ĐẠI HỌC QUỐC GIA HÀ NỘI
TRƯỜNG ĐẠI HỌC CÔNG NGHỆ

NGUYỄN NGỌC TUÂN

ÁP DỤNG KỸ THUẬT KHAI PHÁ DỮ LIỆU DỰ BÁO THUÊ BAO RỜI MẠNG TRONG MẠNG DI ĐỘNG

LUẬN VĂN THẠC SĨ CÔNG NGHỆ THÔNG TIN

Hà Nội - 2016

ĐẠI HỌC QUỐC GIA HÀ NỘI
TRƯỜNG ĐẠI HỌC CÔNG NGHỆ

NGUYỄN NGỌC TUÂN

ÁP DỤNG KỸ THUẬT KHAI PHÁ DỮ LIỆU DỰ BÁO THUÊ BAO RỜI MẠNG TRONG MẠNG DI ĐỘNG

Ngành: Công nghệ Thông tin

Chuyên ngành: Hệ thống Thông tin

Mã số: 60480104

LUẬN VĂN THẠC SĨ CÔNG NGHỆ THÔNG TIN

NGƯỜI HƯỚNG DẪN KHOA HỌC: PGS.TS. NGUYỄN HÀ NAM

Hà Nội - 2016

LỜI CẢM ƠN

Tôi xin gửi lời cảm ơn sâu sắc tới PGS.TS.Nguyễn Hà Nam, Trường Đại học Công nghệ - Đại học Quốc gia Hà Nội, người thầy đã dành nhiều thời gian tận tình chỉ bảo, hướng dẫn, giúp đỡ tôi trong suốt quá trình tìm hiểu, nghiên cứu. Thầy là người định hướng và đưa ra nhiều góp ý quý báu trong quá trình tôi thực hiện luận văn.

Tôi xin chân thành cảm ơn các thầy, cô ở khoa Công nghệ thông tin – Trường Đại học Công nghệ - ĐHQGHN đã cung cấp cho tôi những kiến thức và tạo cho tôi những điều kiện thuận lợi trong suốt quá trình tôi học tập tại trường.

Tôi cũng bày tỏ lòng biết ơn về sự giúp đỡ của lãnh đạo cơ quan, đồng nghiệp đã cung cấp dữ liệu, tài liệu và cho tôi những lời khuyên quý báu. Tôi xin cảm ơn gia đình, người thân, bạn bè và các thành viên trong nhóm nghiên cứu luôn động viên và tạo mọi điều kiện tốt nhất cho tôi.

Tôi xin chân thành cảm ơn!

Hà Nội, tháng 5 năm 2016

Họ và tên

Nguyễn Ngọc Tuân

LỜI CAM ĐOAN

Tôi xin cam đoan đây là đề tài nghiên cứu của riêng tôi, thực hiện dưới sự hướng dẫn của PGS.TS. Nguyễn Hà Nam.

Các kết quả nêu trong luận văn là trung thực và chưa được ai công bố trong bất cứ công trình nào khác.

Hà Nội, tháng 5 năm 2016

Họ và tên

Nguyễn Ngọc Tuân

MỤC LỤC

DANH MỤC CÁC KÝ HIỆU VÀ CHỮ VIẾT TẮT

Chữ viết tắt	Ý nghĩa
KPDL	Khai phá dữ liệu
CSDL	Cơ sở dữ liệu
OLAP	Online analytical processing
MobiFone	Tổng công ty Viễn thông MobiFone
VLR	Visitor Location Register. Tổng đài ghi nhận đăng nhập mạng của thuê bao di động
3K3D_VLR	Chỉ tiêu xác định thuê bao 1 tháng có > 3 ngày nhập mạng VLR và phát sinh doanh thu >3000
GSM	Global System for Mobile Communications Mạng thông tin di động
CDR	Call Data Record
CLV	Giá trị vòng đời khách hàng (custommer lifecycle value)

DANH MỤC CÁC BẢNG

DANH MỤC CÁC HÌNH VẼ, ĐỒ THỊ

LỜI MỞ ĐẦU

Thuê bao rời mạng luôn là vấn đề "đau đầu" của các nhà mạng trong nước cũng như trên thế giới bởi lẽ khách hàng (thuê bao) chính là người mang lại doanh thu và duy trì hoạt động của các nhà mạng. Để duy trì và phát triển hoạt động kinh doanh của mình, các nhà mạng phải tìm mọi cách để phát triển thuê bao mới nhưng đồng thời cũng phải tìm cách để duy trì hoạt động của các thuê bao hiện hữu. Theo các nghiên cứu và thực tế triển khai của các nhà mạng cho thấy, tổng chi phí để phát triển một thuê bao mới cao hơn nhiều so với việc duy trì một thuê bao hiện hữu. Trong khi đó, doanh thu từ các thuê bao hiện hữu (đặc biệt là các thuê bao lâu năm) cao hơn nhiều so với doanh thu của các thuê bao mới (theo thống kê của MobiFone thuê bao lâu năm có doanh thu trung bình cao hơn 48% so thuê bao mới). Chính vì lý do trên, các nhà mạng trên thế giới không ngừng tìm kiếm các giải pháp và nghiên cứu phát triển các ứng dụng để xác định, dự đoán sớm thuê bao có khả năng rời mạng để có biện pháp kịp thời tác động nhằm duy trì thuê bao đó hoạt động.

Xuất phát từ yêu cầu đặt ra đối với đơn vị mình, tôi đã thực hiện đề tài luận văn "**ÁP DỤNG KỸ THUẬT KHAI PHÁ DỮ LIỆU DỰ BÁO THUÊ BAO RỜI MẠNG TRONG MẠNG DI ĐỘNG".** Luận văn đi sâu vào việc áp dụng các kỹ thuật khai phá dữ liệu từ hành vi sử dụng của các thuê bao di động từ đó dự báo thuê bao có khả năng rời mạng.

Luận văn gồm có phần mở đầu, kết luận và 04 chương, cụ thể như sau:

Chương 1: Giới thiệu tổng quan về mạng di động và các vấn đề liên quan

Chương này trình bày về tổng quan mạng di động, thuê bao rời mạng. Phát biểu bài toán và các nghiên cứu liên quan.

Chương 2: Khai phá dữ liệu và các kỹ thuật phân tích dự báo

Nghiên cứu các cơ sở lý thuyết KPDL, các vấn đề liên quan đến KPDL trong bài toán dự báo thuê bao rời mạng. Tìm hiểu các kỹ thuật khai phá dữ liệu sử dụng trong bài toán phân lớp, dự báo (chuẩn bị dữ liệu, lựa chọn thuộc tính, phân tích đặc trưng, cây quyết định, …) và áp dụng kỹ thuật KPDL trên ứng dụng WEKA.

Chương 3: Giải pháp phát hiện thuê bao di động có khả năng rời mạng

Các phương pháp để phát hiện thuê bao di động có khả năng rời mạng bao gồm các kỹ thuật dựa vào đặc trưng thuê bao, phương pháp ứng dụng kỹ thuật khai phá dữ liệu.

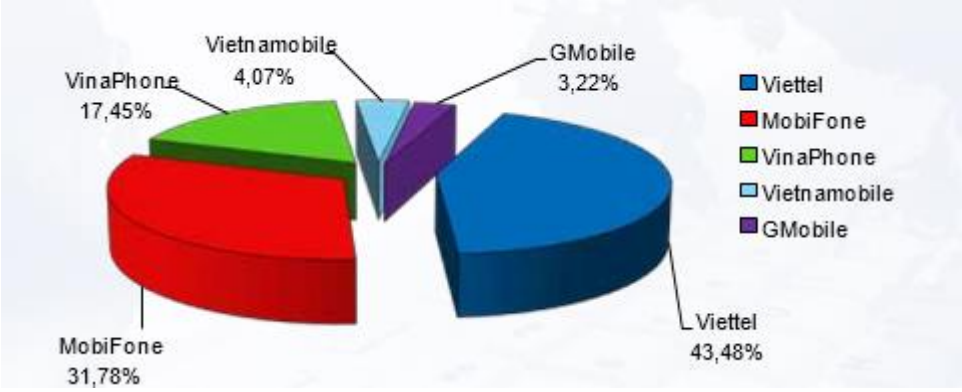
Chương 4: Thực nghiệm và đánh giá kết quả

Nội dung chủ yếu là áp dụng các mô hình đã tìm hiểu vào việc dự báo thuê bao rời mạng. Trước tiên áp dụng các kỹ thuật cơ bản trên toàn bộ dữ liệu đã được chuẩn bị với mô hình phân tích đặc trưng, cây quyết định và đánh giá. Tiếp theo sử dụng các kỹ thuật lựa chọn thuộc tính để cải tiến tốc độ xử lý và đánh giá độ chính xác của mô hình dự báo. Cuối cùng tiến hành đánh giá kết quả dự báo của từng mô hình và đưa ra mô hình dự báo khuyến nghị để áp dụng vào bài toán thực tế.

Chương 1 Giới thiệu tổng quan về mạng di động và các kiến thức cơ sở liên quan

1.1. Giới thiệu về mạng di động

Hai mươi năm qua ghi nhận một thời gian dài phát triển sôi động, vượt bậc của ngành công nghiệp viễn thông nói chung và thông tin di động nói riêng. Được thành lập từ năm 1993, sau 22 năm phát triển, Tổng Công ty Viễn thông MobiFone (tiền thân là Công ty Thông tin di động), từ một doanh nghiệp cỡ vừa và nhỏ, đã phát triển trở thành một doanh nghiệp lớn, hàng đầu Việt Nam, có doanh thu lên tới hơn 40.000 tỷ, lợi nhuận hơn 6.000 tỷ VNĐ hàng năm và cung cấp dịch vụ viễn thông di động cho trên 20 triệu khách hàng sử dụng thường xuyên. Thương hiệu MobiFone cũng trở thành một trong những thương hiệu có uy tín hàng đầu. Bước sang những năm đầu của thập kỷ mới, cùng với sự bão hòa của số lượng thuê bao toàn thị trường, những thách thức đang dần trở nên thực tế hơn, đòi hỏi MobiFone phải có những sự theo dõi sát sao hơn với tình hình kinh doanh. Trong khi ở giai đoạn bùng nổ của thập niên đầu của thế kỷ 21, việc phát triển thuê bao là khá dễ dàng khi gần như cứ bổ sung thêm các trạm mới là có thể có thêm thuê bao, đến những năm đầu thập kỷ thứ hai, khi số lượng thuê bao trên toàn thị trường trở nên bão hòa, mạng lưới cũng đã phủ sóng gần như khắp quốc gia, mỗi khách hàng tiềm năng đều đã có 1 hay nhiều thuê bao, việc phát triển thuê bao mới trở nên khó khăn hơn bao giờ hết. Các nhà cung cấp dịch vụ viễn thông cũng cạnh tranh với nhau gay gắt, quyết liệt, giành giật nhau từng thuê bao, bao gồm cả các thuê bao đang sử dụng dịch vụ của mạng đối thủ. Chính vì vậy, không chỉ việc phát triển thuê bao mà ngay cả việc giữ thuê bao cũng trở nên cấp bách.



Hình 1-1 Thị phần các thuê bao di động tính đến 2013 [2]

1.2. Sơ lược tình hình nghiên cứu trên thế giới

Xu hướng khách hàng ngừng sử dụng sản phẩm/dịch vụ của công ty trong một khoảng thời gian nhất định được định nghĩa là khách hàng rời mạng Chandar, Laha, &Krishna [5].

Các công ty thì luôn muốn có thêm càng nhiều khách hàng càng tốt. Mặc dù vậy, qua thời gian, tỷ lệ khách hàng mới / khách hàng rời mạng có xu hướng tiến tới bằng 1. Vì vậy, tác động của rời mạng ngày càng trở nên mạnh mẽ và cần được quan tâm hơn.

Việc rời mạng thường gắn liền với vòng đời của ngành, khi ngành đang trong giai đoạn phát triển, việc bán hàng tăng trưởng bùng nổ, số khách hàng mới vượt xa số khách hàng rời mạng, nhưng khi ở giai đoạn bão hòa, các công ty sẽ tập trung vào việc giảm tỉ lệ rời mạng.

Thời điểm khách hàng rời mạng sẽ cho biết khách hàng gắn bó với công ty trong bao lâu, giá trị vòng đời của khách hàng (CLV) đối với công ty. CLV được tính bằng tổng doanh thu mà Công ty thu được từ khách hàng trong suốt vòng đời của khách hàng trừ đi tổng chi phí thu hút khách hàng, bán hàng, dịch vụ khách hàng (quy ra tiền).

Các nghiên cứu trước đây đã đưa ra khái niệm khách hàng rời mạng từ nhiều quan điểm khác nhau. Theo Olafsson, Li, và Wu[10], có 2 loại rời mạng khác nhau. Loại thứ nhất là *rời mạng chủ động* (nghĩa là khách hàng chủ động chọn dừng sử dụng dịch vụ). Loại thứ hai là *rời mạng bị động* (nghĩa là khi những khách hàng không còn là khách hàng tốt nữa và công ty lựa chọn dừng mối quan hệ này).

Burez và Van den Poel [9] đã chia rời mạng chủ động thành 2 nhóm: Rời mạng do vấn đề thương mại và rời mạng do vấn đề tài chính. Rời mạng do vấn đề thương mại là trường hợp khách hàng không gia hạn hợp đồng có thời hạn cố định của họ khi hợp đồng hết hạn. Rời mạng do vấn đề tài chính là trường hợp khách hàng ngừng thanh toán trong quá trình thực hiện hợp đồng mà họ đang bị ràng buộc.

Ngày nay, khách hàng rời mạng đã trở thành vấn đề quan tâm chính của các công ty trong tất cả các lĩnh vực và các công ty buộc phải hành động để xử lý vấn đề này.

Xem xét tỷ lệ rời mạng của các ngành khác nhau, có thể nhận thấy ngành viễn thông là một trong những ngành có tỉ lệ rời mạng cao nhất với tỉ lệ rời mạng trung bình hàng năm từ 20% đến 40%. Khách hàng rời mạng trong lĩnh vực viễn thông được hiểu là khách hàng chuyển từ nhà cung cấp này sang nhà cung cấp khác.

Có 2 cách tiếp cận cơ bản đối với việc quản lý rời mạng. Cách tiếp cận thứ nhất là tiếp cận không có mục tiêu dựa vào các sản phẩm nổi trội và truyền thông rộng rãi để tăng lòng trung thành và duy trì khách hàng. Cách tiếp cận thứ hai là tiếp cận có mục tiêu dựa vào việc xác định những khách hàng có khả năng rời mạng, sau đó cung cấp cho họ những giá trị khuyến khích trực tiếp (khuyến mại) hoặc tạo ra các gói dịch vụ phù hợp cho khách hàng để giữ họ ở lại.

Cách tiếp cận có mục tiêu gồm 2 loại: bị động và chủ động. Với cách tiếp cận bị động, công ty chờ cho đến khi khách hàng liên hệ với công ty để hủy dịch vụ, công ty sau đó mới đưa ra cho khách hàng những chính sách khuyến khích, ví dụ khuyến mại giảm giá, để giữ khách hàng ở lại. Với cách tiếp cận chủ động, công ty cố gắng xác định những khách hàng có khả năng rời mạng trong một thời gian ngắn tiếp theo. Sau đó, công ty sẽ thực hiện những chương trình hoặc chính sách đặc biệt để giữ cho khách hàng không rời mạng. Cách tiếp cận chủ động có những lợi ích là chi phí khuyến khích thấp (bởi vì phần khuyến khích đó không cần thiết phải cao như tại thời điểm khách hàng đã quyết định sẽ rời mạng rồi) và bởi vì khách hàng không được chuẩn bị sẵn để thương lượng có được mức khuyến khích tốt hơn với lý do rời mạng. Tuy nhiên, cách tiếp cận này sẽ gây lãng phí nếu việc dự đoán rời mạng là không chính xác, bởi vì sau đó công ty sẽ phải lãng phí một lượng lớn chi phí để khuyến khích những khách hàng thực tế vẫn ở lại với mạng mình.

Để giải quyết vấn đề này, rất nhiều nỗ lực đã thực hiện để có được cái nhìn chính xác hơn về rời mạng. Nhìn chung, các nghiên cứu trong lĩnh vực này đều tập trung về một trong những mục đích chính sau: tìm ra các nhân tố ảnh hưởng đến khách hàng rời mạng, hoặc xây dựng mô hình cho việc dự đoán khách hàng rời mạng.

1.3. Phát biểu bài toán

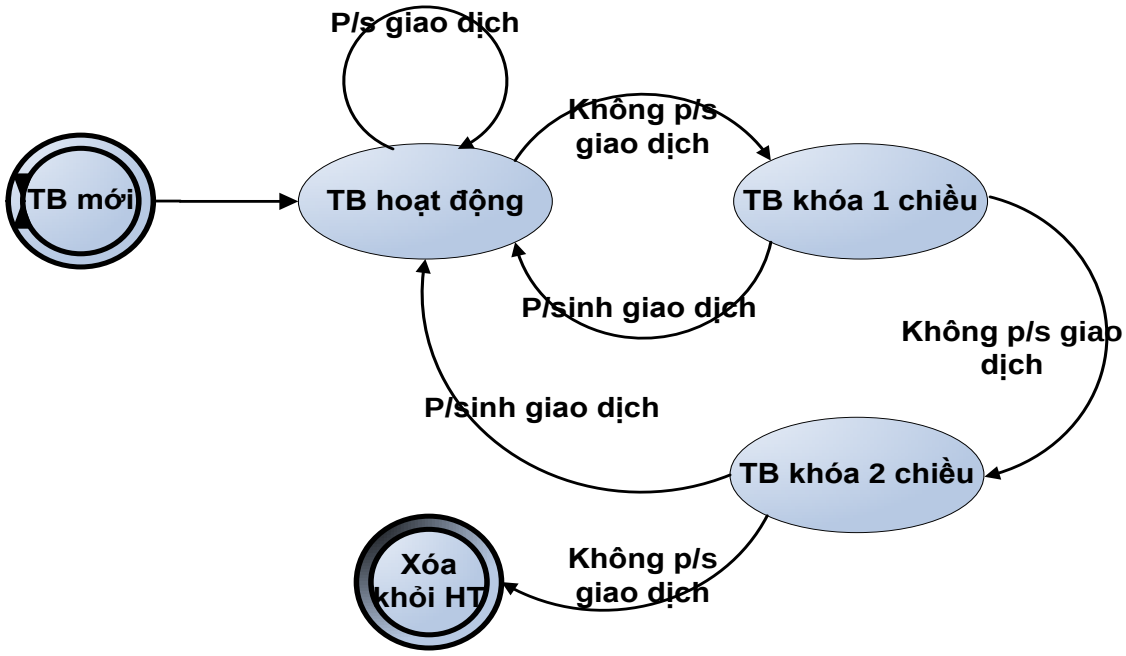
1.3.1. Chu trình của thuê bao di động

Rời mạng thường được phân thành 2 loại: rời mạng chủ động và rời mạng bị động. Rời mạng chủ động là trường hợp những khách hàng chọn để rời mạng, việc rời mạng là do lựa chọn của khách hàng. Ví dụ, khách hàng chuyển sang mạng đối thủ hoặc chuyển đổi sang hợp đồng thuê bao trả sau. Rời mạng bị động là trường hợp khách hàng bị nhà cung cấp ngừng cung cấp dịch vụ, thường là vì lý do gian lận hoặc nợ cước. Rời mạng vì lý do gian lận dường như rất hiếm xảy ra. Rời mạng do nợ cước thì chỉ xảy ra với thuê bao trả sau. Như đã đề cập ở trên, trong nghiên cứu này, chúng tôi chỉ tập trung vào vấn đề rời mạng của thuê bao trả trước. Vì vậy, rời mạng bị động xảy ra khi khách hàng không nạp lại tiền trong một khoảng thời gian đủ dài theo quy định.

Một trong những vấn đề quan trọng nhất của thuê bao trả trước là thiếu một định nghĩa đủ rõ ràng. Khi xem xét rời mạng đối với thuê bao trả sau, ngày thuê bao bị khóa 2 chiều (ngày thuê bao ngừng kết nối với mạng) chính là ngày rời mạng, đây là ngày thuê bao thực sự ngừng sử dụng dịch vụ của nhà cung cấp. Tuy nhiên, trường hợp thuê bao trả trước, ngày khóa 2 chiều cũng không thực sự là ngày rời mạng. Điều này có

thể được nhìn một cách rõ ràng hơn thông qua các giai đoạn khác nhau của thuê bao trả trước. Thuê bao trả trước thông thường sẽ có 4 giai đoạn:

- Giai đoạn 1: Kích hoạt, trở thành thuê bao mới
- Giai đoạn 2: Hoạt động bình thường (thuê bao phát sinh các giao dịch và hoạt động bình thường trên mạng).
- Giai đoạn 3: Khóa 1 chiều (thuê bao chỉ có thể nhận các giao dịch chiều đến, không thực hiện được các giao dịch chiều đi). Trường hợp này, thuê bao nạp tiền trở lại thì sẽ quay lại trạng thái hoạt động bình thường (giai đoạn 2). Trường hợp thuê bao không nạp lại tiền, sau một khoảng thời gian nhất định theo quy định, thuê bao sẽ bị chuyển sang giai đoạn 4 (khóa 2 chiều).
- Giai đoạn 4: Khóa 2 chiều.



