đến tầng thấp hơn, lần lượt xác định phân loại ở mỗi tầng cho đến khi thu được kết quả phân loại chính xác nhất.



Hình 1.5 Ví dụ mô hình phân loại đa cấp

1.3.3. Xác định giới tính sử dụng dữ liệu từ các thông điệp trên twitter bằng phương pháp hồi quy

a. Giới thiệu

Xác định giới tính sử dụng dữ liệu từ các thông điệp Twitter là phương pháp phân loại cho từng bình luận theo đặc trưng dựa trên nội dung bình luận bằng phương pháp hồi quy. Ở bước đầu tiên, từ tập dữ liệu thô là những ý kiến trên Twitter được thu thập theo chủ đề, ta tiến hành tiền xử lý các kí tự đặc biệt của Twitter, các kí tự trùng lặp gần nhau, từ viết tắt, tiếng lóng, biểu tượng cảm xúc, mạng ngữ nghĩa. Nghiên cứu được trình bày bởi Nguyen [14].

b. Ý tưởng

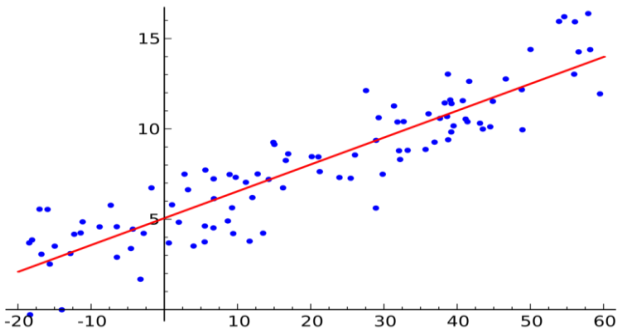
Đọc nội dung twitter của ai đó, trong một số trường hợp người ta phần nào có thể đoán được giới tính của người dùng. Ví dụ, Bạn có thể biết giới tính người dùng phía sau twitter sau đây?

I LIKE PLAYING FOOTBALL <3

Hồi Quy (regression) là một phương pháp học có giám sát (supervised learning) trong Máy Học. Mục tiêu chính là tìm ra mối quan hệ giữa các đặc trưng của một vấn đề nào đó. Cụ thể hơn, từ một tập dữ liệu cho trước, ta xây dựng một mô hình (phương trình, đồ thị, …) khớp nhất với tập dữ liệu, thể hiện được xu hướng biến thiên và mối quan hệ giữa các đặc trưng. Khi có một mẫu dữ liệu mới vào, dựa vào mô hình, chúng ta có thể dự đoán giá trị của mẫu dữ liệu đó.

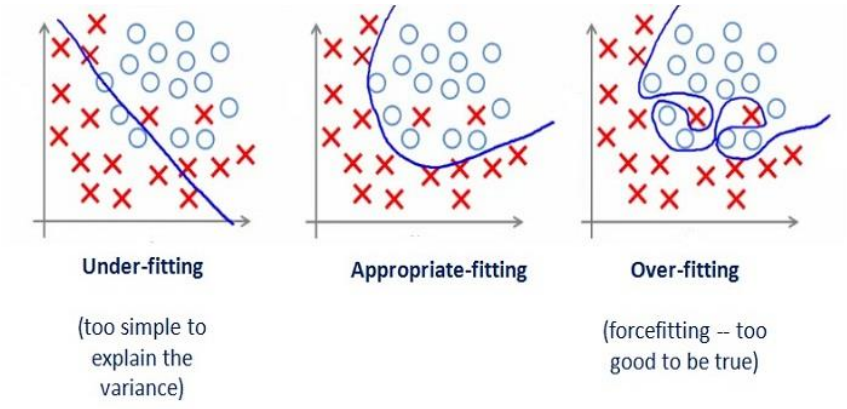
Lấy ví dụ như chúng ta cần dự đoán giới tính của một twitter dựa vào nội dung và đặc trưng viết của twitter đó. Như vậy chúng ta cần tìm mối quan hệ giữa giới tính phụ thuộc vào nội dung và đặc trưng viết. Dựa vào tập dữ liệu (giả sử thu thập nội dung, đặc trưng viết và các ký tự đặc biệt của 100 người dùng twitter), ta xây dựng một phương trình $y=\theta 0+\theta 1x1+\theta 2x2$ trong đó y là giới tính phụ thuộc x_1 (nội dung) và x_2 (đặc trưng viết). Khi có thêm một mẫu dữ liệu của một người dùng mới, chỉ cần áp vào phương trình như vậy ta sẽ dự đoán được giới tính của người đó.

Ta thấy phương trình $y =\theta 0+\theta 1x1+\theta 2x2$ là phương trình của mặt phẳng trong không gian 3 chiều. Những mô hình tương tự như phương trình đường thẳng, phương trình mặt phẳng chính là những mô hình tuyến tính. Hồi quy tuyến tính (linear regression) là một mô hình đơn giản trong bài toán hồi quy, trong đó chúng ta dùng đường thẳng, mặt phẳng, hay phương trình tuyến tính nói chung để dự đoán xu hướng của dữ liệu. Giải bài toán hồi quy tuyến tính chính là đi tìm các tham số $\theta 0$, $\theta 1$, ...để xác định phương trình tuyến tính.



Hình 1.6 Ví dụ về hồi quy tuyến tính

Một trong những vấn đề gặp phải trong khi chạy mô hình Hồi Quy Tuyến Tính chính là hiện tượng quá khớp (overfitting). Overfitting là vấn đề xảy ra khi mô hình ta tạo ra cố gắng quá mức để khớp với các mẫu trong tập huấn luyện. Mô hình tuy rằng khớp với các mẫu huấn luyện nhưng lại không thể hiện được xu hướng của dữ liệu dẫn đến việc mô hình chỉ đúng với các giá trị trong tập huấn luyện và sai hoàn toàn với các giá trị test.



Hình 1.7 Quá trình khớp

Vấn đề quá khớp thường xảy ra khi bộ dữ liệu twitter của ta có nhiều đặc trưng nhưng lại có ít mẫu dữ liệu. Ví dụ như chúng ta muốn tạo ra một mô hình có dạng đường thẳng tức là cần hai đặc trưng x_1, x_2 (đặc trưng tọa độ trong mặt phẳng) nhưng lại chỉ có một mẫu dữ liệu được biểu diễn thành một điểm. Để xác định đường thẳng cần ít nhất hai điểm và nếu chỉ có một điểm thì có vô số mô hình phù hợp với mẫu dữ liệu nhưng trong đó chỉ có một mô hình là thật sự đúng với thực tế.

1.4. Kết luận chương

Kết luận chương:

Chương này đã giới thiệu tổng quan về bài toán xác định giới tính, ứng dụng của bài toán vào thực tiễn và một số phương pháp xác định giới tính và dữ liệu lịch sử liên quan đến việc phân loại giới tính nam hay giới tính nữ. Bên cạnh đó, chương 1 còn đưa ra lý do và thực trạng các hoạt động của người dùng Internet trong luận văn. Ngoài ra cần lưu ý đến yếu tố quan trọng tác động đến kết quả phân loại giới tính đó là phải có một tập dữ liệu lịch sử để huấn luyện chuẩn và đủ lớn để cho thuật

toán học phân loại. 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 qủa phân loại tốt sau khi đã được học. Trong chương 1, luận văn cũng đã giới thiệu một số phương pháp xác định giới tính đã được nghiên cứu trong thời gian gần đây. Những mô tả của chương 1 sẽ làm tiền đề cho việc xác định giới tính người dùng Internet sử dụng dữ liệu lịch sử truy cập trong các chương tiếp theo.

CHƯƠNG 2: DỰ ĐOÁN GIỚI TÍNH NGƯỜI DÙNG INTERNET SỬ DỤNG LỊCH SỬ TRUY CẬP

2.1. Giới thiệu về phương pháp học máy SVM

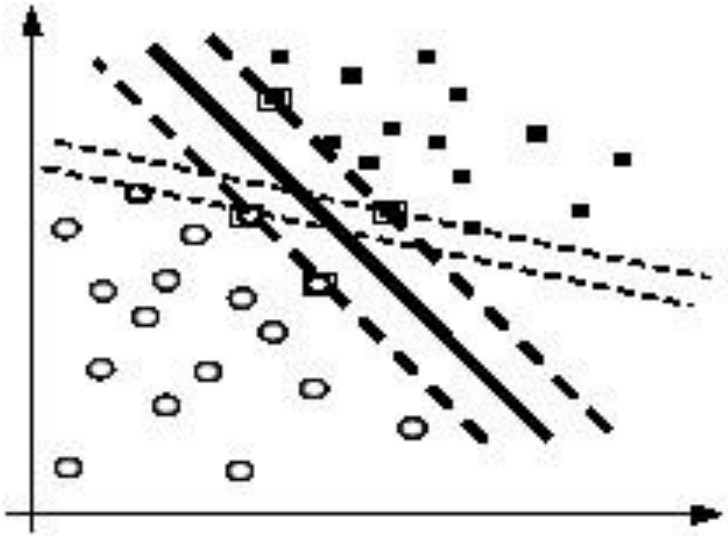
2.1.1. Giới thiệu về SVM

Support Vector Machines (SVM) là một phương pháp phân loại xuất phát từ lý thuyết học thống kê, dựa trên nguyên tắc tối thiểu rủi ro cấu trúc (Structural Risk Minimisation). SVM sẽ cố gắng tìm cách phân loại dữ liệu sao cho có lỗi xảy ra trên tập kiểm tra là nhỏ nhất (Test Error Minimisation). Vào thời kỳ đầu khi SVM xuất hiện, khả năng tính toán của máy tính còn rất hạn chế, nên phương pháp SVM không được lưu tâm. Tuy nhiên, từ năm 1995 trở lại đây, các thuật toán sử dụng cho SVM phát triển rất nhanh, cùng với khả năng tính toán mạnh mẽ của máy tính, đã có được những ứng dụng rất to lớn.

a. Ý tưởng

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 phẳng f quyết định tốt nhất có thể chia các điểm trên không gian này thành hai lớp riêng biệt tương ứng là lớp "+" và lớp "−". Chất lượng của siêu phẳng này được quyết định bởi khoảng cách (gọi là biên) của điểm dữ liệu gần nhất của mỗi lớp đến mặt phẳng này. Khi đó, khoảng cách biên càng lớn thì mặt phẳng quyết định càng tốt, đồng thời việc phân loại càng chính xác. Ý tưởng của nó là ánh xạ (tuyến tính hoặc phi tuyến) dữ liệu vào không gian các vector đặc trưng (space of feature vectors) mà ở đó một siêu phẳng tối ưu được tìm ra để tách dữ liệu thuộc hai lớp khác nhau.

Mục đích của phương pháp SVM là tìm được khoảng cách biên lớn nhất:



Hình 2.1 Mô tả phương pháp SVM

Đường tô đậm là siêu phẳng tốt nhất và các điểm được bao bởi hình chữ nhật là những điểm gần siêu phẳng nhất, chúng được gọi là các vector hỗ trợ (support vector). Các đường nét đứt mà các support vector nằm trên đó được gọi là lề (margin).

b. Cơ sở lý thuyết

SVM thực chất là một bài toán tối ưu, mục tiêu của thuật toán này là tìm được một không gian F và siêu phẳng quyết định f trên F sao cho sai số phân loại là thấp nhất.

Cho tập mẫu $D = \{(x_1, y_1), (x_1, y_1), \dots, (x_l, y_l)\}$ với $x_i \in \mathbb{R}^n$, thuộc vào hai lớp nhãn $y_i \in \{-1, 1\}$ là nhãn lớp tương ứng của các x_i (-1 biểu thị lớp I, 1 biểu thị lớp II).

Ta có, phương trình siêu phẳng chứa vector $\overrightarrow{x_i}$ trong không gian:

$$\overrightarrow{x_i}.\vec{w} + b = 0$$

Đặt:

$$f(\overrightarrow{x_i}) = sign(\overrightarrow{x_i}.\vec{w} + b) = \begin{cases} +1, \overrightarrow{x_i}.\vec{w} + b > 0 \\ -1, \overrightarrow{x_i}.\vec{w} + b < 0 \end{cases}$$

Như vậy, $f(\overrightarrow{x_i})$ biểu diễn sự phân lớp của $\overrightarrow{x_i}$ vào hai lớp như đã nêu.

Ta nói $y_i = +1$ nếu $\overrightarrow{x_i}$ thuộc lớp I và $y_i = -1$ nếu $\overrightarrow{x_i}$ thuộc lớp II.

