PHƯƠNG PHÁP BIỂU DIỄN CÂY CHO DỰ ĐOÁN GIỚI TÍNH KHÁCH HÀNG DỰA TRÊN DỮ LIỆU THƯƠNG MẠI ĐIỆN TỬ

Dương Trần Đức
Học viện Công nghệ Bưu chính Viễn thông

Tóm tắt: Các đặc điểm cá nhân của khách hàng như giới tính, độ tuổi, v.v. cung cấp các thông tin quan trọng cho các nhà cung cấp dịch vụ thương mại điện tử (TMĐT) trong các hoạt động quảng cáo và cá nhân hóa hệ thống. Tuy nhiên, khách hàng trực tuyến thường hạn chế cung cấp thông tin do vấn đề riêng tư. Bài báo này đề xuất một phương pháp dự đoán giới tính của khách hàng dựa trên dữ liệu lịch sử truy cập hệ thống TMĐT. Chúng tôi sử dụng phương pháp học máy trên một tập các đặc trưng được trích xuất từ thông tin xem sản phẩm của người dùng để dự đoán giới tính của họ. Các thực nghiệm được thực hiện trên tập dữ liệu được cung cấp trong khuôn khổ cuộc thi về khai phá dữ liệu trong Hội nghị PAKDD'15. Kết quả có độ chính xác 81.9% trên độ đo chính xác cân bằng và 82.3% trên độ đo macro F1 cho thấy thuật toán học máy và các đặc trưng được đề xuất đã mang lại hiệu quả tốt trong nhận diện giới tính của khách hàng.

Từ khóa: học máy, dữ liệu lớn, dự đoán giới tính.

I. MỞ ĐẦU

Ngày nay, rất nhiều các ứng dụng web như các hệ thống thương mại điện tử (TMĐT), các máy tìm kiếm, các hệ thống quảng cáo trực tuyến, sử dụng các đặc điểm cá nhân hóa để làm gia tăng sự trải nghiệm của người dùng và thúc đẩy hoạt động kinh doanh, bán hàng. Với một dịch vụ được cá nhân hóa tốt, thông tin hiển thị sẽ được tối ưu hóa cho mỗi người dùng cá nhân thay vì giống nhau cho toàn bộ người dùng. Chẳng hạn, một hệ thống TMĐT có thể hiển thị các thông tin khuyến mãi hoặc giới thiệu sản phẩm có liên quan đến từng khách hàng thay vì hiển thị quảng cáo chung hoặc giới thiệu các sản phẩm ngẫu nhiên.

Việc cá nhân hóa thông tin hiển thị dựa trên 2 loại dữ liệu chính: dữ liệu lịch sử (chẳng hạn các mặt hàng trước đó đã xem hoặc đã mua v.v.) và đặc điểm cá nhân của người dùng (chẳng hạn giới tính, độ tuổi, trình độ giáo dục .v.v). Dữ liệu lịch sử chỉ có thể thu thập được nếu người dùng đã sử dụng hệ thống trước đó và đã đăng nhập vào hệ thống. Do đó, các phương pháp cá nhân hóa dựa trên dữ liệu lịch sử không khả thi trong trường hợp khách hàng truy cập lần đầu hoặc khách hàng chưa đăng ký sử dụng hệ thống. Ngược lại, phương pháp cá nhân hóa dựa trên đặc điểm cá nhân của người dùng hữu ích kể cả khi người dùng chưa từng sử dụng hệ thống. Tuy nhiên, các thông tin về đặc điểm cá nhân của người dùng thường khó thu thập được, do người dùng Internet thường không sẵn sàng cung cấp các thông tin cá nhân có tính riêng tư. Vì lý do này, trong nhiều trường hợp, cách duy nhất để có được thông tin đặc điểm cá nhân của người dùng là dự đoán dựa trên các dữ liệu khác mà người dùng để lại trên hệ thống.

Vấn đề dự đoán đặc điểm người dùng dựa trên phân tích văn bản (còn gọi dự đoán đặc điểm tác giả văn bản - author profiling) đã được nghiên cứu trong nhiều thập kỷ, tuy nhiên, trong nhiều trường hợp, người dùng không để lại các văn bản trên hệ thống. Một phương pháp khác có thể được sử dụng để dự đoán đặc điểm người dùng là dựa vào hành vi của họ trên hệ thống, chẳng hạn các hành vi duyệt web ([6], [13]), phân tích lưu lượng web ([3]), hoặc hành vi xem danh mục sản phẩm. Ưu điểm chính của phương pháp tiếp cận này là trong hầu hết các trường hợp, người dùng sẽ thực hành các hành vi trên hệ thống như truy cập vào các trang web, nhấp chuột vào các mặt hàng/mục tin, xem danh mục sản phẩm v.v.

Trong nghiên cứu này, chúng tôi giải quyết vấn đề dự đoán giới tính người dùng dựa trên dữ liệu xem danh mục sản phẩm như thời gian/thời lượng xem, danh sách các sản phẩm/loại sản phẩm đã xem v.v. Tập dữ liệu thực nghiệm được cung cấp bởi Tập đoàn FPT trong cuộc thi về khai phá dữ liệu trong khuôn khổ Hội nghị Quốc tế về Khai phá dữ liệu và Phát hiện tri thức khu vực Châu Á Thái Bình Dương năm 2015 (PAKDD'15). Ý tưởng của phương pháp là khai thác tối đa mối quan hệ giữa các sản phẩm/loại sản phẩm được xem trong cùng 1 lượt xem dựa trên 1 biểu diễn dạng cây của danh sách sản phẩm/loại sản phẩm. Theo đó, bên cạnh các đặc trưng cơ bản như thời gian, tần suất xem, danh sách các sản phẩm/loại sản phẩm riêng rẽ, chúng tôi nghiên cứu đề xuất sử dụng các đặc trưng như chuỗi các sản phẩm/loại sản phẩm được xem liên tiếp, các cặp chuyển tiếp sản phẩm/loại sản