Hình 1-2 Vòng đời thuê bao

Tùy từng loại hình thuê bao mà khoảng thời gian chuyển từ giai đoạn này sang giai đoạn tiếp theo sẽ khác nhau. Ví dụ, đối với thuê bao MobiQ, thuê bao chỉ cần phát sinh một giao dịch thì sẽ luôn có 60 ngày sử dụng ở giai đoạn 2. Trong 60 ngày đó, nếu khách hàng không sử dụng dịch vụ gì, không nạp tiền thì thuê bao sẽ chuyển sang giai đoạn 3 (khóa 1 chiều). Thuê bao sẽ có khoảng 10 ngày ở giai đoạn 2 này, nếu không nạp lại tiền và sử dụng lại dịch vụ thì thuê bao sẽ chuyển sang giai đoạn 3 (khóa 2 chiều). Ở giai đoạn 3, thuê bao sẽ có 30 ngày giữ số trước khi bị xóa hoàn toàn khỏi hệ thống. Như vậy, từ giai đoạn 2 đến hết giai đoạn 4, thuê bao có tới 100 ngày. Thực tế, thuê bao đã có thể rời mạng tại bất cứ thời điểm nào trong 100 ngày này.

1.3.2. Phát biểu bài toán

Vì lý do thời điểm tác động được đến thuê bao quan trọng nên việc xác định thời điểm nào được coi là rời mạng sẽ rất quan trọng trong việc dự đoán rời mạng và thực hiện các tác động để duy trì, ngăn chặn thuê bao rời mạng. Trong phạm vi đề tài này, khái niệm "rời mạng" được xác định là trường hợp khách hàng không phát sinh cước (không phát sinh bất cứ giao dịch nào hoặc không có biến động về tài khoản trong vòng một tháng). Tức là, thuê bao được xác định là rời mạng khi có thể thực tế vẫn đang ở giai đoạn 2. Lý do sử dụng khái niệm rời mạng này như sau:

Theo kinh nghiệm thực tế, thuê bao trả trước chuyển sang giai đoạn 3 (khóa 1 chiều) thì hầu như rất khó liên lạc, thậm chí đã vứt bỏ sim-card ra khỏi điện thoại. Do vậy, việc tác động đến thuê bao ở giai đoạn này hầu như không có hiệu quả.

Mốc "không phát sinh cước" cho phép dự đoán thuê bao rời mạng khi thuê bao vẫn còn đang ở giai đoạn 2, đảm bảo còn đủ thời gian để thực hiện tác động trước khi thuê bao chuyển sang giai đoạn 3.

Mục đích của nghiên cứu: **Phát hiện các thuê bao trả trước lâu năm có khả năng rời mạng để có thể tác động và duy trì thuê bao**

Mục tiêu của nghiên cứu: **Dự báo các thuê bao trả trước lâu năm có khả năng rời mạng khi vẫn còn trong giai đoạn 2 của vòng đời thuê bao tức là không phát sinh cước trong khoảng thời gian 1 tháng.**

1.4. Kết luận chương 1

Chương này giới thiệu về bài toán, những yêu cầu đặt ra cần giải quyết đối với bài toán đồng thời trình bày một số hướng nghiên cứu về thuê bao rời mạng trong mạng di động, hướng tiếp cận của luận văn.

Chương 2 Khai phá dữ liệu và các kỹ thuật phân tích dự báo

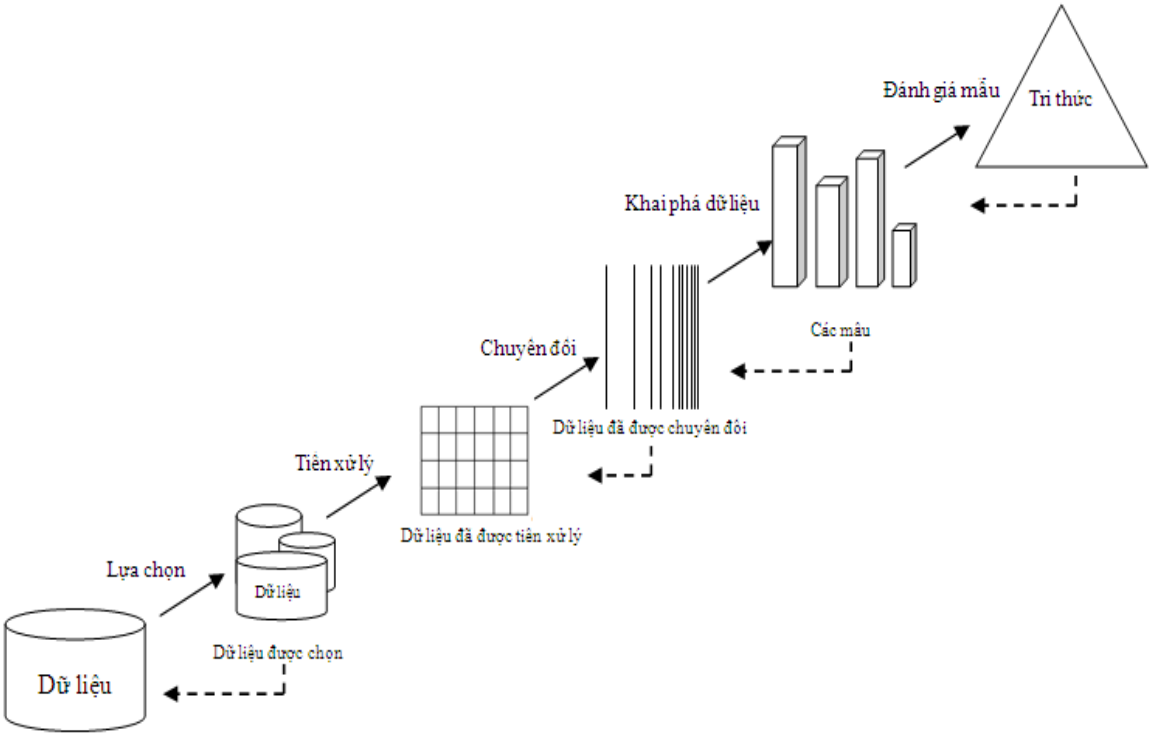
2.1. Khai phá dữ liệu

2.1.1. Khái niệm KPDL

KPDL là quá trình khảo sát và phân tích một khối lượng lớn các dữ liệu được lưu trữ trong các cơ sở dữ liệu, kho dữ liệu…để từ đó trích xuất ra các thông tin quan trọng, có giá trị tiềm ẩn bên trong. Do nhu cầu nghiệp vụ cần có cách nhìn thông tin trên quy mô toàn đơn vị.

Các dữ liệu này được thu từ nhiều nguồn, đa số là từ các phần mềm nghiệp vụ như: phần mềm tài chính, kế toán, các hệ thống quản lý tài nguyên doanh nghiệp, các hệ thống quản lý khách hàng hay từ tác công cụ lưu trữ thông tin trên web…

Đây là những khối dữ liệu khổng lồ nhưng những thông tin mà nó thể hiện ra thì lộn xộn và "nghèo" đối với người dùng. Kích thước của khối dữ liệu khổng lồ đó cũng tăng với tốc độ rất nhanh chiếm nhiều dung lượng lưu trữ. KPDL sẽ giúp trích xuất ra các mẫu điển hình có giá trị và biến chúng thành những tri thức hữu ích. Quá trình này gồm một số bước được thể hiện trong hình sau.



Hình 2-1 Các bước xây dựng một hệ thống KPDL [1]

Ý nghĩa cụ thể của các bước như sau:

- Lựa chọn dữ liệu liên quan đến bài toán quan tâm.

- Tiền xử lý dữ liệu, làm sạch dữ liệu

- Chuyển đổi dữ liệu về dạng phù hợp thuận lợi cho việc khai phá.

- KPDL, trích xuất ra các mẫu dữ liệu (mô hình).

- Đánh giá mẫu.

- Sử dụng tri thức khai phá được.

2.1.2. **Những nhóm bài toán của KPDL**

KPDL có thể được dùng để giải quyết hàng trăm bài toán với những mục đích và nhiệm vụ khác nhau. Dựa trên bản chất tự nhiên của các bài toán đó, người ta có thể nhóm các bài toán đó thành những nhóm sau:

Bài toán phân loại

Bài toán phân loại là một trong những bài toán phổ biến nhất của KPDL, ví dụ như: phân tích xem loại khách hàng nào có khả năng cao nhất sẽ chuyển sang dùng sản phẩm dịch vụ của đối thủ cạnh tranh của công ty (churn analysis), quản lý rủi ro hay lựa chọn ảnh quảng cáo nào sẽ xuất hiện đối với mỗi loại khách hàng…

Phân loại là tổ chức dữ liệu trong các lớp cho trước, còn được gọi là học có quan sát. Phân loại sử dụng các nhãn lớp cho trước để sắp xếp các đối tượng. Trong đó, có một tập huấn luyện gồm các đối tượng đã được kết hợp với các nhãn đã biết. Những thuật toán học có quan sát sẽ được áp dụng cho tập các đối tượng cần phân loại để từ đó mô hình phân loại chúng.

Một số thuật toán dùng trong bài toán phân loại như: cây quyết định, mạng nơ ron, mạng Naïve Bayes.

Bài toán hân cụm

Bài toán phân cụm hay còn gọi là phân đoạn. Điểm khác với bài toán phân loại là ở đây các nhãn lớp chưa biết và không có huấn luyện. Các đối tượng được phân loại dựa trên các thuộc tính tương đồng giữa chúng. Bài toán phân lớp hay còn gọi là học không có giám sát.

Bài toán phân tích luật kết hợp

Bài toán này đôi khi còn gọi là bài toán phân tích giỏ hàng bởi vì nó được sử dụng rộng rãi trong phân tích các giao dịch dữ liệu, các bài toán lựa chọn hàng hóa đi kèm…

Phân tích luật kết hợp khám phá ra các luật kết hợp thể hiện mối liên hệ giữa các thuộc tính dữ liệu thường xuất hiện cùng nhau trong các tập dữ liệu.

Bài toán hồi quy

Bài toán hồi quy cũng tương tự như bài toán phân loại. Điểm khác biệt là hồi quy dự đoán cho các dữ liệu liên tục.

Bài toán dự đoán

Dự đoán là một mảng quan trọng của KPDL. Dự đoán có hai loại chính: một loại là dự đoán về một số giá trị dữ liệu chưa biết hay có xu hướng sắp xảy ra, còn loại kia là dự đoán để phân lớp dựa trên một tập huấn luyện và giá trị thuộc tính của đối tượng. **Trong phạm vi luận văn, tác giả sẽ sử dụng loại dự đoán thứ hai.**

Bài toán phân tích chuỗi

Phân tích chuỗi được sử dụng để tìm ra các mẫu trong một loạt các giá trị hay trạng thái rời rạc. Ví dụ như việc chọn mua hàng của khách hàng có thể mô hình là một chuỗi dữ liệu. Hành động chọn mặt hàng A, sau đó chọn mặt hàng B, C… là một chuỗi các trạng thái rời rạc. Trong khi đó thời gian lại là chuỗi số liên tục.

Phân tích chuỗi và phân tích luật kết hợp giống nhau ở chỗ đều phân tích tập hợp các đối tượng hay trạng thái. Điểm khác nhau là mô hình chuỗi phân tích sự chuyển của các trạng thái, trong khi mô hình luật kết hợp thì coi mỗi một mặt hàng trong giỏ hàng là như nhau và độc lập. Với mô hình chuỗi, việc chọn mặt hàng A trước mặt hàng B hay chọn mặt hàng B trước A sau là khác nhau. Còn ở mô hình kết hợp thì cả hai trường hợp là như nhau.

Bài toán phân tích độ lệch

Bài toán này còn được gọi là phát hiện điểm biên. Điểm biên là những đối tượng dữ liệu mà không tuân theo các đặc tính, hành vi chung. Bài toán phát hiện điểm biên ứng dụng rất nhiều trong các ứng dụng. Ứng dụng quan trọng nhất của bài toán phát hiện điểm biên là bài toán kiểm tra xác nhận thẻ tín dụng…

2.1.3. Các bước xây dựng một giải pháp về KPDL

Có rất nhiều tác giả đưa ra các bước của một hệ thống KPDL, mọi sự phân chia chỉ mang tính chất tương đối và tư tưởng chủ đạo của nó là như sau:

- Bước 1: Xác định mục tiêu bài toán.

- Bước 2: Thu thập dữ liệu.

- Bước 3: Làm sạch dữ liệu và chuyển đổi dữ liệu.

- Bước 4: Xây dựng mô hình.

- Bước 5: Đánh giá mô hình hay đánh giá mẫu.

- Bước 6: Báo cáo.

- Bước 7: Dự đoán.

- Bước 8: Tích hợp vào ứng dụng

- Bước 9: Quản lý mô hình

2.1.4. Ứng dụng KPDL trong viễn thông

Là một ngành công nghiệp với khối lượng dữ liệu cần xử lý rất lớn, ngành công nghiệp viễn thông đã nhanh chóng phát triển từ cung cấp dịch vụ điện thoại cố định để cung cấp nhiều dịch vụ thông tin liên lạc toàn diện khác. Chúng bao gồm di động, điện thoại thông minh, truy cập Internet, email, tin nhắn văn bản, hình ảnh, máy tính và truyền dữ liệu web và các dữ liệu giao thông. Sự hội nhập của viễn thông, mạng máy tính, Internet và nhiều phương tiện truyền thông khác đã được tiến hành, thay đổi bộ mặt của viễn thông và điện toán. Điều này đã tạo ra một nhu cầu lớn về khai thác dữ liệu để giúp hiểu số liệu kinh doanh, xác định mô hình viễn thông, xác định các hoạt động gian lận, sử dụng tốt hơn các nguồn lực và cải thiện chất lượng dịch vụ. Bài toán khai thác dữ liệu trong viễn thông có điểm tương đồng với những người trong ngành công nghiệp bán lẻ. Bài toán thường gặp bao gồm xây dựng kho dữ liệu quy mô lớn, thực hiện biểu diễn đa chiều trực quan, OLAP và phân tích chuyên sâu về các xu hướng, mẫu của khách hàng và các mẫu tuần tự. Các bài toán này góp phần cải thiện kinh doanh, giảm chi phí, duy trì khách hàng, phân tích gian lận và tìm hiểu về đối thủ cạnh tranh. Có rất nhiều bài toán khai thác dữ liệu cùng với các công cụ khai thác dữ liệu cho viễn thông đã được biết đến và sẽ đóng vai trò ngày càng quan trọng trong kinh doanh ngày nay [6]. Một số bài toán cụ thể như sau :

2.1.4.1 Phân tích đa chiều số liệu bán hàng, khách hàng, sản phẩm, theo thời gian, khu vực

Ngành viễn thông nói riêng và dịch vụ bán lẻ nói chung cần nhìn cái nhìn đa chiều về doanh thu bán hàng, lợi nhuận, sản phẩm ... nên việc phân tích dữ liệu đa chiều rất cần thiết giúp cho những người quản lý nắm bắt được tình hình kinh doanh và thị trường cùng các thông tin hữu ích khác.

2.1.4.2 Phân tích hiệu quả của các chương trình khuyến mại:

Hiện các chương trình khuyến mại và sản phẩm của ngành viễn thông liên tục được thay đổi, để đánh giá hiệu quả các chương trình khuyến mại là bài toán rất quan trọng của các mạng viễn thông.

2.1.4.3 Duy trì khách hàng - Phân tích về lòng trung thành của khách hàng

Chúng ta có thể sử dụng thông tin khách hàng trung thành để đăng ký trình tự mua hàng của khách hàng cụ thể. lòng trung thành của khách hàng và xu hướng mua hàng có thể được phân tích một cách hệ thống. Hàng hóa mua tại giai đoạn khác nhau của cùng một khách hàng có thể được nhóm lại thành chuỗi. Tuần tự khai thác mô hình sau đó có thể được sử dụng để điều tra những thay đổi trong tiêu dùng của khách hàng hoặc lòng trung thành và đề nghị điều chỉnh về giá cả và sự đa dạng của hàng hóa để giúp giữ chân khách hàng và thu hút những người mới.

Ngoài ra KPDL còn giúp phân tích dự báo khách hàng rời mạng từ đó đề nghị điều chỉnh về giá cả và sự đa dạng của hàng hóa để giúp giữ chân khách hàng

2.1.4.4 Xây dựng hệ thống gợi ý

Ở ngành viễn thông và bán lẻ nói chung thì việc tư vấn cho khách hàng rất quan trọng nhằm tăng doanh thu. Vấn đề ở đây là phát hiện nhu cầu của khách hàng để có thể tư vấn và gợi ý cho khách hàng đúng nhu cầu, thời gian và địa điểm. Các ứng dụng của việc KPDL ở đây bao gồm việc phân tích hành vi khách hàng, phân lớp, …

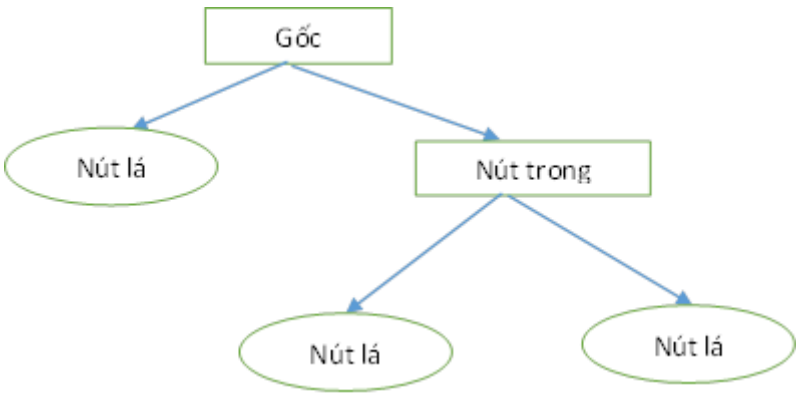
2.1.4.5 Phòng chống gian lận và xác định hành vi sử dụng bất thường

Trong ngành viễn thông thì việc rất quan trọng là phát hiện gian lận và hành vi sử dụng bất thường nhằm giảm thiểu thất thoát doanh thu. Bài toán phát hiện dấu hiệu bất thường và phòng chống gian lận để phát hiện các hành vi gian lận của khách hàng và các bất thường của hành vi sử dụng. Ứng dụng KPDL là sử dụng các công cụ phân lớp, phân tích hành vi.

2.2. Một số kỹ thuật KPDL trong phân lớp, dự báo

2.2.1. Cây quyết định

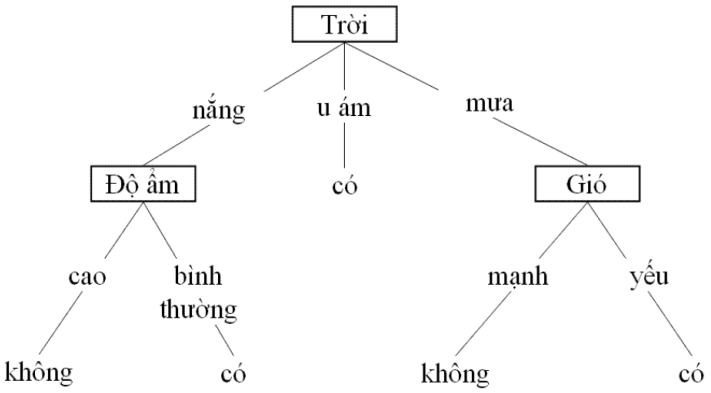
Cây quyết định là một cấu trúc biểu diễn dưới dạng cây. Trong đó, mỗi nút trong biểu diễn một thuộc tính, mỗi nhánh biểu diễn giá trị có thể có của thuộc tính, mỗi lá biểu diễn các lớp quyết định và đỉnh trên cùng của cây gọi là gốc.



Hình 2-2 Biểu diễn cây quyết định cơ bản

Trong lĩnh vực học máy, cây quyết định là một kiểu mô hình dự báo, nghĩa là một ánh xạ từ các quan sát về một sự vật/hiện tượng tới các kết luận về giá trị mục tiêu của sự vật/hiện tượng. Mỗi nút trong tương ứng với một biến, đường nối giữa nó với nút con của nó thể hiện giá trị cụ thể cho biến đó. Mỗi nút lá đại diện cho giá trị dự đoán của biến mục tiêu, cho trước các giá trị dự đoán của các biến được biểu diễn bởi đường đi từ nút gốc tới nút lá đó. Kỹ thuật học máy dùng trong cây quyết định được gọi là học bằng cây quyết định, hay chỉ gọi với cái tên ngắn gọn là cây quyết định.

Ví dụ 2.1: Một người có chơi tennis hay không?



Hình 2-3 Cây quyết định cho việc chơi Tennis

Cây quyết định là một cây phân cấp có cấu trúc được dùng để phân lớp các đối tượng dựa vào dãy các luật. Các thuộc tính của đối tượng (ngoại trừ thuộc tính phân lớp) có thể thuộc các kiểu dữ liệu khác nhau (Binary, Nominal, ordinal, quantitative values) trong khi đó thuộc tính phân lớp phải có kiểu dữ liệu là Binary hoặc Ordinal.

Tóm lại, cho dữ liệu về các đối tượng gồm các thuộc tính cùng với lớp của nó, cây quyết định sẽ sinh ra các luật để dự đoán lớp của các đối tượng chưa biết.

So với các phương pháp KPDL khác, cây quyết định là một trong những hình thức mô tả dữ liệu tương đối đơn giản, trực quan, dễ hiểu đối với người dùng nhưng lại hiệu quả nên được sử dụng nhiều. Trong những năm qua, nhiều mô hình phân lớp dữ liệu đã được các nhà khoa học trong nhiều lĩnh vực khác nhau đề xuất như mạng notron, mô hình thống kê tuyến tính /bậc 2, cây quyết định, mô hình di truyền... Trong số những mô hình đó, cây quyết định được đánh giá là một công cụ mạnh, phổ biến, đặt biệt là thích hợp cho DM nói chung và cho phân lớp dữ liệu nói riêng.

