HỌC VIỆN CÔNG NGHỆ BƯU CHÍNH VIỄN THÔNG

NGUYỄN XUÂN SANG

CẢI TIẾN THUẬT TOÁN SVM VỚI SVM SONG SONG, ỨNG DỤNG VÀO PHÂN LỚP VÀ DỰ BÁO SỐ KHÁCH HÀNG SỬ DỤNG DI ĐỘNG

Chuyên ngành: Hệ Thống thông tin
Mã số: 8.48.01.04

TÓM TẮT LUẬN VĂN THẠC SĨ
(Theo định hướng ứng dụng)

THÀNH PHỐ HỒ CHÍ MINH - NĂM 2021

Luận văn được hoàn thành tại:

HỌC VIỆN CÔNG NGHỆ BƯU CHÍNH VIỄN THÔNG

Người hướng dẫn khoa học: **PGS. TS Nguyễn Đình Thuân**

Phản biện 1: ……..

Phản biện 2: ……..

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

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

Có thể tìm hiểu luận văn tại:

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

MỞ ĐẦU

Dịch vụ thông tin di động ngày càng phát triển mạnh mẽ, trở thành một phần tất yếu trong cuộc sống của mỗi người dân Việt Nam. Quản lý khách hàng ngày càng nhận được sự quan tâm vì việc giữ chân khách hàng hiện tại mang lại lợi nhuận và quan trọng đối với các công ty viễn thông. Chi phí để tìm khách hàng mới lớn hơn nhiều so với chi phí để giữ chân khách hàng hiện tại trong kinh doanh, đặc biệt là trong thị trường viễn thông bão hòa. Hơn nữa, khách hàng dài hạn ít biến động hơn trong thị trường cạnh tranh.

Vì những nhu cầu đặt ra, các công ty viễn thông đang rất chú trọng và đầu tư nhiều hơn vào việc phát triển một mô hình dự báo khách hàng rời mạng. Nhiều phương pháp tiếp cận máy học đã được các nhà nghiên cứu đề xuất để dự báo khách hàng rời mạng, đặc biệt là trong lĩnh vực kinh doanh viễn thông. Các phương pháp tiếp cận máy học như vậy bao gồm các phương pháp phân lớp truyền thống như thuật toán Random Forest (RF), Naive Bayes (NB), Logistic Regression (LR) và Support Vector Machine (SVM).

Xuất phát từ những lý do trên, học viên chọn thực hiện đề tài luận văn tốt nghiệp chương trình đào tạo thạc sĩ có tên "**Cải tiến thuật toán SVM bằng SVM song song, ứng dụng vào phân lớp và dự báo số khách hàng sử dụng di động**".

Nhằm mục đích tìm hiểu về hướng tiếp cận mới này trong lĩnh vực khai thác dữ liệu, cũng như khả năng ứng dụng của vào trong thực tế, luận văn xin trình bày về phương pháp dự báo dữ liệu khách hàng rời mạng kết hợp giữa mô hình Logistic Regression (LR) và Support Vector Machine (SVM), cùng ứng dụng mô hình kết hợp này vào dự báo khách hàng rời mạng tại Viễn Thông Tây Ninh.

Đối tượng nghiên cứu của đề tài tập trung vào các mô hình dự báo dữ liệu khách hàng rời mạng, đặc biệt là mô hình LR, thuật giải SVM và phương pháp kết hợp mô hình LR và SVM trong dự báo dữ liệu khách hàng rời mạng. Bên cạnh đó đề tài còn trình bày kết quả áp dụng các mô hình dự báo dữ liệu khách hàng rời mạng vào trong thực tế dựa trên bộ dữ liệu được thu thập tại Viễn Thông Tây Ninh.

Phạm vi nghiên cứu của đề tài giới hạn trong việc tìm hiểu và ứng dụng các mô hình dự báo dữ liệu khách hàng rời mạng, mô hình LR, thuật giải SVM và mô hình kết hợp LR và SVM.

