HỌC VIỆN CÔNG NGHỆ BƯU CHÍNH VIỄN THÔNG

NGUYỄN NGỌC THƠ

HỆ HỖ TRỢ QUYẾT ĐỊNH PHÂN NHÓM CÁC TRẠM BTS THEO LƯU LƯỢNG

Chuyên ngành: HỆ THỐNG THÔNG TIN

Mã số: 8.48.01.04

TÓM TẮT LUẬN VĂN THẠC SĨ

(Theo định hướng ứng dụng)

TP. HỒ CHÍ MINH – NĂM 2022

Luận văn được hoàn thành tại:

HỌC VIỆN CÔNG NGHỆ BƯU CHÍNH VIỄN THÔNG

Người hướng dẫn khoa học: **TS. NGUYỄN XUÂN SÂM**

Phản biện 1: .....

Phản biện 2: .....

Luận văn sẽ được bảo vệ trước Hội đồng chấm luận văn 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

1. Lý do chọn đề tài

Trong lĩnh vực dịch vụ Viễn thông, các hoạt động đều gắn liền với việc tiếp nhận và xử lý thông tin, do vậy việc ứng dụng công nghệ thông tin có ý nghĩa quan trọng đối với ngành Viễn thông để phát triển bền vững và có hiệu quả cao. Qua quá trình hoạt động, dữ liệu được tích lũy có kích thước ngày càng lớn, trong nó có thể ẩn chứa nhiều thông tin dạng những quy luật chưa được khám phá. Chính vì vậy, một nhu cầu đặt ra là cần tìm cách biến đổi dữ liệu "thô" thành thông tin phục vụ các công tác dự báo, phân loại nhằm mục đích tư vấn và hỗ trợ công việc kinh doanh.

Công nghệ, kỹ thuật dữ liệu đã, đang và sẽ phát triển mạnh mẽ trước những khao khát tri thức của con người, thu hút sự quan tâm các nhà nghiên cứu trong nhiều lĩnh vực khác nhau như học máy, hệ chuyên gia, thống kê... Nhiều phương pháp kỹ thuật phân lớp đã được đề xuất nhưng không có phương pháp tiếp cận phân loại nào là tối ưu và chính xác hơn hẳn những phương pháp khác.

Với mong muốn nghiên cứu về việc xây dựng một hệ thống hỗ trợ ra quyết định để đánh giá, phân nhóm lưu lượng các trạm NodeB/eNodeB từ dữ liệu mạng Vinaphone Viễn thông Tây Ninh, tôi đã chọn đề tài "***Hệ hỗ trợ quyết định phân nhóm các trạm BTS theo lưu lượng***" làm luận văn tốt nghiệp.

2. Tổng quan vấn đề nghiên cứu

Trong những năm gần đây Học máy (Machine Learning - ML) là một trong những công cụ tiềm năng và hứa hẹn nhất để dự báo một loạt các vấn đề phức tạp. Sự phát triển nhanh chóng của ML tương quan trực tiếp với sự phát triển của công nghệ; sự phát triển nhanh chóng của cộng đồng AI có lợi cho sự phát triển của nhiều thư viện và công cụ mã nguồn mở (ví dụ: TensorFlow, Keras, PyTorch, fast.ai), giúp nhiều nhà nghiên cứu trong việc triển khai và triển khai các thuật toán ML.

Công việc trong luận văn này được thực hiện theo hướng dữ liệu, và nó tập trung vào việc tìm hiểu cách sử dụng và biến đổi dữ liệu này thành thông tin[1] phục vụ mục đích sản xuất kinh doanh trong mạng di động; mô tả đặc điểm lưu lượng truy

cập di động của người dùng, việc sử dụng ứng dụng và các kiểu lưu lượng truy cập của họ. Sau đó, cần phân tích số liệu thống kê về thời gian của mạng để xác định lưu lượng từng khu vực. Việc khai thác một lượng lớn thông tin cho phép cải thiện hiệu suất của chính mạng nhưng cũng để giải quyết một loạt vấn đề (ví dụ: phát hiện bất thường) có thể ảnh hưởng đến cơ sở hạ tầng mạng. Công việc bắt đầu từ việc nghiên cứu các bộ dữ liệu đến từ việc triển khai mạng di động thực tế sau đó quyết định tối ưu hóa mạng và ứng phó với vô số các vấn đề mạng như phân bổ tài nguyên, tiết kiệm năng lượng.

3. Mục tiêu nghiên cứu

Nghiên cứu tổng quan về lưu lượng mạng di động, cơ chế hoạt động cũng như các yếu tố tác động đến lưu lượng mạng.

Nghiên cứu các mô hình và thuật toán học máy hỗ trợ việc phân nhóm trạm BTS theo lưu lượng.

Nghiên cứu về công cụ và ngôn ngữ hỗ trợ việc khai phá dữ liệu (như Google Colab, Python), từ đó cài đặt và sử dụng cho đề tài.

4. Đối tượng và phạm vi nghiên cứu

Đối tượng nghiên cứu: hệ hỗ trợ ra quyết định, thuật toán máy học(Machine learning): Cây quyết định, rừng ngẫu nhiên… trong khai phá dữ liệu.

Phạm vi nghiên cứu: Ứng dụng các thuật toán máy học để phân nhóm các trạm BTS theo lưu lượng. Các biểu mẫu, số liệu liên quan đến việc phân nhóm các trạm BTS: Total traffic, Call setup Success rate. Mẫu dữ liệu là danh sách lưu lượng các trạm BTS của mạng Vinaphone khu vực tỉnh Tây Ninh.

5. Phương pháp nghiên cứu

Đề tài này sử dụng phương pháp nghiên cứu lý thuyết kết hợp với xây dựng ứng dụng thực nghiệm:

- Phương pháp nghiên cứu lý thuyết: Tìm hiểu, phân tích, tổng hợp các tài liệu về hệ hỗ trợ ra quyết định, khai phá dữ liệu và đề xuất cải tiến một số thuật toán máy học nhằm đạt được mục tiêu nghiên cứu. Thu thập, tìm

hiểu, nghiên cứu tài liệu; số liệu mạng di động Vinaphone khu vực tỉnh Tây Ninh.

- Phương pháp nghiên cứu thực nghiệm: Phân tích yêu cầu thực tế của công việc, áp dụng lý thuyết, các thuật toán liên quan để xây dựng hệ hỗ trợ ra quyết định; Xây dựng bộ dữ liệu mẫu dùng để kiểm tra, thử nghiệm chương trình và đưa ra đánh giá kết quả.