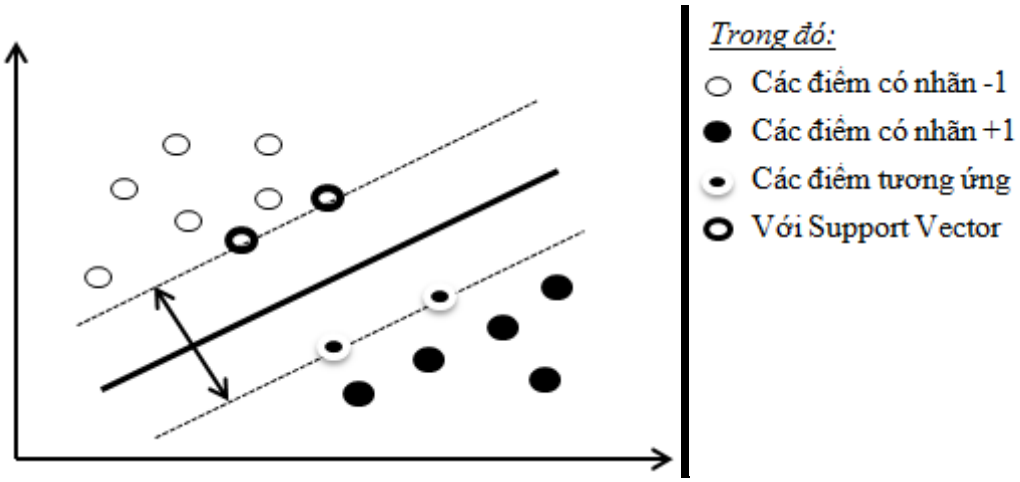
2.1.2. Bài toán phân 2 lớp với SVM

Bài toán đặt ra là: Xác định hàm phân lớp để phân lớp các mẫu trong tương lai, nghĩa là với một mẫu dữ liệu mới x_i thì cần phải xác định x_i được phân vào lớp $+1$ hay lớp -1.

Ta xét 3 trường hợp, mỗi trường hợp sẽ có 1 bài toán tối ưu, giải được bài toán tối ưu đó ta sẽ tìm được siêu phẳng cần tìm.

Trường hợp 1:

Tập D có thể phân chia tuyến tính được mà không có nhiễu (tất cả các điểm được gán nhãn +1 thuộc về phía dương của siêu phảng, tất cả các điểm được gán nhãn -1 thuộc về phía âm của siêu phẳng).



Hình 2.2 Tập dữ liệu được phân chia tuyến tính

Để xác định hàm phân lớp dựa trên phương pháp SVM, ta sẽ tiến hành tìm hai siêu phẳng song song sao cho khoảng cách y giữa chúng là lớn nhất có thể để phân tách hai lớp này ra làm hai phía. Hàm phân tách tương ứng với phương trình siêu phẳng nằm giữa hai siêu phẳng tìm được.

Các điểm mà nằm trên hai siêu phẳng phân tách được gọi là các Support Vector. Các điểm này sẽ quyết định đến hàm phân tách dữ liệu.

Ta sẽ tìm siêu phẳng tách với $w \in \mathbb{R}^n$ là vector trọng số, $b \in \mathbb{R}^n$ là hệ số tự do, sao cho:

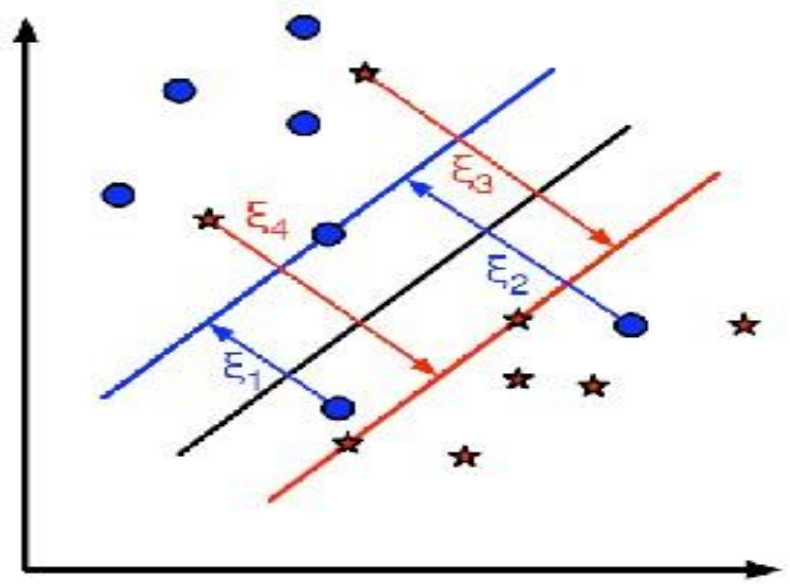
$$f(x_i) = sign(x_i.w^T + b) = \begin{cases} +1, y_i = +1 \\ -1, y_i = -1 \end{cases} \forall (x_i, y_i) \in D$$

Lúc này ta cần giải bài toán tối ưu:

$$\begin{cases} Min\big(L(w)\big) = \frac{1}{2}\|w\|^2 \\ y_i(x_i.w^T + b) \geq 1, i = 1, \dots, l \end{cases}$$

Trường hợp 2:

Tập dữ liệu D có thể phân chia tuyến tính được nhưng có nhiễu. Trong trường hợp này, hầu hết các điểm đều được phân chia đúng bởi siêu phẳng. Tuy nhiên có 1 số điểm bị nhiễu, nghĩa là: Điểm có nhãn dương nhưng lại thuộc phía âm của siêu phẳng, điểm có nhãn âm nhưng lại thuộc phía dương của siêu phẳng.



Hình 2.3 Tập dữ liệu được phân chia nhưng có nhiễu

Trong trường hợp này, ta sử dụng 1 biến mềm $\varepsilon_i \geq 0$ sao cho: $y_i(x_i.w^T + b) \geq 1 - \varepsilon_i, i = 1, \dots, l$

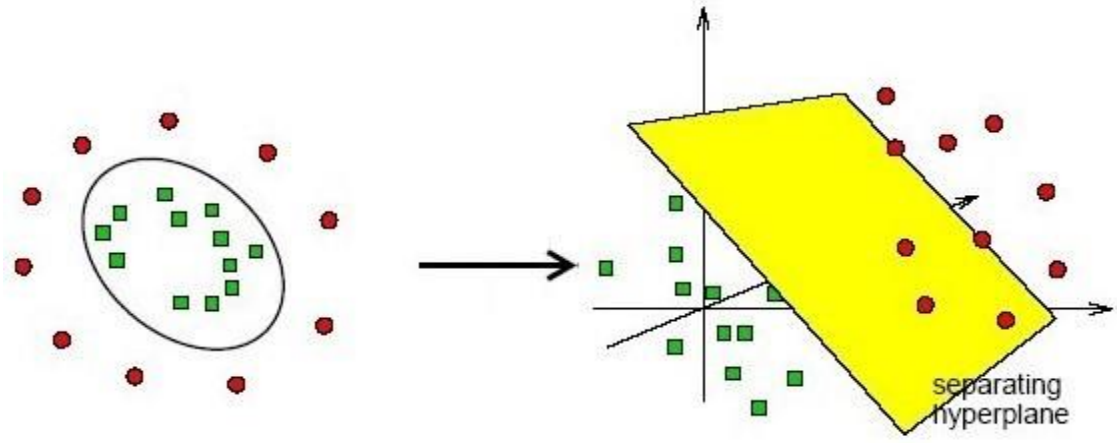
Bài toán tối ưu trở thành:

$$\begin{cases} Min\big(L(w,\varepsilon)\big) = \frac{1}{2}\|w\|^2 + C\sum_{i=1}^{l}\varepsilon_i \\ y_i(x_i.w^T + b) \geq 1 - \varepsilon_i, i = 1, \dots, l \\ \varepsilon_i \geq 0, i = 1, \dots, l \end{cases}$$

Trong đó C là tham số xác định trước, định nghĩa giá trị ràng buộc, C càng lớn thì mức độ vi phạm đối với những lỗi thực nghiệm (là lỗi xảy ra lúc huấn luyện, tính bằng thương số của số phần tử lỗi và tổng số phần tử huấn luyện) càng cao.

Trường hợp 3:

Tập dữ liệu D không thể phân chia tuyến tính được, ta sẽ ánh xạ các vector dữ liệu x từ không gian n chiều vào một không gian m chiều (m>n), sao cho trong không gian m chiều, D có thể phân chia tuyến tính được.



Hình 2.4 Tập dữ liệu không phân chia tuyến tính

Gọi ϕ là một ánh xạ phi tuyến từ không gian $\mathbb{R}^n$ vào không gian $\mathbb{R}^m$.

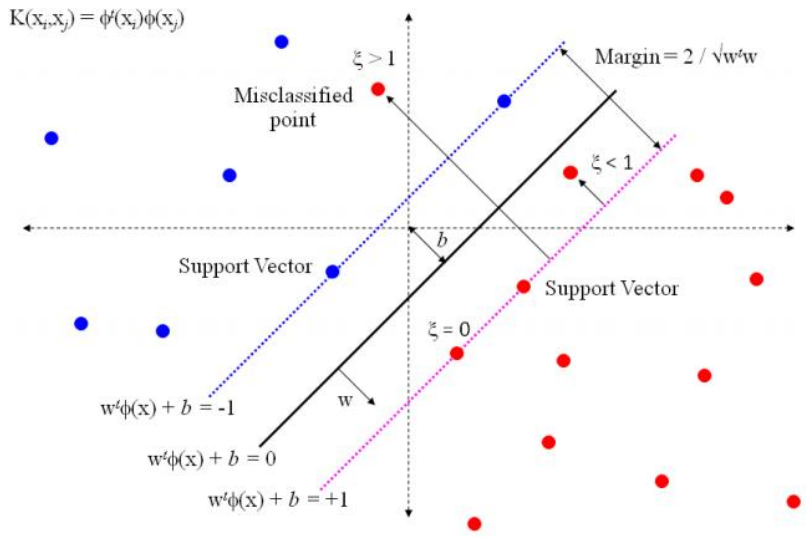
$$\phi: \mathbb{R}^n \to \mathbb{R}^m$$

Bài toán tối ưu trở thành:

$$\begin{cases} Min\big(L(w,\varepsilon)\big) = \frac{1}{2}\|w\|^2 + C\sum_{i=1}^{l}\varepsilon_i \\ y_i(\phi(x_i).w^T + b) \geq 1 - \varepsilon_i, i = 1, \dots, l \\ \varepsilon_i \geq 0, i = 1, \dots, l \end{cases}$$

Ví dụ:

Để dễ hiểu hơn ta xét ví dụ mô tả hình học sau: Xét trong không gian 2 chiều (n=2), tập dữ liệu được cho bởi tập các điểm trên mặt phẳng.



Hình 2.5 Ví dụ biểu diễn tập dữ liệu trên không gian 2 chiều

Bây giờ ta tiến hành tìm siêu phẳng phân lớp dựa trên phương pháp SVM (1). Ta sẽ tìm 2 siêu phẳng song song (nét đứt trong hình …) sao cho khoảng cách giữa chúng là lớn nhất để có thể phân tách lớp này thành 2 phía (Ta gọi là 2 siêu phẳng phân tách). Siêu phẳng (1) nằm giữa 2 siêu phẳng trên (nét đậm trong hình).

Hình trên cho ta tập dữ liệu có thể phân tách tuyến tính. Bây giờ ta xét trường hợp tập dữ liệu không thể phân tách tuyến tính. Bây giờ ta sẽ xử lý bằng cách ánh xạ tập dữ liệu đã cho vào một không gian mới có số chiều lớn hơn không gian cũ (Gọi là không gian đặc trưng) mà trong không gian này tập dữ liệu có thể phân tách tuyến tính. Trong không gian đặc trưng ta sẽ tiếp tục tìm 2 siêu phẳng phân tách như trường hợp ban đầu.

Các điểm nằm trên 2 siêu phẳng phân tách gọi là các vector hỗ trợ (Support vector). Các điểm này quyết định hàm phân tách dữ liệu. Từ đây, chúng ta có thể thấy phương pháp SVM không phụ thuộc vào các mẫu dữ liệu ban đầu, mà chỉ phụ thuộc vào các suport vector (quyết định 2 siêu phẳng phân tách). Cho dù các điểm khác bị xoá thì thuật toán vẫn cho ra các kết quả tương tự. Đây chính là điểm nổi bật

của phương pháp SVM so với các phương pháp khác do các điểm trong tập dữ liệu đều được dùng để tối ưu kết quả.

2.1.3. Các bước chính của phương pháp SVM

- Tiền xử lý dữ liệu: Phương pháp SVM yêu cầu dữ liệu được diễn tả như các vector của các số thực. Như vậy nếu đầu vào chưa phải là số thực thì ta cần phải tìm cách chuyển chúng về dạng số của SVM. Tránh các số quá lớn, thường nên chuẩn hóa dữ liệu để chuyển về đoạn [-1,1] hoặc [0,1].
- Chọn hàm nhân: Cần chọn hàm nhân phù hợp tương ứng cho từng bài toán toán cụ thể để đạt được độ chính xác cao trong quá trình phân lớp.
- Thực hiện việc kiểm tra chéo để xác định các tham số cho ứng dụng.
- Sử dụng các tham số cho việc huấn luyện tập mẫu.
- Kiểm thử tập dữ liệu Test.