Tuy phạm vi nghiên cứu của đề tài giới hạn trong việc tìm hiểu và ứng dụng các mô hình dự báo dữ liệu khách hàng rời mạng nhưng đề tài cũng đã mang lại một số ý nghĩa về khoa học và thực tiễn. Về khoa học, kết quả thực nghiệm của đề tài cũng cố thêm tính đúng đắn của hướng tiếp cận kết hợp các mô hình dự báo dữ liệu khách hàng rời mạng nói chung và mô hình dự báo khách hàng rời mạng kết

hợp LR và SVM nói riêng. Về thực tiễn, kết quả dự báo của mô hình kết hợp LR và SVM giúp ích cho Viễn Thông Tây Ninh dự báo được khách hàng rời mạng để có thể lên kế hoạch tiếp cận và khuyến mãi hợp lý nhằm giữ chân khách hàng.

Luận văn được trình bày thành 4 chương:

Chương 1. Tổng quan: Giới thiệu về khách hàng rời mạng và dự báo khách hàng rời mạng. Trình bày về tình hình nghiên cứu trong và ngoài nước, xác định những vấn đề còn tồn tại trong các mô hình dự khách hàng rời mạng. Xác định mục tiêu, nội dung và phương pháp nghiên cứu của đề tài.

Chương 2: Mô hình kết hợp Logistic Regression và Support Vector Machine: Giới thiệu về mô hình kết hợp Logistic Regression và Support Vector Machine trong dự báo khách hàng rời mạng.

Chương 3: Dự báo tại Viễn Thông Tây Ninh: Giới thiệu về vấn để cần dự báo và ứng dụng mô hình kết hợp Logistic Regression và Support Vector Machine vào dự báo tại Viễn Thông Tây Ninh.

Chương 4: Kết luận và khuyến nghị: Đánh giá về các kết quả đạt được và hướng phát triển tiếp theo của đề tài.

CHƯƠNG 1. TỔNG QUAN

1.1 Khách hàng rời mạng và dự báo khách hàng rời mạng

1.1.1 Khách hàng rời mạng

Trong ngành viễn thông di động, thuật ngữ khách hàng rời mạng (churn customer), còn được gọi là khách hàng tiêu hao hoặc xáo trộn thuê bao, dùng để chỉ hiện tượng mất khách hàng. Quá trình di chuyển từ nhà cung cấp dịch vụ viễn thông này sang nhà cung cấp khác thường xảy ra do giá hoặc dịch vụ tốt hơn, hoặc do các lợi ích khác nhau mà công ty đối thủ cạnh tranh cung cấp.

Để thu hút thuê bao mới, các mạng di động phải thi nhau khuyến mại liên tục các tháng trong năm. Tuy nhiên, sau khi kết thúc mỗi đợt khuyến mại, số lượng thuê bao sử dụng hết tài khoản ngay lập tức rời mạng, tạm ngưng hoặc chuyển sang mạng khác lại tăng lên đáng kể, số thuê bao rời mạng nhiều hơn số thuê bao hòa mạng mới. Số lượng thuê bao đang hoạt

động tăng giảm bất thường, doanh thu không tăng theo tốc độ phát triển của số lượng thuê bao. Đây là kiểu cạnh tranh đang đi ngược lại với xu thế hội nhập của ngành thông tin di động Việt Nam. Ở góc độ quản lý vĩ mô, thực trạng trên cho thấy tiêu cực thị trường và gây lãng phí nguồn lực của ngành.

1.1.2 Dự báo khách hàng rời mạng

Trong dự báo khách hàng rời mạng, những giá trị trong quá khứ được thu thập và phân tích để tìm ra các mô hình phù hợp. Giá trị tương lai của khách hàng rời mạng được dự báo từ các mô hình đó. Do đó, dữ liệu trong quá khứ ảnh hưởng rất lớn đến quá trình xây dựng mô hình và cải thiện kết quả dự báo của mô hình.

1.2 Tình hình dự báo khách hàng rời mạng

Chính vì có nhiều ý nghĩa quan trọng nên từ lâu đã có nhiều nhà khoa học tìm hiểu, nghiên cứu và mô hình hóa khách hàng rời mạng để ứng dụng trong

phân tích, dự báo. Trong những năm gần đây nhiều mô hình, phương pháp được đề xuất để cải thiện kết quả, tăng độ chính xác cho dự báo dữ liệu khách hàng rời mạng nhưng nhìn chung các mô hình, phương pháp dự báo dữ liệu khách hàng rời mạng tập trung vào các hướng nghiên cứu