Có rất nhiều giải thuật đã được cài đặt sẵn như: CART (Breiman [4]), C4.5 (Quinlan [7])

2.2.1.1 Thuật toán ID3

Thuật toán ID3 (Interative Dichotomizer 3) được đề ra bởi Quinlan [8] và những dạng biểu diễn của nó được sử dụng rộng rãi trong thuật toán cây quyết định.

Đây cũng được gọi là thuật toán tham lam (greedy algorithm) vì thuật toán ID3 tìm kiếm những mô hình "tham lam" mà trong đó các thuộc tính đạt được tối đa lượng thông tin cho việc xác định nhãn lớp của các mẫu trong tập huấn luyện. Kết quả là ta thu được một Cây quyết định mà gán nhãn đúng cho mỗi mẫu trong tập huấn luyện.

Thuật toán xây dựng cây ID3 sử dụng Entropy làm cơ sở đo độ đồng nhất của tập dữ liệu. Trên cơ sở Entropy, thuật toán tính độ tăng thông tin như mức tăng độ đồng nhất, từ đây xác định thuộc tính tốt nhất tại mỗi nút.

Đầu vào: Một tập các mẫu. Mỗi mẫu bao gồm các thuộc tính rời rạc, mô tả một tình huống, hay một đối tượng nào đó, và một giá trị phân loại của nó.

- S là tập huấn luyện
- c_1, c_2, …, c_m là các nhãn lớp phân loại
- S_1, S_2, …, S_m là tập con của S thuộc các lớp c_1, …, c_m tương ứng
- $S = \bigcup_{i=1}^{m} S_i$ và $S_i \cap S_j = \emptyset \ \forall \mathrm{i} \neq \mathrm{j}$

Các độ đo:

- **p_i:** xác suất để một phần tử bất kỳ trong S thuộc về lớp c_i.

$$p_i = \frac{|S_i|}{|S|} \tag{2.1}$$

- **Entropy** của tập dữ liệu S

$$H(S) = -\sum_{i=1}^{m} p_i log_2(p_i) = -\sum_{i=1}^{m} \frac{|S_i|}{|S|} log_2 (\frac{|S_i|}{|S|}) \tag{2.2}$$

  H(S) đạt giá trị cực đại là $\log_2(m)$ khi các nhãn c_1, c_2, …, c_m có xác suất như nhau và giá trị nhỏ nhất của H(S) là 0 khi tất cả các đối tượng có chung một nhãn.

- **Độ tăng thông tin (Information Gain, ký hiệu IG)** là chỉ số đánh giá độ tốt của thuộc tính trong việc phân chia tập dữ liệu thành những tập con đồng nhất. IG được tính trên độ tăng thông tin theo công thức sau:

$$IG(S, A) = H(S) - \sum_{v \in values(A)} \frac{|S_V|}{|S|} H(S_V) \tag{2.3}$$

Trong đó:

- S là tập dữ liệu ở nút hiện tại
- A là thuộc tính được sử dụng để đánh giá độ tốt phân chia
- Values(A) là tập các giá trị của A
- S_v là tập mẫu con của S có các giá trị thuộc tính A bằng v

- |S| và $|S_v|$ là số phần tử của các tập |S|, $|S_v|$ tương ứng

Thuật toán ID3 sử dụng độ tăng thông tin để lựa chọn thuộc tính tốt nhất tại mỗi nút, thuộc tính được lựa chọn là thuộc tính mang lại độ tăng thông tin lớn nhất.

2.2.1.2 Thuật toán C4.5

C4.5 là thuật toán dùng để xây dựng cây quyết định được đề xuất bởi Quinlan [7] năm 1993 , là mở rộng của ID3. Đặc điểm của C4.5:

- Cho phép dữ liệu đầu vào ở các thuộc tính là liên tục
- Cho phép thao tác với các thuộc tính có dữ liệu không xác định (do bị mất mát dữ liệu)
- Đưa ra phương pháp cắt tỉa cây và giản lược các luật để phù hợp với những bộ dữ liệu lớn
- C4.5 sử dụng cơ chế lưu trữ dữ liệu thường trú trong bộ nhớ, chính đặc điểm này làm C4.5 thích hợp với những cơ sở dữ liệu nhỏ, và cơ chế sắp xếp lại dữ liệu tại mỗi nút trong quá trình phát triển cây quyết định.
- C4.5 còn chứa kỹ thuật cho phép biểu diễn lại cây quyết định dưới dạng một danh sách sắp thứ tự các luật **if-then** (một dạng quy tắc phân lớp dễ hiểu). Kỹ thuật này cho phép làm giảm bớt kích thước tập luật và đơn giản hóa các luật mà độ chính xác so với nhanh tương ứng cây quyết định là tương đương.

C4.5 có cơ chế sinh cây quyết định hiệu quả và chặt chẽ bằng việc sử dụng độ đo lựa chọn thuộc tính tốt nhất là Information Gain. Các cơ chế xử lý với giá trị lỗi, thiếu và tránh quá phù hợp của dữ liệu của C4.5 cùng với cơ chế cắt tỉa cây đã tạo nên sức mạnh của C4.5. Thêm vào đó, mô hình phân lớp C4.5 còn có phần chuyển đổi từ cây quyết định sang luật dạng if-then, làm tăng độ chính xác và tính dễ hiểu của kết quả phân lớp. Đây là tiện ích rất có ý nghĩa đối với người sử dụng

Độ đo GainRatio được sử dụng trong thuật toán C4.5 là cải tiến của thuật toán ID3 và được xây dựng bởi Quinlan là thước đo cải tiến của chỉ số Information Gain.

- Độ đo này giải quyết vấn đề thuộc tính có nhiều giá trị: Trong tập mẫu huấn luyện bài toán chơi Tennis ở trên, nếu sử dụng thêm thuộc tính ngày thì thuộc tính này có 14 giá trị khác nhau và thuộc tính này có IG cao nhất, và khi đó Ngày được chọn làm gốc với 14 nhánh và không phân loại được các ngày tiếp theo. Do đó cần tránh thuộc tính có nhiều giá trị. Thành phần thông tin chia (Split Information-SI) được bổ sung để phạt các thuộc tính có nhiều giá trị:

$$SI(S,A) = -\sum_{i=1}^{m} \frac{|S_i|}{|S|} log_2(\frac{|S_i|}{|S|}) \quad (2.4)$$

- Tiêu chuẩn đánh giá thuộc tính GainRatio được xác định bằng cách chia độ tăng thông tin cho thông tin chia (khi SI(S,A)=0 thì cần có phương án xử lý riêng)

$$GR(S,A) = \frac{GI(S,A)}{SI(S,A)} \quad (2.5)$$

2.2.2. Phân lớp Naïve Bayes

Định lý Bayes cho phép tính xác suất xảy ra của một sự kiện ngẫu nhiên A khi biết sự kiện liên quan B đã xảy ra. Xác suất này được ký hiệu là P(A|B), và đọc là "xác suất của A nếu có B". Đại lượng này được gọi xác suất có điều kiện vì nó được rút ra từ giá trị được cho của B hoặc phụ thuộc vào giá trị đó.

Theo định lí Bayes, xác suất xảy ra A khi biết B sẽ phụ thuộc vào 3 yếu tố:

- Xác suất xảy ra A của riêng nó, không quan tâm đến B, kí hiệu là P(A) và đọc là xác suất của A. Đây được gọi là xác suất biên duyên hay xác suất tiên nghiệm (prior), nó là "tiên nghiệm" theo nghĩa rằng nó không quan tâm đến bất kỳ thông tin nào về B.
- Xác suất xảy ra B của riêng nó, không quan tâm đến A, kí hiệu là P(B) và đọc là "xác suất của B". Đại lượng này còn gọi là hằng số chuẩn hóa, vì nó luôn giống nhau, không phụ thuộc vào sự kiện A đang muốn biết.
- Xác suất xảy ra B khi biết A xảy ra, kí hiệu là P(B|A) và đọc là "xác suất của B nếu có A". Đại lượng này gọi là khả năng (likelihood) xảy ra B khi biết A đã xảy ra. Chú ý không nhầm lẫn giữa khả năng xảy ra A khi biết B và xác suất xảy ra A khi biết B.

Khi biết ba đại lượng trên, xác suất của A khi biết B cho bởi công thức:

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)} = \frac{likelihood * prior}{normallisingconstant} \quad (1)$$

Từ đó dẫn tới:

$$P(A|B)P(B) = P(A \cap B) = P(A)P(B|A) \quad (2)$$

Khi có n giả thuyết thì:

$$P(A_i|B) = \frac{P(B|A_i)P(A_i)}{\sum_{j=1}^{n} P(B|A_j)P(A_j)} \quad (3)$$

Phương pháp Naïve Bayes phù hợp các bài toán có yêu cầu về chi phí xuất hiện của các giá trị thuộc tính.

- Thiết kế hệ thống phân lớp thường dễ dàng hơn so với các phương pháp khác.

- Các thuộc tính trong tập mẫu học phải độc lập với điều kiện.

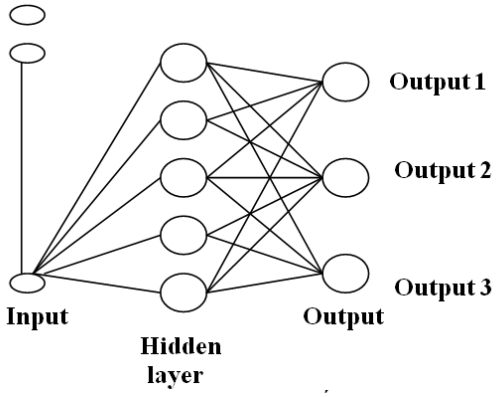
- Độ chính xác thuật toán phân lớp phụ thuộc nhiều vào tập dữ liệu học ban đầu.

2.2.3. Mạng nơ ron nhân tạo

Nơ ron nhân tạo là sự mô phỏng đơn giản của nơ ron sinh học. Mỗi nơ ron nhân tạo thực hiện hai chức năng: chức năng tổng hợp đầu vào và chức năng tạo đầu ra. Mỗi nơ ron có một giá trị ngưỡng, chức năng đầu vào chính là tổng có trọng số các tín hiệu vào kết hợp với ngưỡng để tạo ra tín hiều đầu vào. Chức năng tạo đầu ra được thực hiện bằng hàm truyền đạt. Hàm này sẽ nhận tín hiệu đầu vào và tạo tín hiệu đầu ra của nơ ron.

Mạng nơ ron là một hệ thống gồm nhiều phần tử xử lý hoạt động song song. Chức năng của nó được xác định bởi cấu trúc mạng, độ lớn của các liên kết và quá trình xử lý tại mỗi nút hoặc đơn vị tính toán.

Mạng nơ ron sử dụng mạng đa tầng bao gồm ba lớp tế bào thần kinh. Các lớp này là một lớp đầu vào, một lớp ẩn và một lớp đầu ra. Trong một mạng nơ ron, mỗi nơ ron nhận được một hoặc nhiều đầu vào và sản xuất một hoặc nhiều kết quả đầu ra. Mỗi đầu ra là một hàm phi tuyến đơn giản của tổng các đầu vào cho các nơ ron.



Hình 2-4 Mô hình mạng nơron nhiều lớp

Có ba loại nơ ron trong một mạng nơ ron được tạo ra với thuật toán mạng nơ ron:

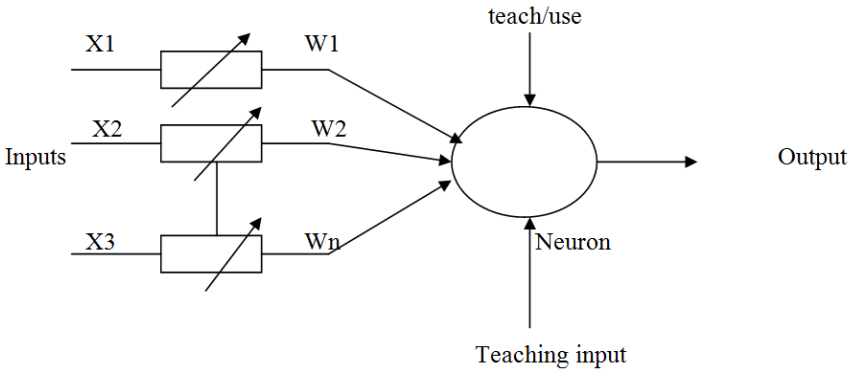
Nơ ron đầu vào: Nơ ron đầu vào cung cấp các giá trị thuộc tính đầu vào cho các mô hình khai thác dữ liệu. Đối với thuộc tính đầu vào rời rạc, một nơ ron đầu vào thường đại diện cho một trạng thái đơn nhất từ các thuộc tính đầu vào. Một thuộc tính đầu vào liên tục tạo ra hai nơ ron đầu vào: một nơ ron cho một trạng thái bị thiếu, một nơ ron cho giá trị của chính thuộc tính liên tục đó. Nơ ron đầu vào cung cấp đầu vào cho một hoặc nhiều nơ ron ẩn.

Nơ ron ẩn: nơ ron ẩn nhận đầu vào từ các nơ ron đầu vào và cung cấp đầu ra cho các nơ ron đầu ra.

Nơ ron đầu ra: Nơ ron đầu ra đại diện cho các giá trị của thuộc tính dự đoán của mô hình KPDL. Đối với các thuộc tính đầu ra là rời rạc, một nơ ron đầu ra đại diện duy nhất cho một trạng thái dự đoán của thuộc tính dự đoán, bao gồm cả giá trị thiếu. Nếu các thuộc tính dự đoán liên tục tạo ra hai nơ ron đầu ra: một nơ ron cho một trạng thái bị thiếu, một nơ ron cho chính các giá trị của thuộc tính liên tục đó.

Mỗi đầu vào có một giá trị được gán cho nó có trọng số là w, mô tả sự liên quan giữa đầu vào đến các nơ ron ẩn hoặc các nơ ron đầu ra.

Mạng nơ ron xây dựng lại cấu trúc bộ não có khả năng nhận biết dữ liệu thông qua tiến trình học, với các thông số tự do của mạng có thể thay đổi liên tục bởi những thay đổi của môi trường và mạng nơ ron ghi nhớ giá trị đó.



Hình 2-5 Tiến trình học

Trong quá trình học, giá trị đầu vào được đưa vào mạng và theo dòng chảy trong mạng tạo thành giá trị ở đầu ra.

Tiếp đến là quá trình so sánh giá trị tạo ra bởi mạng nơ ron với giá trị ra mong muốn. Nếu hai giá trị này giống nhau thì không thay đổi gì cả. Tuy nhiên, nếu có một sai lệch giữa hai giá trị này vượt quá giá trị sai số mong muốn thì đi ngược mạng từ đầu ra về đầu vào để thay đổi một số kết nối.

Đây là một quá trình lặp liên tục và có thể không dừng khi không tìm các giá trị w sao cho đầu ra tạo bởi mạng nơ ron bằng đúng đầu ra mong muốn. Do đó trong thực tế người ta phải thiết lập tiêu chuẩn dựa trên một giá trị sai số nào đó của hai giá trị này, hay dựa trên một số lần lặp xác định.

2.2.4. Luật kết hợp

Khai phá luật kết hợp: Là tìm các mẫu phổ biến, sự kết hợp, sự tương quan, hay các cấu trúc nhân quả giữa các tập đối tượng trong các cơ sở dữ liệu giao tác, cơ sở dữ liệu quan hệ và những kho thông tin khác [3].

Các ứng dụng: Luật kết hợp có ứng dụng trong nhiều lĩnh vực khác nhau của đời sống như: khoa học, hoạt động kinh doanh, tiếp thị, thương mại, phân tích thị trường chứng khoán, tài chính và đầu tư,...

Ví dụ về luật kết hợp:
Bia => Lạc (0,5%; 60%)

Luật này có nghĩa: **Nếu** mua bia **thì** mua lạc trong 60% trường hợp. Bia và lạc được mua chung trong 0.5% tổng giao dịch.

Thu nhập= 60.000.000_max => ***Tài khoản tiết kiệm***= yes (20% ; 100%)

Luật này có nghĩa: **Nếu** thu nhập lớn hơn hoặc bằng 60 triệu một năm **thì** khách hàng có tài khoản tiết kiệm với độ tin cậy là 100%.

Từ các luật kết hợp được trích rút từ chính các cơ sở dữ liệu giao dịch, cơ sở dữ liệu khách hàng mà các siêu thị, các ngân hàng sẽ có chiến lược kinh doanh (sắp xếp các mặt hàng, số lượng các mặt hàng,..), chiến lược tiếp thị, quảng cáo,… để từ đó thúc đẩy hoạt động kinh doanh của mình.

Một số định nghĩa, khái niệm cơ bản:

Cho I = $\{i_1, i_2, i_3, \ldots, i_n\}$ là tập bao gồm n mục (Item – còn gọi là thuộc tính - attribute). X $\subseteq$ I được gọi là tập mục (itemset).

T = $\{t_1, t_2, \ldots, t_m\}$ là tập gồm m giao tác (Transaction – còn gọi là bản ghi - record).

R là một quan hệ nhị phân trên I và T (hay R $\subseteq$ IxT). Nếu giao tác t có chứa mục i thì ta viết $(i,t) \in R$ (hoặc iRt). Ta sẽ ký hiệu D = (T,I,R) là dữ liệu để khai thác. Về mặt hình thức, D chính là một quan hệ dạng bảng. Về ý nghĩa, một cơ sở dữ liệu là một tập các giao tác (hay giao dịch), mỗi giao dịch t chứa một tập mục X $\subseteq$ I.

Độ hỗ trợ của tập mục X

Cho dữ liệu D = (T, I, R); X $\subseteq$ I. Gọi T(X) là tập giao tác chứa X.

Độ hỗ trợ (support) của tập mục X, ký hiệu support(X) là tỷ số của số lượng giao tác trong cơ sở dữ liệu D chứa X trên tổng số các giao tác trong cơ sơ dữ liệu D. Hay

$$\text{Support(X)} = \text{Card (T(X))} / \text{Card (T)} = \frac{|T(X)|}{|T|} \quad (4)$$

Tập phổ biến

Cho D = (T, I, R); minsup $\in$ (0,1].

Tập mục X $\subseteq$ I được gọi là một tập phổ biến theo ngưỡng minsup (gọi tắt là tập phổ biến) nếu support(X) >= minsup.

Ký hiệu FX(T, I, R, minsup) là tập hợp các tập phổ biến theo ngưỡng minsup:

FX(T, I, R, minsup) = { X $\subseteq$ I | support(X) $\geq$ minsup}

Luật kết hợp X => Y

Cho D = (T, I, R) là dữ liệu để khai thác. X, Y $\subseteq$I là các tập mục thỏa mãn điều kiện $X \cap Y = \varnothing$.

Luật kết hợp của X và Y, ký hiệu X=>Y, đây là luật chỉ khả năng xuất hiện Y khi X xuất hiện. Luật kết hợp có hai độ đo gắn với nó là: độ hỗ trợ và độ tin cậy (confidence) của luật.

Độ hỗ trợ của luật kết hợp X => Y

Độ hỗ trợ của luật kết hợp X => Y, ký hiệu support(X => Y) là tỷ số của số các giao tác trong D có chứa X $\cup$Y trên số tất cả giao tác trong D.

Hay

Support(X => Y) = card (T(X $\cup$ Y))/card(T) = $\frac{|T(X \cup Y)|}{|T|}$; (5)

Trong đó T(X) là tập giao tác chứa tập mục X.

Độ tin cậy của luật kết hợp X => Y

Độ tin cậy (confidence) của luật X => Y, ký hiệu: confidence(X => Y) là tỷ số các giao tác trong D có chứa X $\cup$ Y trên số các giao tác chứa X. Hay

Confidence(X => Y) = card(T(X $\cup$ Y))/card(T(X)) = $\frac{|T(X \cup Y)|}{|T(X)|}$; (6)

Về mặt xác suất, độ tin cậy confidence(X => Y) của một luật là xác suất (có điều kiện) xảy ra Y với điều kiện đã xảy ra X.

Confidence(X => Y) = P(Y | X) (7)

Luật kết hợp tin cậy: Một luật được xem là tin cậy nếu độ tin cậy confidence của nó lớn hơn hoặc bằng một ngưỡng minconf $\in (0,1]$ nào đó do người dùng xác định. Ngưỡng minconf phản ánh mức độ xuất hiện của Y khi cho trước X.

Luật kết hợp cần tìm là luật kết hợp thỏa minsup và minconf cho trước. Chúng ta chỉ quan tâm đến các luật có độ hỗ trợ lớn hơn độ hỗ trợ tối thiểu và độ tin cậy lớn hơn độ tin cậy tối thiểu.

Hầu hết các thuật toán khai phá luật kết hợp thường chia thành hai pha:

- Pha 1: Tìm tất cả các tập mục phổ biến từ cơ sở dữ liệu D tức là tìm tất cả các tập mục X thỏa mãn support(X) >= minsup.
- Pha 2: Sinh các luật tin cậy từ các tập phổ biến đã tìm thấy ở pha 1.

Cho minconf; X,Y là các tập mục phổ biến tìm thấy trong pha 1 luật kết hợp được sinh từ X, Y có dạng: X => Y và confidence(X=>Y) >= minconf.