6. Cấu trúc luận văn

Ngoài phần mở đầu, mục lục, kết luận và kiến nghị, danh mục hình vẽ, danh mục bảng biểu, tài liệu tham khảo, phụ lục, phần chính của luận văn gồm 4 chương như sau:

Chương 1: TỔNG QUAN LƯU LƯỢNG MẠNG DI ĐỘNG CÁC TRẠM BTS

Chương 2: CƠ SỞ LÝ THUYẾT VÀ CÁC NGHIÊN CỨU LIÊN QUAN

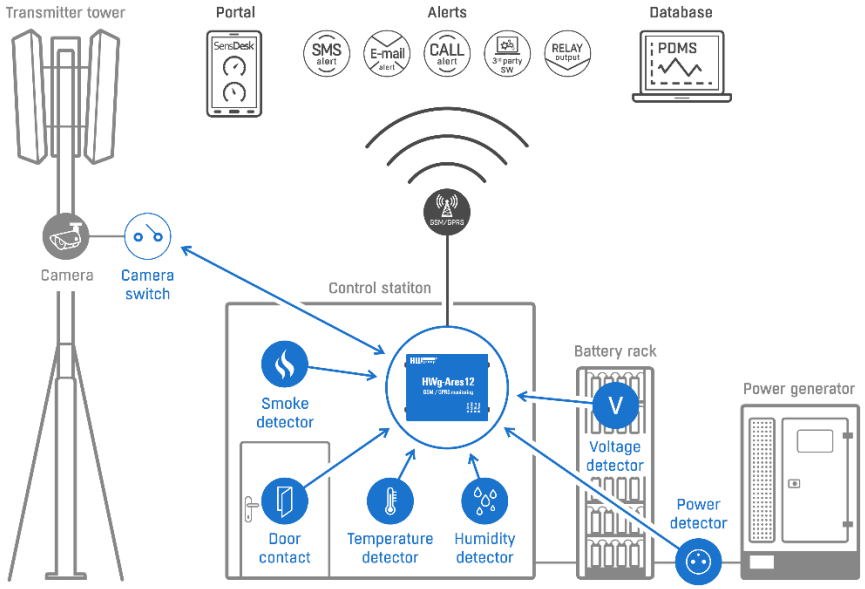
Chương 3: ĐÁNH GIÁ ĐỀ XUẤT VÀ TRIỂN KHAI ỨNG DỤNG

Chương 4: THỰC NGHIỆM TRÊN MÔI TRƯỜNG GOOGLE COLAB VÀ ĐÁNH GIÁ KẾT QUẢ

Tóm tắt luận văn

CHƯƠNG 1. TỔNG QUAN VỀ LƯU LƯỢNG MẠNG DI ĐỘNG CÁC TRẠM BTS

1.1 Giới thiệu mô hình tổng quát



Hình 1.1: Cấu tạo trạm BTS

Trạm thu phát gốc (BTS) là một thiết bị hỗ trợ giao tiếp không dây giữa thiết bị người dùng (UE) và mạng. UE là các thiết bị như điện thoại di động (thiết bị cầm tay), điện thoại WLL, máy tính có kết nối Internet không dây. Mạng có thể là mạng của bất kỳ công nghệ truyền thông không dây nào như GSM, CDMA, vòng lặp cục bộ không dây, Wi-Fi, WiMAX hoặc công nghệ mạng diện rộng (WAN) khác. BTS còn được gọi là nút B (trong mạng 3G) hay đơn giản hơn là trạm gốc (BS). Để thảo luận về tiêu chuẩn LTE, chữ viết tắt eNB cho nút phát triển B được sử dụng rộng rãi và GNodeB cho 5G.

Mặc dù thuật ngữ BTS có thể áp dụng cho bất kỳ tiêu chuẩn truyền thông không dây nào, nhưng nó thường được kết hợp với các công nghệ thông tin di động như GSM và CDMA. Về vấn đề này, BTS là một phần của sự phát triển của hệ thống con trạm gốc (BSS) để quản lý hệ thống. Nó cũng có thể có thiết bị để mã hóa và giải mã thông tin liên lạc, các công cụ lọc phổ (bộ lọc băng thông), v.v. Anten cũng có thể được coi là thành phần của BTS theo nghĩa chung vì chúng tạo điều kiện thuận lợi cho hoạt động của BTS. Thông thường, một trạm BTS sẽ có một số bộ thu phát

(TRX) cho phép nó phục vụ một số tần số khác nhau và các cung khác nhau của tế bào (trong trường hợp các trạm gốc được phân chia). Một BTS được điều khiển bởi bộ điều khiển trạm gốc thông qua chức năng điều khiển trạm gốc (BCF). BCF được thực hiện như một đơn vị rời rạc hoặc thậm chí được kết hợp trong TRX trong các trạm gốc nhỏ gọn. BCF cung cấp kết nối vận hành và bảo trì (O&M) với hệ thống quản lý mạng (NMS), đồng thời quản lý các trạng thái hoạt động của từng TRX, cũng như xử lý phần mềm và thu thập cảnh báo. Cấu trúc và chức năng cơ bản của trạm BTS vẫn giữ nguyên bất kể công nghệ không dây nào.

Một trạm BTS cơ bản bao gồm:

- Một trạm thu phát (TRX) có nhiệm vụ truyền và nhận tín hiệu, gửi và nhận các tín hiệu từ các phần tử mạng cao hơn;
- Một bộ tổ hợp sẽ kết hợp nguồn cấp dữ liệu từ một số trạm thu phát để được gửi đi thông qua một ăng-ten duy nhất do đó làm giảm số lượng ăng-ten cần cài đặt;
- Một bộ khuếch đại công suất giúp khuếch đại tín hiệu từ trạm thu phát để truyền thông tin qua ăng-ten;

Một bộ song công được sử dụng để tách việc gửi và nhận tín hiệu từ các ăng-ten hoặc từ một ăng-ten là một phần bên ngoài của BTS.

1.2 Cơ chế vận hành mạng

Các thiết bị di động của người dùng truy cập Internet sẽ đưa yêu cầu đến các trạm thu phát sóng di động(BTS). Sau đó các trạm BTS tập trung về thiết bị RNC vào mạng Core VNPT ra IntraNet.

Từ đó người quản lý có thể thống kê lưu lượng các trạm BTS qua mạng Intranet để thống kê lưu lượng hàng ngày của trạm thu phát gốc đó.

1.3 Tổng quan về lưu lượng mạng

1.3.1 Giới thiệu về lưu lượng mạng

Lưu lượng mạng di động hoặc mạng di động là mạng truyền thông trong đó liên kết đến và đi từ các nút cuối là không dây. Mạng được phân phối trên các vùng đất được gọi là cell (tạm dịch là tế bào), mỗi vùng được phục vụ bởi ít nhất một bộ thu phát vị trí cố định (thường là ba điểm di động hoặc trạm thu phát cơ

sở). Các trạm gốc này cung cấp cho tế bào phạm vi phủ sóng mạng có thể được sử dụng để truyền thoại, dữ liệu và các loại nội dung khác. Một tế bào thường sử dụng một tập hợp tần số khác với các lưu lượng lân cận, để tránh nhiễu và cung cấp chất lượng dịch vụ đảm bảo trong mỗi lưu lượng. Khi kết hợp với nhau, các tế bào này cung cấp vùng phủ sóng vô tuyến trên một khu vực địa lý rộng. Điều này cho phép nhiều bộ thu phát di động (ví dụ: điện thoại di động, máy tính bảng và máy tính xách tay được trang bị modem băng thông rộng di động, máy nhắn tin, v.v.) giao tiếp với nhau và với các bộ thu phát và điện thoại cố định ở bất kỳ đâu trong mạng, thông qua các trạm gốc, ngay cả khi một số máy thu phát đang di chuyển qua nhiều tế bào trong quá trình truyền.