Tác giả liên hệ: Dương Trần Đức,
Email: duongtranduc@gmail.com
Đến tòa soạn: 2/2018 , chỉnh sửa: 4/2018 , chấp nhận đăng: 5/ 2018

phẩm khác nhau trong cùng 1 lượt xem v.v. (gọi chung là các đặc trưng nâng cao). Với cấu trúc phân cấp nhiều cấp độ của danh mục sản phẩm/loại sản phẩm, chúng tôi sử dụng một phương pháp biểu diễn dạng cây để cung cấp khung nhìn tốt hơn về mối quan hệ giữa các sản phẩm/loại sản phẩm so với biểu diễn dạng liệt kê. Sau khi xây dựng được tập dữ liệu huấn luyện, một số thuật toán học máy phổ biến như Rừng ngẫu nhiên (Random Forest-RF), Máy véc tơ hỗ trợ (Support Vector Machine-SVM), và Mạng Bayes (Bayesian Network-BN) được sử dụng để xây dựng mô hình phân loại kết hợp với các kỹ thuật hỗ trợ để xử lý vấn đề không cân bằng lớp như Tái chọn mẫu (Resampling), Học nhạy cảm chi phí (Cost-Sensitive Learning-CSL). Ngoài ra, do số lượng đặc trưng sử dụng là khá lớn cùng với tính chất thưa của dữ liệu xem danh mục sản phẩm, các phương pháp lựa chọn đặc trưng (feature selection) được thử nghiệm và áp dụng nhằm nâng cao kết quả dự đoán và giảm độ phức tạp của mô hình. Cuối cùng, thuật toán phân loại được tối ưu tham số và kết hợp với thuật toán boosting để cải tiến kết quả dự đoán. Các kết quả thực nghiệm cho thấy độ chính xác nhận diện tốt trên tập đặc trưng có tính tổng quát và có thể dễ dàng áp dụng sang các hệ thống TMĐT khác nhau. Bài báo này cũng là phiên bản mở rộng của nghiên cứu đã được báo cáo tại Hội nghị Quốc tế Kỹ nghệ tri thức và hệ thống năm 2016 (Knowledge and System Engineering - KSE 2016), trong đó các vấn đề về xây dựng tập đặc trưng, lựa chọn đặc trưng, và tối ưu tham số thuật toán đã được nghiên cứu và cải tiến.

Bài báo có cấu trúc như sau. Phần II trình bày về các nghiên cứu liên quan trong lĩnh vực dự đoán đặc điểm người dùng. Phần III mô tả phương pháp tiếp cận và hoạt động của hệ thống. Phần IV trình bày về các kết quả và thảo luận. Cuối cùng, các kết luận sẽ được trình bày trong phần V của bài báo.

II. TỔNG QUAN VỀ DỰ ĐOÁN ĐẶC ĐIỂM NGƯỜI DÙNG

Vấn đề dự đoán đặc điểm người dùng đã được nghiên cứu trong thời gian dài trước đây. Trong giai đoạn đầu, các nhà nghiên cứu trong lĩnh vực này tập trung nghiên cứu về vấn đề xác định đặc điểm tác giả văn bản. Đó là việc xác định hoặc dự đoán đặc điểm của người dùng dựa trên phân tích các văn bản được tạo ra bởi người đó. Các phương pháp được sử dụng trong các nghiên cứu này chủ yếu là dựa trên phân tích phong cách viết với các đặc trưng đa dạng như dựa trên các dùng từ vựng, ngữ pháp, các đặc trưng dựa trên nội dung [9]. Các nghiên cứu trước đây chủ yếu tập trung vào các loại văn bản chính thống như các bài báo, tiểu thuyết, bài luận v.v. Gần đây, do sự phát triển mạnh mẽ của Internet và các kênh truyền thông trực tuyến, các nghiên cứu trong lĩnh vực này chuyển sang thực hiện trên các loại văn bản truyền thông trực tuyến như email, bài viết blogs, bài viết diễn đàn v.v. De Vel và các cộng sự [4] sử dụng 221 đặc trưng để xác định tác giả các emails. Argamon và các cộng sự [1] nghiên cứu sự khác biệt giữa phong cách viết của nam và nữ trong 604 tài liệu từ kho ngữ liệu Anh Quốc (British National Corpus). Argamon và các cộng sự [2] khảo sát việc sử dụng các đặc trưng dựa theo phong cách và nội dung để dự đoán giới tính và tuổi của các tác giả bài viết blogs trên tập dữ liệu gồm hơn 71.000 bài viết từ trang blogger.com. Mô hình này cho kết quả dự đoán có độ chính xác 80% cho giới tính và 76% cho độ tuổi. Iqbal và các cộng sự [7] đề xuất một phương pháp tính một giá trị được gọi là "vân chữ viết" (write print) dựa trên các mẫu xuất hiện thường xuyên được trích chọn từ các emails để dự đoán đặc điểm người dùng. Nguyen và các cộng sự [14] thực hiện nghiên cứu về dự đoán giới tính và độ tuổi của các tác giả bài viết trên mạng xã hội twitter và bài viết diễn đàn tiếng Hà Lan sử dụng phương pháp hồi quy tuyến tính và cho độ chinh xác dự đoán khoảng 80%.