2.2.5. Đánh giá độ chính xác thuật toán

Giả sử ta có bài toán phân lớp với đầu ra là 02 lớp Tốt/Xấu (hoặc Đúng/Sai), kết quả phân lớp trên tập mẫu so với thực tế có 4 khả năng thể hiện**Error! Reference source not found.**. Bảng này được gọi là ma trận sai số (confusion matrix).

		Lớp dự đoán (predicted class)	
		Đúng	***Sai***
Lớp thực tế (actual class)	***Đúng***	True Positive (TP)	False Negative (FN)
	Sai	False Positive (FP)	True Negative (TN)

True Positive thể hiện khả năng dự đoán phân lớp phân lớp đúng của một mẫu dự đoán thuộc phân lớp Đúng, False Positive thể hiện khả năng dự đoán phân lớp sai của một mẫu được dự đoán thuộc phân lớp Đúng.

False Negative thể hiện khả năng dự đoán phân lớp phân lớp đúng của một mẫu dự đoán thuộc phân lớp Sai, True Negative thể hiện khả năng dự đoán phân lớp sai của một mẫu được dự đoán thuộc phân lớp Sai.

Ta có các độ đo đánh giá hiệu quả của kết quả phân loại như sau:

Tên độ đo	**Công thức**	**Diễn giải**
Độ chính xác	$Accuracy = \frac{TP + TN}{TP + TN + FP + FN}$	Tỷ lệ các mẫu được phân lớp đúng trên toàn bộ tập mẫu
Tỷ lệ lỗi	$Error\_Rate = \frac{FP + FN}{TP + TN + FP + FN}$	Tỷ lệ các mẫu được phân lớp sai trên toàn bộ tập mẫu
Độ bao phủ	$Recall = \frac{TP}{TP + FN}$	Tỷ lệ các mẫu phân lớp Tốt đúng trên toàn bộ các mẫu thực sự thuộc phân lớp Tốt
Độ hiệu dụng	$Precision = \frac{TP}{TP + FP}$	Tỷ lệ các mẫu phân lớp Tốt đúng trên toàn bộ các mẫu dự đoán phân lớp Tốt

Bảng 2-1 Các độ đo chính xác

2.3. Giới thiệu về công cụ weka

Weka (Waikato Environment for Knowledge Analysis) [11] là một phần mềm mã nguồn mở viết bằng Java, được phát triển bởi Đại học Waikato, New Zealand.

Phiên bản Weka đến thời điểm tháng 07/2015 là 3.7.12. Weka chứa một tập các công cụ mô hình và thuật toán cho việc phân tích dữ liệu và mô hình dự đoán, cùng với giao diện đồ họa cho người sử dụng dễ dàng truy cập vào các chức năng. Weka được sử dụng cho nhiều ứng dụng khác nhau, phục vụ cho mục đích nghiên cứu.

Những ưu điểm của Weka:

- Miễn phí cho người sử dụng;
- Hỗ trợ trên nhiều nền tảng hệ điều hành;
- Là một tập hợp xử lý dữ liệu và kỹ thuật mô hình;
- Hỗ trợ đồ họa.

Weka cung cấp một số công cụ khai phá dữ liệu chuẩn, xử lý dữ liệu, phân cụm, phân lớp, hồi quy, .v.v. Tất cả các kỹ thuật của Weka được xây dựng dựa trên các giả định là dữ liệu đơn hoặc là dữ liệu quan hệ, khi mỗi điểm dữ liệu được miêu tả bởi một số xác định các thuộc tính (thông thường, là dạng liên tục hoặc các thuộc tính được chuẩn hóa, một vài dạng thuộc tính khác cũng được hỗ trợ). Weka cung cấp cách truy cập cơ sở dữ liệu bằng các sử dụng kết nối Java Database Connectivity (JDBC) và có thể xử lý kết quả trả về bởi một truy vấn cơ sở dữ liệu.

Weka hỗ trợ 4 chế độ kiểm thử:

- Sử dụng tập huấn luyện (Use Training Set): đánh giá khả năng dự đoán lớp của các mẫu được huấn luyện
- Tập hỗ trợ kiểm thử (Supplied Test Set): đánh giá khả năng dự đoán lớp của một tập kiểm tra có đầy đủ giá trị thuộc tính Out-put
- Đánh giá chéo (Cross-Validation): giống như mô tả tại mục 2
- Tách theo tỷ lệ (Percentage Split): Chia tập dataset theo tỷ lệ quy định trước về số mẫu học, số mẫu kiểm tra

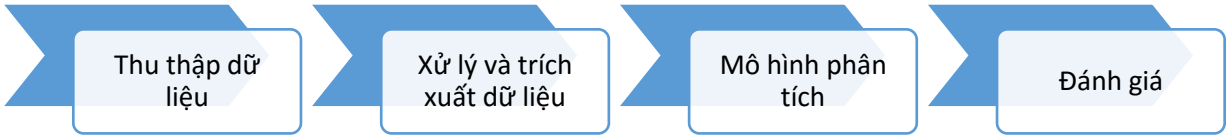
2.4. Kết luận chương 2

Chương 2 trình bày một số khái niệm cơ bản và các kiến thức có liên quan về các thuật toán thường được áp dụng trong KPDL ở dạng phân lớp, dự báo đó là: cây quyết định, mạng nơron nhân tạo, phân lớp Naïve Bayes và luật kết hợp.

Chương 3 Giải pháp phát hiện thuê bao di động có khả năng rời mạng

3.1. Giải pháp chung:

Đối với các bài toán về dự đoán thuê bao rời mạng nói chung thì đều sử dụng mô hình như sau:



Hình 3-1 Giải pháp chung cho bài toán dự báo thuê bao rời mạng

Trong tất cả các giải pháp trình bày tiếp theo đều sử dụng chung giải pháp này và chỉ khác nhau trong khối mô hình phân tích.

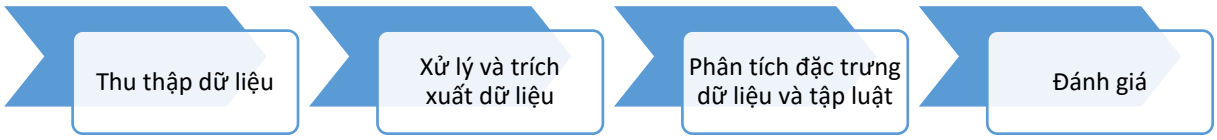
Giải thích các chức năng của các khối như sau:

- Thu thập dữ liệu: Có nhiệm vụ thu thập dữ liệu từ các nguồn liên quan đến bài toán, đối với bài toán này thì dữ liệu được thu thập từ các nguồn sẵn có của MobiFone như dữ liệu CDR từ tổng đài, dữ liệu thông tin khách hàng, dữ liệu lịch sử mua hàng và khuyến mại, dữ liệu lịch sử khiếu nại … Các dữ liệu này đều được các hệ thống tác nghiệp của MobiFone xử lý và đã được lưu trữ trong CSDL.
- Xử lý và trích xuất dữ liệu: Bước này thực hiện tiền xử lý và trích xuất dữ liệu để đưa vào mô hình phân tích. Trong khối dữ liệu khổng lồ của mạng di động thì tôi đã dựa vào kinh nghiệm làm việc gần 10 năm tại mạng di động MobiFone về các công việc liên quan đến xử lý, quản lý dữ liệu khách hàng cũng như tham khảo các chuyên gia trong lĩnh vực chăm sóc khách hàng để chọn lọc ra những dữ liệu có liên quan nhất với bài toán.
- Mô hình phân tích: Bước này sẽ thực hiện mô hình hóa các dữ liệu đã được xử lý từ bước trước để xây dựng các mô hình dữ liệu.
- Đánh giá: Đánh giá kết quả thực hiện

Việc thực hiện giải pháp trên là quá trình cải tiến liên tục sau khi đánh giá thì lấy kết quả phân tích nhằm cải tiến giải pháp nếu cần.

3.2. Giải pháp hiện tại của mạng MobiFone

Giải pháp hiện tại của mạng MobiFone như sau:

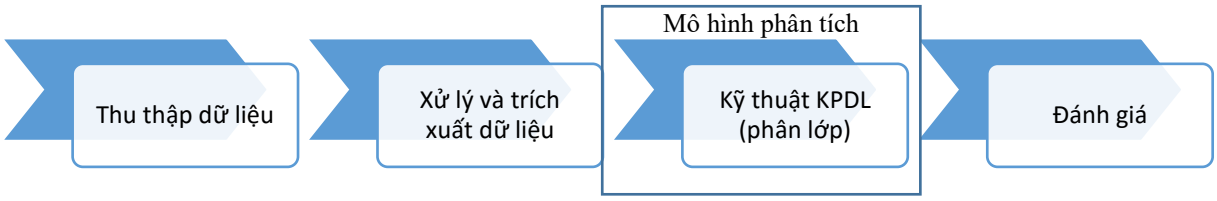


Hình 3-2 Giải pháp hiện tại MobiFone đang áp dụng

Trong giải pháp này thì các chuyên gia MobiFone trong lĩnh vực chăm sóc khác hàng đã sử dụng phương pháp phân tích đặc trưng dữ liệu từ đó tìm ra đặc trưng của các thuê bao có khả năng rời mạng để áp dụng dự báo. Phương pháp này sẽ sử dụng các tập luật để phân lớp dữ liệu. Phương pháp này có ưu điểm là đơn giản, rất dễ thực hiện để phân lớp, tuy nhiên để xác định chính xác các tập luật đòi hỏi phải rất hiểu được dữ liệu nên không thể tự động hóa được một khi dữ liệu thay đổi đòi hỏi phải có chuyên gia phân tích lại dữ liệu nên sẽ tốn rất nhiều thời gian để đưa ra được tập luật chính xác. Hơn nữa, đối với những loại dữ liệu phức tạp thì phương pháp này cho độ chính xác không cao.

3.3. Giải pháp đề xuất

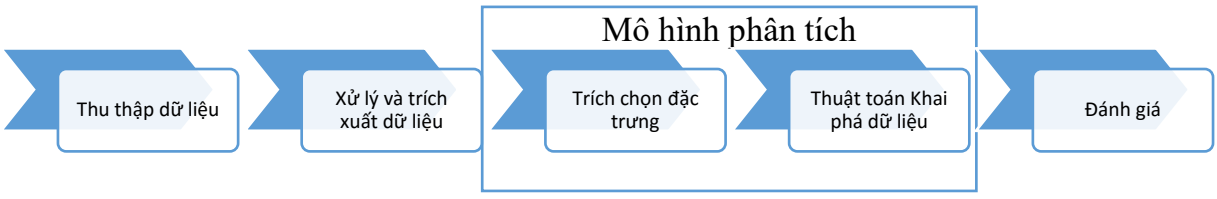
3.3.1. Giải pháp đề xuất dùng kỹ thuật khai phá dữ liệu



Hình 3-3 Giải pháp đề xuất dùng kỹ thuật khai phá dữ liệu

Giải pháp này đề xuất sử dụng để giải quyết bài toán. Trong bước mô hình hóa sẽ sử dụng các thuật toán KPDL cụ thể các thuật toán phân lớp để ứng dụng dự báo thuê bao rời mạng. Phương pháp này có ưu điểm là có thể tự động hóa được tuy nhiên thời gian xử lý dữ liệu thường chậm hơn phương pháp dựa vào đặc trưng.

3.3.2. Giải pháp đề xuất sau khi cải tiến



Hình 3-4 Giải pháp đề xuất cải tiến

Giải pháp này là cải tiến của giải pháp dùng thuật toán khai phá dữ liệu. Để làm giảm thời gian xử lý dữ liệu đối thì trước khi đưa số liệu vào xử lý bằng thuật toán khai phá dữ liệu ta sẽ sử dụng các thuật toán trích chọn đặc trưng để tìm ra những thuộc tính

phù hợp nhất với các thuật toán. Áp dụng phương pháp này thời gian xử lý dữ liệu sẽ nhanh hơn giải pháp đề cập ở mục 3.3.1 tuy nhiên độ chính xác của kết quả dự đoán có giảm nhưng không đáng kể. Do dữ liệu mạng di động thời rất lớn mà bài toán này thì đòi hỏi thời gian xử lý nhanh do vậy ưu tiên về thời gian xử lý hơn.

Các thuật toán trích chọn thuộc tính:

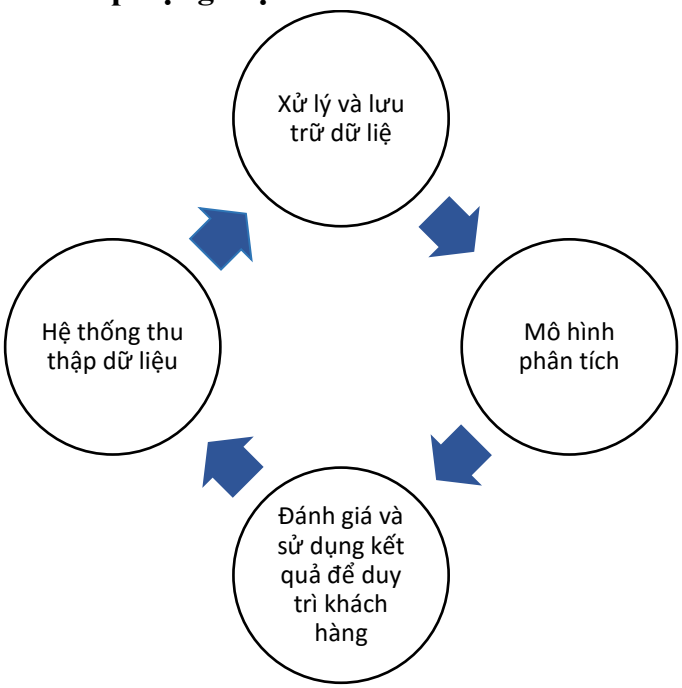
- Trích chọn theo Correlation-based: Đánh giá giá trị của một tập hợp các thuộc tính bằng cách xem xét các khả năng tiên đoán riêng rẽ của mỗi thuộc tính cùng với mức độ dư thừa giữa chúng. (M. A. Hall (1998). Correlation-based Feature Subset Selection for Machine Learning.) (Trích chọn dựa vào tương quan)
- Trích chọn theo độ đo GainRatio: Đánh giá giá trị của một thuộc tính bằng cách đo GainRatio của thuộc tính trong quá trình phân lớp:

  GainR(Class, Attribute) = (H(Class) - H(Class | Attribute)) / H(Attribute).

- Trích chọn theo độ đo InfoGain: Đánh giá giá trị của một thuộc tính bằng cách đo InfoGain liên quan đến các phân lớp.

  InfoGain(Class,Attribute) = H(Class) - H(Class | Attribute).
- Trích chọn theo phương pháp PCA (principal components analysis – Phân tích thành phần chính)

3.4. Mô hình đề xuất áp dụng thực tế



Hình 3-5 Mô hình đề xuất áp dụng thực tế

Trong đó Hệ thống thu thập dữ liệu là hệ thống có chức năng lấy dữ liệu CDR, thông tin khách hàng, lịch sử khiếu nại của khách hàng

Hệ thống xử lý và lưu trữ dữ liệu: Đây chính là hệ thống Kho dữ liệu của MobiFone có chức năng xử lý dữ liệu đã thu thập và lưu trữ dưới dạng database.

Mô hình phân tích: Sử dụng kết nối từ WEKA đến cơ sở dữ liệu của hệ thống Kho dữ liệu để lấy thông tin phân tích và đầu ra là kết quả dự báo thuê bao rời mạng

Đánh giá và sử dụng: Đánh giá kết quả dự đoán và đưa số liệu dự đoán để sử dụng thực tế. Qua việc phản hồi kết quả đánh giá sẽ tác động ngược lại quá trình thực hiện ban đầu để tối ưu mô hình dự đoán. Do đặc thù của kinh doanh luôn thay đổi ảnh hưởng hành vi sử dụng của khách hàng do không có mô hình dự đoán nào luôn đúng. Trong quá trình thực hiện đánh giá nhằm cải tiến mô hình để phù hợp nhất với tình hình số liệu thực tế.

Chương 4 Thực nghiệm và đánh giá kết quả

Theo các phương pháp dự báo đã trình bày trong chương 3, áp dụng vào số liệu thực tế bài toán: *dự báo thuê bao di động rời mạng*, luận văn trình bày các kết quả thực nghiệm đạt được trên các phương pháp đã nêu.

4.1. Chuẩn bị dữ liệu

- Yêu cầu thu thập dữ liệu: Dữ liệu được sử dụng trong luận văn này là dữ liệu thực tế của mạng di động MobiFone đã được biến đổi khi công bố trong luận văn nhằm đảm bảo tính bảo mật của dữ liệu.

Đối tượng: Thuê bao MobiFone lâu năm (kích hoạt/hòa mạng từ 2 năm trở lên) không phát sinh cước trong tháng 8/2015.

Khoảng thời gian thu thập dữ liệu: thu thập dữ liệu lịch sử của những thuê bao này trong vòng 3 tháng, từ tháng 5 đến tháng 7/2015.

Chi tiết dữ liệu cần thu thập: 86 trường dữ liệu chi tiết, gồm: (chi tiết tại Phụ lục đính kèm):

- Gói cước
- Thời gian kích hoạt
- Loại khách hàng (cá nhân/doanh nghiệp)
- Trạng thái thuê bao tại thời điểm xuất dữ liệu
- Doanh thu tài khoản chính và tài khoản khuyến mại (có chia dịch vụ thoại, sms, data, khác) trong 3 tháng gần nhất (tháng 5, 6 và 7)
- Số lượng, sản lượng cuộc gọi đến/đi, sms đến/đi trong 3 tháng gần nhất (tháng 5, 6 và 7)
- Số ngày cập nhật VLR, số ngày phát sinh cước trong 3 tháng gần nhất (tháng 5, 6 và 7).
- Số tiền nạp/số lần nạp thẻ trong 3 tháng gần nhất.
- Số chương trình khuyến mại tham gia trong 3 tháng gần nhất.
- Và 1 số trường thông tin khác.

Dữ liệu sau khi làm sạch và lấy mẫu gồm gần 2 triệu bản ghi với 86 trường dữ liệu chi tiết.

- Tiền xử lý dữ liệu: Dữ liệu đã được xử lý bởi các hệ thống của MobiFone và đã chuẩn hóa các thông tin chi tiết như sau:

1. Loại bỏ những bản ghi có thông tin thiếu.

2. Loại bỏ những cuộc gọi bất thường

3. Chuẩn hóa các trường dữ liệu theo định dạng quy định

4.2. Mô tả dữ liệu thực nghiệm

Bộ dữ liệu tháng 8/2015, toàn mạng MobiFone có 1.622.229 thuê bao kích hoạt từ 2 năm trở lên và không phát sinh cước. Trong đó, thuê bao khách hàng cá nhân chiếm đa số với 98% (tương đương 1.610.136 thuê bao) và thuê bao khách hàng doanh nghiệp chiến 2% (12.093 thuê bao).

Nhằm phân tích sâu hơn và tìm ra những đặc điểm đặc trưng của thuê bao trước khi rời mạng, nhóm chuyên gia MobiFone đã thực hiện phân tập thuê bao không phát sinh cước trong tháng 8/2015 thành 5 nhóm để phân tích, cụ thể:

- Nhóm I: Thuê bao trả sau.
- Nhóm II: Thuê bao trả trước có tiêu dùng tài khoản chính (TKC) nhưng không đạt 3k3d_vlr trong tháng 7/2015.
- Nhóm III: Thuê bao trả trước chỉ tiêu dùng tài khoản khuyến mại (TKKM) trong tháng 7/2015.
- Nhóm IV: Thuê bao trả trước không phát sinh cước trong tháng 7/2015.
- Nhóm V: Thuê bao trả trước có đạt chuẩn 3k3d_vlr trong tháng 7/2015.