2.1.4. Ưu điểm phương pháp SVM trong phân lớp dữ liệu

Như đã biết, phân lớp dữ liệu là một tiến trình đưa các dữ liệu chưa biết nhãn vào các lớp dữ liệu đã biết nhãn tương ứng. Mỗi nhãn được xác định bởi một số tập dữ liệu mẫu của nhãn đó. Để thực hiện quá trình phân lớp, các phương pháp huấn luyện được sử dụng để xây dựng tập phân lớp từ các bản ghi mẫu, sau đó dùng tập phân lớp này để dự đoán lớp của những bản ghi mới chưa biết nhãn.

Chúng ta có thể thấy các thuật toán phân lớp hai lớp như SVM đều có đặc điểm chung là yêu cầu dữ liệu phải được biểu diễn dưới dạng vector đặc trưng, tuy nhiên các thuật toán khác đều phải sử dụng các ước 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, Joachims [20] đã so sánh SVM với Naïve Bayesian, k-Nearest Neighbour, Rocchio, và C4.5 và đến năm 2003, Joachims đã chứng minh rằng SVM làm việc rất tốt cùng với các đặc tính được đề cập trước

đây của tập dữ liệu. 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 [22] 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.

2.2. Một số thuật toán học máy khác

Hiện nay trên thế giới, ngoài phương pháp học máy SVM (Support Vector Machine) đã có nhiều phương pháp, thuật toán học máy được nghiên cứu để phân lớp dữ liệu. Một số thuật toán cần kể đến là: Naïve Bayes, Random Tree ... Trong mỗi phương pháp đều có cách tính toán khác nhau, tuy nhiên các phương pháp này đều phải thực hiện một số bước chung như: mỗi phương pháp sẽ dựa vào thông tin về sự xuất hiện của các thông tin và tần số để biểu diễn thành dạng vector, sau đó tùy từng bài toán cụ thể mà chúng ta sẽ quyết định chọn áp dụng phương pháp nào, công thức tính toán nào cho phù hợp để phân loại tập dữ liệu dựa trên tập các vector đã xây dựng được ở bước trên, nhằm mục đích đạt được kết quả phân loại tốt nhất.

Thuật toán Random Tree

a. Giới thiệu

Cây ngẫu nhiên là một phương pháp học tập toàn bộ để xây dựng một cây mà được coi là k-ngẫu nhiên chọn thuộc tính tại mỗi nút. Phương pháp cây ngẫu nhiên phát triển một cây quyết định dựa trên việc chọn ngẫu nhiên dữ liệu và lựa chọn ngẫu nhiên các biến. Nó cung cấp lớp biến phụ thuộc dựa trên cây.

b. Ý tưởng

Một cây ngẫu nhiên là một tập hợp của cây phân loại hoặc hồi quy được tạo ra bởi một thủ tục bootstrap. Cây được phát triển từ một mẫu thử nghiệm khởi động độc lập cho đến khi tất cả các nút chứa các quan sát không quá kích thước nút tối đa được xác định trước.

Thuật toán Naïve Bayes

a. Giới thiệu

Naïve Bayes (NB) là phương pháp phân loại dựa vào xác suất được sử dụng rộng rãi trong lĩnh vực máy học (Mitchell trình bày năm 1996, Joachims trình bày năm 1997 và Jason năm 2001) được sử dụng lần đầu tiên trong lĩnh vực phân loại bởi Maron vào năm 1961, sau đó trở nên phổ biến dùng trong nhiều lĩnh vực như trong các công cụ tìm kiếm (được mô tả năm 1970 bởi Rijsbergen), các bộ lọc mail (mô tả năm 1998 bởi Sahami).

b. Ý tưởng

Ý tưởng cơ bản của cách tiếp cận Naïve Bayes là sử dụng xác suất có điều kiện giữa từ và chủ đề để dự đoán xác suất chủ đề của một văn bản cần phân loại. Điểm quan trọng của phương pháp này chính là ở chỗ giả định rằng sự xuất hiện của tất cả các từ trong văn bản đều độc lập với nhau. Với giả định này NB không sử dụng sự phụ thuộc của nhiều từ vào một chủ đề, không sử dụng việc kết hợp các từ để đưa ra phán đoán chủ đề và do đó việc tính toán NB chạy nhanh hơn các phương pháp khác với độ phức tạp theo hàm số mũ.

c. Thuật toán

Công thức

- Dựa trên định lý Bayes:

$$P(H|E) = \frac{P(E|H).P(H)}{P(E)}$$

- Trong đó:
  - H (Hypothesis) là giả thuyết và E (Evidence) là chứng cứ hỗ trợ cho giả thuyết H
  - P(E|H): xác xuất E xảy ra khi H xảy ra (xác suất có điều kiện, khả năng của E khi H đúng).
  - P(H|E): xác suất hậu nghiệm của H nếu biết E.

Áp dụng trong bài toán phân loại

- Các dữ kiện cần có:

- Một tập dữ liệu D, trong đó mỗi ví dụ học x đã được vector hóa dưới dạng một vector n chiều (x_1, x_2, ..., x_n)
- Một tập xác định các nhãn lớp: C={c_1, c_2, ..., c_m}

- Với một ví dụ (mới) z*, z* sẽ được phân vào lớp nào?

  → Áp dụng định lý Bayes để xác định phân lớp phù hợp nhất đối với z:

$$c_{MAP} = \text{argmax } P(c_i|z) = \text{argmax } P(c_i \mid z_1, z_2, ..., z_n) = \text{argmax } \frac{P(z_1,z_2,...,z_n|c_i).P(c_i)}{P(z_1,z_2,...,z_n)}$$

Vì $P(z_1, z_2, \ldots, z_n)$ là như nhau với các lớp nên chỉ cần tìm phân lớp với:

$$c_{MAP} = \text{argmax } P(z_1, z_2, ..., z_n \mid c_i).P(c_i)$$

Giả sử trong phương pháp phân loại Naïve Bayes, các thuộc tính là độc lập có điều kiện đối với các lớp thì:

$$P(z_1,z_2,...,z_n|c_i) = \prod_{j=1}^{n} P(z_j|c_i)$$

→ $c_{MAP} = \text{argmax } P(c_i).\prod_{j=1}^{n} P(z_j|c_i)$

2.3. Giới thiệu về dữ liệu sử dụng

Dữ liệu được sử dụng trong luận văn để dự đoán giới tính là tập dữ liệu có sẵn PAKDD'15 do Công ty Cổ phần FPT cung cấp (www.fpt.com.vn). Dữ liệu được lấy từ lịch sử các hoạt động xem trang Web sản phẩm của người dùng. Nhiệm vụ trong luận văn này là tái tạo lại thông tin về giới tính của người dùng từ các bản ghi xem sản phẩm. Từ thông tin người dùng đã biết giới tính và các thông tin về lịch sử truy cập web ta huấn luyện máy nhận biết giới tính của những người dùng khác khi chỉ biết lịch sử truy cập các trang web và dữ liệu danh mục của họ. Nội dung dữ liệu đã thu thập được bao gồm các đặc trưng về thời gian và các danh mục chủng loại sản phẩm.

***Định dạng dữ liệu*:** Dữ liệu đã cho được chia thành tập huấn luyện (*trainningData.csv*) và tập thử nghiệm (*testData.csv*) riêng biệt. Mỗi file này chứa 15.000 bản ghi tương ứng với nhật ký xem sản phẩm của người dùng Internet. Mỗi

bản ghi hoạt động truy cập bao gồm bốn loại thông tin được phân cách bằng dấu phẩy (,). Loại thông tin đầu tiên là ID nhật ký hoạt động. Loại thông tin thứ hai và thứ ba tương ứng với thời gian bắt đầu truy cập và thời gian kết thúc truy cập. Thông tin cuối cùng là danh sách các danh mục, sản phẩm được người dùng truy cập trong một phiên hoạt động. Thứ tự truy cập danh mục, sản phẩm trong một phiên được phân cách nhau bởi dấu chấm phẩy (;). Ngoài ra còn có 2 file nhãn Label là trainningLabel và testLabel tương ứng với tập huấn luyện và tập thử nghiệm chứa hai loại thông tin về giới tính là nam và nữ.

Ví dụ minh họa trong tập dữ liệu, mỗi bản ghi gồm có các thông tin như sau:

• Session ID

• Start time (thời gian bắt đầu phiên)

• End time (thời gian kết thúc phiên)

• Danh sách các ID sản phẩm

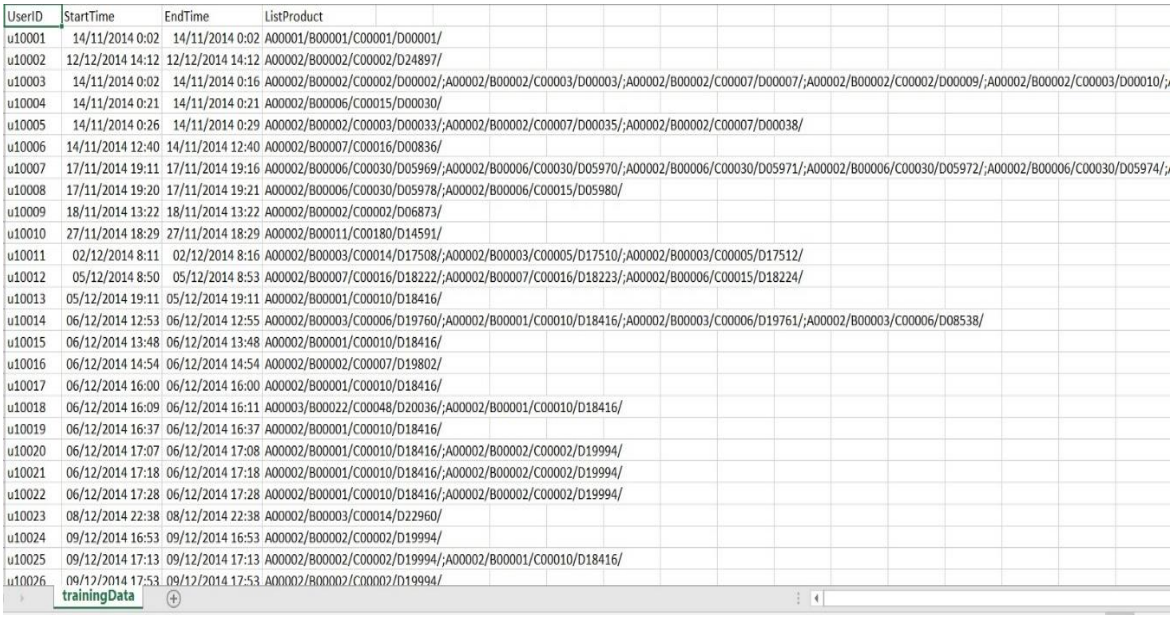
Ví dụ về một bản ghi lịch sử truy cập:

u10008, 2014/11/17 19:20:06, 2014/11/17 19:21:54,

A00001 / B00001 / C00001 / D00001 /; A00001 / B00002 / C00002 / D00002

Từ tập dữ liệu cung cấp các thông tin ở trên, ta chia thông tin thành hai loại đặc trưng chính: Đặc trưng theo mốc thời gian và đặc trưng về các danh mục chủng loại sản phẩm.