Bên cạnh việc nhận diện người dùng thông qua phân tích văn bản, gần đây, nhiều nhà nghiên cứu trong lĩnh vực khoa học máy tính đã mở rộng sang phân tích nhận diện đặc điểm người dùng dựa trên hành vi của họ, chẳng hạn như các hành vi duyệt website [6, 14], hành vi trong mạng di động [5], hành vi xem sản phẩm trong hệ thống thương mại điện tử v.v. Khác với vấn đề xác định đặc điểm tác giả văn bản, các đặc trưng hành vi của người dùng trên các hệ thống là đa dạng hơn nhiều. Do vậy, các nghiên cứu trong lĩnh vực này đã sử dụng các tập đặc trưng khác nhau và phụ thuộc vào các hệ thống cụ thể. Phương pháp nhận diện chủ yếu sử dụng kỹ thuật học máy.

Hu và các cộng sự [6] đề xuất một phương pháp để giải quyết vấn đề dự đoán giới tính và độ tuổi của người dùng Internet thông qua phân tích hành vi duyệt web của họ. Hu sử dụng các thông tin xem trang web của người dùng như là các biến đầu vào để suy diễn thông tin đặc điểm cá nhân của họ. Thuật toán SVM đã được sử dụng trên tập đặc trưng bao gồm các đặc trưng dựa trên nội dung (các từ trong trang web) và dựa trên phân loại (theo các mục trong cấu trúc của trang web). Kết quả thực nghiệm đạt độ chính xác 79.7% khi dự đoán giớ tính và 60.3% khi dự đoán tuổi. Kabbur và các cộng sự [8] cũng thực hiện 1 nghiên cứu sử dụng học máy để dự đoán đặc điểm người dùng website dựa trên thông tin về nội dung và cấu trúc siêu liên kết.

Nghiên cứu của Dong và các cộng sự [5] có mục tiêu suy diễn ra thông tin cá nhân của người dùng dựa trên các mẫu giao tiếp hàng ngày trên mạng di động. Nghiên cứu được thực hiện trên một mạng di động thực với hơn 7.000.000 người dùng và hơn 1 tỷ bản ghi giao dịch mỗi ngày. Các đặc trưng được sử dụng bao gồm các đặc trưng cá nhân, bạn bè, đặc trưng tuần hoàn v.v. và đạt kết quả dự đoán 80% cho giới tính và 70% cho độ tuổi. Ying và các cộng sự [15] đề xuất một phương pháp dự đoán thông tin cá nhân người dùng dựa trên phân tích hành vi và môi trường. Nghiên cứu cũng phát triển một phương pháp mới là mô hình phân loại nhiều cấp độ (multi-level classification model) để giải quyết vấn đề không cân bằng trong dữ liệu.

Phuong và các cộng sự [13] giải quyết vấn đề dự đoán giới tính người dùng thông qua hành vi duyệt website. Nghiên cứu sử dụng phương pháp phân loại học máy và dùng các đặc trưng thu được từ dữ liệu lưu trữ thông tin duyệt web. Các đặc trưng cơ bản được sử dụng cũng tương tự nghiên cứu của Hu và các cộng sự [6], nhưng nhóm tác giả sử dụng thêm nhiều loại đặc trưng khác như các đặc trưng dựa trên chủ đề, đặc trưng thời gian, đặc trưng kế tiếp v.v. qua đó làm tăng đáng kể kết quả dự đoán.

Nghiên cứu của Lu và các cộng sự [12] cũng giải quyết vấn đề tương tự như nghiên cứu này. Lu sử dụng 1 tập đặc trưng bao gồm các đặc trưng về tần suất, thời gian, các sản phẩm/loại sản phẩm được xem và thuật toán phân loại Gradient Boosting Decision Trees. Sau đó, Lu thực hiện việc cập nhật nhãn để nâng cao độ chính xác bằng cách đưa các thông tin về sản phẩm được xem vào tính toán làm mượt (tổng số lượng nam/nữ xem sản phẩm). Kết quả cuối cùng cho độ chính xác F1 trung bình của 2 lớp phân loại là 80.6.

Bài báo này nghiên cứu một phương pháp dự đoán giới tính của người dùng dựa trên dữ liệu xem sản phẩm của họ trên hệ thống TMĐT. Theo khảo sát của chúng tôi, hiện chỉ có nghiên cứu của Lu và các cộng sự [12] là nghiên cứu chính thức được thực hiện và công bố trong lĩnh vực này.

III. PHƯƠNG PHÁP

A. Tổng quan về hệ thống

Trong nghiên cứu này, chúng tôi phát triển một hệ thống có thể nhận dữ liệu từ các file lưu trữ thông tin xem sản phẩm của các khách hàng đã biết giới tính, trích chọn các đặc trưng và nhãn phân loại để tạo ra 1 tập dữ liệu huấn luyện. Mô hình dự đoán sẽ được xây dựng dựa trên tập dữ liệu huấn luyện tạo được sử dụng một phương pháp phân loại và sau đó có thể sử dụng để dự đoán giới tính của các khách hàng chưa biết dựa trên hành vi xem sản phẩm của họ.

File dữ liệu huấn luyện chứa các bản ghi tương ứng với các thông tin lưu trữ về hành vi xem sản phẩm của người dùng. Một bản ghi lưu trữ chứa các thông tin về hành vi xem sản phẩm của 1 người dùng, như thời gian bắt đầu xem, kết thúc xem, danh sách các sản phẩm và loại sản phẩm đã xem. Nhãn phân loại cho mỗi dữ liệu mẫu là male/female (nam/nữ). Do vậy, vấn đề cần giải quyết là một vấn đề phân loại nhị phân với 2 nhãn tương ứng.

Phần tiếp theo sẽ mô tả chi tiết hơn về các đặc trưng và các kỹ thuật được sử dụng để dự đoán.