1.3 Những vấn đề còn tồn tại

Thứ nhất, mỗi một mô hình, phương pháp dự báo khách hàng rời mạng đều chỉ phù hợp với một số dạng dữ liệu đặc thù, mà chưa có một mô hình nào có thể dự báo tốt được cho tất cả các dạng dữ liệu, ví dụ như những mô hình dựa trên xác xuất thống kê như mô hình hồi quy Logistic Regression chỉ phù hợp để dự báo cho các dữ liệu dạng tuyến tính, còn các mô hình máy học như SVM lại chỉ phù hợp để dự báo cho các dạng dữ liệu phi tuyến tính. Mặt khác, dữ liệu trong thực tế đa số đều tính tuyến tính và phi tuyến tính, nên việc chỉ sử dụng một mô hình, phương pháp

để dự báo dữ liệu khách hàng rời mạng thường chưa mang lại kết quả như mong đợi.

Thứ hai, với tình hình thị trường viễn thông thay đổi nhanh chóng hiện nay, mọi thứ có thể khác rất nhanh chỉ trong một đêm. Vấn đề đặt ra cần xây dựng một mô hình tối ưu về thời gian để có thể đáp ứng ngay lập tức nhu cầu của viễn thông hiện nay.

1.4 Mục tiêu, nội dung, phương pháp nghiên cứu

Mục tiêu của đề tài nhằm tìm hiểu và áp dụng kết hợp mô hình Logistic Regression và SVM song song trong dự báo dữ liệu khách hàng rời mạng. Ứng dụng mô hình này vào dự báo số khách hàng sử dụng dịch vụ viễn thông của Viễn Thông Tây Ninh. Lý do đề tài lựa chọn mô hình Logistic Regression và phương pháp SVM song sing để kết hợp dự báo vì:

- Mô hình LR và phương pháp SVM trong ước lượng hồi quy đều là những mô hình, phương pháp dự báo khách hàng rời mạng cho kết quả dự báo

tương đối tốt. Tùy thuộc vào đặc tính của dữ liệu khách hàng rời mạng mà mô hình LR và phương pháp SVM thường được lựa chọn để thực hiện dự báo. Mô hình LR được chọn để dự báo cho thành phần tuyến tính của dữ liệu khách hàng rời mạng, còn phương pháp SVM thường được chọn để dự báo cho thành phần phi tuyến tính của dữ liệu khách hàng rời mạng.

- Thực tế đã có những nghiên cứu và ứng dụng cho thấy hiệu quả của phương pháp kết hợp LR và SVM trong dự báo như Ứng dụng mô hình kết hợp LR và SVM trong dự báo tín dụng. Mô hình kết hợp LR và SVM trong dự báo các chứng bệnh tim mạch trong y tế. Tất cả các nghiên cứu và ứng dụng trên đều cho thấy kết quả dự báo của mô hình kết hợp LR và SVM hiệu quả hơn so với các mô hình, phương pháp dự báo đơn lẻ.

- Tuy nhiên với hạn chế về độ phức tạp và thời gian của SVM, mô hình sẽ rất tốn tài nguyên khi sử dụng SVM truyền thống. Chính vì vậy việc cài đặt

sẽ sử dụng SVM song song thay thế cho SVM truyền thống. SVM song song sử dụng các GPUs nhằm tăng tốc độ tính toán nhưng vẫn đạt được độ chính xác tương đương với SVM truyền thống.

- Mô hình LR và phương pháp SVM đều là những mô hình, phương pháp dự báo dữ liệu khách hàng rời mạng hiệu quả và đã được nghiên cứu từ lâu.

Nội dung nghiên cứu của đề tài bao gồm:

- Tìm hiểu các mô hình dự báo dữ liệu khách hàng rời mạng, tập trung tìm hiểu về mô hình LR, mô hình SVM và mô hình kết hợp LR với SVM.

- Tiền xử lý dữ liệu để biến đổi dữ liệu về dạng phù hợp với các mô hình dự báo.

- Tiến hành cài đặt và thử nghiệm các mô hình dự báo dựa trên tập dữ liệu được thu thập từ dữ liệu của Viễn Thông Tây Ninh.

- So sánh, đánh giá kết quả dự báo của các mô hình với nhau và với dữ liệu thực tế.

Phương pháp nghiên cứu của đề tài:

- Tìm hiểu các mô hình, phương pháp trong dự báo khách hàng rời mạng.

- Tìm hiểu mô hình LR.

- Tìm hiểu về SVM và SVM song song.

- Tìm hiểu phương pháp kết hợp mô hình LR và SVM để tăng độ chính xác kết quả dự báo.