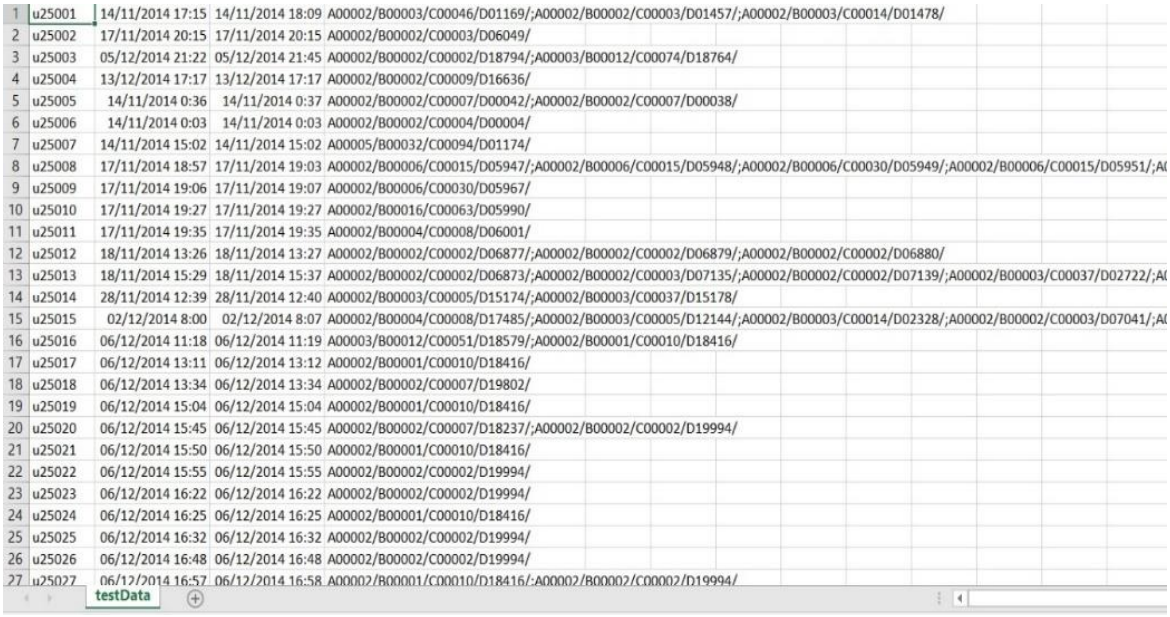
- Đặc trưng về thời gian được đại diện bằng hai loại thông tin là thời gian bắt đầu truy cập và thời gian kết thúc truy cập bao gồm các thuộc tính như Ngày, tháng, năm, giờ, phút.
- Đặc trưng về danh mục, sản phẩm trong mỗi bản ghi nhật ký hoạt động được phân thành 4 cấp theo các chữ cái bắt đầu là A, B, C, D. Thứ tự các cấp được bảo toàn trong mỗi lượt xem. Mỗi bản ghi, danh mục lớn nhất bắt đầu bằng chữ cái A, các danh mục con tiếp theo bắt đầu bằng B, C và sản phẩm xem là D.



UserID	StartTime	EndTime	ListProduct
u10001	14/11/2014 0:02	14/11/2014 0:02	A00001/B00001/C00001/D00001/
u10002	12/12/2014 14:12	12/12/2014 14:12	A00002/B00002/C00002/D24897/
u10003	14/11/2014 0:02	14/11/2014 0:16	A00002/B00002/C00002/D00002/;A00002/B00002/C00003/D00003/;A00002/B00002/C00007/D00007/;A00002/B00002/C00002/D00009/;A00002/B00002/C00003/D00010/;
u10004	14/11/2014 0:21	14/11/2014 0:21	A00002/B00006/C00015/D00030/
u10005	14/11/2014 0:26	14/11/2014 0:29	A00002/B00002/C00003/D00033/;A00002/B00002/C00007/D00035/;A00002/B00002/C00007/D00038/
u10006	14/11/2014 12:40	14/11/2014 12:40	A00002/B00007/C00016/D00836/
u10007	17/11/2014 19:11	17/11/2014 19:16	A00002/B00006/C00030/D05969/;A00002/B00006/C00030/D05970/;A00002/B00006/C00030/D05971/;A00002/B00006/C00030/D05972/;A00002/B00006/C00030/D05974/;
u10008	17/11/2014 19:20	17/11/2014 19:21	A00002/B00006/C00030/D05978/;A00002/B00006/C00015/D05980/
u10009	18/11/2014 13:22	18/11/2014 13:22	A00002/B00002/C00002/D06873/
u10010	27/11/2014 18:29	27/11/2014 18:29	A00002/B00011/C00180/D14591/
u10011	02/12/2014 8:11	02/12/2014 8:16	A00002/B00003/C00014/D17508/;A00002/B00003/C00005/D17510/;A00002/B00003/C00005/D17512/
u10012	05/12/2014 8:50	05/12/2014 8:53	A00002/B00007/C00016/D18222/;A00002/B00007/C00016/D18223/;A00002/B00006/C00015/D18224/
u10013	05/12/2014 19:11	05/12/2014 19:11	A00002/B00001/C00010/D18416/
u10014	06/12/2014 12:53	06/12/2014 12:55	A00002/B00003/C00006/D19760/;A00002/B00001/C00010/D18416/;A00002/B00003/C00006/D19761/;A00002/B00003/C00006/D08538/
u10015	06/12/2014 13:48	06/12/2014 13:48	A00002/B00001/C00010/D18416/
u10016	06/12/2014 14:54	06/12/2014 14:54	A00002/B00002/C00007/D19802/
u10017	06/12/2014 16:00	06/12/2014 16:00	A00002/B00001/C00010/D18416/
u10018	06/12/2014 16:09	06/12/2014 16:11	A00003/B00022/C00048/D20036/;A00002/B00001/C00010/D18416/
u10019	06/12/2014 16:37	06/12/2014 16:37	A00002/B00001/C00010/D18416/
u10020	06/12/2014 17:07	06/12/2014 17:08	A00002/B00001/C00010/D18416/;A00002/B00002/C00002/D19994/
u10021	06/12/2014 17:18	06/12/2014 17:18	A00002/B00001/C00010/D18416/;A00002/B00002/C00002/D19994/
u10022	06/12/2014 17:28	06/12/2014 17:28	A00002/B00001/C00010/D18416/;A00002/B00002/C00002/D19994/
u10023	08/12/2014 22:38	08/12/2014 22:38	A00002/B00003/C00014/D22960/
u10024	09/12/2014 16:53	09/12/2014 16:53	A00002/B00002/C00002/D19994/
u10025	09/12/2014 17:13	09/12/2014 17:13	A00002/B00002/C00002/D19994/;A00002/B00001/C00010/D18416/
u10026	09/12/2014 17:53	09/12/2014 17:53	A00002/B00002/C00002/D19994/

trainingData

Hình 2.6 Bộ huấn luyện – TranningData



1	u25001	14/11/2014 17:15	14/11/2014 18:09	A00002/B00003/C00046/D01169/;A00002/B00002/C00003/D01457/;A00002/B00003/C00014/D01478/
2	u25002	17/11/2014 20:15	17/11/2014 20:15	A00002/B00002/C00003/D06049/
3	u25003	05/12/2014 21:22	05/12/2014 21:45	A00002/B00002/C00002/D18794/;A00003/B00012/C00074/D18764/
4	u25004	13/12/2014 17:17	13/12/2014 17:17	A00002/B00002/C00009/D16636/
5	u25005	14/11/2014 0:36	14/11/2014 0:37	A00002/B00002/C00007/D00042/;A00002/B00002/C00007/D00038/
6	u25006	14/11/2014 0:03	14/11/2014 0:03	A00002/B00002/C00004/D00004/
7	u25007	14/11/2014 15:02	14/11/2014 15:02	A00005/B00032/C00094/D01174/
8	u25008	17/11/2014 18:57	17/11/2014 19:03	A00002/B00006/C00015/D05947/;A00002/B00006/C00015/D05948/;A00002/B00006/C00030/D05949/;A00002/B00006/C00015/D05951/;A0
9	u25009	17/11/2014 19:06	17/11/2014 19:07	A00002/B00006/C00030/D05967/
10	u25010	17/11/2014 19:27	17/11/2014 19:27	A00002/B00016/C00063/D05990/
11	u25011	17/11/2014 19:35	17/11/2014 19:35	A00002/B00004/C00008/D06001/
12	u25012	18/11/2014 13:26	18/11/2014 13:27	A00002/B00002/C00002/D06877/;A00002/B00002/C00002/D06879/;A00002/B00002/C00002/D06880/
13	u25013	18/11/2014 15:29	18/11/2014 15:37	A00002/B00002/C00002/D06873/;A00002/B00002/C00003/D07135/;A00002/B00002/C00002/D07139/;A00002/B00003/C00037/D02722/;A0
14	u25014	28/11/2014 12:39	28/11/2014 12:40	A00002/B00003/C00005/D15174/;A00002/B00003/C00037/D15178/
15	u25015	02/12/2014 8:00	02/12/2014 8:07	A00002/B00004/C00008/D17485/;A00002/B00003/C00005/D12144/;A00002/B00003/C00014/D02328/;A00002/B00002/C00003/D07041/;A0
16	u25016	06/12/2014 11:18	06/12/2014 11:19	A00003/B00012/C00051/D18579/;A00002/B00001/C00010/D18416/
17	u25017	06/12/2014 13:11	06/12/2014 13:12	A00002/B00001/C00010/D18416/
18	u25018	06/12/2014 13:34	06/12/2014 13:34	A00002/B00002/C00007/D19802/
19	u25019	06/12/2014 15:04	06/12/2014 15:04	A00002/B00001/C00010/D18416/
20	u25020	06/12/2014 15:45	06/12/2014 15:45	A00002/B00002/C00007/D18237/;A00002/B00002/C00002/D19994/
21	u25021	06/12/2014 15:50	06/12/2014 15:50	A00002/B00001/C00010/D18416/
22	u25022	06/12/2014 15:55	06/12/2014 15:55	A00002/B00002/C00002/D19994/
23	u25023	06/12/2014 16:22	06/12/2014 16:22	A00002/B00002/C00002/D19994/
24	u25024	06/12/2014 16:25	06/12/2014 16:25	A00002/B00001/C00010/D18416/
25	u25025	06/12/2014 16:32	06/12/2014 16:32	A00002/B00002/C00002/D19994/
26	u25026	06/12/2014 16:48	06/12/2014 16:48	A00002/B00002/C00002/D19994/
27	u25027	06/12/2014 16:57	06/12/2014 16:58	A00002/B00001/C00010/D18416/;A00002/B00002/C00002/D19994/

testData

Hình 2.7 Bộ thử nghiệm –TestData

Về cơ bản, tập huấn luyện TrainningData và tập thử nghiệm TestData có cùng định dạng dữ liệu giống nhau để phù hợp với công việc xử lý dữ liệu và dự đoán.

Hình 2.8 Các nhãn trong tập dữ liệu

Với tập dữ liệu ở trên được phân loại thành 2 nhãn **male** và **female** (**nam** và **nữ**), từ các thử nghiệm thực tế cho thấy, phương pháp SVM có khả năng phân loại khá tốt đối với bài toán phân loại 2 lớp và đa lớp cũng như trong nhiều ứng dụng khác (như phân loại văn bản theo chủ đề, phát hiện mặt người trong các ảnh, ước lượng hồi quy, dự đoán lỗi phần mềm...). So sánh với các phương pháp phân loại khác, khả năng phân loại của SVM là tương đương hoặc tốt hơn đáng kể. Vì những lý do đó luận văn đã chọn phương pháp này cho việc dự đoán giới tính của người dùng Internet, các thực nghiệm và đánh giá sẽ được trình bày trong chương sau.

2.4. Các dạng đặc trưng sẽ dùng trong phân lớp

Trong khuôn khổ nội dung nghiên cứu, tác giả phát triển một hệ thống nhằm lấy dữ liệu từ các bản ghi xem sản phẩm của người dùng với giới tính đã được biết đến, với các đặc trưng và nhãn lớp để tạo ra tập dữ liệu huấn luyện. Từ đó xây dựng mô hình từ các tập dữ liệu huấn luyện và sử dụng phương pháp phân loại để có thể dự đoán giới tính của người dùng chưa biết là nam hay nữ dựa trên các hoạt động xem sản phẩm của người đã biết. Trong các phần tiếp theo, tác giả mô tả các đặc trưng và kỹ thuật sử dụng để phân loại giới tính.

2.4.1. Dạng đặc trưng theo mốc thời gian

Trong tập dữ liệu PAKDD'15, đặc trưng về thời gian được biểu diễn thành hai loại thông tin là thời gian bắt đầu truy cập và thời gian kết thúc truy cập. Thời gian trong ngày, ngày trong tháng, tháng trong năm, thời gian bắt đầu xem, thời gian xem trong một bản ghi lịch sử truy cập, ... là những yếu tố có thể được sử dụng để dự đoán giới tính của một người dùng Internet.

StartTime	EndTime
14/11/2014 0:02	14/11/2014 0:02
12/12/2014 14:12	12/12/2014 14:12
14/11/2014 0:02	14/11/2014 0:16
14/11/2014 0:21	14/11/2014 0:21
14/11/2014 0:26	14/11/2014 0:29
14/11/2014 12:40	14/11/2014 12:40
17/11/2014 19:11	17/11/2014 19:16
17/11/2014 19:20	17/11/2014 19:21
18/11/2014 13:22	18/11/2014 13:22
27/11/2014 18:29	27/11/2014 18:29

Hình 2.9 Thông tin về thời gian truy cập

Các đặc điểm của một bản ghi có chứa thông tin về thời gian được cụ thể như sau:

- Ngày truy cập: Ngày người dùng bắt đầu và kết thúc 1 phiên truy cập, được hiển thị từ ngày số 01 cho đến ngày số 31 trong một tháng.
- Tháng: Hiển thị tháng người dùng truy cập (12 tháng trong một năm).
- Năm: Hiển thị năm truy cập.
- Giờ: Hiển thị thông tin giờ truy cập trong một ngày (24 giờ/ngày).
- Phút: Hiển thị thông tin phút truy cập trong một giờ (60 phút/giờ).
- Thời gian trung bình một phiên truy cập: Thời gian trung bình của một phiên truy cập từ lúc bắt đầu cho đến lúc kết thúc.