B. Các đặc trưng phân loại

Các đặc trưng được sử dụng trong nghiên cứu này được chia làm 2 loại, được gọi là các đặc trưng cơ bản và các đặc trưng nâng cao.

1) Đặc trưng cơ bản

Các đặc trưng cơ bản bao gồm các đặc trưng liên quan đến thời gian, tần suất xem sản phẩm và các đặc trưng về các sản phẩm/loại sản phẩm riêng rẽ. Các thông tin như thời gian xem trong ngày, ngày trong tuần, ngày nghỉ/ngày lễ, thời lượng xem, số sản phẩm xem, thời gian trung bình khi xem 1 sản phẩm v.v. là các nhân tố có thể được dùng để dự đoán giới tính của người xem. Tổng cộng có 98 đặc trưng nhị phân và 3 đặc trưng số được sử dụng và được mô tả chi tiết hơn như trong bảng 1.

Bảng 1. Các đặc trưng cơ bản

Đặc trưng	Mô tả
Day	Ngày trong tháng (31 đặc trưng)
Month	Tháng trong năm (12 đặc trưng)
DayOfWeek	Ngày trong tuần (7 đặc trưng)
StartTime/EndTime	Giờ (24 đặc trưng)/ Giờ (24 đặc trưng)
Duration	Tổng thời gian xem (1 đặc trưng)
NumberOfProducts	Số sản phẩn xem (1 đặc trưng)
AverageTimePerProduct	Thời gian trung bình xem 1 sản phẩm (1 đặc trưng)

Đặc trưng về các sản phẩm/loại sản phẩm bao gồm tất cả các sản phẩm và loại sản phẩm có trong hệ thống. Để xây dựng danh mục các đặc trưng này, chúng tôi thực hiện trích từ trong tập dữ liệu ra các mã sản phẩm/mã phân loại và sử dụng chúng như các đặc trưng dạng số. Với mỗi sản phẩm/loại sản phẩm, chúng tôi thực hiện đếm số lần người dùng xem sản phẩm/loại sản phẩm đó trong lượt xem và sử dụng con số này làm giá trị của đặc trưng tương ứng. Do mỗi mã sản phẩm đầy đủ được hình thành từ 4 mã khác nhau, bao gồm mã loại sản phẩm ở mức chung nhất (bắt đầu bằng ký tự "A"), các mã loại sản phẩm ở mức tiếp theo (bắt đầu bằng ký tự "B" và "C"), và cuối cùng là mã sản phẩm cụ thể (bắt đầu bằng ký tự "D"), có 4 loại đặc trưng thuộc dạng này với tổng cộng 8.035 đặc trưng như trong bảng 1. Lưu ý rằng do số lượng mã sản phẩm cụ thể là rất lớn và nhiều sản phẩm xuất hiện ở tập dữ liệu huấn luyện nhưng không xuất hiện ở tập dữ liệu kiểm tra và ngược lại, chúng tôi chỉ lựa chọn các mã sản phẩm có tần suất xuất hiện từ 3 lần trở lên và bổ sung thêm các sản phẩm có tần suất thấp hơn nhưng xuất hiện ở cả 2 tập dữ liệu. Ngoài ra, do một sản phẩm có thể thuộc về nhiều hơn 1 phân loại, các sản phẩm này sẽ tạo ra nhiều hơn 1 đặc trưng, tương ứng với các phân loại.

Bảng 2. Các đặc trưng về sản phẩm/loại sản phẩm riêng rẽ

Đặc trưng	Mô tả
Loại sản phẩm mức chung nhất	Mã bắt đầu là A (11 đặc trưng)
Loại sản phẩm mức 2	Mã bắt đầu là B (60 đặc trưng)
Loại sản phẩm mức 3	Mã bắt đầu là C (186 đặc trưng)
Sản phẩm cụ thể	Mã bắt đầu là D (7.778 đặc trưng)

2) Các đặc trưng nâng cao

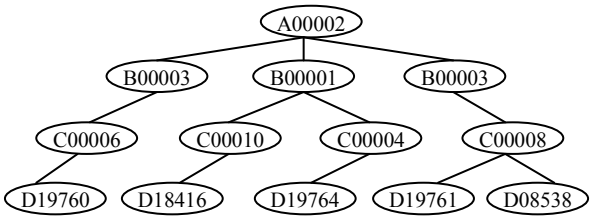
Bên cạnh các đặc trưng sản phẩm/loại sản phẩm riêng rẽ, chúng tôi đặt giả thiết rằng mối quan hệ giữa các sản phẩm/loại sản phẩm được xem trong cùng 1 lượt xem cũng là một yếu tố có thể dùng để dự đoán giới tính người dùng. Chẳng hạn người dùng nam thường chỉ xem ít loại sản phẩm trong 1 lượt xem trong khi người dùng nữ có thể xem liên tiếp nhiều loại sản phẩm khác nhau. Trong file dữ liệu, danh sách

các sản phẩm/loại sản phẩm đã xem trong 1 lượt xem được biểu thị dưới dang danh sách liệt kê như dưới đây:

A00002/B00003/C00006/D19760/; A00002/B00001/C00010/D18416; A00002/B00001/C00004/D19764/; A00002/B00003/C00008/D19761/; A00002/B00003/C00008/D08538/

Việc sử dụng danh sách liệt kê này có thể gây khó khăn cho việc trích chọn hiệu quả tất cả các thông tin về mối quan hệ giữa các sản phẩm/loại sản phẩm trong 1 lượt xem, chúng tôi đề xuất một biểu diễn dạng cây nhằm cung cấp 1 khung nhìn tốt hơn về các quan hệ này. Theo biểu diễn này, loại sản phẩm ở mức chung nhất sẽ nằm ở gốc của cây, các sản phẩm cụ thể nằm ở phần lá của cây và các loại sản phẩm ở cấp độ trung gian nằm ở các tầng giữa của cây. Theo đó, danh mục sản phẩm/loại sản phẩm được biểu diễn dưới dạng danh sách liệt kê ở trên được chuyển đổi sang biểu diễn cây như trong hình 1.