Loại TB	Thuê bao trả sau	Thuê bao trả trước				**Tổng**
		Không đạt 3k3d_vlr trong tháng trước			Đạt 3k3d_vlr trong tháng trước	
Nhóm	Nhóm I	Nhóm II	Nhóm III	Nhóm IV	Nhóm V	
Thuê bao Fastconnect	17.244	3.096	2.690	14.578	2.154	39.762
Thuê bao thường	34.821	161.527	219.993	1.067.414	105.703	1.589.458
Số lượng TB	52.065	164,623	222.683	1.081.992	107.857	1.629.220
Tỷ lệ	3.20%	10.10%	13.67%	66.41%	6.62%	100.00%

Bảng 4-1 Nhóm thuê bao theo đặc trưng

Tuy nhiên trong mục đích nghiên cứu này ta chỉ tập trung vào thuê bao trả trước đồng thời qua phân tích dữ liệu thì nhóm 2 sẽ gồm 2 nhóm có đặc trưng khác nhau do đó ta chia tiếp nhóm II thành 2 nhóm nhỏ (nhóm 1 – tiêu dùng TKC dưới 3.000đ và nhóm 2 – tiêu dùng TKC từ 3.000đ trở lên)

Từ bộ dữ liệu thực tế trên để ta lấy mẫu dữ liệu để xử lý như sau:

Nhóm	Rời mạng		Không rời mạng lấy mẫu (thuê bao)	Tổng (thuê bao)
	Số mẫu (thuê bao)	Số lượng thực tế (thuê bao)		
Nhóm 1	20.000	93.962	136.000	156.000
Nhóm 2	10.000	67.565	68.000	78.000
Nhóm 3 (nhóm III)	20.002	219.993	136.000	156.002
Nhóm 4 (nhóm IV)	25.001	1.061.729	170.000	195.001
Nhóm 5 (nhóm V)	20.000	105.703	136.000	156.000
Tổng	**95.003**	**1.548.952**	**646.000**	**741.003**

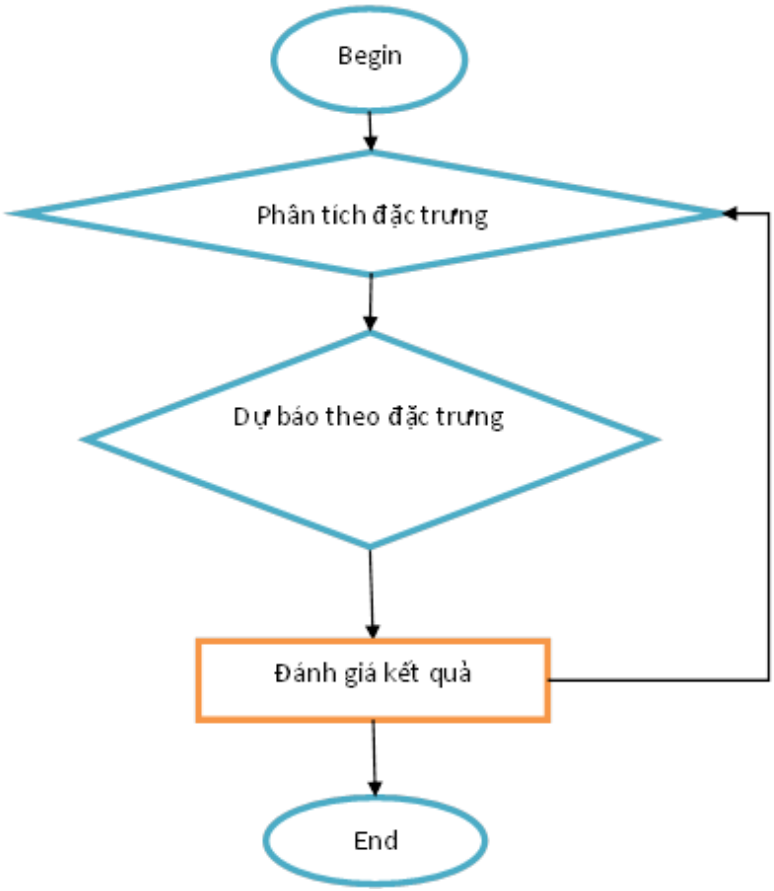
Bảng 4-2 Các nhóm dữ liệu mẫu

Trong đó số liệu không rời mạng là nhóm thuê bao có phát sinh cước trong tháng 8.

4.3. Kết quả thực nghiệm theo phương pháp hiện tại

Kết quả phân tích các nhóm dữ liệu đã được các chuyên gia MobiFone tìm ra được những đặc điểm của thuê bao trước khi rời mạng. Sau khi áp dụng các đặc trưng này vào bộ dữ liệu kiểm nghiệm để đánh giá thì kết quả như sau. Như đã đề cập ở chương 1, trong phạm vi đề tài này chỉ tập trung vào nhóm thuê bao trả trước mà không tập trung vào nhóm thuê bao trả sau. Do vậy, trong phần này, chỉ tập trung vào việc sử dụng các đặc điểm đã phân tích để dự báo cho các thuê bao trả trước

Cách thức thực hiện như sau:



- Bước 1: Sử dụng tập dữ liệu thuê bao rời mạng trong tháng 8 để thực hiện phân tích đặc trưng của các thuê bao rời mạng theo từng nhóm thuê bao đã mô tả tại mục 4.2.
- Bước 2: Dự báo thuê bao rời mạng trên 1 tập dữ liệu lấy mẫu để đánh giá kết quả dự báo
- Bước 3: Đánh giá kết quả theo phương pháp đánh giá đã nêu tại chương 2, để lựa chọn các đặc trưng tối ưu. Sau khi đánh giá quay lại bước phân tích nếu chưa được kết quả tối ưu.

Việc thực hiện các bước trên nhiều lần cùng với kiến thức chuyên gia về lĩnh vực CSKH của mạng MobiFone đã đưa ra các đặc trưng (tập luật) để thực hiện phân lớp (chi tiết đặc trưng xem phụ lục 2). Sau khi thực hiện dùng tập luật để dự đoán thì kết quả như sau:

Nhóm 1:

- Kết quả phân lớp

		Kết quả dự đoán		
		Rời mạng	*Không rời mạng*	***Tổng***
Thực tế	*Rời mạng*	16.706	3.294	20.000
	Không rời mạng	8.461	127.539	136.000
	Tổng	25.167	130.833	156.000

Nhóm 2:

- Kết quả phân lớp (dự báo):

		Kết quả dự đoán		
		Rời mạng	*Không rời mạng*	***Tổng***
Thực tế	*Rời mạng*	9.530	470	10.000
	Không rời mạng	2.375	65.625	68.000
	Tổng	11.905	66.095	78.000

Nhóm 3:

- Kết quả phân lớp (dự báo):

		Kết quả dự đoán		
		Rời mạng	*Không rời mạng*	***Tổng***
Thực tế	*Rời mạng*	19.047	955	20.002
	Không rời mạng	913	135.087	136.000
	Tổng	19.960	136.042	156.002

Nhóm 4:

- Kết quả phân lớp (dự báo):

		Kết quả dự đoán		
		Rời mạng	*Không rời mạng*	***Tổng***
Thực tế	*Rời mạng*	22.724	3.953	26.677
	Không rời mạng	2.218	167.782	170.000
	Tổng	24.942	171.735	196.677

Nhóm 5:

- Kết quả phân lớp (dự báo):

		Kết quả dự đoán		
		Rời mạng	*Không rời mạng*	*Tổng*
Thực tế	*Rời mạng*	9.861	10.139	20.000
	Không rời mạng	23.559	112.441	136.000
	Tổng	33.420	122.580	156.000

Tổng hợp kết quả dự báo dữ liệu thực dựa vào các đặc điểm đặc trưng của thuê bao trước khi rời mạng như sau:

Tên độ đo	**Nhóm 1**	**Nhóm 2**	**Nhóm 3**	**Nhóm 4**	**Nhóm 5**
Accuracy	92,5%	96,4%	98,7%	96,9%	78,4%
Error_Rate	7,5%	3,6%	1,3%	3,1%	21,6%
Recall	83,5%	80,1%	95,1%	91,1%	49,3%
Precision	66,4%	95,3%	95,1%	85,2%	29,5%

Bảng 4-3 Tổng hợp kết quả của phương pháp hiện tại

Nhận xét: Từ kết quả tổng hợp cũng như phân tích ta thấy các nhóm từ 1,2, 3 và 4 có độ chính xác tương đối cao tuy nhiên nhóm này theo như phân tích lại không phải là đối tượng chính của việc dự báo do những nhóm này là các thuê bao có dấu hiệu tương đối rõ là đã rời mạng. Đối với các thuê bao ở nhóm 5 đây là nhóm thuê bao trong tháng trước là các thuê bao bình thường thì đây là đối tượng cần quan tâm nhất thì phương pháp này cho kết quả không cao.

4.4. Kết quả thực nghiệm dựa trên khai phá dữ liệu

4.4.1. Kết quả thực nghiệm dựa trên giải pháp khai phá dữ liệu

Trong phần này, ta sẽ sử dụng tập dữ liệu thuê bao chưa được phân lớp (chưa được xác định là rời mạng hay không) làm đầu vào cho mô hình. Sử dụng ứng dụng WEKA và thuật toán cây quyết định để thực nghiệm. Lý do sử dụng thuật toán cây quyết định vì đây thuật toán có thời gian chạy dữ liệu nhanh và đối với dữ liệu này cho kết quả cao hơn so với phương pháp đang thực hiện tại MobiFone nên phù hợp với yêu cầu của MobiFone.

Tập dữ liệu chạy được trích xuất từ bộ dữ liệu thực tế như mô tả tại mục 4.1. Trong tập dữ liệu trên ta chia tập train (luyện tập) và tập test (kiểm chứng) theo nguyên tắc như sau:

Tập luyện tập: Là tập chứa 66% số liệu của bộ dữ liệu đầu vào.

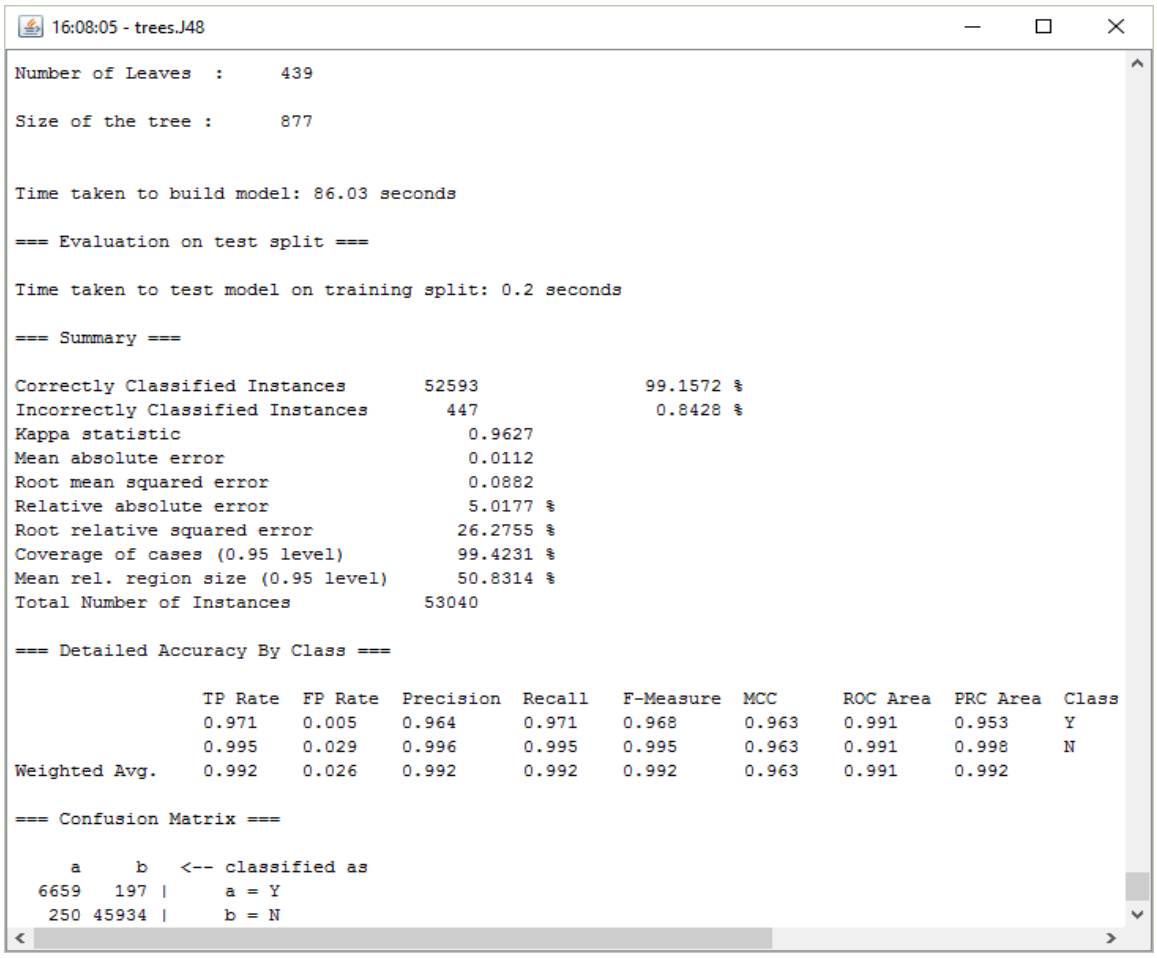
Tập kiểm chứng: Là tập chứa 34% số liệu của bộ dữ liệu đầu vào còn lại.

Thuật toán phân lớp: Cây quyết định theo thuật toán C4.5 được chương trình hóa trong mục phân lớp trên công cụ Weka là J48.

Kết quả chi tiết như sau:

a. Nhóm 1

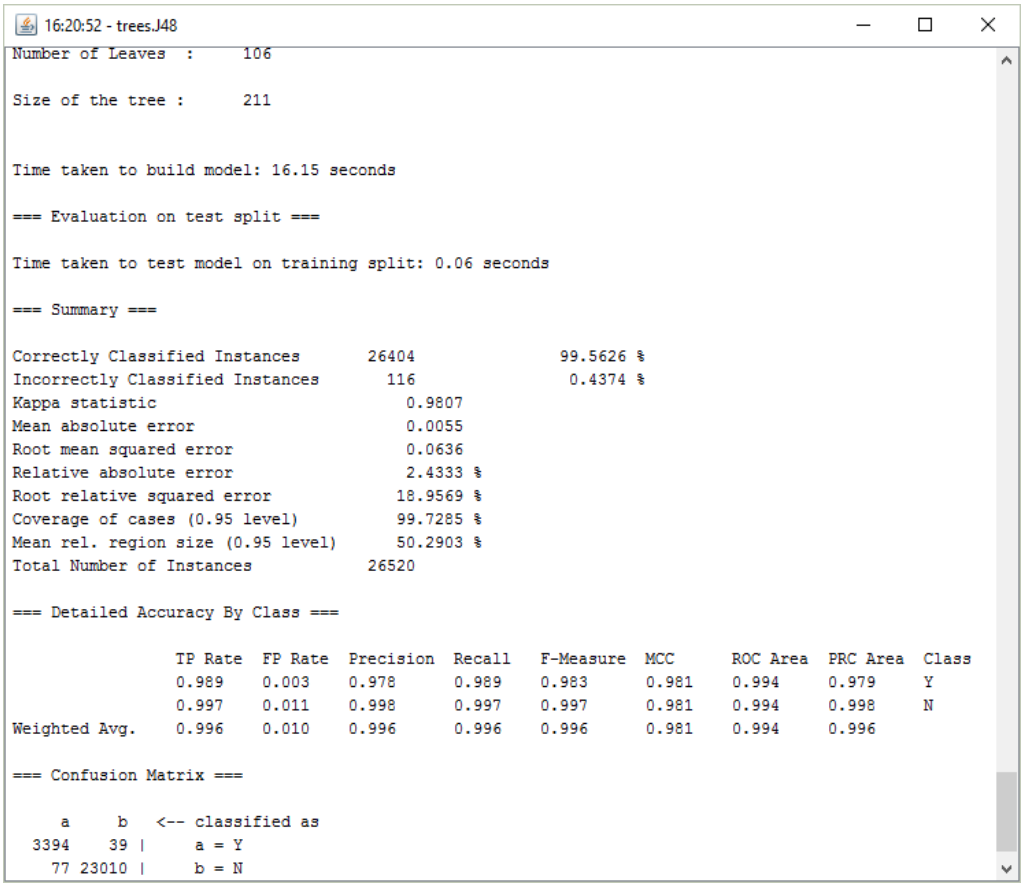
Tập dữ liệu chạy bao gồm 156.000 TB, trong đó: 20.000 TB rời mạng, 136.000 TB không rời mạng



Hình 4-1 Kết quả nhóm 1 của giải pháp đề xuất

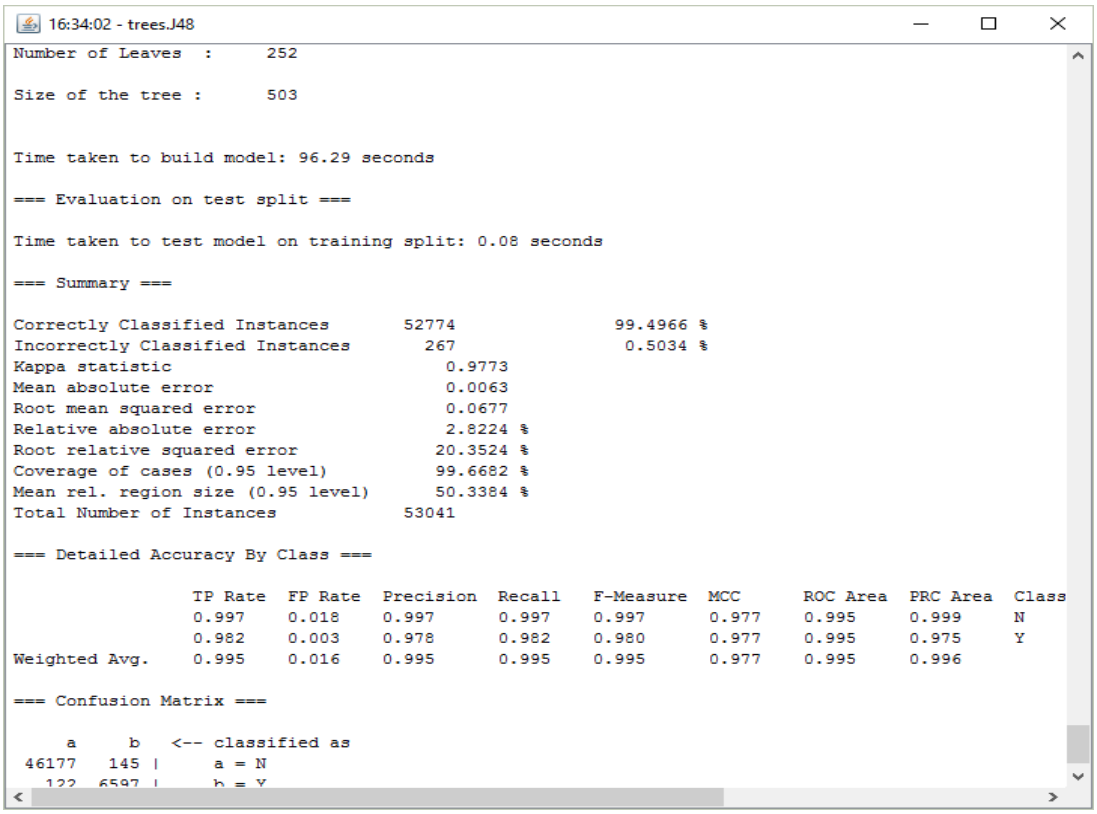
b. Nhóm 2:

Tập dữ liệu chạy bao gồm 156.000 TB, trong đó: 20.000 TB rời mạng, 136.000 TB không rời mạng



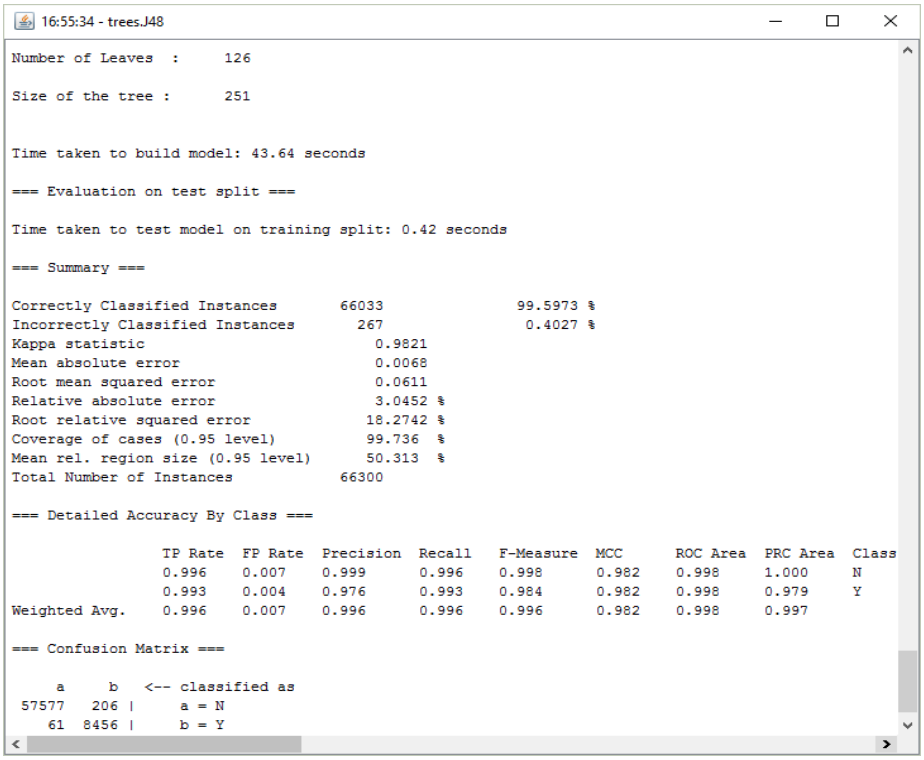
Hình 4-2 Kết quả nhóm 2 của giải pháp đề xuất

c. Nhóm 3:



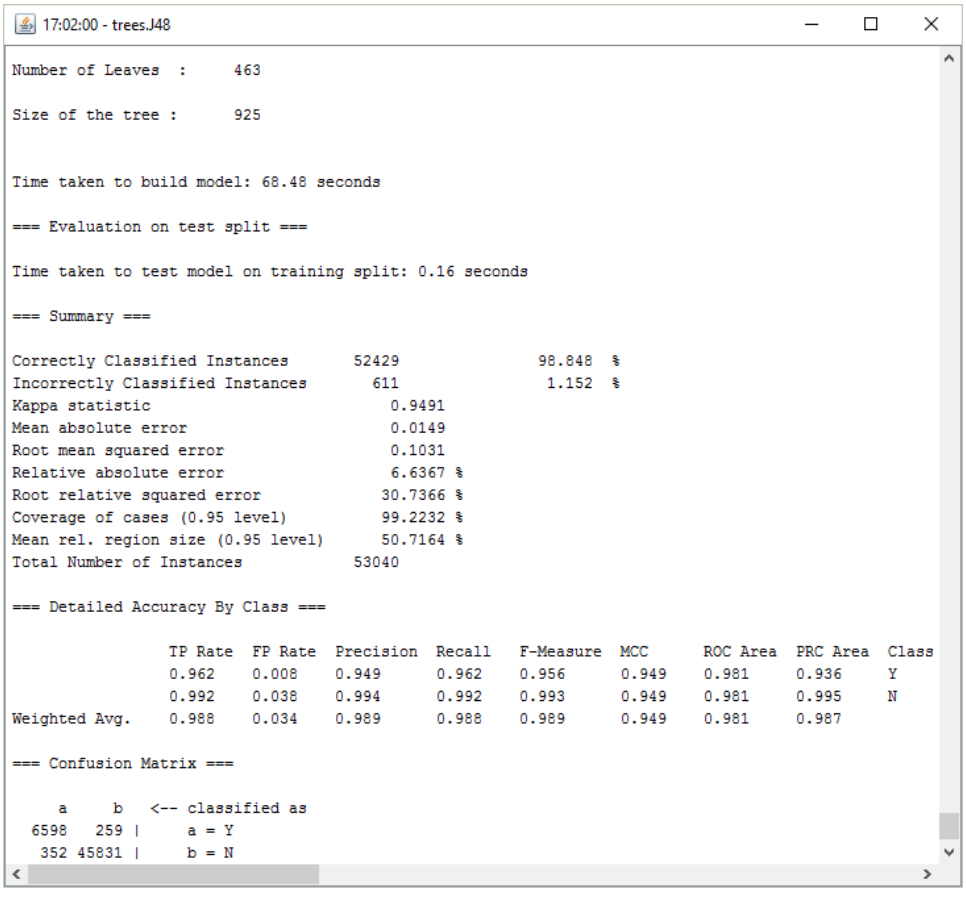
Hình 4-3 Kết quả nhóm 3 của giải pháp đề xuất

d. Nhóm 4



Hình 4-4 Kết quả nhóm 4 của giải pháp đề xuất

e. Nhóm 5:



Hình 4-5 Kết quả nhóm 5 của giải pháp đề xuất

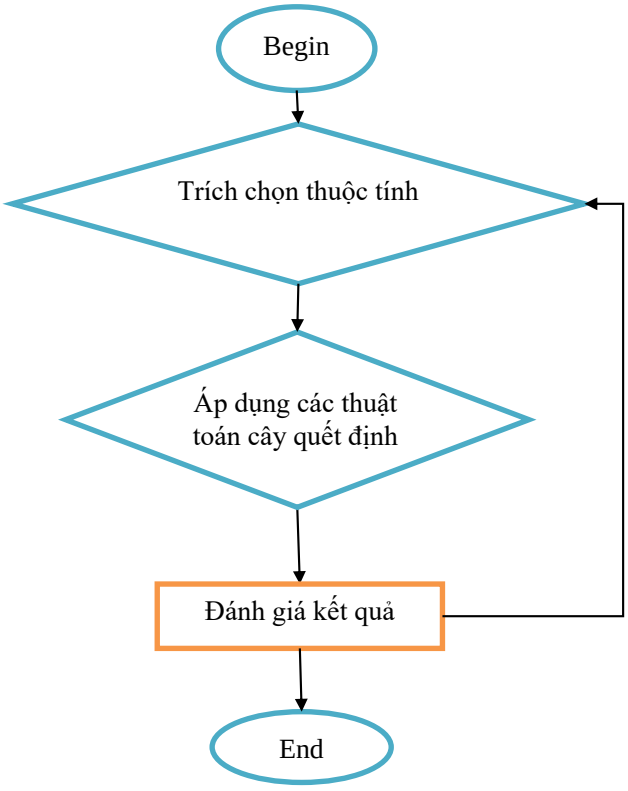
f. Tổng hợp kết quả và đánh giá:
Bảng tổng hợp đánh giá

Tên độ đo	Nhóm 1	Nhóm 2	Nhóm 3	Nhóm 4	Nhóm 5
Accuracy	99.2%	99.6%	99.5%	99.6%	98.8%
Error_Rate	0.8%	0.4%	0.5%	0.4%	1.2%
Recall	97.1%	98.9%	99.7%	99.6%	96.2%
Precision	96.4%	97.8%	99.7%	99.9%	94.9%
Thời gian xây dựng mô hình (giây)	86	16	96	44	68
Thời gian dự báo (giây)	0.2	0.06	0.08	0.42	0.16

Bảng 4-4 Bảng tổng hợp kết quả theo giải pháp khai phá dữ liệu

4.4.2. Kết quả thực nghiệm dựa trên khai phá dữ liệu đã cải tiến

Trong phần này, ta sẽ sử dụng tập dữ liệu thuê bao đã được chọn mẫu tại mục 4.1 để thực nghiệm. Các bước của quá trình thực nghiệm như sau:



❖ Bước 1:

- Chạy thuật toán để trích chọn đặc trưng các thuộc tính. Các thuật toán đã thực nghiệm gồm các thuật toán đã nêu tại chương 3. Qua thực nghiệm trên bộ số liệu thực tế thì thuật toán GainRatioAttributeEval phù hợp nhất với các bộ số liệu thực nghiệm. Sau khi thực nghiệm theo thuật toán này thì sẽ ra được danh sách đánh theo thứ tự quan trọng của các thuộc tính và ta tiếp tục phải tìm số lượng thuộc tính tối ưu.

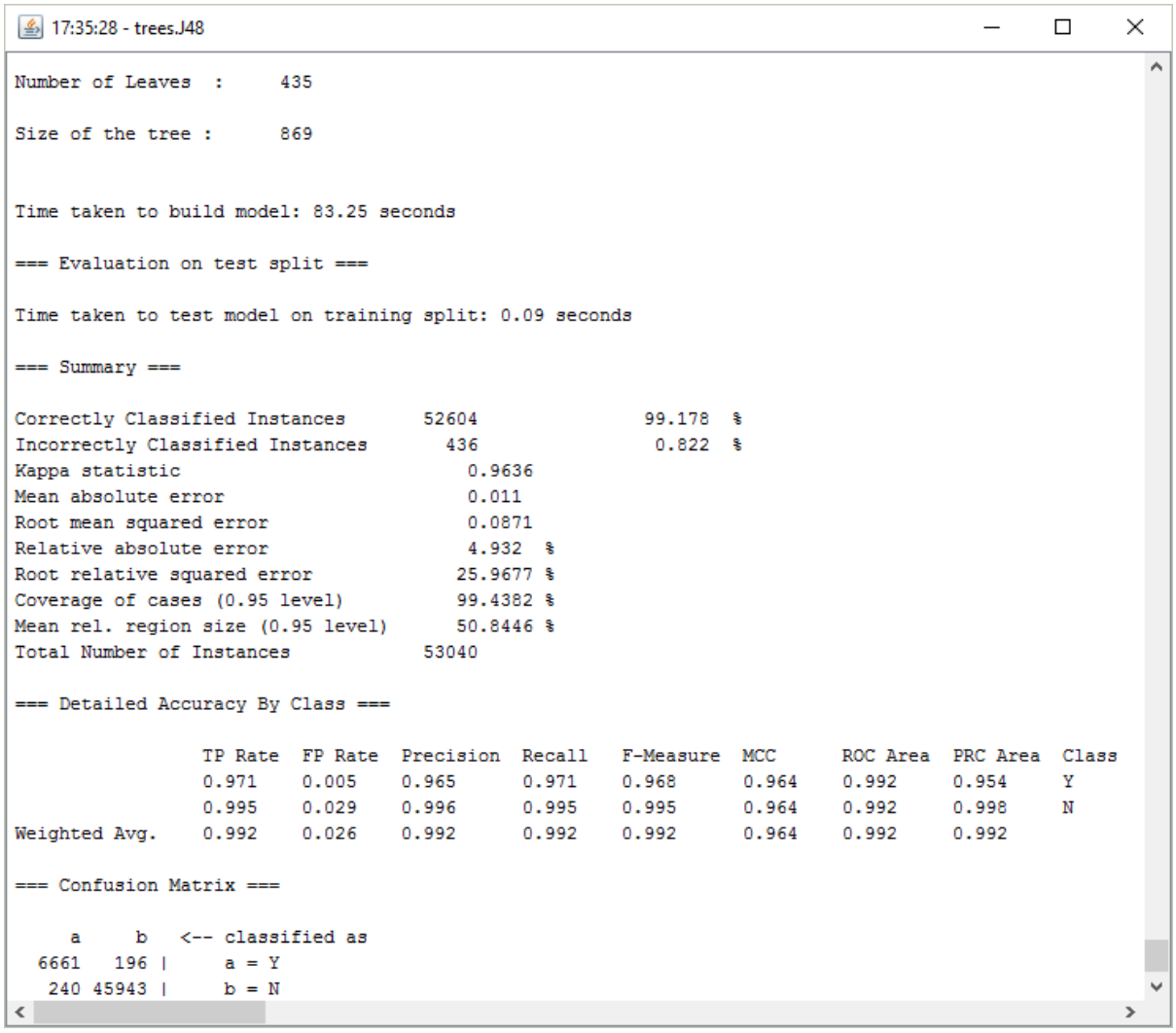
- Lựa chọn số thuộc tính trích chọn: Lựa chọn càng nhiều thuộc tính thì kết quả càng chính xác tuy nhiên thời gian xây dựng mô hình lại càng lâu. Để đảm bảo phù hợp với yêu cầu của MobiFone là xử lý dữ liệu nhanh với độ chính xác chấp nhận được nên ta phải lựa chọn số lượng thuộc tính vừa phải không làm giảm độ chính xác quá nhiều nhưng thời gian xây dựng mô hình nhanh.

❖ Bước 2: Chạy kỹ thuật Cây quyết định (Kỹ thuật J48 trên mô hình WEKA) để xây dựng mô hình phân lớp tập dữ liệu sau khi trích chọn và dự báo đánh giá mô hình đã xây dựng.

4.4.2.1 Nhóm 1:

Tập Training: 156.000 TB, trong đó: 20.000 TB rời mạng, 136.000 TB không rời mạng

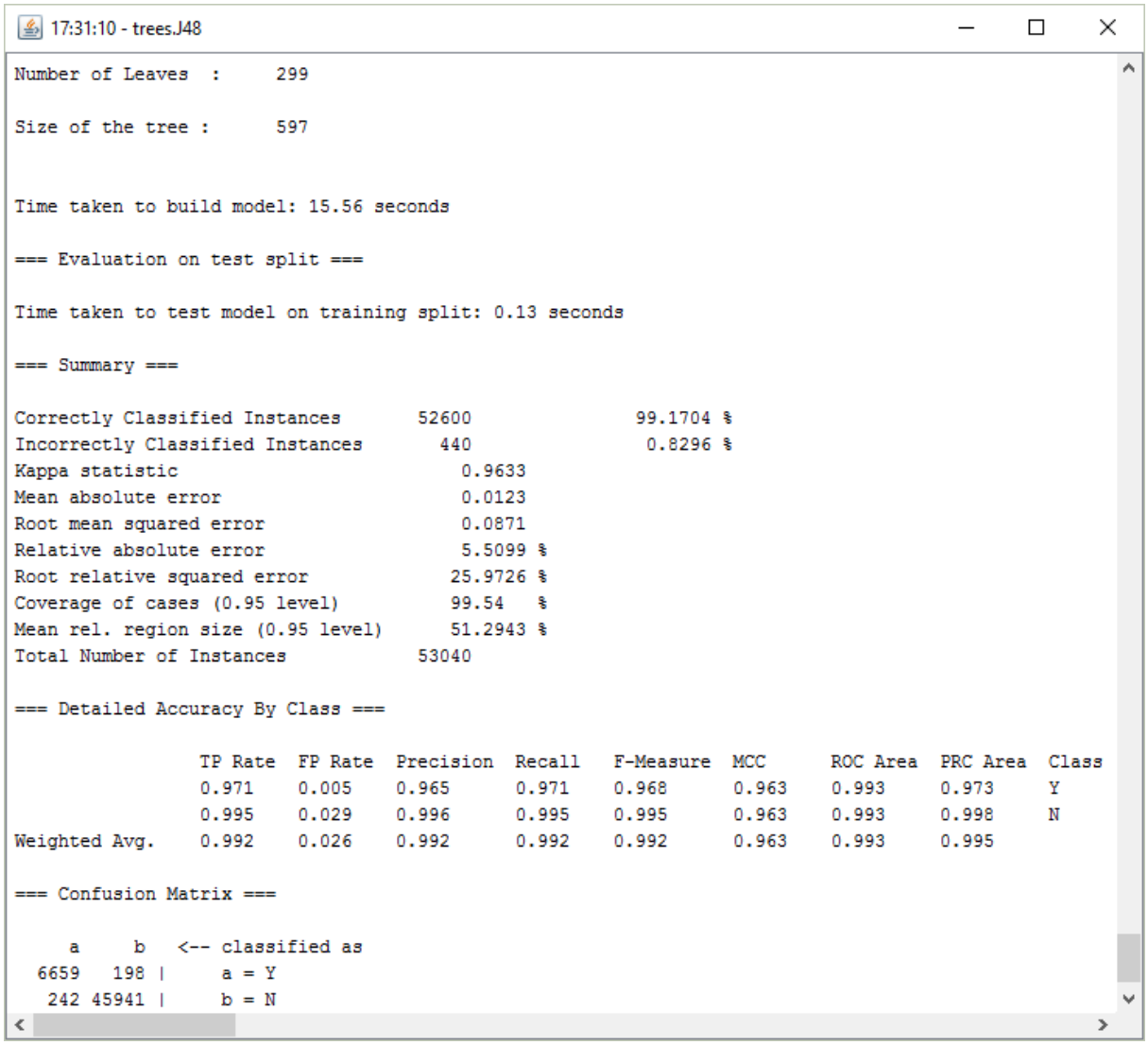
Kết quả sau khi sắp sếp lại các trường dữ liệu theo xếp hạng của thuật toán GainRatioAttributeEval



Hình 4-6 Kết quả nhóm 1 của giải pháp đề xuất cải tiến thử nghiệm

Kết quả cho thấy độ chính xác của thuật toán tăng so với trước khi sắp xếp. Thời gian xây dựng mô hình giảm 3s so với bộ dữ liệu chưa xếp hạng. Như vậy có nhận xét thuật toán cây quyết định bị ảnh hưởng bởi thứ tự trường dữ liệu đầu vào.

Lựa chọn số thuộc tính từ xếp hạng trên với 15 thuộc tính được xếp hạng cao nhất kết quả thuật toán C4.5 như sau:



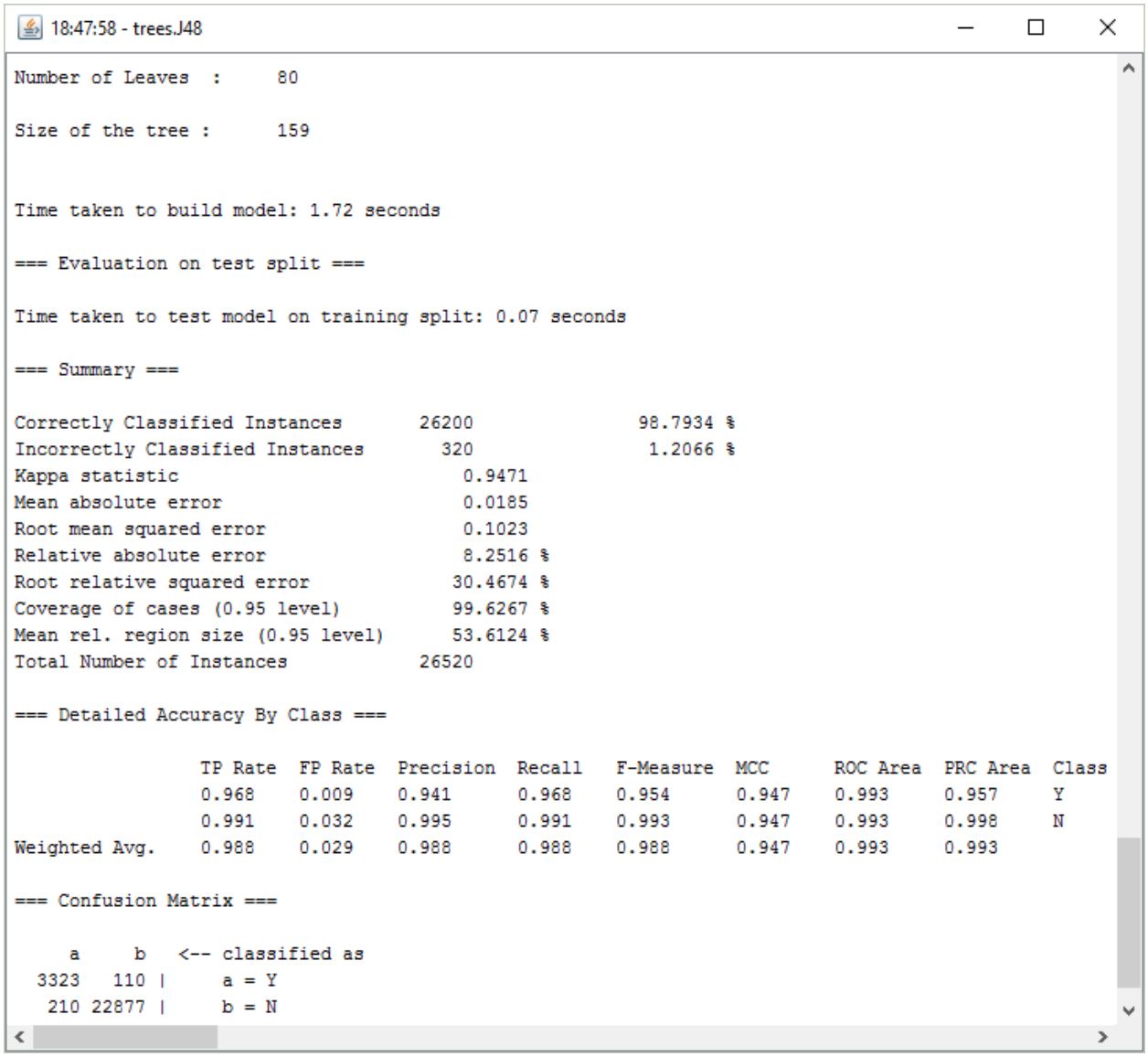
Hình 4-7 Kết quả nhóm 1 của giải pháp đề xuất cải tiến sau tối ưu

Đánh giá đối với lựa chọn này thì kết quả độ chính xác thuật toán vẫn tương đương so với bộ dữ liệu đã xếp hạng. Thời gian chạy số liệu giảm từ 83s xuống còn 16s

4.4.2.2 Nhóm 2:

Tập dữ liệu: 78.000 TB, trong đó: 10.000 TB rời mạng, 68.000 TB không rời mạng

Kết quả sau khi thực nghiệm áp dụng thuật toán GainRatioAttributeEval và trích chọn được 10 trường dữ liệu tối ưu như sau:



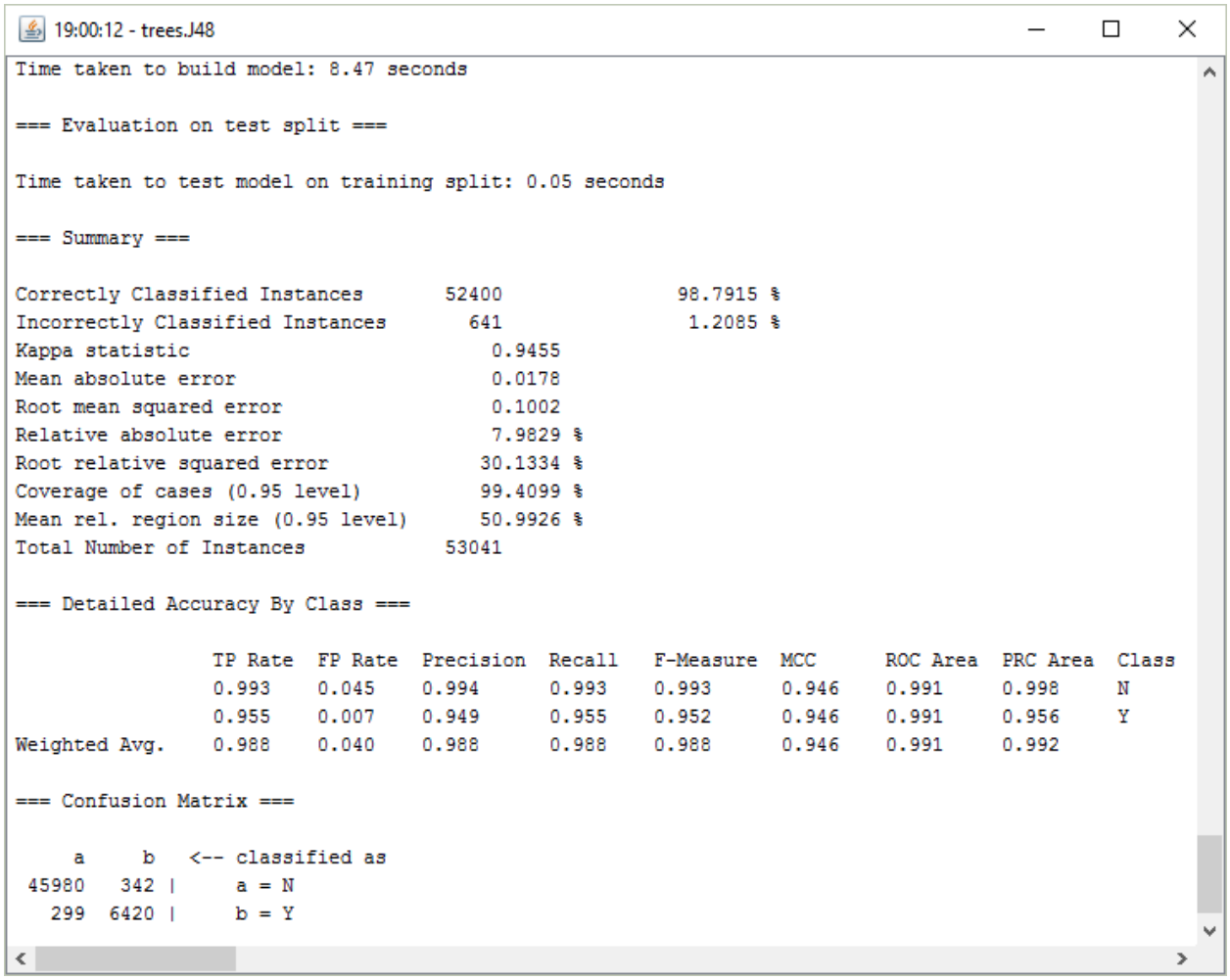
Hình 4-8 Kết quả nhóm 2 của giải pháp đề xuất cải tiến

Đánh giá Thời gian dựng mô hình giảm rất nhiều so với giải pháp đề xuất nhưng độ chính xác gần tương đương.