Các đặc trưng theo mốc thời gian là quá trình theo dõi thời gian truy cập của người dùng mạng. Thông thường, thời gian truy cập của phiên hoạt động cao sẽ cho thấy người dùng đang mất nhiều thời gian hơn để xem các danh mục và chủng loại sản phẩm và các sản phẩm liên quan. Yếu tố này thường xảy ra với người dùng là nữ giới bởi các thói quen mua sắm và tìm hiểu thông tin của mình.

Trong tập dữ liệu ta có thể thấy các đặc trưng theo mốc thời gian là một trong những mảnh ghép trong bức tranh toàn cảnh cho việc phân loại giới tính dựa trên lịch sử truy cập.

Ngoài ra việc sử dụng các đặc trưng về thời gian dựa trên giả định rằng những nam giới và nữ giới có những hoạt động truy cập với thời điểm khác nhau khi lướt web. Ví dụ, nữ giới thường phải chuẩn bị bữa ăn tối trong thời gian 17h00 – 20h00, vì vậy hoạt động truy cập của nữ giới có thể bắt đầu sau nam giới tại thời điểm từ lúc 20h00 trở đi.

2.4.2. Dạng đặc trưng về danh mục và chủng loại sản phẩm

Dữ liệu thu thập được của tập huấn luyện gồm có 15000 bản ghi nhật ký hoạt động truy cập.

Các bản ghi này đều ghi lại quá trình truy cập sản phẩm của người dùng mạng. Quá trình này được bắt đầu từ danh mục cấp 1, tiếp đến là các danh mục con cấp 2, từ danh mục con cấp 2 chuyển đến danh mục con cấp 3 và từ danh mục con cấp 3 truy cập đến đích cuối là sản phẩm. Dựa vào tập dữ liệu và nhật ký hoạt động của người dùng, ta có thể phân cấp quá trình truy cập thành 4 loại đặc trưng của danh mục, sản phẩm theo các cấp tương ứng là A, B, C, D. Trong đó số lượng Danh mục A (cấp 1) là 11, số lượng danh mục B (cấp 2) là 86, số lượng danh mục C (cấp 3) là 383 và Sản phẩm D là 21881.

***Bảng 2.1.* Tóm tắt các đặc trưng dựa trên danh mục & sản phẩm**

Tên đặc trưng	Miêu tả
Danh mục A_ID	ID Danh mục cấp một
Danh mục B_ID	ID Danh mục cấp hai
Danh mục C_ID	ID Danh mục cấp ba
Sản phẩm D_ID	ID Sản phẩm theo các danh mục

Các danh mục chủng loại sản phẩm chứa các đặc trưng và yếu tố quan trọng cho việc phân loại giới tính của người dùng. Phần lớn các danh mục sản phẩm được

người dùng truy cập thể hiện sự quan tâm, sở thích dựa trên các yếu tố về giới tính. Ví dụ như các sản phẩm liên quan đến thời trang, mỹ phẩm thường được xem bởi người dùng là giới tính nữ, các sản phẩm liên quan đến thể thao, công nghệ thì hay được truy cập bởi nam giới.

Từ tập dữ liệu đã cho, tác giả đã phân tích số lượt truy cập của nam giới và nữ giới đối với các danh mục, sản phẩm thu được từ các nhật ký xem sản phẩm của họ. Số liệu thống kê truy cập được phân chia theo từng cấp danh mục chủng loại sản phẩm được thể hiện trong hình 2.10.

Danh mục A (cấp 1)

proid	female	male	total
A00001	557	1792	2349
A00002	8726	1109	9835
A00003	2210	228	2438
A00005	271	57	328
A00010	35	40	75
A00008	15	21	36
A00004	74	133	207
A00007	27	13	40
A00006	139	30	169
A00009	28	16	44
A00011	54	22	76

Danh mục B (cấp 2)

proid	female	male	total
B00001	1544	775	2319
B00002	3553	336	3889
B00006	265	22	287
B00007	881	114	995
B00011	73	37	110
B00003	2030	267	2297
B00022	495	60	555
B00005	450	45	495
B00051	20	7	27
B00004	800	400	1200
B00044	31	21	52

Danh mục C (cấp 3)

proid	female	male	total
C00001	23	132	155
C00002	854	84	938
C00012	186	117	303
C00011	47	7	54
C00004	335	56	391
C00003	813	74	887
C00007	1619	133	1752
C00015	152	13	165
C00016	349	40	389
C00030	132	12	144
C00180	17	6	23

Sản phẩm D (cấp 4)

proid	female	male	total
D00001	1	1	2
D24897	5	0	5
D00009	1	0	1
D00007	1	0	1
D00003	1	0	1
D00002	1	0	1
D00017	1	0	1
D00010	1	0	1
D00014	1	0	1
D00011	1	0	1
D00012	1	0	1

Hình 2.10 Số liệu thống kê truy cập theo các cấp danh mục chủng loại sản phẩm

Từ các thống kê theo các cấp danh mục, sản phẩm trong tập dữ liệu cho thấy sự phân loại một cách rõ rệt giữa nam giới và nữ giới thông qua hoạt động truy cập các danh mục, sản phẩm dựa theo mức độ và tỉ lệ truy cập. Điều này cho thấy các đặc trưng về danh mục và chủng loại sản phẩm là loại đặc trưng chính trong việc xác định giới tính.

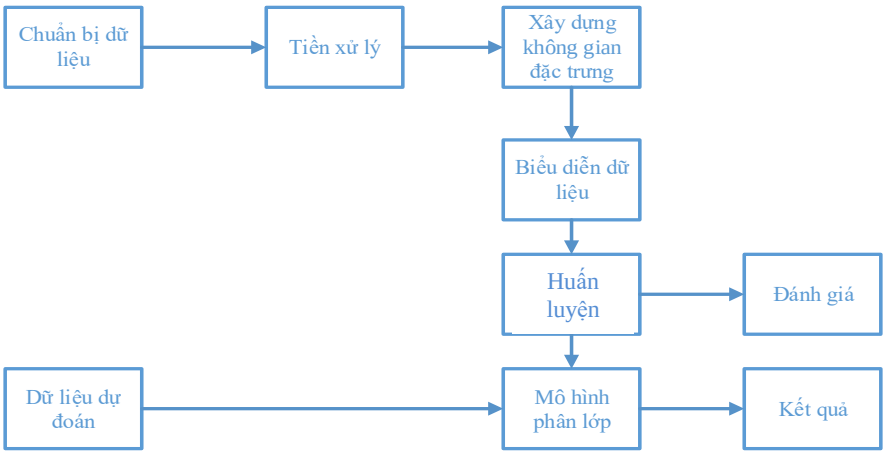
Từ đó ta sẽ xây dựng các mô hình huấn luyện cho bộ dữ liệu này và sử dụng các đặc trưng này để dự đoán kết quả đầu ra cho bộ dữ liệu thử nghiệm. Với việc dữ liệu đầu vào chứa nhiều danh mục chủng loại sản phẩm, ta dự đoán kết quả cho từng

cấp riêng biệt sau đó kết hợp các cấp và thuộc tính lại với nhau để tính kết quả cuối cùng.

2.5. Xây dựng mô hình dự đoán giới tính dựa trên học máy có giám sát

Đối với bài toán xác định giới tính thì mỗi phiên truy cập của người dùng Internet sẽ là cơ sở để phân loại giới tính là nam hay nữ.

Trong luận văn, với mục đích sử dụng phương pháp SVM để phân loại giới tính do đã được nhiều công trình đánh giá có độ chính xác cao trong phân lớp văn bản. Tuy nhiên, để đánh giá với đặc điểm của dạng dữ liệu ngắn, tác giả sẽ sử dụng phương pháp đếm số lần xuất hiện của đặc trưng có trong từng bàn ghi để phân loại, từ đó so sánh và rút ra kết luận cuối cùng về lựa chọn phương pháp phân loại cho bài toán.



Hình 2.11 Mô hình phân loại dự đoán giới tính người dùng Internet

Đối với mô hình trên, việc đánh giá mô hình huấn luyện là rất quan trọng, nó được dùng làm căn cứ để hiệu chỉnh lại số liệu huấn luyện, xây dựng không gian đặc trưng nhằm tăng độ chính xác, cải thiện tốc độ tính toán.

2.5.1. Tiền xử lý dữ liệu

Đây là giai đoạn "làm sạch" dữ liệu, bao gồm các bước sau:

- Loại bỏ trường không cần thiết
- Loại bỏ các kí tự đặc biệt biệt ([ ], [,], [:], [;], [/])

- Tách các đặc trưng

2.5.2. Biểu diễn dữ liệu

Để có thể tiến hành thực nghiệm và đánh giá kết quả phân loại giới tính dựa trên tập dữ liệu lịch sử truy cập ta cần phải chuẩn hóa dữ liệu và áp dụng phương pháp cho tập dữ liệu đã có nhằm tiết kiệm không gian lưu trữ và gia tăng tốc độ xử lý. Ví dụ minh họa được trình bày ở dưới đây

Ta có 4 bản ghi hoạt động truy cập của người dùng lấy từ tập dữ liệu lịch sử truy cập:

1. u10150,07/12/2014 20:30,07/12/2014 20:32, A00002/B00003/C00006/D06023/;A00002/B00003/C00006/D15152/,female
2. u10151,07/12/2014 20:53,07/12/2014 20:53, A00002/B00003/C00006/D02909/;A00002/B00002/C00007/D10813/,female
3. u10152,20/12/2014 22:43,20/12/2014 22:43, A00002/B00003/C00006/D13680/,male
4. u10153,14/11/2014 0:36,14/11/2014 0:39, A00002/B00004/C00018/D00043/;A00002/B00004/C00018/D00047/,male

Bước 1: Ta loại bỏ thuộc tính id vì thuộc tính này không dùng trong mô hình.
Bước 2: Tạo một danh sách các thuộc tính theo thứ tự bắt đầu từ các thuộc tính ngày, tháng, năm, giờ truy cập, phút sau đó là các thuộc tính phân cấp các danh mục A, B, C, D.

STT	Thuộc tính
1	Ngày
2	Tháng
3	Năm
4	Giờ bắt đầu truy cập
5	Thời gian truy cập
6	A0002
7	B0002
8	B0003

9	B0004
10	C0006
11	C0007
12	C00018
13	D00043
14	D00047
15	D02909
16	D06023
17	D10813
18	D13680
19	D15152

Bảng 2.2 Thứ tự các thuộc tính

Bước 3: Định dạng thuộc tính thời gian và đếm số lần xuất hiện của các thuộc tính danh mục, sản phẩm có trong bản ghi (Giá trị, vị trí)

7,1 12,2 2014,3 20,4 2,5 2,6 2,8 2,10 1,16 1,19 female

7,1 12,2 2014,3 20,4 0,5 2,6 2,8 1,10 1,11 1,15 1,17 female

20,1 12,2 2014,3 22,4 0,5 1,6 1,8 1,10 1,18 male

14,1 11,2 2014, 3 0,4 3,5 2,6 2,9 2,12 1,13 1,14 male

Các bước tiến hành xử lý và huấn luyện dữ liệu cụ thể đối với mô hình dự đoán giới tính người dùng Internet với dữ liệu lịch sử truy cập sẽ được trình bày chi tiết trong Chương 3.

2.6. Kết luận chương

Chương này tác giả đã giới thiệu chi tiết thuật toán Máy vector hỗ trợ SVM và giới thiệu về tập dữ liệu PAKDD'15 sử dụng trong luận văn. Ngoài ra, tác giả mô tả cụ thể các dạng đặc trưng có trong dữ liệu. Dạng đặc trưng thứ nhất là dạng được trưng theo dấu thời gian truy cập của người dùng Internet. Dạng đặc trưng thứ hai là dạng đặc trưng theo danh mục và sản phẩm. Từ hai dạng đặc trưng này ta có thể khai phá và phân loại giới tính cho tập học và xây dựng mô hình dự đoán giới tính dựa trên phương pháp học máy SVM.

CHƯƠNG 3: THỬ NGHIỆM VÀ ĐÁNH GIÁ

3.1. Mô tả dữ liệu

Dữ liệu PAKDD'15 do Tập đoàn FPT cung cấp đã được giới thiệu từ Chương 2. Dữ liệu được chia thành các bộ huấn luyện và kiểm tra. Mỗi bộ chứa 15.000 bản ghi tương ứng với hoạt động truy cập xem sản phẩm của người dùng Internet. Để bắt đầu quá trình phân loại, ta cần xây dựng một tập huấn luyện theo đúng định dạng.