Từ biểu diễn dạng cây này, chúng ta có thể dễ dàng chuyển đổi ngược trở lại biểu diễn dạng danh sách liệt kê bằng cách duyệt cây theo chiều sâu và từ trái sang phải. Ngoài ra, từ biểu diễn cây, chúng ta có thể rút ra được các thông tin về quan hệ giữa các sản phẩm/loại sản phẩm bằng cách khai thác các thuộc tính của cây như các nút, các tầng, đường đi, nút kề v.v.



Hình 1. Biểu diễn dạng cây của danh mục sản phẩm/loại sản phẩm được xem

Trong vấn đề hiện tại, chúng ta có thể sử dụng các thuộc tính sau của cây để làm đặc trưng về mối quan hệ:

- Số các nút tại mỗi tầng: Tương ứng với số sản phẩm/loại sản phẩm được xem trong mỗi lượt xem.
- Chuỗi các nút liên tiếp trên cùng 1 tầng: Tương ứng với các chuỗi sản phẩm/loại sản phẩm được xem liền nhau trong cùng một lượt xem. Từ chuỗi các nút liên tiếp trên cùng tầng, chúng tôi trích ra tất cả các chuỗi con k nút và chọn các chuỗi con có tần suất cao nhất làm đặc trưng chuỗi.
- Cặp nút chuyển đổi tại các tầng khác nhau: Đặc trưng này phản ánh thói quen xem sản phẩm của 1 người dùng khi chuyển từ 1 loại sản phẩm này sang 1 loại khác ở tầng khác nhau.

Chẳng hạn, với biểu diễn cây như ở hình 1.1, một số thuộc tính như ở trên có thể được trích ra như sau:

- Số lượng nút tại mỗi tầng: {1, 3, 4, 5}
- Chuỗi các nút liên tiếp trên cùng 1 tầng có thể là {B00001, B00003, B00001}, {B00001, B00003}, {C00006, C00010}, {D19760, D18416, D19764}, v.v.
- Các cặp nút chuyển đổi tại các tầng khác nhau có thể là {D19760, B00001}, {D18416, C00004}, v.v.

Với số lượng lớn các sản phẩm và phân loại sản phẩm, tổng số lượng các chuỗi nút và các cặp nút chuyển đổi có thể rất lớn. Do đó, tương tư nhue cách xây dựng tập đặc trưng cho các sản phẩm đơn lẻ, chúng tôi chỉ lựa chọn các chuỗi nút và các cặp nút chuyển đổi có tần suất xuất hiện ít nhất 3 lần và hoặc tần suất ít hơn nhưng xuất hiện trong cả 2 tập dữ liệu. Theo đó, danh sách và số lượng các đặc trưng nâng cao được liệt kê trong bảng 3.

Bảng 3. Các đặc trưng nâng cao

Đặc trưng	Mô tả
Số lượng nút tại mỗi tầng	4 đặc trưng
Các chuỗi nút có tần suất xuất hiện cao nhất	2.277 đặc trưng
Các cặp nút chuyển đổi có tần suất xuất hiện cao nhất	465 đặc trưng

C. Các phương pháp phân loại

Trong nghiên cứu này, chúng tôi sử dụng 3 thuật toán học máy để xây dựng mô hình phân loại như đã nói ở trên. Đó là Random Forest (RF), Support Vector Machine (SVM), và Bayesian Network (BN). RF là một thuật toán học kết hợp sử dụng các tập con của dữ liệu và tập con đặc trưng để xây dựng nên các cây quyết định. RF xây dựng nhiều cây quyết định như vậy và kết hợp chúng để cho kết quả phân loại cuối cùng có độ chính xác cao hơn. Do thuật toán này lựa chọn ngẫu nhiên các tập con đặc trưng để xây dựng cây quyết định nên phù hợp với các vấn đề có tập đặc trưng lớn và thưa như vấn đề hiện tại. SVM là phương pháp phân loại dựa trên lý thuyết học thống kê được đề xuất bởi Vapnik năm 1995. SVM là thuật toán học máy có ưu điểm là có thể xử lý số lượng lớn các đặc trưng phân loại và không cần đến việc giảm bớt số lượng đặc trưng nhằm tránh vấn đề quá khớp (over-fitting). Đặc điểm này rất hữu ích khi xử lý các vấn đề có số chiều lớn. BN là một mô hình xác suất dạng đồ thị biểu thị sự phụ thuộc thống kê trên một tập hợp các biến ngẫu nghiên. Đây cũng là thuật toán được sử dụng khá phổ biến trong xây dựng các mô hình học máy.

Bên cạnh các thuật toán học máy, do tập dữ liệu huấn luyện có đặc điểm không cân bằng giữa các lớp (khoảng 80% là nữ và chỉ 20% nam), một số kỹ thuật hỗ trợ như Resampling, Cost-Sensitive Learning (CSL) được áp dụng để nâng cao độ chính xác cho lớp thiểu số. Resampling là một phương pháp được sử dụng phổ biến để xử lý các trường hợp không cân bằng trong dữ liệu huấn luyện. Ý tưởng cơ bản của phương pháp này là thêm vào hoặc bớt đi 1 số mẫu để làm cho tập dữ liệu trở nên cân bằng hơn. Ngoài ra,

cũng có thể đặt lại trọng số cho các mẫu của mỗi lớp để giúp cân bằng tổng trọng số của mỗi lớp [10]. Trong khi resampling là một phương pháp ở mức dữ liệu thì CSL là một phương pháp ở mức thuật toán dùng để giải quyết vấn để phân loại không cân bằng. Theo Ling và các cộng sự [11], CSL là một phương pháp có tính đến chi phí phân loại sai, nghĩa là nó xem xét các phân loại sai của các lớp khác nhau là khác nhau, nhờ đó có thể cân bằng độ chính xác giữa 2 lớp khi xây dựng mô hình phân loại.