Vùng phủ sóng lớn hơn so với một máy phát trên mặt đất, vì các tháp di động bổ sung có thể được thêm vào vô thời hạn và không bị giới hạn bởi đường chân trời.

1.3.2 Lịch sử mạng di động

Mạng di động thương mại đầu tiên, thế hệ 1G, được Nippon Telegraph and Telephone (NTT) ra mắt tại Nhật Bản vào năm 1979, ban đầu ở khu vực thủ đô Tokyo. Trong vòng 5 năm, mạng NTT đã được mở rộng đến toàn bộ dân số Nhật Bản và trở thành mạng 1G đầu tiên trên toàn quốc. Đó là một mạng không dây tương tự. Hệ thống Bell đã phát triển công nghệ di động từ năm 1947 và có mạng di động hoạt động ở Chicago và Dallas trước năm 1979, nhưng dịch vụ thương mại đã bị trì hoãn do sự tan rã của Hệ thống Bell, với các tài sản di động được chuyển giao cho các Công ty điều hành Bell khu vực.

Cuộc cách mạng không dây bắt đầu vào đầu những năm 1990, dẫn đến sự chuyển đổi từ mạng tương tự sang kỹ thuật số. Điều này đã được kích hoạt bởi những tiến bộ trong công nghệ MOSFET. MOSFET, ban đầu được phát minh bởi Mohamed M. Atalla và Dawon Kahng tại Bell Labs vào năm 1959, đã được điều chỉnh cho các mạng di động vào đầu những năm 1990, với việc áp dụng rộng rãi MOSFET công suất, LDMOS (bộ khuếch đại RF), và Thiết bị RF CMOS (mạch RF) dẫn đến sự phát triển và phổ biến của mạng di động không dây kỹ thuật số.

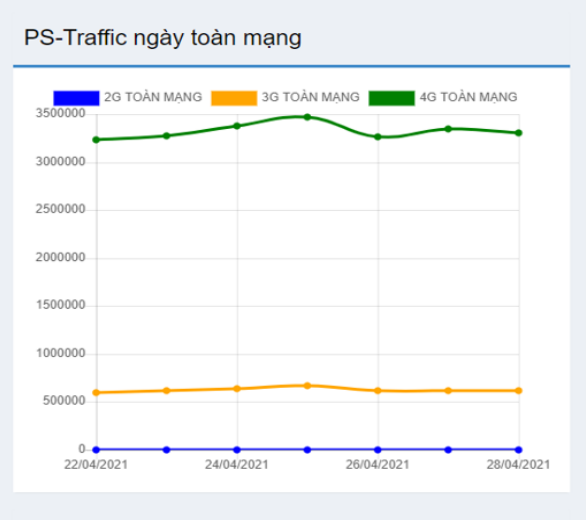
Mạng di động kỹ thuật số thương mại đầu tiên, thế hệ 2G, được ra mắt vào năm 1991. Điều này đã gây ra sự cạnh tranh trong lĩnh vực này khi các nhà khai thác mới thách thức các nhà khai thác mạng tương tự 1G đương nhiệm.

1.3.3 Các yếu tố ảnh hưởng đến lưu lượng mạng

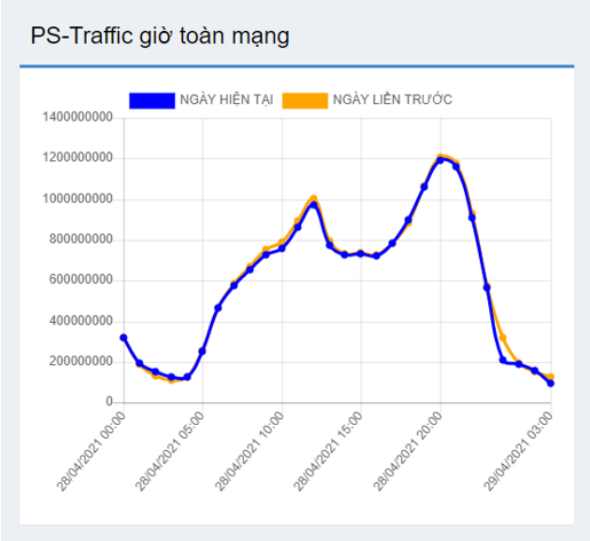
Có rất nhiều yếu tố có thể gây ảnh hưởng đến lưu lượng mạng trong quá trình sử dụng. Một số trong những yếu tố này không thể tránh được và phải có biện pháp để cố gắng giảm thiểu các ảnh hưởng tiêu cực mà chúng tác động lên hiệu suất mạng, tuy nhiên một số yếu tố khác có thể được khắc phục hoàn toàn qua việc nâng cấp thiết bị hay quy hoạch mạng lưới tốt.

1.4 Khảo sát hệ thống nguồn số liệu

Trước khi phân tích, thiết kế và xây dựng hệ thống trợ giúp quyết định, cần chuẩn bị hạ tầng kĩ thuật và tư liệu cho hệ thống:



Hình 1.2: Thống kê lưu lượng theo ngày



Hình 1.3: Thống kê lưu lượng theo giờ

1.4.1 Chuẩn bị dữ liệu

Trong giai đoạn chuẩn bị dữ liệu cần phân tích, thiết kế và xây dựng cơ sở dữ liệu về các Cell của trạm. Cơ sở dữ liệu này được xem như cơ sở dữ liệu về lưu lượng. Trong thời gian đầu, cơ sở dữ liệu lưu lượng Cell có ý nghĩa đối với bài toán thống kê, chưa thực sự giúp cho người quản lý phân nhóm theo các trạm theo lưu lượng.

1.4.2 Nhu cầu về ra quyết định

Trên hệ thống thông tin với cơ sở dữ liệu lưu lượng đã được giai đoạn 1 tạo nên, người quản lý cần ra các quyết định đầu tư thêm trạm hay tối ưu lưu lượng. Việc ra quyết định chính là công tác của nhà quản lý tối ưu trạm. Hệ thống trợ giúp quyết định cho phép thực hiện các trợ giúp người quản lý ra quyết định. Các trợ giúp có ý nghĩa cùng nhà quản lí đưa ra quyết định cuối cùng. Trong trường hợp này, phần mềm máy tính là công cụ giúp cho con người ra quyết định quản lí.

1.5 Kết luận chương

Chương 1 đã trình bày tổng quan về các vấn đề nghiên cứu như lưu lượng mạng di động cũng như các yếu tố gây ảnh hưởng đến lưu lượng và chất lượng dịch vụ mạng di động. Dựa vào cơ chế vận hành mạng, bộ dữ liệu về lưu lượng từ một nhà mạng ở Việt Nam được thu thập để có thể thực hiện các mục tiêu mà luận văn đã đề ra.