Sau các bước tiền xử lý, lịch sử truy cập được biểu diễn dưới dạng:

<$label_i$> <$index_1$>:<$value_1$> <$index_2$>:<$value_2$> ... <$index_n$>:<$value_n$>

Trong đó:

- $label_i$ là giá trị đích của tập huấn luyện. Đối với việc phân loại, nó là một số nguyên xác định một lớp, nhãn. Trong bài toán dự đoán giới tính ở đây thì label sẽ có hai giá trị là -1 nếu là nữ và 1 nếu là nam.
- $index_i$ là một số nguyên bắt đầu từ 1. Cụ thể trong bài toán phân loại nó đại diện cho các đặc trưng.
- $value_i$ là trọng số của index. Nếu value = 0 thì không cần phải ghi.

Ví dụ một bản ghi nhật ký hoạt động:

"14/11/2014 0:02 14/11/2014 0:02 A00001/B00001/C00001/D00001 female" sẽ được chuyển thành như sau:

- 1 1:14 2:10 3:114 4:0 5:0 6:1 17:1 103:1 486:1

3.2. Các tiêu chuẩn đánh giá

Việc đánh giá một giải thuật học máy cho bộ dữ liệu là rất quan trọng, nó cho phép đánh giá được độ chính xác của các kết quả phân lớp và so sánh các giải thuật học máy khác nhau.

Các tiêu chuẩn đánh giá thường phụ thuộc vào các yếu tố như sau:

- Tập dữ liệu càng lớn thì độ chính xác càng tốt.
- Tập kiểm thử cần lớn thì việc đánh giá càng chính xác.

- Vấn đề là rất khó (ít khi) có thể có được các tập dữ liệu (rất) lớn.

Để đánh giá một giải thuật máy học một số chỉ số thông dụng được sử dụng. Giả sử như bộ phân lớp có 2 lớp là lớp âm (negative) và lớp dương (positive) thì các chỉ số được định nghĩa như sau:

- Số đúng dương (TP- True positive): số phần tử dương được phân loại dương
- Số sai âm (FN - False negative): số phần tử dương được phân loại âm
- Số đúng âm (TN- True negative): số phần tử âm được phân loại âm
- Số sai dương (FP - False positive): số phần tử âm được phân loại dương
- Độ chính xác dùng cho đo lường (Precision) = TP/(TP + FP)
- Độ bao phủ (Recall) = TP/(TP + FN)
- Độ đo F-Score = 2*Precision*Recall/(Precision + Recall)
- Độ chính xác dùng cho kết quả (Accuracy) = (TP + TN)/(TP+FP+TN+FN)

Precision và recall

Hay còn gọi là **Độ chính xác** và **Độ bao phủ**

Precision: trong tập tìm được thì bao nhiêu cái (phân loại) đúng. Dành cho việc đo lường.

Recall: trong số các tồn tại, tìm ra được bao nhiêu cái (phân loại).

Thường được sử dụng để đánh giá các hệ thống phân loại

- Precision đối với lớp c_i

$$\text{Precision} = \frac{tp}{tp + fp}$$

Tổng số các ví dụ thuộc lớp c_i được phân loại chính xác chia cho tổng số các ví dụ được phân loại vào lớp c_i

- Recall đối với lớp c_i

$$\text{Recall} = \frac{tp}{tp + fn}$$

Tổng số các ví dụ thuộc lớp c_i được phân loại chính xác chia cho tổng số các ví dụ thuộc lớp c_i

Recall cũng được gọi là True Positive Rate hay Sensitivity (độ nhạy), và precision cũng được gọi là Positive predictive value (PPV); ngoài ra, ta có các độ đo khác như True Negative Rate và Accuracy (Độ chính xác dành cho kết quả) . True Negative Rate cũng được gọi là Specificity.

F-Score

Tiêu chí đánh giá F_1 là sự kết hợp của 2 tiêu chí đánh giá *Precision* và *Recall*

$$F = 2 \cdot \frac{\mathrm{precision} \cdot \mathrm{recall}}{\mathrm{precision} + \mathrm{recall}}$$

F-Score là một **trung bình điều hòa (harmonic mean)** của các tiêu chí *Precision* và *Recall*

- F-Score có xu hướng lấy giá trị gần với giá trị nào nhỏ hơn giữa 2 giá trị Precision và Recall
- F-Score có giá trị lớn nếu cả 2 giá trị Precision và Recall đều lớn

Độ chính xác kết quả của thuật toán

Chúng ta không thể khẳng định một phương pháp xác định giới tính cụ thể nào là chính xác hoàn toàn. 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 độ chính xác của thuật toán, từ đó áp dụng thuật toán đó vào việc phân lớp nhãn.

Độ chính xác có thể được tính theo công thức:

Công thức đánh giá:

$$\text{Độ } chính\ xác = \frac{Số\ bản\ ghi\ dự\ đoán\ đúng}{Tổng\ số\ bản\ ghi}$$

3.3. Phương pháp thực nghiệm

Trong phạm vi của luận văn, tác giả sử dụng các bộ công cụ phân lớp để cài đặt mô hình thực nghiệm và đánh giá kết quả. Các bước tiến hành được bắt đầu từ

việc xử lý dữ liệu kết xuất và đưa ra các định dạng của công cụ đó, sau đó huấn luyện và kiểm thử để đưa ra kết quả phân loại giới tính.

3.3.1 Công cụ dùng để phân lớp

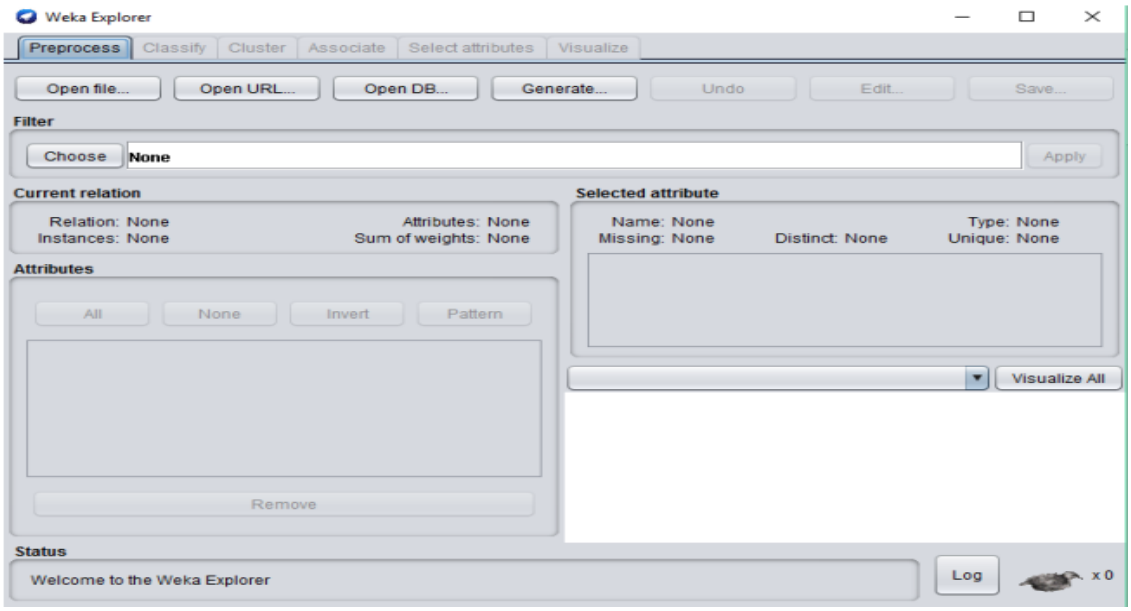
Để dự đoán giới tính bằng kỹ thuật học máy SVM, tác giả sử dụng bộ thư viện hỗ trợ phân lớp LibSVM_Tool [11] và bộ phần mềm học máy Weka [21] và chuẩn hóa bộ dữ liệu huấn luyện và kiểm thử theo định dạng của hai bộ công cụ này.

LibSVM_Tool:

LibSVM là bộ công cụ hỗ trợ phân lớp theo phương pháp SVM (Support Vector Machines). Đây là một bộ công cụ đơn giản, dễ sử dụng và hiệu quả.

Weka:

Weka – Waikato Environment for Knowledge Analysis, là bộ phần mềm học máy, mã nguồn mở, do đại học Waikato phát triển bằng Java, nhằm phục vụ cho các nhiệm vụ chuyên về khai phá dữ liệu. Weka chứa các công cụ phục vụ cho tiền xử lý dữ liệu, phân loại, hồi quy, phân cụm, các luật liên quan và trực quan hóa. Nó cũng phù hợp cho việc phát triển, xây dựng các mô hình học máy và có khả năng chạy được trên nhiều hệ điều hành khác nhau như Windows, Mac, Linux.



Hình 3.2 Bộ công cụ Weka

Những tính năng vượt trội trong Weka có thể kể đến là:

- Hỗ trợ các thuật toán học máy (machine learning) và khai phá dữ liệu
- Trực quan hóa, dễ dàng xây dựng các ứng dụng thực nghiệm
- Do sử dụng JVM nên Weka độc lập với môi trường

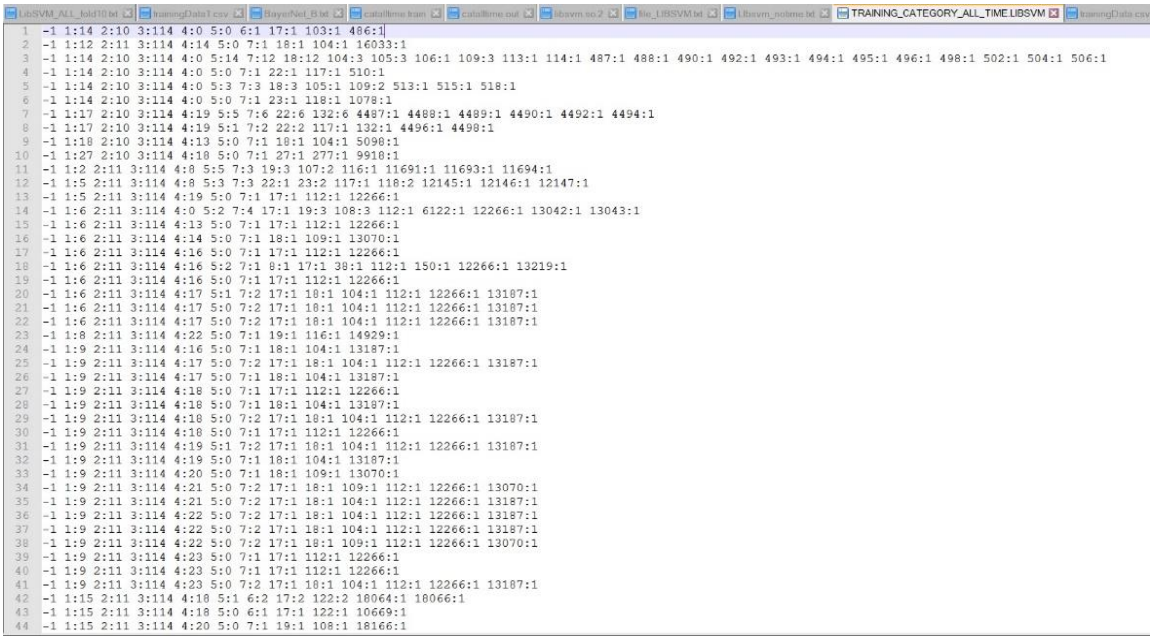
3.3.2 Xây dựng dữ liệu huấn luyện và kiểm tra

Tập dữ liệu huấn luyện và kiểm thử dự đoán giới tính chứa các thông tin: Ngày, tháng, năm, giờ truy cập, thời gian xem, danh mục cấp A, danh mục con cấp B, danh mục con cấp C và sản phẩm D. Để định dạng dữ liệu, chúng ta cần biết LibSVM_Tool và Weka học thế nào. Trong học máy nó thường được gọi là "Bộ thuộc tính". Trong trường hợp phân loại giới tính chúng ta xem mỗi danh mục sản phẩm và thời gian truy cập như một thuộc tính và được sắp xếp theo thứ tự bắt đầu từ 1 cho đến thuộc tính cuối cùng.