Ngoài ra, do số lượng các đặc trưng lớp và dữ liệu thưa, các kỹ thuật lựa chọn đặc trưng được nghiên cứu, áp dụng để giảm bớt độ phức tạp và loại bỏ đi các đặc trưng ít liên quan đến quá trình phân loại. Trong nghiên cứu này, chúng tôi thử nghiệm một số độ đo như Độ lợi thông tin (Information Gain), Khi-bình phương (Chi-Square), Tương quan (Correlation) để chọn ra phương pháp và số lượng đặc trưng phù hợp nhất.

IV. THỰC NGHIỆM

A. Dữ liệu và phương pháp đánh giá

Trong nghiên cứu này, chúng tôi sử dụng các tập dữ liệu được cung cấp bởi tập đoàn FPT cho cuộc thi về khai phá dữ liệu và phát hiện tri thức trong khuôn khổ hội nghị PAKDD'15. Dữ liệu được chia thành 2 tập là tập huấn luyện và tập kiểm chứng. Mỗi tập dữ liệu chứa 15.000 bản ghi, tương ứng với các bản lưu trữ về thông tin xem sản phẩm của mỗi người dùng.

Về phương pháp đánh giá, như đã trình bày ở trên, do vấn đề không cân bằng của các lớp dự đoán, độ đo chính xác cân bằng được sử dụng để đánh giá mô hình. Độ đo chính xác cân bằng được định nghĩa là độ chính xác trung bình của mỗi lớp và việc sử dụng độ đo này có thể tránh được các dự báo hiệu suất giả tạo trong các tập dữ liệu không cân bằng lớp.

$$balanced\ accuracy\ (BAC) = \frac{0.5*tp}{tp+fn} + \frac{0.5*tn}{tn+fp}$$

Trong đó *tp (true positive)* là số các mẫu mang nhãn "dương" được phân đúng vào lớp "dương", *tn (true nagative)* là số các mẫu mang nhãn "âm" được phân đúng vào lớp "âm", *fp (false positives)* là số các mẫu mang nhãn "âm" được phân sai vào lớp "dương", và *fn (false negative)* là số các mẫu mang nhãn "dương" được phân sai vào lớp "âm".

Đây cũng là độ đo được sử dụng để đánh giá các kết quả trong cuộc thi PAKDD'15 Data Mining Competition. Trong nghiên cứu này, chúng tôi sử dụng độ đo này cũng với độ đo Macro F1 để tiện so sánh với các nghiên cứu trước đây.

B. Kết quả và đánh giá

Nhằm đánh giá hiệu quả của các đặc trưng cơ bản và nâng cao, chúng tôi thực hiện các thí nghiệm trên các tập đặc trưng khác nhau, bao gồm tập đặc trưng cơ bản và tập đặc trưng cơ bản kết hợp nâng cao. Theo cách phân loại tập đặc trưng, các đặc trưng nâng cao chỉ mang tính bổ sung, nếu sử dụng riêng rẽ sẽ không hiệu quả. Do đó, chúng tôi không tiến hành thí nghiệm trên tập đặc trưng nâng cao riêng rẽ trong nghiên cứu này.

Mỗi tập đặc trưng sẽ được thử nghiệm trên 3 thuật toán học máy và các kỹ thuật hỗ trợ như đã nói ở trên, trong đó Resampling sử thuật toán tái cân bằng lớp dựa trên kỹ thuật đặt lại trọng số Class Balancer (CB). Công cụ thực nghiệm sử dụng bộ công cụ học máy WEKA (Waikato Environment for Knowledge Analysis). Đây là một tập hợp các thuật toán học máy và các công cụ xử lý dữ liệu được phát triển bởi nhóm nghiên cứu tại Đại học Waikato, New Zealand. Công cụ này được viết bằng ngôn ngữ Java và được phân phối dưới dạng mã nguồn mở. Kết quả thực nghiệm cuối cùng cho thấy khi thuật toán học máy kết hợp với kỹ thuật tái cân bằng lớp theo phương pháp đặt lại trọng số cho các lớp ClassBalancer và kỹ thuật học nhạy cảm chi phí CostSensitiveClassifier cho kết quả BAC tốt nhất. Bảng 4 cho thấy kết quả cụ thể của các thực nghiệm khi chưa áp dụng các thuật toán lựa chọn đặc trưng và tối ưu tham số học máy.

Bảng 4. Kết quả thực nghiệm khi sử dụng CSL kết hợp CB

	Đặc trưng cơ bản		**Đặc trưng cơ bản + nâng cao**	
	BAC	***Macro F1***	***BAC***	***Macro F1***
RF	77.3	75.5	**81.0**	**78.5**
SVM	76.6	74.4	79.5	76.7
BN	76.0	74.4	78.5	76.0