CHƯƠNG 2. CƠ SỞ LÝ THUYẾT VÀ CÁC NGHIÊN CỨU LIÊN QUAN

2.1 Giới thiệu học máy

Ngày nay, Machine Learning (ML): Học máy đã có mặt trong cuộc sống hàng ngày của chúng ta và là một phần thiết yếu của nhiều sản phẩm và dịch vụ mà chúng ta sử dụng thường xuyên. Các công ty sử dụng Học máy để tạo ra các dịch vụ mới tuyệt vời, làm cho các sản phẩm và dịch vụ hiện có của họ tốt hơn và giải quyết một loạt các vấn đề kinh doanh. Khi các công ty nhanh chóng sử dụng Học máy để làm lợi thế của họ, họ tập trung phần lớn nỗ lực chuyển đổi và ngân sách vào việc sử dụng các công nghệ này để kích hoạt tăng trưởng.

2.2 Độ đo đánh giá mô hình

2.2.1 Độ chính xác

Accuracy (độ chính xác) là chỉ số đánh giá thường được sử dụng để đánh giá độ chính xác của mô hình dự đoán. Độ chính xác là tỉ lệ giữa số điểm dữ liệu được dự đoán đúng và tổng số điểm dữ liệu. Nếu $\hat{y}_i$ là giá trị dự đoán của mẫu thứ $i-th$ và y_i là giá trị thực tương ứng, thì phần dự đoán độ chính xác trên các ví dụ được định nghĩa là

$$accuracy(y, \hat{y}) = \frac{1}{n_{samples}} \sum_{i=0}^{n_{samples}-1} 1(\hat{y}_i = y_i) \qquad (2.1)$$

Trong đó 1(x) là hàm chỉ thị

2.2.2 Độ đo mất mát

Độ chính xác của dự báo là một thước đo, thể hiện hiệu suất của mô hình dự báo. Nó là một giá trị ngược lại với độ đo của sai số dự báo. Có nhiều lựa chọn cũng như cách tính toán cho độ đo sai số dự báo. Mỗi một độ đo thể hiện một chút thông tin khác nhau và nó được biểu thị bằng độ lệch của giá trị dự đoán và giá trị thực tế. Một vài độ đo sai số thường được sử dụng trong các bài toán dự báo:

Mean absolute percentage error (MAPE)

Mean absolute percentage error (MAPE) thường được sử dụng như một hàm tổn thất cho các bài toán hồi quy và trong đánh giá mô hình, vì cách giải thích rất trực quan về sai số tương đối

$$\text{MAPE} = \frac{1}{N}\sum_{t=1}^{N}\frac{|y(t)-\hat{y}(t)|}{y(t)}.100\% \quad (2.2)$$

Root Mean squared error (RMSE)

Root Mean Square Error (RMSE) là độ lệch chuẩn của các phần dư (sai số dự đoán). Phần dư là thước đo khoảng cách từ các điểm dữ liệu đường hồi quy; RMSE là thước đo mức độ lan tỏa của những phần dư này. Nói cách khác, nó cho bạn biết mức độ tập trung của dữ liệu xung quanh dòng phù hợp nhất. Sai số bình phương trung bình gốc thường được sử dụng trong khí hậu học, dự báo và phân tích hồi quy để xác minh kết quả thực nghiệm.

$$\text{RMSE} = \sqrt{\frac{1}{N}\sum_{t=1}^{N}(y(t)-\hat{y}(t))^2} \quad (2.3)$$

Mean square error (MSE)

MSE là tổn thất bình phương trung bình cho mỗi ví dụ trên toàn bộ tập dữ liệu. Để tính toán MSE, hãy tính tổng tất cả các tổn thất bình phương cho các mẫu riêng lẻ và sau đó chia cho số lượng, ví dụ

$$\text{MSE} = \frac{1}{N}\sum_{(x,y)\in D}^{N}(y-\hat{y}(x))^2 \quad (2.4)$$

Mean absolute error (MAE)

Trong thống kê, sai số tuyệt đối trung bình (MAE) là một thước đo sai số giữa các quan sát được ghép nối biểu hiện cùng một hiện tượng. Ví dụ về Y so với X bao gồm so sánh dự đoán so với quan sát, thời gian tiếp theo so với thời điểm ban đầu và một kỹ thuật đo lường so với một kỹ thuật đo lường thay thế. MAE được tính như sau:

$$\text{MAE} = \frac{1}{N}\sum_{t=1}^{N}|y(t)-\hat{y}(t)| \quad (2.5)$$

Sum of squared errors (SSE)

SSE là tổng của sự khác biệt bình phương giữa mỗi quan sát và trung bình của nhóm của nó. Nó có thể được sử dụng như một thước đo sự thay đổi trong một cụm. Nếu tất cả các trường hợp trong một cụm đều giống nhau thì SSE sẽ bằng 0.

$$\text{SSE} = \sum_{t=1}^{N}(y(t) - \hat{y}(t))^2 \quad (2.6)$$

Logloss

Đây là hàm mất mát được sử dụng trong hồi quy logistic (đa thức) và các phần mở rộng của nó, chẳng hạn như mạng nơ-ron, được định nghĩa là khả năng log âm của một mô hình logistic trả về xác suất y_pred cho dữ liệu huấn luyện y_true của nó. Mất nhật ký chỉ được xác định cho hai hoặc nhiều nhãn. Đối với một mẫu đơn có nhãn đúng $y \in \{0,1\}$ và ước lượng xác suất $p = \Pr(y = 1)$, công thức logloss là:

$$L_{\log}(y, p) = -(y \log(p) + (1 - y) \log(1 - p)) \quad (2.7)$$

2.3. Công trình liên quan

- Merima Kulin, Tarik Kazaz, Eli De Poorter, Ingrid Moerman, "A Survey on Machine Learning-Based Performance Improvement of Wireless Networks: PHY, MAC and Network Layer," 29 January 2021.

- Fengli Xu, Yong Li, Senior Member, IEEE, Huandong Wang, Pengyu Zhang, and Depeng Jin, Member, IEEE, "Understanding Mobile Traffic Patterns of Large Scale Cellular Towers in Urban Environment," 2016.

- Hoang Duy Trinh, Angel Fernandez Gambiny, Lorenza Giupponi, Michele Rossiy and Paolo Dini, "Mobile Traffic Classification through Physical Control Channel Fingerprinting: a Deep Learning Approach," 2020.

- Sébastien Dujardin, Damien Jacques, Jessica Steele and Catherine Linard, "Mobile Phone Data for Urban Climate Change Adaptation: Reviewing Applications, Opportunities and Key Challenges," 11 December 2020.

- P. Muñoz, R. Barco, E. Cruz, A. Gómez-Andrades, E. J. Khatib1 and N. Faour, "A method for identifying faulty cells using a classification tree-based UE diagnosis in LTE," 2017.

- Hoang Duy Trinh, Nicola Bui, Joerg Widmer, Lorenza Giupponi, Paolo Dini, "Analysis and Modeling of Mobile Traffic Using Real Traces," 2017.