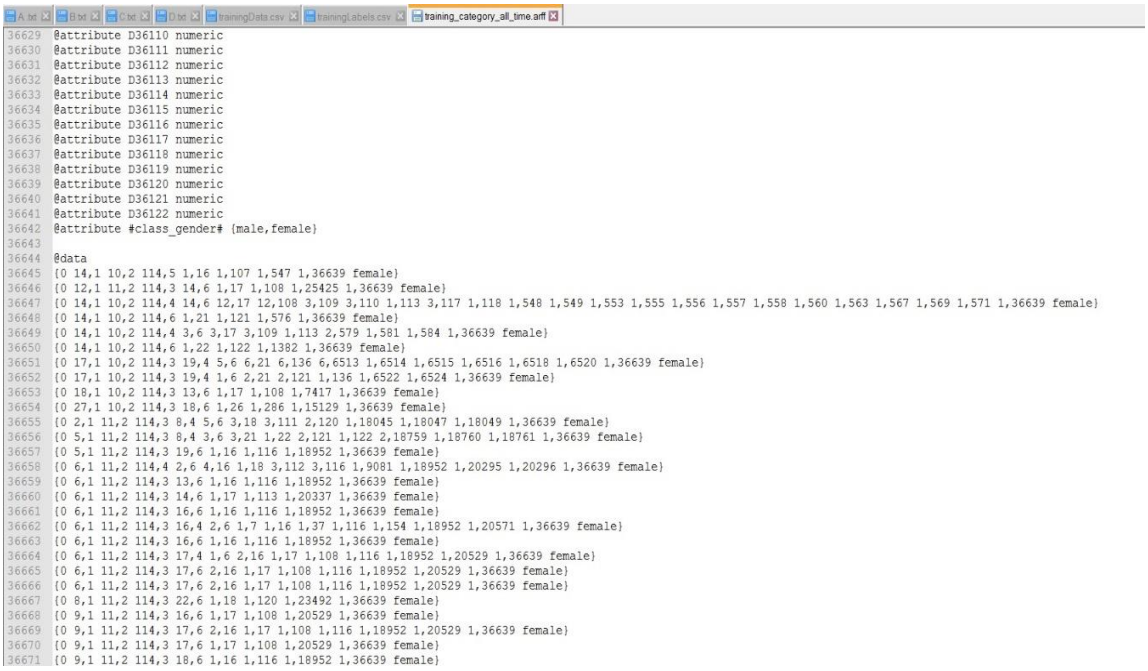
Tác giả khai thác những đặc điểm này và đếm số lần xuất hiện của thuộc tính trong mỗi bản ghi tương ứng với giới tính đã cho, sau đó đưa ra tập dữ liệu huấn luyện theo định dạng dữ liệu của hai bộ công cụ. Để đưa ra các tập dữ liệu đã được xử lý tác giả tạo 1 project Java tên là **Gender_Prediction_Network**.

Input: Là tập dữ liệu huấn luyện (*trainningData.csv*) và tập dữ liệu thử nghiệm (*TestData.csv*).

Output: Là file định dạng *.arff (theo Weka) và file *.txt (theo LibSVM_Tool) có chứa tập dữ liệu huấn luyện và tập dữ liệu thử nghiệm kèm theo nhãn (xóa các dấu cách thừa, dấu phẩy ",", dấu chấm phẩy ";" và dấu gạch ngang "/") với mỗi dòng là một bản ghi lịch sử truy cập.

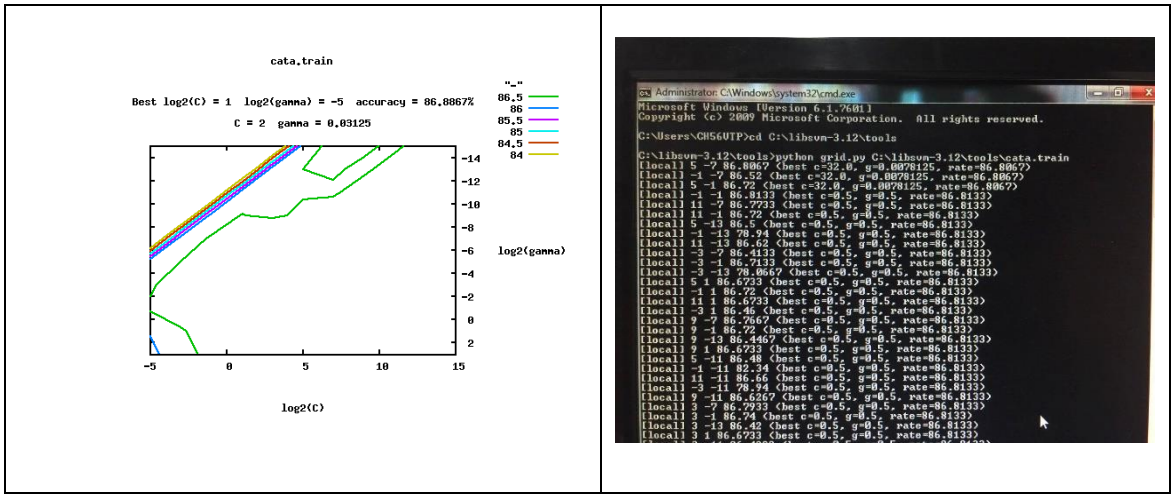


Hình 3.3 Dữ liệu theo định dạng LibSVM_Tool



Hình 3.4 Dữ liệu theo định dạng Weka

Sau khi xử lý dữ liệu ta bắt đâu tiến hành huấn luyện dữ liệu. Đầu tiên, để đánh giá hiệu quả của phương pháp SVM, tác giả sử dụng phương thức kiểm tra chéo (10-fold cross-validation) trên tập học cùng với công cụ **grid.py** trong LibSVM. Công cụ này sẽ tìm ra hai tham số tối ưu sao cho kết quả phân loại khi áp dụng hai tham số này sẽ đưa ra tỉ lệ cao nhất.



Hình 3.5 Sử dụng grid.py tool lựa chọn tham số tối ưu cho C-SVM classification sử dụng Kernel RBF

Khi sử dụng grid.py ta tìm được hai tham số tối ưu Tham số C (cost) và gamma cho các mô hình.

Bảng 3.1 **Hai tham số tối ưu cho các mô hình huấn luyện**

Mô hình\ Tham số	C	Gamma
A	2.0	0.03125
B	32.0	0.0078125
C	8.0	0.03125
D	8.0	0.03125
ALL	32.0	0.0078125

Hai tham số tối ưu này sẽ được kết hợp trong quá trình huấn luyện với Cross-Validation. Một tập con sẽ được giữ lại để làm tập dữ liệu kiểm tra, còn 9 tập còn lại sẽ được sử dụng để huấn luyện SVM, sau đó SVM này sẽ được dùng để dự đoán trên tập dữ liệu kiểm tra. Quá trình này sẽ được lặp đi lặp lại 10 lần sao cho tất cả các tập con đều sẽ được chọn làm tập dữ liệu kiểm tra. Trong quá trình thực nghiệm ta chia ra thành 4 tập dữ liệu rời rạc theo các cấp danh mục và chủng loại sản phẩm A, B, C, D và 1 tập dữ liệu chính lấy tên là ALL bao gồm tất cả các đặc trưng của 4 tập dữ liệu rời rạc để huấn luyện để so sánh kết quả phân loại giới tính giữa các mô hình đặc trưng rời rạc và mô hình đặc trưng tổng thể. Các tiêu chuẩn đánh giá sẽ được tính

trung bình từ các giá trị có được từ 10 lần lặp đó. Kết quả phân loại giới tính của tập dữ liệu lịch sử truy cập theo các tập dữ liệu được trình bày trong mục 3.4.

3.4. Kết quả thực nghiệm

Kết quả thực nghiệm sử dụng 5 mô hình theo tập dữ liệu đã chia và 4 tiêu chuẩn đánh giá để đưa ra hiệu quả mô hình học máy SVM cho việc phân loại giới tính. Các mô hình được đánh giá được ghi lại sau khi sử dụng phương pháp Cross validation hay còn gọi là k-fold Cross validation để thực nghiệm và huấn luyện. Kết quả thu được cho thấy khả năng phân loại có độ chính xác cao nhưng giảm dần với các tập dữ liệu rời rạc từ tập dữ liệu danh mục A cho đến chủng loại sản phẩm D.

Lý do bởi vì độ nhiễu dữ liệu khá lớn đối với các mô hình theo tập dữ liệu rời rạc, tập dữ liệu càng nhiều đặc tính thì độ nhiễu càng lớn. Kết quả cụ thể với các mô hình theo tập dữ liệu rời rạc được thu thập trong các bảng dưới đây:

Bảng 3.2 **Kết quả thu được với tập dữ liệu A**

NHÃN	SVM Với A			
	Precision	**Recall**	**F-Score**	**Accuracy**
Nam	77.4 %	55.3 %	64.5 %	**86.51 %**
Nữ	88.2 %	95.4 %	91.7 %	
Weighted Avg	85.8 %	86.5 %	85.6 %	

Bảng 3.3 **Kết quả thu được với tập dữ liệu B**

NHÃN	SVM Với B			
	Precision	**Recall**	**F-Score**	**Accuracy**
Nam	75.8 %	37.5 %	50.2 %	**83.48 %**
Nữ	84.4 %	96.6 %	90.1 %	
Weighted Avg	82.5 %	83.5 %	81.2 %	

Bảng 3.4 **Kết quả thu được với tập dữ liệu C**

NHÃN	SVM Với C			
	Precision	**Recall**	**F-Score**	**Accuracy**
Nam	73.9 %	40.6 %	52.4 %	**83.63 %**
Nữ	85 %	95.9 %	90.1 %	
Weighted Avg	82.5 %	83.6 %	81.7 %	

Bảng 3.5 **Kết quả thu được với tập dữ liệu D**

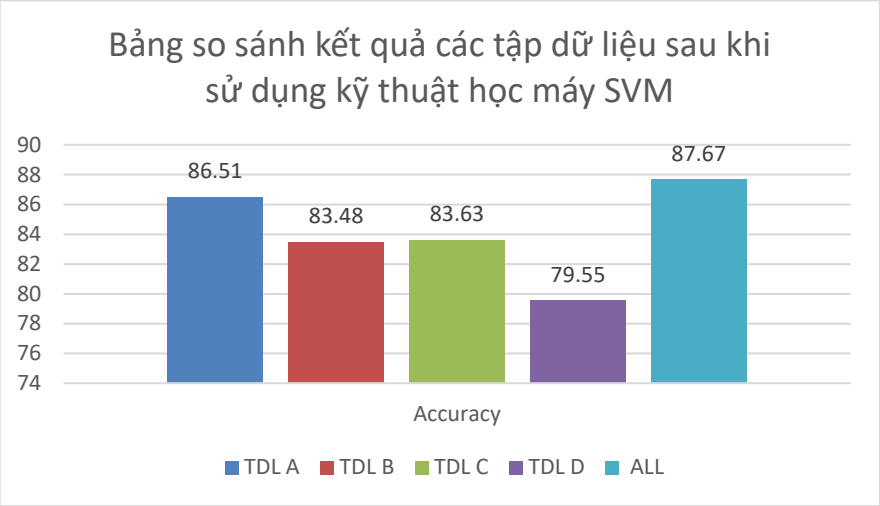
NHÃN	SVM Với D			
	Precision	**Recall**	**F-Score**	**Accuracy**
Nam	82.7 %	9.9 %	17.7 %	**79.55 %**
Nữ	79.5 %	99.4 %	88.3 %	
Weighted Avg	80.2 %	79.5 %	72.6 %	

Tại Bảng 3.6 là bảng kết quả chính thu được từ mô hình học máy SVM khi sử dụng và kết hợp tất cả các đặc trưng cùng với mô hình rời rạc của tập dữ liệu đã chuẩn hóa và đưa ra các tiêu chí đánh giá. So với kết quả của 4 tập dữ liệu rời rạc ở trên, tỉ lệ dự đoán khi kết hợp các đặc trưng lại với nhau mang đến tỉ lệ chính xác là 87.67 %. Từ các thực nghiệm trên cho thấy, SVM có độ phân lớp chính xác rất cao có thể đáp ứng được yêu cầu mà bài toán dự đoán giới tính đề ra.

Bảng 3.6 **Kết quả thu được từ tập dữ liệu ALL**

NHÃN	SVM với tất cả các đặc trưng (ALL)			
	Precision	**Recall**	**F-Score**	**Accuracy**
Nam	79.4 %	59.3 %	67.9 %	**87.67 %**
Nữ	89.3 %	95.7 %	92.4 %	
Weighted Avg	87.1 %	87.7 %	87 %	

Biểu đồ thể hiện độ chính xác của các tập dữ liệu sau khi sử dụng SVM:



3.5. So sánh với một số phương pháp khác

Để đánh giá thêm hiệu suất của mô hình dự đoán, luận văn đã tiến hành huấn luyện tập dữ liệu trên các mô hình học máy phổ biến khác là NaiveBayes và RandomTree, kết quả cụ thể được đưa ra trong bảng 3.7, 3.8.