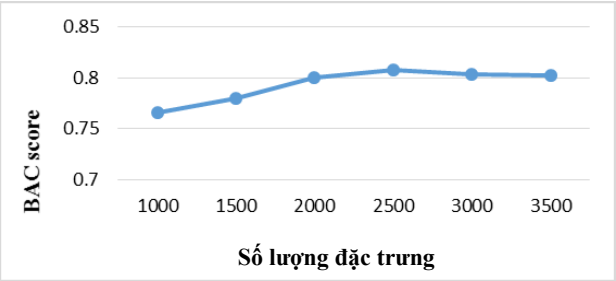
Có thể thấy, các đặc trưng nâng cao khi được sử dụng kết hợp với các đặc trưng cơ bản đã cải tiến kết quả đáng kể khi so sánh với việc chỉ sử dụng đặc trưng cơ bản. Mặc dù vậy, trong tập dữ liệu được cung cấp, có khá nhiều lượt xem chỉ có một sản phẩm được xem (khoảng 30%) và các đặc trưng nâng cao không có hiệu quả với các trường hợp này (do không có nhiều sản phẩn được xem trong cùng lượt để khai thác mối quan hệ giữa chúng). Trên thực tế, số lượng người dùng xem nhiều sản phẩm trong 1 lượt xem sẽ nhiều hơn và do đó việc sử dụng các đặc trưng nâng cao sẽ đem lại hiệu quả cao hơn khi áp dụng trong các trường hợp này.

So sánh kết quả của các thuật toán học máy, thuật toán RF có kết quả vượt trội so với các thuật

toán SVM và BN. Thuật toán RF thực hiện học kết hợp thông qua việc lựa chọn nhiều tập con đặc trưng và dữ liệu để xây dựng nên 1 tập các cây quyết định, do đó phù hợp với bài toán có số lượng đặc trưng lớn và thưa như bài toán hiện tại. Một điểm thú vị khác là phương pháp biểu diễn đặc trưng được sử dụng trong nghiên cứu này cũng có cấu trúc dạng cây. Tuy nhiên, kết quả vẫn có thể tiếp tục được cải tiến thông qua việc lựa chọn đặc trưng và tối ưu tham số.

C. Lựa chọn đặc trưng và tối ưu tham số

Mặc dù thuật toán RF đã tiến hành lựa chọn tập đặc trưng tốt trong quá trình học thông qua việc lựa chọn ngẫu nhiên các đặc trưng tại các bước xây dựng cây quyết định, tuy nhiên vẫn có thể cải tiến độ chính xác bằng việc thực hiện các thuật toán lựa chọn đặc trưng dựa trên các độ đo thống kê. Trong nghiên cứu này, chúng tôi thử nghiệm 3 phương pháp lựa chọn đặc trưng là Information Gain, Chi-Square, và Correlation. Information Gain sử dụng cách đo độ quan trọng của mỗi đặc trưng trong việc phân biệt các lớp phân loại và đã được ứng dụng trong nhiều nghiên cứu trước đây và cho kết quả tốt. Chi-Square là phép thử có thể đánh giá sự độc lập của 2 biến trong thống kê, và được sử dụng để đo mức độ độc lập giữa 1 đặc trưng và lớp phân loại. Trong khi đó, phương pháp Correlation sử dụng độ đo tương tự giữa các đặc trưng với nhau và với lớp phân loại để đánh giá tập đặc trưng tốt. Kết quả thử nghiệm cho thấy Information Gain là phương pháp phù hợp nhất cho vấn đề hiện tại với số lượng tối ưu được lựa chọn là 2.500 đặc trưng. Hình 2 cho thấy kết quả phân loại tốt dần với các số lượng đặc trưng thấp và đạt đỉnh tại mức 2.500 đặc trưng.



Hình 2. Kết quả phân loại với các số lượng đặc trưng được lựa chọn khác nhau

Ngoài ra, các thực nghiệm ở phần trước được thực hiện trên tập tham số mặc định của thuật toán. Các kết quả có thể được cải tiến thông qua việc tối ưu các tham số. Thuật toán RF có 3 tham số có thể ảnh hưởng tới độ chính xác phân loại. Đó là số lượng đặc trưng tối đa được lựa chọn khi xây dựng các cây quyết định, số lượng cây được xây dựng (số vòng lặp), kích thước lá tối thiểu của cây. Các tham số này được tối ưu sử dụng thuật toán Grid Search để chọn ra các tham số cho kết quả tốt nhất với thời gian tính toán phù hợp. Bảng 6 cho biết kết quả phân loại sau khi thực hiện lựa chọn đặc trưng và tối ưu tham số cho thuật toán RF.

Bảng 5. Kết quả phân loại sau khi lựa chọn đặc trưng và tối ưu tham số

	BAC	***Macro F1***
Kết quả ban đầu	81.0	78.5
Áp dụng lựa chọn đặc trưng với Information Gain	81.2	78.8
Tối ưu tham số cho thuật toán RF (1000 cây, với số đặc trưng 13)	81.7	79.3

D. Đánh giá

Kết quả cơ sở của các nghiên cứu về dự đoán giới tính tác giả văn bản là hơn 80% (độ đo chính xác thông thường accracy và độ đo F1). Mặc dù so sánh các kết quả của các nghiên cứu trên các tập dữ liệu khác nhau không thực sự hợp lý, tuy nhiên, với cùng mục đích dự đoán giới tính người dùng, kết quả của nghiên cứu này có thể xem là có nhiều triển vọng. Với các nghiên cứu có độ tương tự cao hơn như [6], [13] khi dự đoán giới tính người dùng thông qua hành vi duyệt website, kết quả Marco F1 của nghiên cứu này cũng tương đương, trong khi hành vi duyệt website tạo ra nhiều dữ liệu có ý nghĩa hơn. Ngoài ra, các trang web còn chứa các văn bản, do vậy có thể tạo ra nhiều loại đặc trưng hơn. So sánh với các giải pháp khác của các nhóm tham gia cuộc thi PAKDD'15 Data Mining Competition, giải pháp trong nghiên cứu này trong top 10 trên 150 nhóm tham dự. Kết quả của nhóm cao nhất là 87.9% và các nhóm trong top 10 có kết quả từ 81%. Tuy nhiên, ưu điểm của giải pháp của nghiên cứu này là sử dụng một cấu trúc đặc trưng đơn giản, nhưng vẫn đạt được các kết quả đáng kể. Cấu trúc đặc trưng này có tính tổng quát, không chứa các đặc trưng mang tính đặc thù, do vậy có thể dễ dàng áp dụng sang các hệ thống khác. So sánh với nghiên cứu được thực hiện trên cùng tập dữ liệu và được công bố chính thức của Lu và các cộng sự [12], nghiên cứu này có kết quả tốt hơn, mặc dù không sử dụng bước cập nhật nhãn.