- Leo Tisljaric, Dominik Cvetek, Martin Gregurić, Zuzanna Kurowska, "Classification of Travel Modes from Cellular Network Data Using Machine Learning Algorithms," October 2021.

- Yan Sun, Chengxi Liu, and Chen Zhang, "Mobile Technology and Studies on Transport Behavior: Literature Analysis, Integrated Research Model, and Future Research Agenda," 25 October 2021.

- Hoang Duy Trinh, Lorenza Giupponi and Paolo Dini, "Urban Anomaly Detection by processing Mobile Traffic Traces with LSTM Neural Networks," 2019.

- Dehai Zhang, Linan Liu, Cheng Xie, Bing Yang and Qing Liu, "Citywide Cellular Traffic Prediction Based on a Hybrid Spatiotemporal Network," 8 January 2020.

- Shuai Zhao, Xiaopeng Jiang, Guy Jacobson, Rittwik Jana, Wen-Ling Hsu, Raif Rustamov, Manoop Talasila, Syed Anwar Aftab, Yi Chen, Cristian Borcea, "Cellular Network Traffic Prediction Incorporating Handover: A Graph Convolutional Approach," in 2020 17th Annual IEEE International Conference on Sensing, Communication, and Networking (SECON), 2020.

- Razan M. AlZoman, Mohammed J. F. Alenazi , "A Comparative Study of Traffic Classification Techniques for Smart City Networks," 08 July 2020.

- Yi Zhao, Jianbo Li, Xin Miao, Xuan Ding, "Urban Crowd Flow Forecasting Based on Cellular Network," 19 May 2019.

- QINGTIAN ZENG, QIANG SUN, GENG CHEN, HUA DUAN, CHAO LI, AND GE SONG, "Traffic Prediction of Wireless Cellular Networks Based on Deep Transfer Learning and Cross-Domain Data," 18 Sep 2020.

- Amin Azari, Fateme Salehi, Panagiotis Papapetrouy, Cicek Cavdar, "Energy and Resource Efficiency by User Traffic Prediction and Classification in Cellular Networks," 02 Nov 2021.

- Carolina Gijón, Matías Toril, Marta Solera, Salvador Luna-Ramírez, Luis Roberto Jiménez, "Encrypted Traffic Classification Based on Unsupervised Learning in Cellular Radio Access Networks," vol. 8, 09 Sep 2020.

- Amin Azari, Panagiotis Papapetrou, Stojan Denic, and Gunnar Peters, "Cellular Traffic Prediction and Classification: a comparative evaluation of LSTM and ARIMA," 03 Jun 2019.

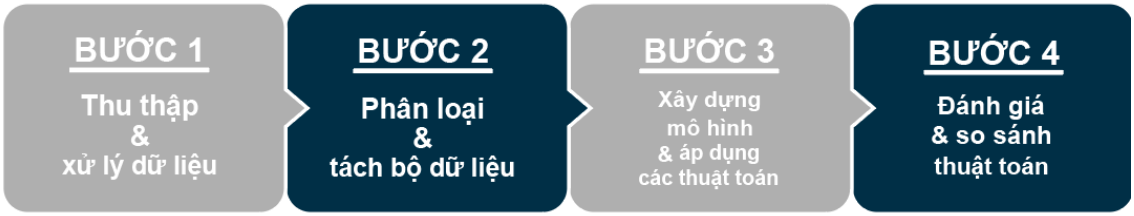
CHƯƠNG 3. ĐÁNH GIÁ ĐỀ XUẤT VÀ TRIỂN KHAI ỨNG DỤNG

3.1. Mô hình nghiên cứu

Luận văn này sử dụng mô hình Decision Forest (DF), là mô hình từ nền tảng mã nguồn mở dành cho việc xây dựng mô hình học máy – Tensorflow. DF gồm tập hợp các thuật toán ML hiện đại để giải quyết các bài toán phân lớp có giám sát (supervised classification), hồi quy (regression) và xếp hạng (ranking). Các thuật toán được sử dụng phổ biến nhất trong tập hợp DF là Random Forests (RF) và Gradient Boosted Decision Trees. Hai thuật toán trên đều là các thuật toán kết hợp sử dụng nhiều "cây quyết định" (decision trees), tuy nhiên mỗi thuật toán có các kĩ thuật thực hiện riêng.

Các bước xây dựng và đề xuất mô hình phân nhóm các trạm BTS dựa trên lưu lượng gồm:

- Bước 1: Thu thập, xử lý và làm sạch dữ liệu lưu lượng mạng di động.
- Bước 2: Phân loại nhãn đại diện cho bốn trạm A, B, C, D dựa trên trường thông tin về lưu lượng tải lên Traffic_Volume_UL_GB sau đó tiến hành tách bộ dữ liệu thành các tập training và testing với tỉ lệ 70%, 30% tương ứng.
- Bước 3: Áp dụng lần lượt từng thuật toán Random Forest, Gradient Boosted Decision Trees vào mô hình.
- Bước 4: Tiến hành chạy mô hình nhiều lần với hai thuật toán, sau đó so sánh và đánh giá kết quả dựa trên các độ đo đánh giá hiệu quả mô hình như độ chính xác, độ mất mát.



Hình 3.1: Các bước thực nghiệm

3.2 Thuật toán RandomForest và Gradient Boosted Decision Trees

3.2.1 Random Forest (RF)

RF [9] là một trong các thuật toán học có giám sát, thường được sử dụng cho các bài toán về phân lớp (classfication) và hồi quy (regression) và đồng thời được sử dụng để dự đoán cho các mô hình và kỹ thuật học máy, hay nói cách khác, RF là tập hợp của thuật toán Decision Tree (DF). Nó là một phần mở rộng của tập hợp bootstrap (đóng gói - bagging) các cây quyết định và có thể được sử dụng cho các bài toán phân loại và hồi quy. Trong bagging, một số cây quyết định được tạo trong đó mỗi cây được tạo từ một mẫu bootstrap khác nhau của tập dữ liệu huấn luyện. Mẫu bootstrap là một mẫu của tập dữ liệu đào tạo trong đó một mẫu có thể xuất hiện nhiều lần trong mẫu, được gọi là lấy mẫu có thay thế.

3.2.2 Gradient Boosted Decision Trees (GBDT) [14]

Cây quyết định được tăng cường độ dốc là một kỹ thuật máy học để tối ưu hóa giá trị dự đoán của một mô hình thông qua các bước liên tiếp trong quá trình học tập. Mỗi lần lặp lại của cây quyết định liên quan đến việc điều chỉnh các giá trị của hệ số, trọng số hoặc độ lệch được áp dụng cho từng biến đầu vào được sử dụng để dự đoán giá trị mục tiêu, với mục tiêu giảm thiểu hàm mất mát (thước đo chênh lệch giữa giá trị được dự đoán và giá trị mục tiêu thực tế). Gradient là sự điều chỉnh gia tăng được thực hiện trong mỗi bước của quy trình; boost là một phương pháp đẩy nhanh việc cải thiện độ chính xác của dự đoán đến một giá trị đủ tối ưu.