- Cài đặt thử nghiệm các mô hình, phương pháp dự báo dữ liệu khách hàng rời mạng.

Chương 2: MÔ HÌNH KẾT HỢP LOGISTIC REGRESSION VÀ SUPPORT VECTOR MACHINE

2.1 Mô hình Logistic Regression

Mô hình LR là một mô hình được sử dụng nhiều trong số các mô hình dự báo dữ liệu khách hàng rời mạng. Trong mục này sẽ trình bày về mô hình LR và giới thiệu mô hình LR.

2.1.1 Giới thiệu

Trong thống kê, mô hình logistic (hay mô hình logit) được sử dụng để lập mô hình xác suất của một lớp hoặc sự kiện nhất định đang tồn tại như đạt / không đạt, thắng / thua, sống / chết hoặc khỏe mạnh / bệnh. Điều này có thể được mở rộng để mô hình hóa một số lớp sự kiện như xác định xem một hình ảnh có chứa mèo, chó, sư tử, v.v. Mỗi đối tượng được phát hiện trong hình ảnh sẽ được gán một xác suất từ 0 đến 1, với tổng là 1.

Logistic Regression là một mô hình thống kê ở dạng cơ bản sử dụng một hàm logistic để mô hình hóa một biến phụ thuộc nhị phân. Trong hồi quy logistic, ước lượng các tham số của mô hình logistic.

2.1.2 Mô hình Logistic

Xét một mô hình logistic với các tham số cho trước, sau đó xem cách các hệ số có thể được ước tính từ dữ liệu. Hãy xem xét một mô hình có hai yếu tố dự báo: x_1 và x_2 và một biến nhị phân Bernoulli Y với tham số p = P(Y = 1). Ta giả định mối quan hệ tuyến tính giữa các biến dự báo và tỷ lệ logit là Y = 1.

2.1.3 Hàm Sigmoid

Hàm sigmoid là một hàm toán học có đường cong hình chữ "S" hoặc đường cong sigmoid đặc trưng.

2.1.4 Hàm mất mát và phương pháp tối ưu

Hàm logistic là một hàm sigmoid, nhận bất kỳ đầu vào thực tế nào và xuất ra giá trị từ 0 đến 1. Đối

với logit, điều này được hiểu là lấy tỷ lệ logit đầu vào và có xác suất đầu ra.

2.2 Support Vector Machine

Support Vector Machine (SVM) là một thuật giải quan trọng và được biết đến nhiều trong lĩnh vực máy học.

2.2.1 Giới thiệu

Trong không gian 2 chiều, ta biết rằng khoảng cách từ một điểm có toạ độ (x_0, y_0) tới *đường thẳng* có phương trình $w_1x + w_2y + b = 0$ được xác định bởi:

$$\frac{|w_1x_0 + w_2y_0 + b|}{\sqrt{w_1^2 + w_2^2}}$$

Trong không gian 3 chiều, khoảng cách từ một điểm có toạ độ (x_0, y_0, z_0) tới một *mặt phẳng* có phương trình $w_1x + w_2y + w_3z + b = 0$ được xác định bởi:

$$\frac{|w_1 x_0 + w_2 y_0 + w_3 z_0 + b|}{\sqrt{w_1^2 + w_2^2 + w_3^2}}$$

2.2.2 Độ rộng của margin

Nếu ta định nghĩa độ thõa mãn của một lớp tỉ lệ thuận với khoảng cách gần nhất từ một điểm của lớp đó tới đường/mặt phân chia, thì ở Hình 2.2 trái, lớp tròn đỏ sẽ không thõa mãn vì đường phân chia gần nó hơn lớp vuông xanh rất nhiều. Chúng ta cần một đường phân chia sao cho khoảng cách từ điểm gần nhất của mỗi lớp (các điểm được khoanh tròn) tới đường phân chia là như nhau. Khoảng cách như nhau này được gọi là margin.

3.2.3 Tìm kiếm siêu phẳng tối ưu

Giả sử rằng các cặp dữ liệu của training *set* là $(x_1, y_1), (x_2, y_2), \dots, (x_N, y_N)$ với vector $x_i \in \mathbb{R}^d$ thể hiện *đầu vào* của một điểm dữ liệu và y_i là *nhãn* của điểm dữ liệu đó, d là số chiều của dữ liệu và N là số điểm dữ liệu. Giả sử rằng *nhãn* của mỗi điểm dữ liệu

được xác định bởi $y_i = 1$ (lớp 1) hoặc $y_i = -1$ (lớp 2) giống như trong PLA.