4.4.2.3 Nhóm 3:

Tập dữ liệu chạy: 156.000 TB, trong đó: 20.002 TB rời mạng, 136.000 TB không rời mạng

Kết quả sau khi thực nghiệm áp dụng thuật toán trích chọn được 12 thuộc tính:

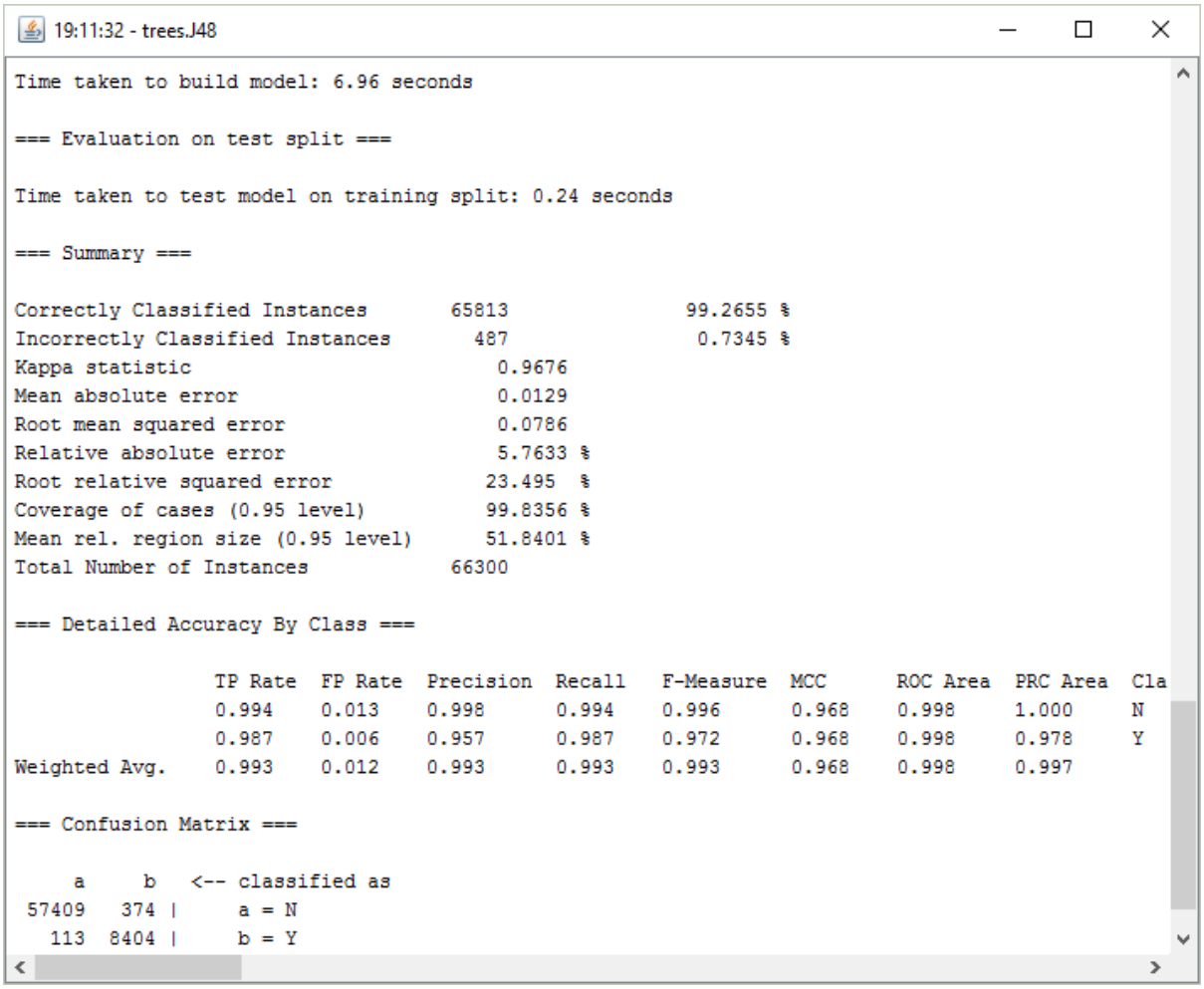


Hình 4-9 Kết quả nhóm 3 của giải pháp đề xuất cải tiến

4.4.2.4 Nhóm 4:

Tập Training: 195.000 TB, trong đó: 25.001 TB rời mạng, 170.000 TB không rời mạng

Kết quả với 20 thuộc tính lựa chọn

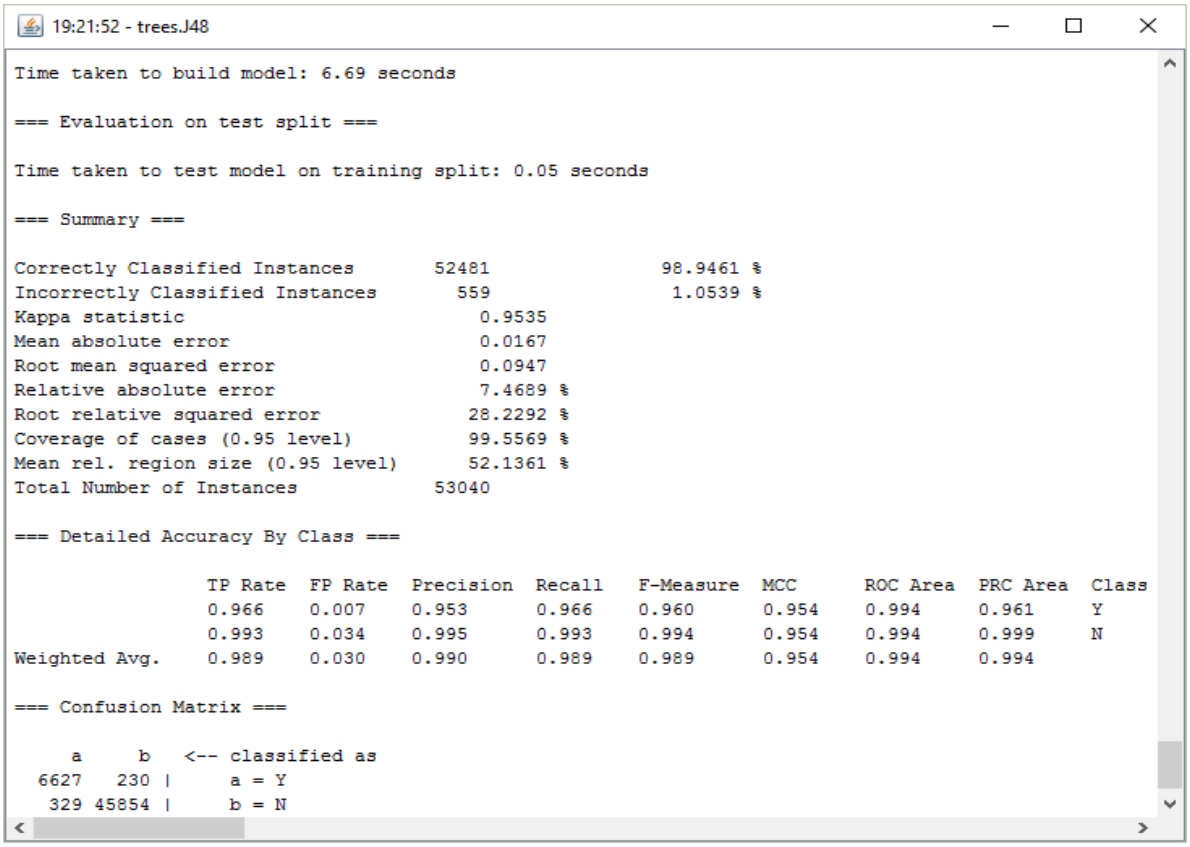


Hình 4-10 Kết quả nhóm 4 của giải pháp đề xuất sau cải tiến

4.4.2.5 Nhóm 5:

Tập dữ liệu: 156.000 TB, trong đó: 20.000 TB rời mạng, 136.000 TB không rời mạng

Kết quả với trích chọn 7 thuộc tính.



Hình 4-11 Kết quả nhóm 5 của giải pháp đề xuất cải tiến

4.4.2.6 Đánh giá mô hình

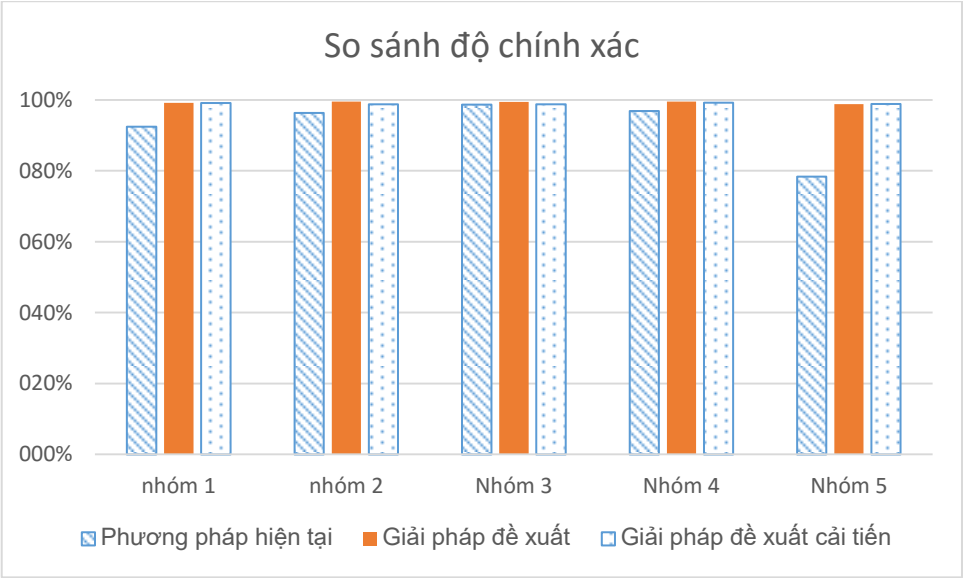
Từ kết quả kiểm nghiệm ở mục 3.3, ta đã tính toán độ chính xác của dự đoán đối với từng nhóm thuê bao dựa vào công thức đã đưa ra như sau:

Tên độ đo	Nhóm 1	Nhóm 2	Nhóm 3	Nhóm 4	Nhóm 5
Accuracy	99.2%	98.8%	98.8%	99.3%	98.9%
Error_Rate	0.8%	1.2%	1.2%	0.7%	1.1%
Recall	96.5%	98.8%	99.3%	99.4%	96.6%
Precision	97.1%	94.1%	99.4%	99.8%	95.3%
Thời gian xây dựng mô hình (giây)	16	2	8	7	7
Thời gian dự báo (giây)	0.13	0.07	0.05	0.24	0.05

Bảng 4-5 Tổng hợp độ chính xác của giải pháp đề xuất cải tiến

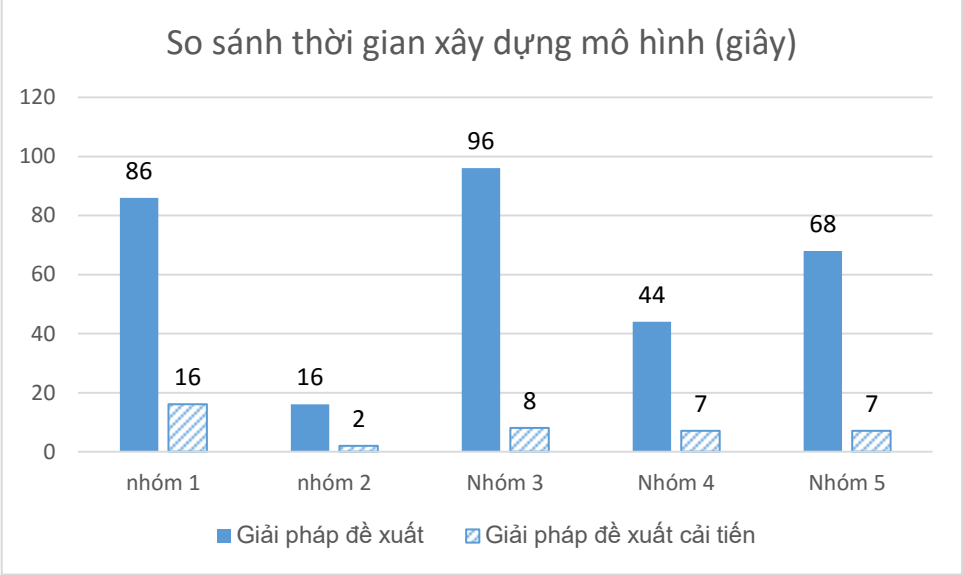
4.5. So sánh đánh giá kết quả

Để đánh giá kết quả đã thực nghiệm ta so sánh độ chính xác (độ đo Accuracy) của 3 giải pháp đã thực hiện như sau:

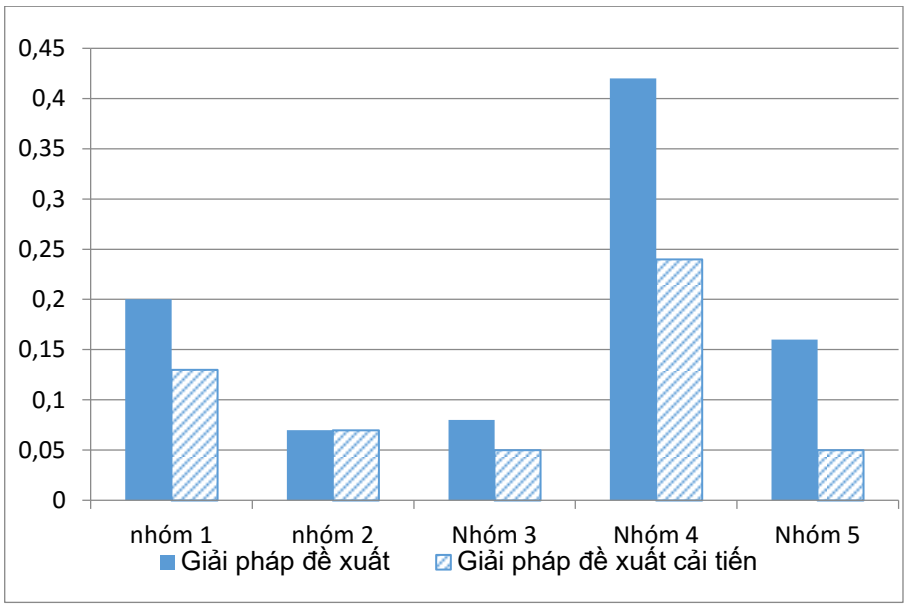


Biểu đồ 4-1 So sánh độ đo Accuracy của ba giải pháp

Nhận xét: giải pháp hiện tại MobiFone đang sử dụng có độ chính xác thấp nhất, đối với các nhóm 3 và 4 độ chính xác nhỏ hơn nhưng gần tương đương với độ chính xác của các giải pháp dùng thuật toán khai phá dữ liệu, nhưng đối với các nhóm 1,2 và 5 thì độ chính xác thấp hơn đáng kể so với các giải pháp dùng khai phá dữ liệu đặc biệt nhóm 5 thấp hơn rất nhiều (78,4% so với 98,8%). Qua việc so sánh độ chính xác của các giải pháp ta nhận thấy được các đặc trưng của các nhóm thuê bao trong đó các nhóm 3,4 là các nhóm đã được nhận định là các thuê bao rời mạng, nhóm 5 nhận định là nhóm mục tiêu chính để dự báo thì kết quả giải pháp hiện tại của MobiFone chưa cao trong khi đó giải pháp đề xuất cho kết quả rất cao. Độ chính xác của giải pháp đề xuất và giải pháp đề xuất sau khi cải tiến tương đương nhau. Để đánh giá tiếp về giải pháp trước và sau cải tiến ta tiếp tục đánh giá thời gian xây dựng mô hình và thời gian dự báo giữa hai giải pháp này kết quả như biểu đồ dưới:



Biểu đồ 4-2 So sánh thời gian xây dựng mô hình của giải pháp đề xuất và đề xuất cải tiến



Biểu đồ 4-3 So sánh thời gian dự báo

Qua biểu đồ ta thấy rõ thời gian xây dựng mô hình và thời gian dự báo của giải pháp sau khi cải tiến ít hơn rất nhiều so với giải pháp trước cải tiến. Do vậy giải pháp đề xuất cải tiến đã đạt được mục tiêu đề ra ban đầu là đưa ra giải pháp với độ chính gần tương đương (có thể giảm) tuy nhiên thời gian xây dựng mô hình và thời gian dự báo sẽ nhanh hơn.

4.6. Kết luận chương 4

Chương này trình bày kết quả thực nghiệm dựa trên các phương pháp đã thực hiện ở chương 3. Qua thực nghiệm ta có nhận xét như sau: Kết quả dùng thuật toán khai phá dữ liệu cao hơn so với phương pháp hiện tại đang thực hiện.

KẾT LUẬN

Từ việc nghiên cứu những yêu cầu cấp thiết đặt ra trong công tác duy trì và phát triển thuê bao của mạng di động, luận văn đã đạt được một số kết quả chính sau đây:

- Xây dựng mô hình dự báo áp dụng kỹ thuật khai phá dữ liệu để phát hiện nhanh chính xác các thuê bao di động có khả năng rời mạng từ đó áp dụng các giải pháp để duy trì thuê bao.

- Triển khai mô hình đề xuất, áp dụng trên dữ liệu thực tế, so sánh với các giải pháp đã sử dụng được áp dụng. Các kết quả đạt được đã cho thấy được tiềm năng áp dụng phương pháp đề xuất vào thực tiễn

Trong thời gian tới chúng tôi sẽ nghiên cứu tích hợp các kỹ thuật này vào các chương trình hỗ trợ kinh doanh của MobiFone đồng thời cải tiến thời gian dự báo cũng như kết quả dự báo. Trong thời gian tới tôi sẽ tiếp tục cập nhật mô hình với dữ liệu của MobiFone để kết quả dự đoán được cải thiện hơn.

TÀI LIỆU THAM KHẢO

Tiếng Việt

[1] Phan Xuân Hiếu (2013),*Bài giảng môn học KPDL và kho dữ liệu,* Trường Đại học Công nghệ, Đại học Quốc gia Hà Nội.

[2] Bộ Thông tin và Truyền thông (2014),*Sách Trắng về Công nghệ thông tin và Truyền thông (CNTT-TT) Việt Nam 2014,* Nhà xuất bản Thông tin và Truyền thông, Hà nội.

[3] Hà Quang Thụy (2010),*Bài giảng môn học Kho dữ liệu và KPDL,* Trường Đại học Công nghệ, Đại học Quốc gia Hà Nội.

Tiếng Anh

[4] Leo Breiman, Jerome Friedmen, and Charles J. Stone (1984),*Classification and Regression Trees,* Wadsworth International Group.

[5] M. Chandar, Laha, A., & Krishna, P. (2006),*Modeling churn behavior of bank customers using predictive data mining techniques,* National conference on soft computing techniques for engineering applications.

[6] Jiawei Han and Micheline Kamber (2012),*Data Mining Concepts and Techniques,* Third Edition ed Elsevier Inc.

[7] John Ross Quinlan (1993),*C4.5: Programs for Machine Learning,* Morgan Kaufmann Publishers.

[8] John Ross Quinlan, *Induction of decision trees*, in *Machine Learning 1*. 1986. p. 81-106.

[9] J. Burez, & Van den Poel, D. (2009), *Handling class imbalance in customer churn prediction,* Expert System with Applications,36, 4626-4636.

[10] S. Olafsson, Li, X., & Wu, S. (2008), *Operations research and data mining,* European Journal of Operational Research,187, 2592-1448.

[11] *Weka 3 - Data Mining with Open Source Machine Learning Software in Java*. Available from: http://www.cs.waikato.ac.nz/ml/weka/.