V. KẾT LUẬN

Trong nghiên cứu này, chúng tôi trình bày một phương pháp dự đoán giới tính người dùng dựa trên dữ liệu thu thập từ hệ thống TMĐT. Phương pháp tiếp cận sử dụng các đặc trưng cơ bản như thời gian, tần suất xem sản phẩm, cùng với các đặc trưng nâng cao như các chuỗi sản phẩm/loại sản phẩm hoặc các cặp sản phẩm/loại sản phẩm chuyển tiếp trong lượt xem. Phương pháp này sử dụng một biểu diễn dạng cây của danh sách các sản phẩm/loại sản phẩm và sử dụng các

thuộc tính của cây như số nút, chuỗi các nút cùng tầng, cặp nút chuyển khác tầng v.v. làm đặc trưng phân loại. Thiết kế tập đặc trưng này cho kết quả tốt nhất trên thuật toán Random Forest cùng với các kỹ thuật hỗ trợ như Cost Sensitive Learning và Class Balancing. Ngoài ra, kết quả cũng được cải tiến thông qua một số kỹ thuật như lựa chọn đặc trưng, tối ưu tham số thuật toán.

Hướng phát triển tiếp theo của nghiên cứu có thể liên quan đến việc khai thác các đặc trưng rút trích từ cây biểu diễn danh sách sản phẩm/loại sản phẩm. Ngoài ra, cũng có thể thu thập thêm các dữ liệu bổ sung và mở rộng sang dự đoán các đặc điểm khác của người dùng như độ tuổi, nghề nghiệp v.v.

TÀI LIỆU THAM KHẢO

[1] S. Argamon, M. Koppel, J. Fine, and A. Shimoni, “Gender, genre, and writing style in formal written texts,” Text 23(3), August 2003.

[2] S. Argamon, M. Koppel, J. Pennebaker, and J. Schler, “Automatically profiling the author of an anonymous text,” Communications of the ACM , v.52 n.2, February 2009.

[3] J. C. A. Culotta, N. R. Kumar, and J. Cutler, “Predicting the demographics of twitter users from website traffic data, ” Proceedings of the 29th AAAI Conference on Artificial Intelligence, Jan 2015.

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

[5] Y. Dong, Y. Yang, J. Tang, Y. Yang, and N. V. Chawla, “Inferring user demographics and social strategies in mobile social networks.” In: KDD’14. ACM. p. 15–24, 2014.

[6] J. Hu, H. J. Zeng, H. Li, C. Niu, and Z. Chen, “Demographic prediction based on user’s browsing behavior,” Proceedings of the 16th international conference on World Wide Web, pp. 151-160, 2007.

[7] F. Iqbal, M. Debbabi, B. C. M. Fung, and L. A. Khan, “E-mail authorship verification for forensic investigation,” Proceedings of the 2010 ACM Symposium on Applied Computing, ser. SAC '10. New York, NY, USA: ACM, pp. 1591-1598, 2010.

[8] S. Kabbur, E. H. Han, and G. Karypis, “Content-based methods for predicting web-site demographic attributes,” Proceedings of ICDM, pp. 863-868, 2010.

[9] M. Koppel, S. Argamon, and A. R. Shimoni, “Automatically categorizing written texts by author gender,” Literary and Linguistic Computing, 17(4), pp : 401-412, 2002.

[10] S. Kotsiantis, D. Kanellopoulos, and P. Pintelas, “Handling unbalanced datasets: A review,” GESTS International Transactions on Computer Science and Engineering 30 (1), pp. 25-36, 2006.

[11] C. X. Ling, and V. S. Sheng, “Cost-sensitive learning and the class imbalance problem.” In: Sammut C (ed) Encyclopedia of machine learning. Springer, Berlin, 2008.

[12] S. Lu, Z. Meng, Z. Hui, Z. Chen, W. Wei, and W. Hao, "GenderPredictor: A Method to Predict Gender of Customers from E-commerce Website," In Web Intelligence and Intelligent Agent Technology (WI-IAT), 2015 IEEE/WIC/ACM International Conference on, vol. 3, pp. 13-16. 2015.

[13] T. M. Phuong, and D. V. Phuong, “Gender prediction using browsing history,” Proceedings of the Fifth International Conference KSE 2013, Volume 1. pp. 271-283, 2013.

[14] D. Nguyen, R. Gravel, D. Trieschnigg, and T. Meder, "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, 2013.

[15] J. J. C. Ying, Y. J. Chang, C. M. Huang, and V. S. Tseng, “ Demographic prediction based on users mobile behaviors,” In Nokia Mobile Data Challenge, 2012.

Dương Trần Đức Tốt nghiệp Đại học Khoa học Tự nhiên, Đại học Quốc gia Hà Nội ngành Công nghệ thông tin năm 1999. Tốt nghiệp Thạc sỹ chuyên ngành Hệ thống thông tin tại Đại học Tổng hợp Leeds, Vương Quốc Anh năm 2004. Hiện đang công tác tại Khoa Công nghệ Thông tin, Học viện Công nghệ Bưu chính Viễn thông.