2.2.4 Phương pháp Lagrange multipliers

Để tìm nghiệm theo công thức chúng ta sẽ dùng bài toán đối ngẫu Lagrange, công thức Lagrange được biểu diễn như sau:

$$\lambda = arg\max_{\lambda} g(\lambda), với: \lambda \geqslant 0 \text{ và } \sum_1^N \lambda_n y_n = 0 \quad (2\text{-}21)$$

Trong đó $g(\lambda) = -\frac{1}{2}\lambda^T K\lambda + 1^T\lambda, với\ K = V^T V$, K là ma trận nữa xác định dương, V là ma trận kết hợp của hai tập dữ liệu đầu vào.

Để giải bài toán tối ưu này ta sử dụng phương pháp Lagrange multipliers, hàm Lagrange được biểu diễn như sau

$$\begin{aligned}\mathcal{L}(x, y, \lambda) &= f(x, y) - \lambda . g(x, y) \\ &= 2 - x^2 - 2y^2 \\ &- \lambda(x + y - 1)\end{aligned} \tag{3.32}$$

2.2.5 Soft Margin và Kernel

Soft Margin: Hình 3.15 là một ví dụ về trường hợp phân lớp dữ liệu trong đó có 2 điểm dữ liệu nhiễu là x_i và x_j. Trong trường hợp này nếu xem hai điểm dữ liệu nhiễu này là các điểm dữ liệu bình thường và áp dụng thuật giải SVM sẽ dẫn đến kết quả là không tìm được một siêu phẳng tối ưu nào để phân lớp dữ liệu.

Kernel: Trong thực tế có rất nhiều dữ liệu không tuyến tính, dữ liệu có thể không được biểu diễn trong không gian vector. Trong khi đó, hàm phân lớp tuyến tính thì đơn giản và thuận lợi hơn nhiều. Điều này đã đặt ra yêu cầu phân lớp mở rộng cho phi tuyến.

2.2.6 SVM song song và bộ công cụ ThunderSVM

Sử dụng lợi thế của GPUs, giới thiệu một bộ công cụ gọi là ThuderSVM dùng để khai thác GPUs và CPUs đa nhân. Nhiệm vụ của bộ công cụ này là để giúp người dùng có thể dễ dàng ứng dụng SVMs một cách hiệu quả để giải quyết các bài toán. Từ đó chỉ ra rằng có thể huấn luyện SVM nhanh hơn bằng cách sử dụng xấp xỉ kernel SVM và tìm một phép biến đổi sao cho dữ liệu ban đầu là không tuyến tính được biến sang không gian mới.

Chương 3. DỰ BÁO KHÁCH HÀNG RỜI MẠNG TẠI VIỄN THÔNG TÂY NINH

3.1 Giới thiệu về công ty và bài toán dự báo

Đã gần khoảng 30 năm, kể từ khi Vinaphone - mạng di động đầu tiên của Việt Nam chính thức đi vào hoạt động. Tại thời điểm đó, di động còn là khái niệm xa lạ với đa số người tiêu dùng, số lượng thuê bao của mạng di động không nhiều vì vùng phủ sóng hạn chế và giá cước cũng như thiết bị đầu cuối còn cao. Mục đích của nghiên cứu: Phát hiện các thuê bao trả trước lâu năm có khả năng rời mạng bằng cách phân lớp kho thuê bao này với nhẫn gán trước là "rời mạng" và "không rời mạng" để có thể tác động và duy trì thuê bao.

Mục tiêu của nghiên cứu: Dự đoán các thuê bao trả trước dài hạn có khả năng rời mạng khi vẫn đang ở giai đoạn đang sử dụng của vòng đời thuê bao, tức là không phát sinh cước trong thời gian 30 ngày. Sau khi có mô hình phân tích tốt và chính xác cho dữ liệu, do

kho dữ liệu rất lớn nên sẽ tìm giải pháp đẩy nhanh quá trình huấn luyện nhằm tối ưu thời gian chạy mô hình.