Phụ lục 1 THÔNG TIN BỘ DỮ LIỆU

1. Thông tin các trường dữ liệu

STT	Tên trường	Ý nghĩa
1	GRAND_PACKAGE_ID	Mã gói cước thuê bao sử dụng
2	STA_DATETIME	Ngày bắt đầu sử dụng của thuê bao
3	CUS_TYPE	Loại khách hàng (cá nhân /doanh nghiệp)
4	STATUS_ID	Trạng thái thuê bao
5	IS_MEMBER	Mã thể hiện có tham gia chương trình kết nối dài lâu
6	RANK_ID	Hạng thuê bao trong chương trình kết nối dài lâu
7	AREA_SK_PSC_N	Khu vực phát sinh cước tháng n-1
8	AREA_SK_PSC_N1	Khu vực phát sinh cước tháng n-2
9	AREA_SK_PSC_N2	Khu vực phát sinh cước tháng n-3
10	DTKC_THOAI_NOI_N	Doanh thu tài khoản chính của dịch vụ thoại nội mạng tháng n-1
11	TKKM_THOAI_NOI_N	Doanh thu tài khoản khuyến mại của dịch vụ thoại nội mạng tháng n-1
12	DTTKC_THOAI_LIEN_MANG_N	Doanh thu tài khoản chính của dịch vụ thoại liên mạng tháng n-1
13	DTTKM_THOAI_LIEN_MANG_N	Doanh thu tài khoản khuyến mại của dịch vụ thoại liên mạng tháng n-1
14	DTTKC_SMS_NOI_N	Doanh thu tài khoản chính của dịch vụ sms nội mạng tháng n-1
15	DTTKM_SMS_NOI_N	Doanh thu tài khoản khuyến mại của dịch vụ sms nội mạng tháng n-1
16	DTTKC_SMS_LIEN_MANG_N	Doanh thu tài khoản chính của dịch vụ sms liên mạng tháng n-1
17	DTTKM_SMS_LIEN_MANG_N	Doanh thu tài khoản khuyến mại của dịch vụ sms liên mạng tháng n-1
18	DTTKC_DATA_N	Doanh thu tài khoản chính của dịch vụ data tháng n-1
19	DTTKM_DATA_N	Doanh thu tài khoản khuyến mại của dịch vụ data tháng n-1
20	DTTKC_KHAC_N	Doanh thu tài khoản chính của dịch vụ khác tháng n-1
21	DTTKM_KHAC_N	Doanh thu tài khoản khuyến mại của dịch vụ khác tháng n-1
22	DTKC_THOAI_NOI_N1	Doanh thu tài khoản chính của dịch vụ thoại nội mạng tháng n-2
23	TKKM_THOAI_NOI_N1	Doanh thu tài khoản khuyến mại của dịch vụ thoại nội mạng tháng n-2
24	DTTKC_THOAI_LIEN_MANG_N1	Doanh thu tài khoản chính của dịch vụ thoại liên mạng tháng n-2
25	DTTKM_THOAI_LIEN_MANG_N1	Doanh thu tài khoản khuyến mại của dịch vụ thoại liên mạng tháng n-2
26	DTTKC_SMS_NOI_N1	Doanh thu tài khoản chính của dịch vụ sms nội mạng tháng n-2
27	DTTKM_SMS_NOI_N1	Doanh thu tài khoản khuyến mại của dịch vụ sms nội mạng tháng n-2
28	DTTKC_SMS_LIEN_MANG_N1	Doanh thu tài khoản chính của dịch vụ sms liên

		mạng tháng n-2
29	DTTKM_SMS_LIEN_MANG_N1	Doanh thu tài khoản khuyến mại của dịch vụ sms liên mạng tháng n-2
30	DTTKC_DATA_N1	Doanh thu tài khoản chính của dịch vụ data tháng n-2
31	DTTKM_DATA_N1	Doanh thu tài khoản khuyến mại của dịch vụ data tháng n-2
32	DTTKC_KHAC_N1	Doanh thu tài khoản chính của dịch vụ khác tháng n-2
33	DTTKM_KHAC_N1	Doanh thu tài khoản khuyến mại của dịch vụ khác tháng n-2
34	DTKC_THOAI_NOI_N2	Doanh thu tài khoản chính của dịch vụ thoại nội mạng tháng n-3
35	TKKM_THOAI_NOI_N2	Doanh thu tài khoản khuyến mại của dịch vụ thoại nội mạng tháng n-3
36	DTTKC_THOAI_LIEN_MANG_N2	Doanh thu tài khoản chính của dịch vụ thoại liên mạng tháng n-3
37	DTTKM_THOAI_LIEN_MANG_N2	Doanh thu tài khoản khuyến mại của dịch vụ thoại liên mạng tháng n-3
38	DTTKC_SMS_NOI_N2	Doanh thu tài khoản chính của dịch vụ sms nội mạng tháng n-3
39	DTTKM_SMS_NOI_N2	Doanh thu tài khoản khuyến mại của dịch vụ sms nội mạng tháng n-3
40	DTTKC_SMS_LIEN_MANG_N2	Doanh thu tài khoản chính của dịch vụ sms liên mạng tháng n-3
41	DTTKM_SMS_LIEN_MANG_N2	Doanh thu tài khoản khuyến mại của dịch vụ sms liên mạng tháng n-3
42	DTTKC_DATA_N2	Doanh thu tài khoản chính của dịch vụ data tháng n-3
43	DTTKM_DATA_N2	Doanh thu tài khoản khuyến mại của dịch vụ data tháng n-3
44	DTTKC_KHAC_N2	Doanh thu tài khoản chính của dịch vụ khác tháng n-3
45	DTTKM_KHAC_N2	Doanh thu tài khoản khuyến mại của dịch vụ khác tháng n-3
46	REMAIN_CREDIT	Số tiền còn lại trong tài khoản chính tháng trước
47	REMAIN_BONUS	Số tiền còn lại trong tài khoản thưởng tháng trước
48	MONTH_3K3D_NEAREST	Tháng đạt chỉ tiêu 3k3d gần nhất
49	NUM_OG_CALLS	Số lượng cuộc gọi thoại chiều đi tháng n-1
50	SUM_DURATION_OG	Số lượng lưu lượng thoại chiều đi tháng n-1
51	SUM_DURATION_IC	số lượng lưu lượng thoại đến tháng n-1
52	NUM_SMO	Số lượng sms chiều đi tháng n-1
53	NUM_SMT	Số lượng sms chiều đến tháng n-1
54	NUM_OG_CALLS_N1	Số lượng cuộc gọi thoại chiều đi tháng n-2
55	SUM_DURATION_OG_N1	Số lượng lưu lượng thoại chiều đi tháng n-2
56	SUM_DURATION_IC_N1	số lượng lưu lượng thoại đến tháng n-2
57	NUM_SMO_N1	Số lượng sms chiều đi tháng n-2
58	NUM_SMT_N1	Số lượng sms chiều đến tháng n-2
59	NUM_OG_CALLS_N2	Số lượng cuộc gọi thoại chiều đi tháng n-3
60	SUM_DURATION_OG_N2	Số lượng lưu lượng thoại chiều đi tháng n-3

61	SUM_DURATION_IC_N2	số lượng lưu lượng thoại đến tháng n-3
62	NUM_SMO_N2	Số lượng sms chiều đi tháng n-3
63	NUM_SMT_N2	Số lượng sms chiều đến tháng n-3
64	SCR_AMOUNT_N	Số tiền nạp thẻ tháng n-1
65	SCR_COUNT_N	Số lượng thẻ nạp tháng n-1
66	SCR_AMOUNT_N1	Số tiền nạp thẻ tháng n-2
67	SCR_COUNT_N1	Số lượng thẻ nạp tháng n-2
68	SCR_AMOUNT_N2	Số tiền nạp thẻ tháng n-3
69	SCR_COUNT_N2	Số lượng thẻ nạp tháng n-3
70	DEBIT_N	Nợ tiền nợ cước tháng n-1
71	MONTH_DEBIT_N	Số tháng nợ cước tính đến tháng n-1
72	DEBIT_N1	Nợ tiền nợ cước tháng n-2
73	MONTH_DEBIT_N1	Số tháng nợ cước tính đến tháng n-2
74	DEBIT_N2	Nợ tiền nợ cước tháng n-3
75	MONTH_DEBIT_N2	Số tháng nợ cước tính đến tháng n-3
76	NUM_VLR_N	Số ngày cập nhật VLR tháng n-1
77	NUM_VLR_N1	Số ngày cập nhật VLR tháng n-2
78	NUM_VLR_N2	Số ngày cập nhật VLR tháng n-3
79	NUM_PSC_N	Số ngày phát sinh cước tháng n-1
80	NUM_PSC_N1	Số ngày phát sinh cước tháng n-2
81	NUM_PSC_N2	Số ngày phát sinh cước tháng n-3
82	PHONE_MODEL	Nhãn hiệu máy khách hàng đang dùng
83	PROM_CODE_N	Số chương trình khuyến mại tham gia tháng n-1
84	PROM_CODE_N1	Số chương trình khuyến mại tham gia tháng n-2
85	PROM_CODE_N2	Số chương trình khuyến mại tham gia tháng n-3
86	RM	Trường đánh dấu rời mạng

PHỤ LỤC 2 PHÂN TÍCH ĐẶC TRƯNG BỘ DỮ LIỆU THỰC NGHIỆM

1. Nhóm 1

MobiFone đưa ra tiêu chí thuê bao 3k3d_vlr để nhằm thống kê thuê bao sử dụng thật và thuê bao ảo (thuê bao nằm trên kênh phân phối, thuê bao được khách hàng mua dùng để nhắn tin rác, …). Theo đó, thuê bao đạt chuẩn 3k3d_vlr là thuê bao thỏa mãn đồng thời 2 tiêu chí sau: (1) tiêu dùng tối thiểu 3.000đ từ tài khoản chính và (2) cập nhật vlr (mở máy) tối thiểu 3 ngày. Do vậy, mốc tiêu dùng 3.000đ từ tài khoản chính và mốc cập nhật vlr 3 ngày là 2 mốc rất quan trọng để đánh giá thuê bao.

Trong phần này, nhóm chuyên gia đã thực hiện phân tích nhóm thuê bao không phát sinh cước trong tháng 8/2015 nhưng có tiêu dùng tài khoản chính và không đạt 3k3d_vlr trong tháng trước (tháng 7/2015). Nhóm này gồm những thuê bao thỏa mãn một trong 2 điều kiện sau:

- Tiêu dùng tài khoản chính dưới 3.000đ trong tháng 7/2015. Hoặc:
- Tiêu dùng tài khoản chính từ 3.000đ trở lên nhưng cập nhật VLR dưới 3 ngày trong tháng 7/2015.

Sau khi phân tích nhóm đã rút ra được các đặc điểm đặc trưng nhất của tập thuê bao này trước khi rời mạng (không phát sinh cước) như sau:

- *Tiêu dùng tài khoản chính dưới 3.000đ trong tháng trước.*
- *Chỉ đạt 3k3d_vlr một đến 2 lần trong 6 tháng liên tục.*
- *Không phát sinh giao dịch hoặc chỉ phát sinh 1 đến 2 giao dịch thoại chiều đi trong tháng trước.*
- *Không phát sinh giao dịch thoại chiều đến trong tháng trước.*
- *Có xu hướng giảm dần số lượng giao dịch thoại chiều đến trong 3 tháng liên tục hoặc giảm sản lượng thoại chiều đến trong 3 tháng liên tục.*
- *Hầu như không phát sinh giao dịch nạp thẻ trong tháng liền trước.*
- *Có xu hướng giảm tần suất nạp thẻ trong 3 tháng liên tiếp.*

2. Nhóm 2

Nhóm này có 67.565 thuê bao, trong đó chiếm tỷ trọng cao nhất là MobiQ (51,07%), MobiCard (26,84%), Zone+ (9,55%). Trong cơ cấu thuê bao của MobiFone hiện tại, MobiQ chiếm khoảng 49%, MobiCard chiếm khoảng 12%, Zone+ chiếm khoảng 12%.

Như vậy, rõ ràng nhóm thuê bao MobiCard đang có tỉ lệ rời mạng cao hơn các nhóm khác.

Sau khi phân tích nhóm chuyên gia đã rút ra được các đặc điểm đặc trưng nhất của tập thuê bao này trước khi rời mạng (không phát sinh cước) như sau:

- *Tiêu dùng TKC ≥ 3.000đ nhưng chỉ tiêu dùng TKKM < 3.000đ trong tháng trước.*
- *Có mức tiêu dùng TKKM giảm dần trong 3 tháng liên tục.*
- *Không phát sinh giao dịch hoặc chỉ phát sinh 1 đến 2 giao dịch thoại chiều đi trong tháng trước.*
- *Không phát sinh giao dịch hoặc chỉ phát sinh 1 đến 2 giao dịch sms chiều đi trong tháng trước.*
- *Không phát sinh giao dịch thoại chiều đến trong tháng trước.*
- *Có xu hướng giảm dần số lượng giao dịch thoại chiều đi trong 3 tháng liên tục.*
- *Có xu hướng giảm dần số lượng giao dịch sms chiều đi trong 3 tháng liên tục.*
- *Có xu hướng giảm dần số lượng giao dịch thoại chiều đến trong 3 tháng liên tục.*
- *Hầu như không phát sinh giao dịch nạp thẻ trong tháng liền trước và có tần suất nạp thẻ giảm dần trong 3 tháng liên tục.*
- *Có xu hướng giảm dần số ngày cập nhật VLR trong 3 tháng liên tục.*

3. Nhóm 3

Nhóm này gồm 219.993 thuê bao, trong đó MobiQ chiếm tỷ trọng cao nhất với 82,16%. MobiQ là loại thuê bao có ngày sử dụng linh hoạt, do vậy rất dễ dàng cho khách hàng duy trì số thuê bao mà hầu như không cần tiêu dùng đến TKC.

Sau khi phân tích nhóm chuyên gia rút ra rằng phần lớn thuê ***bao thuộc nhóm này có thể đang nằm trên kênh hoặc nằm trong ngăn kéo*** *(KH đã mua nhưng không sử dụng nữa). Đặc điểm của nhóm như sau:*

- *Chủ yếu là loại hình thuê bao MobiQ.*
- *Hầu như không đạt 3k3d_vlr trong nhiều tháng liên tục (chỉ có 7,5% thuê bao thuộc nhóm có đạt 3k3d_vlr ít nhất 1 lần trong 7 tháng liên tiếp).*
- *Hầu như không tiêu dùng tài khoản chính trong 3 tháng liên tục (tỉ lệ đạt trên 90%).*

- *Có mức tiêu dùng TKKM thấp < 3.000đ/tháng trong 3 tháng liên tục (tỉ lệ đạt trên 90%).*
- *Hầu như không phát sinh giao dịch thoại chiều đi hoặc chỉ phát sinh dưới 3 giao dịch (chiếm trên 95% trong 3 tháng liên tiếp).*
- *Hầu như không phát sinh giao dịch sms chiều đi hoặc chỉ phát sinh dưới 3 giao dịch (chiếm trên 94% trong 3 tháng liên tiếp).*
- *Hầu như không phát sinh giao dịch thoại chiều đến (chiếm trên 96% trong 3 tháng liên tiếp).*
- *Hầu như không phát sinh hoặc chỉ phát sinh dưới 3 sms chiều đến (chiếm trên 91% trong 3 tháng liên tiếp).*
- *Theo xu hướng giảm sản lượng thoại đến trong 3 tháng liên tiếp (khoảng 97%).*
- *Chỉ mở máy dưới 3 ngày (chiếm trên 91% trong 3 tháng liên tiếp).*
- *Hầu như không nạp thẻ trong tháng trước khi rời mạng (chỉ có 0,44% nạp thẻ).*
- *Theo xu hướng giảm tần suất nạp thẻ trong 3 tháng liên tiếp (khoảng 98% thuê bao thể hiện xu hướng này).*

4. Nhóm 4

Bản chất nhóm này đã rời mạng (không phát sinh cước) từ tháng 7/2015.

Nhóm này gồm có 1.067.414 thuê bao, trong đó chiếm tỉ trọng cao nhất là MobiCard (37,69%), MobiQ (27,30%) và Mobi365 (21,02%). Trong khi đó, xét trong tỉ trọng thuê bao PSC trong tháng 8/2015 của toàn mạng, MobiCard chỉ chiếm 12,23%, Mobi365 chỉ chiếm 2,64% và MobiQ chiếm tới 48,81%.

Sau khi phân tích nhóm chuyên gia có nhận xét về đặc điểm nhận dạng của nhóm thuê bao này như sau:

- *Tiêu dùng TKC ít (dưới 3000đ) trong 3 tháng liên tiếp.*
- *Tiêu dùng TKKM ít (dưới 3000đ) trong 3 tháng liên tiếp.*
- *Có xu hướng tiêu dùng TKC và TKKM giảm dần trong 3 tháng liên tiếp (95% thuê bao thể hiện rõ xu hướng này).*
- *Không phát sinh hoặc chỉ phát sinh dưới 3 giao dịch thoại chiều đi trong 3 tháng liên tiếp.*
- *Không phát sinh hoặc chỉ phát sinh dưới 3 giao dịch sms chiều đi trong 3 tháng liên tiếp.*

- *Có xu hướng giảm dần số lượng giao dịch thoại chiều đi trong 3 tháng liên tiếp (93,26% thuê bao thể hiện xu hướng này).*
- *Có xu hướng giảm dần số lượng giao dịch sms chiều đi trong 3 tháng liên tiếp (93,9% thể hiện xu hướng này).*
- *Có xu hướng giảm dần số lượng giao dịch sms chiều đến trong 3 tháng liên tiếp (91,41% thuê bao thể hiện xu hướng này).*
- *Có xu hướng giảm dần số lượng giao dịch thoại chiều đến trong 3 tháng liên tiếp (91,41% thuê bao thể hiện xu hướng này).*
- *Có xu hướng giảm dần số ngày cập nhật VLR trong 3 tháng liên tiếp (88,94% thể hiện xu hướng này).*
- *Hầu như không nạp thẻ vào tháng liền trước và có xu hướng giảm tần suất nạp thẻ trong 3 tháng liên tiếp.*

5. Nhóm 5

Nhóm này gồm 105.703 thuê bao, trong đó chiếm tỷ trọng cao nhất là MobiQ (56,68%), MobiCard (15,32%) và Zone+ (10,87%). Trong số này, 10,17% thuê bao hội viên chương trình Kết nối dài lâu.

Sau khi phân tích nhóm chuyên gia có nhận xét về đặc điểm nhận dạng của nhóm thuê bao này như sau:

- *Nhóm thuê bao này hầu như vẫn hoạt động rất bình thường trên mạng: Vẫn nạp thẻ, phát sinh các giao dịch thoại đi/đến với tỉ lệ cao, số ngày cập nhật VLR nhiều (≥3 ngày).*
- *Số tiền còn lại trong TKC thấp (trên 60% thuê bao có số tiền trong TKC dưới 3.000đ).*
- *Doanh thu TKC từ dịch vụ sms (nội mạng và liên mạng) thấp (dưới 3.000đ/tháng).*
- *Doanh thu TKKM từ dịch vụ sms (nội mạng và liên mạng) thấp (dưới 3.000đ/tháng).*

Phụ lục 3 Đặc trưng phân lớp của giải pháp MobiFone đang áp dụng

1. **Nhóm 1**

- Tiêu chí phân lớp: Thuê bao trả trước thỏa mãn đồng thời các tiêu chí sau:
    - Tiêu dùng tài khoản chính dưới 3.000đ trong tháng trước.
    - Phát sinh dưới 3 giao dịch thoại chiều đi trong tháng trước.
    - Giảm dần số lượng giao dịch thoại chiều đến trong 3 tháng liên tục liền trước (số giao dịch của tháng sau nhỏ hơn hoặc bằng số giao dịch của tháng trước).
    - Giảm dần số lượng giao dịch thoại chiều đến trong 3 tháng liên tục liền trước.

2. **Nhóm 2:**

- Tiêu chí phân lớp:
    - Tiêu dùng TKC ≥ 3.000đ trong tháng trước.
    - Cập nhật VLR < 3 ngày trong tháng trước.
    - Phát sinh dưới 3 giao dịch thoại chiều đi trong tháng trước.
    - Không phát sinh giao dịch thoại chiều đến trong tháng trước.

3. **Nhóm 3:**

- Tiêu chí phân lớp:
    - Không tiêu dùng TKC trong tháng trước (Tiêu dùng TKC= 0).
    - Tiêu dùng TKKM > 0 trong tháng trước.
    - Không nạp thẻ trong tháng trước.
    - Có tần suất nạp thẻ giảm dần trong 3 tháng liên tục liền trước (số lần nạp thẻ của tháng sau nhỏ hơn hoặc bằng số lần nạp thẻ trong tháng trước).
    - Có xu hướng giảm dần số lượng giao dịch thoại chiều đến (hoặc giảm dần sản lượng gọi chiều đến) trong 3 tháng liên tục liền trước.

4. **Nhóm 4:**

- Tiêu chí phân lớp:
    - Không tiêu dùng TKC trong tháng trước (Tiêu dùng TKC= 0).

- Không tiêu dùng TKKM trong tháng trước (Tiêu dùng TKKM= 0).
- Có số ngày cập nhật VLR giảm dần trong 3 tháng liên tiếp liền trước (số ngày cập nhật VLR tháng sau nhỏ hơn hoặc bằng số ngày cập nhật trong tháng trước).
- Không nạp thẻ trong tháng trước.
- Có tần suất nạp thẻ giảm dần trong 3 tháng liên tục liền trước (số lần nạp thẻ của tháng sau nhỏ hơn hoặc bằng số lần nạp thẻ trong tháng trước).

5. Nhóm 5:

- Tiêu chí phân lớp:
  - Tiêu dùng TKC ≥ 3.000đ trong tháng trước.
  - Cập nhật VLR ≥ 3 ngày trong tháng trước.
  - Có số tiền còn lại trong TKC tại thời điểm cuối tháng trước < 5.000đ.
  - Tiêu dùng TKC < 3.000đ cho dịch vụ sms (nội mạng và liên mạng) trong tháng trước.
  - Tiêu dùng TKKM < 3.000đ cho dịch vụ sms (nội mạng và liên mạng) trong tháng trước