Giống như các phương pháp thúc đẩy khác, tăng cường độ dốc kết hợp những "người học" yếu thành một người học mạnh duy nhất theo kiểu lặp đi lặp lại. Điều này dễ giải thích nhất trong cài đặt hồi quy bình phương nhỏ nhất, trong đó mục tiêu là "dạy" một mô hình F để dự đoán các giá trị của biểu mẫu $\hat{y} = F(x)$ bằng cách giảm thiểu sai số bình phương trung bình $\frac{1}{n}\sum(\hat{y_i} - y_i)^2$ trong đó *i* lập chỉ mục trên một số tập hợp kích thước đào tạo *n* các giá trị thực của biến đầu ra *y*

- $\hat{y_i}$: *giá trị dự đoán* $F(x)$
- y_i: giá trị quan sát được
- n: số lượng mẫu trong y

3.3 Kết luận chương

Chương này đã đề xuất các bước xây dựng mô hình Decision Forest và các bước nghiên cứu của đề tài. Trong đó, các thuật toán được sử dụng cho đề tài gồm có Random Forest và Gradient Boosted Decision Tree. Trong chương tiếp theo, luận văn sẽ trình bày quá trình xây dựng mô hình và thực nghiệm trên môi trường Google Colaboratory với bộ dữ liệu được lấy từ một nhà mạng ở Việt Nam.

CHƯƠNG 4. THỰC NGHIỆM TRÊN MÔI TRƯỜNG GOOGLE COLAB VÀ ĐÁNH GIÁ KẾT QUẢ

4.1 Cài đặt môi trường

Môi trường thực nghiệm sử dụng Google Colab và bộ thư viện hỗ trợ các thuật toán học máy là Tensorflow. Ngoài ra một số thư viện hỗ trợ tính toán khác của python được liệt kê như sau: Pandas, Numpy

4.2 Dữ liệu thực nghiệm

4.2.1 Giới thiệu bộ dữ liệu

Bộ dữ liệu về lưu lượng mạng có tổng cộng 24 trường và 1000 dòng được dùng trong thực nghiệm để đánh giá hiệu quả của mô hình sử dụng thuật toán Random Forest. Trong đó, các trường dữ liệu liên quan đến lưu lượng như Traffic_Volume_UL_GB, Traffic_Volumn_DL_GB,… được sử dụng để đánh trọng số và lấy nhãn phục vụ cho mô hình. Thông tin về bộ dữ liệu được rút gọn một số trường và mô tả chi tiết trong bảng 4.1.

Bảng 4.1: Tập dữ liệu lưu lượng mạng

TT	**Tên viết tắt**	**Tên gốc**	**Ý nghĩa**
1	IRHS	Inter_RAT_HO_SR	Tỉ lệ chuyển giao sang mạng di động khác thành công
2	HSRP	Handover_Success_Rate_via_Per	Tỉ lệ chuyển giao di động thành công
3	UDATK	User_Downlink_Average_Throughput_Kbps	Thông lượng trung bình của đường xuống của người dùng Kbps
4	TVU	Traffic_Volume_UL_GB	Lưu lượng đường lên(GB)
5	TVD	Traffic_Volumn_DL_GB	Lưu lượng đường xuống(GB)
6	CellUpMax	Cell_PDCP_Uplink_Max_Throughput	Thông lượng tối đa của đường lên Cell_PDCP

7	EUTRAN	EUTRAN_Initial_Context_Setup_Success_Ratio_being_Subject_for_CS_Fallback_Per	EUTRAN Thiết lập ban đầu Tỷ lệ thành công là Đối tượng cho CS Dự phòng
8	CellDownAvg	Cell_PDCP_Downlink_Average_Throughput	Thông lượng trung bình đường xuống của cell PDCP
9	IRHPSR	Inter_RAT_HO_Preparation_Success_Ratio	Tỷ lệ chuyển giao Fallback về mạng 2G/3G thành công
10	IRTHS	Inter_RAT_Total_HO_SR	Tỉ lệ cuộc gọi chuyển giao sang công công nghệ vô tuyến từ eNodeB(4G) sang 3G thành công
11	IeHS	Intra_eNB_HO_SR_total	Tỉ lệ cuộc gọi chuyển giao 4G thành công
12	UUAT	User_Uplink_Average_Throughput_Kbps	Thông lượng trung bình của đường lên PDCP của tế bào
13	CellUpAvg	Cell_PDCP_Uplink Average_Throughput	Thông lượng trung bình của đường lên Cell PDCP
14	IRHL	Inter_RAT_HOSR_LTE_to_WCDMA_Per	Tỉ lệ cuộc gọi chuyển giao sang công công nghệ vô tuyến từ eNodeB(4G) sang 3G thành công
15	TDTV	Total_Data_Traffic_Volume_GB	Tổng khối lượng lưu lượng dữ liệu GB
16	Downlink Latency	Downlink_Latency	Độ trễ đường xuống
17	CellDownMax	Cell_PDCP_Downlink_Max_Throughput	Thông lượng tối đa của đường xuống của Cell PDCP

4.2.2 Xử lý dữ liệu

Bộ dữ liệu trước khi được đưa vào mô hình để huấn luyện cần trải quá các bước làm sạch dữ liệu, bao gồm việc rút trích và chọn ra các trường dữ liệu cần thiết,

thay thế các ô dữ liệu rỗng hoặc có giá trị gây nhiễu. Đối với các mô hình học máy khác, việc chuẩn hóa dữ liệu sẽ hỗ trợ cho quá trình huấn luyện và mang lại kết quả tốt và khả quan hơn, tuy nhiên việc chuẩn hóa dữ liệu không yêu cầu đối với mô hình sử dụng thuật toán Random Forest. Dữ liệu sau quá trình xử lí sẽ giảm được các trường dữ liệu và chỉ giữ lại những trường liên quan đến đề tài nghiên cứu. Thông tin tóm tắt bộ bộ dữ liệu được mô tả trong bảng sau:

Bảng 4.2: Thông tin tóm tắt bộ dữ liệu

	mean	**std**	**min**	**25%**	**50%**	**75%**	**max**
IRATHO_SR	87.92	30.85	0.00	97.96	99.82	100.00	100.00
HSRate_via_Per	98.50	8.65	0.00	99.36	99.79	99.94	100.00
UDAT_Kbps	30821.87	8696.75	0.00	25535.90	31321.61	36454.98	64943.29
TraVol_UL_GB	2.24	2.90	0.00	0.71	1.42	2.68	38.01
TraVol_DL_GB	26.08	26.90	0.00	9.86	18.74	31.84	246.75
CMax_Throughput	31922.03	18243.68	0.00	16298.50	31392.50	46583.75	69771.00
EUTRAN	99.20	8.91	0.00	100.00	100.00	100.00	100.00
CDown_Avg_Throughput	20.43	4.61	0.00	17.56	20.47	23.21	39.62
IRHPS_Ratio	88.61	30.77	0.00	99.34	100.00	100.00	100.00
IRTHS	87.15	30.82	0.00	96.46	99.35	100.00	100.00
IeHS_total	96.68	17.37	0.00	99.94	100.00	100.00	100.00
UUAT_Kbps	2414.84	945.76	0.00	1718.90	2392.82	3067.18	10840.40
CUp_Avg_Throughput	2.05	0.88	0.00	1.41	2.00	2.62	9.43
IRHL_toWPer	87.92	30.85	0.00	97.96	99.82	100.00	100.00
TDTV_GB	28.32	29.58	0.00	10.57	20.19	34.61	284.76
Downlink_Latency	21.18	12.00	0.00	15.77	18.61	23.01	169.26