3.2 Chuẩn bị và tiền xử lý dữ liệu

Giai đoạn chuẩn bị và xử lý dữ liệu ban đầu luôn là một giai đoạn quan trọng trong quy trình khai thác dữ liệu. Dữ liệu là một trong hai thành phần của phân lớp dữ liệu. Truy cập dữ liệu thực hiện việc trích xuất và thu thập dữ liệu cần thiết cho việc phân lớp mô hình thuê bao gửi đi.

Dữ liệu thu thập là dữ liệu di động trả trước vinaphone tại đơn vị tổng hợp từ nhiều nguồn như: thông tin thuê bao, lịch sử nạp thẻ, lịch sử gọi, SMS, … trong 6 tháng từ tháng 07/2019 đến 12/2019. Chi tiết các trường dữ liệu được mô tả trong Bảng 4.2. Từ bảng dữ liệu này sẽ tiến hành làm sạch dữ liệu bằng cách loại bỏ các dòng dữ liệu có trường trống hoặc null, các trường dữ liệu xuất hiện nhiều giá trị 0 có ảnh hưởng đến quá trình chạy mô hình.

3.3 Dự báo

Mô hình kết hợp Logistic và Support Vector Machine thực hiện hai bước dự báo. Bước đầu tiên sử dụng mô hình LR để dự báo thành phần tuyến tính. Bước thứ hai sử dụng phương pháp SVM để dự báo thành phần phi tuyến.

3.3.1 Dự báo thành phần tuyến tính bằng mô hình LR

Như đã trình bày trong chương 2, mô hình LR có thể dự báo cho dữ liệu tuyến tính. Đầu tiên xây dựng mô hình LR để dự báo cho phần tuyến tính của dữ liệu.

3.3.2 Dự báo thành phần phi tuyến bằng SVM

Sau khi dự báo thành phần tuyến tính bằng mô hình LR, tiếp tục huấn luyện dự báo thành phần phi tuyến bằng SVM.

3.4 Kết quả dự báo và đánh giá

Bảng 3.3: Kết quả dự báo của các mô hình

Mô hình	Độ chính xác	Thời gian(s)
Logistic Regression	0.830	10.2433
SVM	0.828	36.0965
Logistic Regression + Naïve Bayes	0.831	10.3104
Logistic Regression + Random Forest	0.787	11.8493
Logistic Regression + SVM	0.914	50.4576
Logistic Regression + ThunderSVM	0.914	9.36385

Từ kết quả dự báo này có thể thấy mô hình kết hợp LR và Support Vector Machine cho kết quả dự báo tốt nhất trên cùng một tập dữ liệu so với các mô hình khác là mô hình LR, mô hình SVM, mô hình kết hợp LR và NB, mô hình kết hợp LR và RF.

Chương 4. KẾT LUẬN VÀ KHUYẾN NGHỊ

4.1 Kết luận

Kết quả dự báo số khách hàng rời mạng củng cố thêm tính đúng đắn của hướng tiếp cận kết hợp các mô hình dự báo dữ liệu khách hàng nói chung và mô hình dự báo dữ liệu khách hàng rời mạng kết hợp Logistic Regression và Support Vector Machine nói riêng.

4.2 Khuyến nghị

Trong hầu hết các nghiên cứu hay ứng dụng về mô hình kết hợp Logistic Regression cho dữ liệu tuyến tính và các phương pháp máy học như Support Vector Machine cho dữ liệu phi tuyến, sự kết hợp này chỉ dừng lại ở việc tổng hợp các kết quả dự báo của các mô hình đơn lẻ lại với nhau để cho ra kết quả dự báo cuối cùng, nên giữa hai mô hình này không có liên kết gì với nhau. Do đó để kết quả dự báo hiệu quả

hơn cần có sự kết hợp chặt chẽ giữa các mô hình sao cho các mô hình này có thể hỗ trợ cho nhau trong việc dự báo. Chính vì vậy mà vấn đề làm thế nào để kết hợp chặt chẽ các phương pháp dự báo trong các mô hình kết hợp cũng là một hướng phát triển của để tài.

TÀI LIỆU THAM KHẢO

[1] Bộ Thông tin và Truyền thông. *Sách Trắng về Công nghệ thông tin và Truyền thông (CNTT-TT) Việt Nam 2018.* Hà Nội: Nhà xuất bản Thông tin và Truyền thông, 2018, tr.31-33.

[2] B. Huang, M. T. Kechadi, and B. Buckley, "*Customer churn prediction in telecommunications*," Expert Syst. Appl., vol. 39, pp. 1414–1425, Jan. 2012.