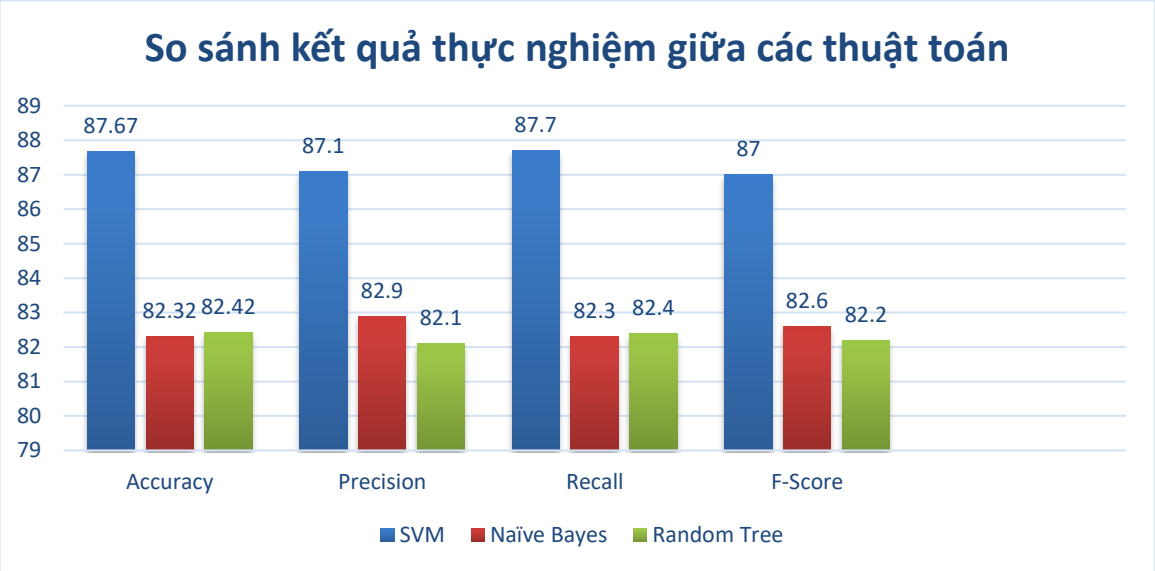
Bảng 3.7 **Kết quả thu được từ mô hình Naïve Bayes**

NHÃN	NaiveBayes			
	Precision	**Recall**	**F-Score**	**Accuracy**
Nam	59 %	64.3 %	61.5 %	82.32 %
Nữ	89.7 %	87.4 %	88.5 %	
Weighted Avg	**82.9 %**	**82.3 %**	**82.6 %**	

Bảng 3.8 **Kết quả thu được từ mô hình Random Tree**

NHÃN	Random Tree			
	Precision	**Recall**	**F-Score**	**Accuracy**
Nam	60.7 %	57 %	58.8 %	82.42 %
Nữ	88.1 %	89.6 %	88.8 %	
Weighted Avg	**82.1 %**	**82.4 %**	**82.2 %**	

Để dễ hình dung hơn thì chúng ta xem biểu đồ sau:



Nhận xét: Dựa vào bảng 3.6, 3.7, 3.8 tổng hợp kết quả phân loại giới tính trên các mô hình SVM, NaiveBayes, RandomTree ta nhận thấy được NaiveBayes cho kết quả thấp nhất khi phân loại mặc dù khả năng đưa ra độ chính xác khá cao với Accuracy = 82.32 % nhưng thực tế vẫn chưa tối ưu. Random Tree khá hơn nhưng tỉ lệ phân loại cũng chỉ nhiều hơn so với NaiveBayes là 0,1 %. Với SVM, tỉ lệ phân loại chính xác cao nhất so với 2 mô hình còn lại Accuracy = 87.67 %, ngoài ra các thông số Precision, Recall, F-Score cũng đều đưa ra tỉ lệ vượt trội. Kết quả này cho phép ta tin tưởng vào khả năng xử lý hiệu quả của mô hình học máy SVM cho vấn đề phân loại và xác định giới tính với dữ liệu có số chiều lớn.

3.6. Kết luận chương

Trong chương này, tác giả đã nêu ra cách thức mô tả dữ liệu và chuẩn hóa dữ liệu Dữ liệu PAKDD'15 được sử dụng trong luận văn. Biểu diễn đặc trưng về danh mục sản phẩm và đặc trưng về thời gian truy cập của người dùng Internet để tạo ra dữ liệu huấn luyện để đưa vào các bộ công cụ hỗ trợ phân lớp cụ thể là LibSVM_Tool và Weka. Kết quả thực nghiệm được thể hiện trong 4 mô hình phân loại nhỏ và 1 mô hình phân loại tổng thể kết hợp với 4 tiêu chí đánh giá để đưa ra mức độ phù hợp của kỹ thuật học máy SVM khi áp dụng vào bài toán.

Do hạn chế về mặt thời gian, nên việc so sánh giữa các mô hình kỹ thuật học máy khác tác giả chỉ đưa ra mô hình SVM với tất cả các đặc trưng và 2 mô hình huấn luyện là Naïve Bayes và Random Tree. Các kết quả thu được thể hiện trong Bảng 3.6, Bảng 3.7 và Bảng 3.8.

Kết quả thử nghiệm và đánh giá được tiến hành sau khi đã huấn luyện bộ dữ liệu theo 3 mô hình. Riêng trường hợp mô hình SVM thì có thêm công cụ grid.py giúp lựa chọn các tham số tối ưu. Kết quả thu được cho thấy SVM cho kết quả phân loại tốt hơn so với NaiveBayes và Random Tree với độ chính xác đạt trên 87 %.

KẾT LUẬN

1. Kết quả đạt được

Luận văn tiến hành nghiên cứu giải quyết bài toán dự đoán giới tính người dùng Internet dựa trên lịch sử truy cập. Từ việc giải quyết bài toán này sẽ giúp cho chúng ta tiến gần hơn đến sự thông minh của thế giới ảo, giúp quản lý tốt hơn hệ thống thông tin ngập tràn những nội dung không mong muốn. Bài toán là nền tảng cho nhiều ứng dụng quan trọng thực tế như quảng cáo nhắm mục tiêu, các hệ thống cung cấp tiếp thị dịch vụ thương mại tới đúng người dùng.

Những kết quả chính mà luận văn đạt được:

- Trình bày một cách khái quát, tổng quan nhất và nêu lên ý nghĩa, vai trò quan trọng của bài toán xác định giới tính người dùng Internet dựa trên lịch sử truy cập.
- Khảo sát nghiên cứu 3 phương pháp dự đoán giới tính đã có
- Đưa ra các đặc trưng của tập dữ liệu lịch sử cho bài toán phân loại giới tính.
- Nghiên cứu và tìm hiểu về thuật toán Support Vector Machine trên hai lớp và nhiều lớp
- Nghiên cứu và làm thực nghiệm khi áp dụng Support Vector Machine để xác định giới tính của tập dữ liệu đã có.
- So sánh và phân tích các kết quả thực nghiệm với các mô hình thuật toán khác và đưa ra được trường hợp cho kết quả tốt nhất.

2. Hạn chế:

- Nghiên cứu dựa trên dữ liệu có sẵn, tập dữ liệu có sự mất cân bằng giới tính khi số lượng nữ nhiều hơn số lượng nam giới.
- Kết quả thực nghiệm đạt được vẫn chưa thực sự tốt so với kỳ vọng.

- Tốc độ xử lý dữ liệu vẫn chậm khi tập dữ liệu lớn

3. Hướng phát triển

- Thu thập bộ dữ liệu lớn hoàn chỉnh, phong phú về các lịch sử truy cập của người dùng Internet.
- Dựa trên nhiều đặc trưng để góp phần cải thiện khả năng phân loại và xác định giới tính người dùng áp dụng cho các bài toán thực tiễn
- Cải thiện hiệu xuất, tăng tốc độ xử lý dữ liệu
- Ngoài ra tác giả cũng sẽ nghiên cứu và thử nghiệm với một số mô hình thuật toán khác để tìm ra thuật toán phù hợp với bài toán xác định giới tính người dùng Internet.

DANH MỤC TÀI LIỆU THAM KHẢO

[1] Do Viet Phuong and Tu Minh Phuong. “Gender Prediction Using Browsing History”. KSE (1) 2013: 271-283.

[2] Hu, J., Zeng, H.-J., Li, H., Niu, C., Chen, Z. (2007) “Demographic prediction based onuser’s browsing behavior”, Proceedings of the 16th international conference on World Wide Web, Banff, Alberta, Canada. [viewed 05.09.2016] Available from:

http://wwwconference.org/www2007/papers/paper686.pdf

[3] Kabbur, S., Han, E.-H., Karypis, G. (2010) “Content-based methods for predicting website demographic attributes”, University of Minnesota Supercomputing Institute Research Report UMSI 2010/98 [viewed 06.09.2016]

[4] Available from: http://www.dtc.umn.edu/publications/reports/2010_01.pdf

[5] Speltdoorn, S. (2010) “Predicting demographic characteristics of web users using semisupervised classification techniques” Master’s dissertation, Ghent University, Faculty of Economucs and Business Administration. [viewed 14.09.2016] Available from:

http://lib.ugent.be/fulltxt/RUG01/001/459/756/RUG01001459756_2011_0001_AC.pdf

[6] Quanzeng You, Sumit Bhatia, Tong Sun, Jiebo Luo (2014) “The eyes of the beholder: Gender prediction using images posted in Online Social Networks”. Available from:

http://www.cs.rochester.edu/u/qyou/papers/gender_classification.pdf

[7] Yuxiao Dong, Yang Yang, Jie Tang, Yang Yang, Nitesh V. Chawla (2014) “Inferring User Demographics and Social Strategies in Mobile Social

Networks". Available from: http://www3.nd.edu/~ydong1/papers/KDD14-Dong-et-al-WhoAmI-demographic-prediction.pdf

[8] Yan, X., Yan, L.: Gender classification of weblogs authors. In: Proceedings of the AAAI Spring Symposium on Computational Approaches for Analyzing Weblogs, Stanford, CA, March 27-29, pp. 228–230 (2006). Available from: http://aaaipress.org/Papers/Symposia/Spring/2006/SS-06-03/SS06-03-046.pdf

[9] Ying, J.J.C., Chang, Y.J., Huang, C.M., Tseng, V.S. (2012). Demographic prediction based on users mobile behaviors. Mobile Data Challenge. Available from:http://www.idiap.ch/project/mdc/publications/files/mdc-final241ying.pdf

[10] Nguyen, D., Gravel, R., Trieschnigg, D., and Meder, T. (2013). "How old do you think i am?"; a study of language and age in twitter. Proceedings of the Seventh International AAAI Conference on Weblogs and Social Media. Available from:

http://www.dongnguyen.nl/publications/nguyen-icwsm2013.pdf

[11] Zhang, C., Zhang, P. (2010). Predicting gender from blog posts. Technical report, Technical Report. University of Massachusetts Amherst, USA.

[12] Chang, C.C., Lin, C.J, 2001. LIBSVM – a library for support vector machines http://www.csie.ntu.edu.tw/~cjlin/libsvm

[13] PENG Qiu-fang, LIU Yang – Research of gender prediciton based on SVM with E-commerce data. Available from:

http://lxbwk.njournal.sdu.edu.cn/EN/abstract/abstract3503.shtml

[14] Dong Nguyen, Rilana Gravel, Theo Meder, Dolf Trieschnigg – TweetGenie: Automatic Age Prediction From Tweets. Available from: http://dolf.trieschnigg.nl/papers/SIGWEB.2013.nguyen.pdf

[15] Josh Jia-Ching Ying, Yao-Jen Chang, Chi-Min Huang and Vincent S. Tseng (2012) – Demographic Prediction Based on User's Mobile Behaviors. Available from: http://www.idiap.ch/project/mdc/publications/files/mdc-final241-ying.pdf

[16] Zhang, C., Zhang, P. (2010) – Predicting gender from blog posts. Technical report, Technical Report. University of Massachusetts Amherst, USA.

[17] Mendenhall's studies of word-length distribution in the works of Shakespeare and Bacon. Available from:

https://academic.oup.com/biomet/article-abstract/62/1/207/220350/Mendenhall-s-studies-of-word-length-distribution

[18] De Vel, O., Anderson, A., Corney, M., Mohay, G. M. (2001). Mining e-mail content for author identification forensics. SIGMOD Record 30(4), pp. 55-64.

[19] Argamon, S., Koppel, M., Fine, J. and Shimoni, A. (2003). Gender, Genre, and Writing Style in Formal Written Texts, Text 23(3), August.

[20] Argamon, S., Koppel, M., Pennebaker, J. and Schler, J. (2008). Automatically Profiling the Author of an Anonymous Text, Communications of the ACM.

[21] Making Large-Scale SVM Learning Practical - Thorsten Joachims. Available from: https://www.cs.cornell.edu/people/tj/publications/joachims_99a.pdf

[22] Weka - Available from: http://www.cs.waikato.ac.nz/ml/weka/

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

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