[3] M. Owczarczuk, "*Churn models for prepaid customers in the cellular telecommunication industry using large data marts*," Expert Systems with Applications, vol. 37,no. 6, pp. 4710 – 4712, 2010.

[4] G. Li and X. Deng, "*Customer churn prediction of china telecom based on cluster analysis and decision tree algorithm*," in Emerging Research in Artificial Intelligence and Computational Intelligence (J. Lei, F. Wang, H. Deng, and D. Miao, eds.), Communications in Computer and Information Science, pp. 319–327, Springer Berlin Heidelberg, 2012.

[5]SAGAR S. NikAM. "*A Comparative Study of Lópifcation Techniques in Data Mining Algorithms*".Department of Computer Science, K.K.Wagh College of Agriculture, Nashik, India. April 10, 2015

[6] Lingling Yang, Dongyang Li, Yao Lu, "*Prediction Modeling and Analysis for Telecom Customer Churn in Two Months*" School of Data and Computer Science, Sun Yat-Sen University, Guangzhou, China Guangdong Province Key Laboratory of Computational Science, 135 Xingang Xi Road, Guangzhou, China, January. 2019

[7] XIA Guo-en, JIN Wei-dong "*Model of Customer Churn Prediction on Support Vector Machine*", Volume 28, Issue 1, January 2008

[8] Hossein Abbasimehr, Mostafa Setak, M. J. Tarokh. "*A Neuro-Fuzzy Lópifier for Customer Churn Prediction*", K. N. Toosi University of Tech Tehran, Iran, Volume 19– No.8, April 2011

[9] Hemlata Jain, Ajay Khunteta, Sumit Srivastava. "*Churn Prediction in Telecommunication using Logistic Regression and Logit Boost*". Computer Science, School of Basic and Applied Sciences, Poornima University Jaipur-303905, India, January 2020.

[10] Mamdouh A. M. Abdelsalam & Doaa Akl Ahmed. "*Combining Forecasts from Linear and Nonlinear Models Using Sophisticated Approaches*". Astley Clarke Building, University of Leicester, Leicester. October 25, 2015.

[11] ALEX J. SMOLA and BERNHARD SCHOLKOPF. "*A tutorial on support vector regression*". RSISE, Australian National University, Canberra 0200, Australia. November 2003.

[12] Tony Van Gestel, Bart Baesens, Peter Van Dijcke, Johan A. K. Suykens, Joao Garcia, Thomas Alderweireld. "*Linear and non-linear credit scoring by combining logistic regression and support vector machines*". Group Risk Management, Dexia Group, Square Meeus 1, B-1000 Brussels, Belgium. Volume 1/Number 4, Fall 2005.

[13] Mythili T., Dev Mukherji, Nikita Padalia, and Abhiram Naidu. "*A Heart Disease Prediction Model using SVM-Decision Trees-Logistic Regression (SDL)".* School of Computing Sciences and Engineering, VIT University Vellore – 632014, Tamil Nadu, India. Volume 68– No.16, April 2013.

[14] Zeyi Wen, Jiashuai Shi, Qinbin Li, Bingsheng He, Jian Chen. "*ThunderSVM: a fast SVM library on*

GPUs and CPUs". The Journal of Machine Learning ResearchVolume 19Issue 1January 2018 pp 797–801.

[15] S. I. Gallant. "*Perceptron-Based Learning Algorithms*," in IEEE transaction on neural. February 1990.

[16] S. Boyd and L. Vandenberghe. "*Convex Optimization*". Cambridge University Press, 2014.

[17] Univesity of Florida. The Foundation for The Gator Nation (Nov. 2016), http://www.cise.ufl.edu/lớp/cis4930sp11dtm/notes/intro_svm_new.pdf

[18] Z. Wen et al. "*ThunderSVM*." Internet: https://github.com/Xtra-Computing/thundersvm

[19] Yuan-chin Ivan Chang. "*Boosting SVM Lớpifiers with Logistic Regression*". Institute of Statistical Science, Academia Sinica, Taipei, Taiwan. June 2003.

[20] Eviyana Atmanegara, Taly Purwa. "*Hybrid Support Vector Machine and Logistic Regression for*

Multilớp Lớpification: A Case Study on Wine Dataset". Indonesian Journal of Data ScienceVol.1, No.1, April 2021, pp. 1~7.