CPDMax_Throughput	97.44	27.50	0.00	81.28	97.52	113.84	195.32
IFHPer	99.20	4.81	0.00	99.49	99.81	99.94	100.00
SD_all_Service	0.18	0.43	0.00	0.07	0.12	0.19	10.16
eSSRas_Per	99.80	3.19	0.00	99.91	99.96	99.98	100.00
RCESR_All_Service	99.83	3.16	0.00	99.92	99.98	100.00	100.02
CSSRC_Per	99.73	3.19	0.00	99.83	99.93	99.97	100.00
INTRA_HOSR_ATT	497.96	731.72	0.00	112.00	286.50	571.25	9784.00
RBURD_Per	6.75	8.80	0.00	2.47	4.30	7.54	79.61

Dựa trên thông tin tóm tắt dữ liệu từ bảng 4.2, tiến hành chọn ra trường dữ liệu quan trọng và liên quan để đánh nhãn, sau đó xây dựng, phân tích và đánh giá hiệu quả mô hình sử dụng.

4.3 Kết quả thực nghiệm

4.3.1 Xây dựng tập train và test cho mô hình

Trước khi xây dựng tập dữ liệu train và test, mô hình cần phải chọn nhãn phù hợp để huấn luyện mô hình. Nhãn sử dụng cho mô hình cần phải qua bước chuyển đổi kiểu dữ liệu về kiểu số nguyên cho phù hợp với mô hình. Dữ liệu thực nghiệm gồm 24 đặc trưng sẽ được chia thành hai tập là dữ liệu huấn luyện (training data), và dữ liệu thử nghiệm (testing data), trong đó dữ liệu huấn luyện chiếm 70% và còn lại là dữ liệu thử nghiệm.

4.3.2 Xây dựng mô hình và đánh giá

Bài toán phân loại trạm BTS dựa trên lưu lượng được mô tả như sau:

Dữ liệu đầu vào là tập dữ liệu huấn luyện của mô hình, trong đó có 70% từ tập dữ liệu gốc (701 dòng dữ liệu) với 24 đặc trưng khác nhau. Trong 24 loại đặc trưng, không có đặc trưng đầu vào nào được chỉ định. Do đó, tất cả các cột sẽ được sử dụng làm đặc điểm đầu vào ngoại trừ nhãn. Đặc trưng được sử dụng bởi mô hình được hiển

thị trong lịch sử huấn luyện (training logs) và trong bản tóm tắt mô hình (model.summary).

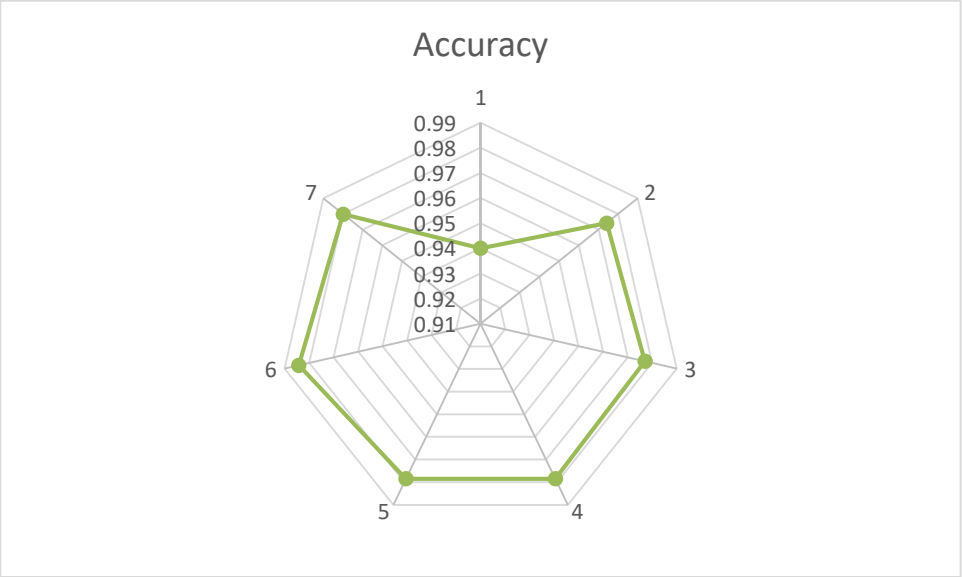
Mô hình DF sử dụng các đặc trưng dạng số, đặc trưng phân loại nguyên bản và các giá trị bị thiếu (missing-values). Các đặc trưng số không cần phải được chuẩn hóa. Các giá trị chuỗi phân loại không cần được mã hóa.

Tính hiệu quả của mô hình được đánh giá dựa trên độ chính xác (accuracy) và độ mất mát (loss). Đối với accuracy, mô hình có hiệu năng tốt khi giá trị càng gần 1 và ngược lại khi giá trị gần về 0 thì khả năng dự đoán của mô hình chưa tốt. Tương tự như vậy, độ mất mát của mô hình đại diện cho sự dự đoán chuẩn xác của mô hình, dự đoán càng chính xác khi giá trị càng gần 0 và ngược lại. Với số lượng cây thay đổi lần lượt K = {1, 51, 151, 201, 251, 300} và độ chính xác cũng như độ mất mát được lấy trung bình qua 5 lần chạy, kết quả được liệt kê như sau:

Bảng 4.3: Kết quả chạy mô hình với thuật toán RF

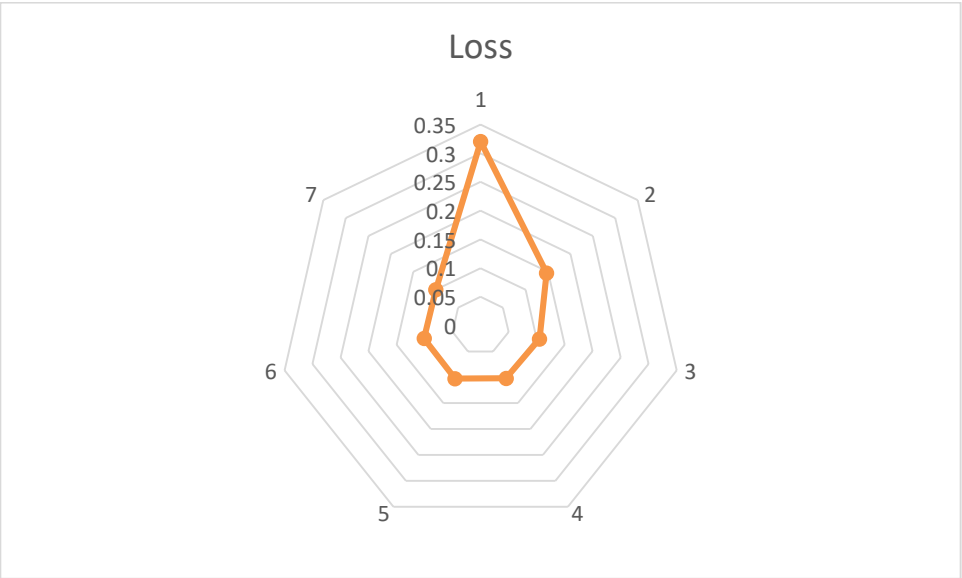
STT	Số cây	Độ chính xác (Accuracy)	Độ mất mát (Loss)
1	1	0.94	0.320204
2	51	0.974212	0.146839
3	101	0.977077	0.105082
4	151	0.97851	0.101884
5	201	0.97851	0.10259
6	251	0.984241	0.101035
7	300	0.979943	0.09969

Dựa vào bảng 4.3, ta thấy qua mỗi lần thay đổi cây số lượng cây, mô hình RF cho kết quả với độ chính xác cao ở lần thực nghiệm đầu tiên, đạt 94% ở cây thứ nhất và tăng thêm 3% (đạt 97%) ở cây thứ 300. Tương tự như vậy, độ mất mát của mô hình cũng có sự cải thiện đáng kể, giảm 2.2% từ 3.2% ở cây quyết định đầu tiên còn 0.9% ở cây cuối cùng.



Hình 4.1: Độ chính xác của mô hình RF trong lần thực nghiệm đầu tiên

Hình 4.1 và 4.2 biểu diễn độ chính xác và độ mất mát của mô hình. Như hình vẽ biểu diễn, các độ đo tăng dần theo từng lớp, càng gần tâm thì độ đo có giá trị càng thấp và ngược lại. Theo như hình 4.1 mô tả, ở lần thực nghiệm đầu tiên, độ chính xác của mô hình đạt khoảng 94% và tăng dần trong những lần tiếp theo, đến lần cuối cùng đạt được gần 98% (97.99%).



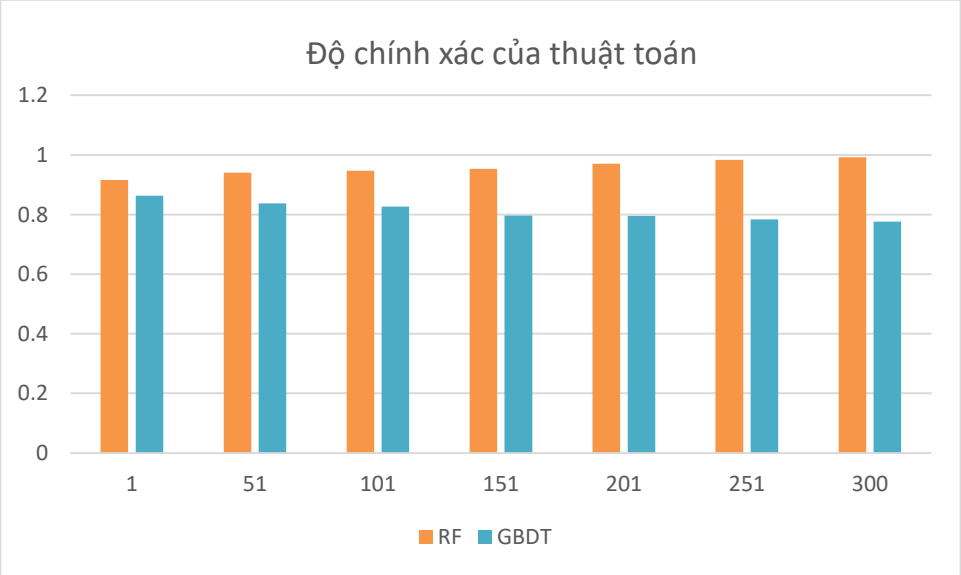
Hình 4.2: Độ mất mát của mô hình RF trong lần thực nghiệm đầu tiên

Trong bảng 4.4, kết quả phân loại trạm của hai mô hình với các tham đầu vào cố định. Đối với mô hình sử dụng thuật toán GBDT, tiến hành chọn lại các đặc trưng có độ quan trọng cao hơn các đặc trưng còn lại khi tham gia xây dựng mô hình, ở đây đó chính là các đặc trưng CellUpMax, TVU, TDVT.

Trong bảng 4.4 và hình 4.3, ta thấy mô hình RF cho kết quả dự đoán chính xác cao hơn mô hình GBDT ở hầu hết các lần thay đổi số lượng cây. Gần như ở tất cả các lần thực nghiệm, độ chính xác của mô hình RF luôn giữ được sự ổn định ở mức trung bình là khoảng 94%, sau đó tăng dần trong những cây quyết định tiếp theo. Ở lần thực nghiệm thứ tư, mô hình GBDT đạt được độ chính xác đáng kể khoảng 88%, tuy nhiên sau đó độ chính xác có xu hướng giảm dần và đến cây quyết định cuối cùng chỉ còn khoảng 72%.

Bảng 4.4: So sánh độ chính xác của hai thuật toán

STT	Thuật toán	K						
		1	51	101	151	201	251	300
1	RF	0.9855	0.9841	0.9826	0.9841	0.9855	0.9768	0.9918
	GBDT	0.8333	0.8182	0.7879	0.8030	0.8182	0.8030	0.8030
2	RF	0.9262	0.9826	0.9897	0.9916	0.9922	0.993	0.9943
	GBDT	0.7932	0.8135	0.8143	0.8208	0.8265	0.8417	0.8548
3	RF	0.9456	0.9521	0.955	0.9555	0.9731	0.9815	0.9852
	GBDT	0.8521	0.8337	0.8282	0.819	0.8081	0.8057	0.803
4	RF	0.9927	0.9844	0.9717	0.9711	0.9709	0.9567	0.9514
	GBDT	0.8849	0.8647	0.8061	0.8056	0.7589	0.7533	0.7246
5	RF	0.9059	0.9061	0.9272	0.9651	0.9751	0.9807	0.9888
	GBDT	0.8339	0.8304	0.8289	0.8254	0.8163	0.812	0.8007
6	RF	0.9208	0.923	0.951	0.9573	0.9698	0.9753	0.9895
	GBDT	0.839	0.8238	0.823	0.8198	0.7918	0.7786	0.7711
7	RF	0.9156	0.941	0.9473	0.953	0.9711	0.9837	0.9916
	GBDT	0.8635	0.8375	0.8267	0.7969	0.7952	0.7837	0.7761



Hình 4.3: So sánh độ chính xác của hai thuật toán ở lần chạy thứ 7

4.4 Kết luận chương

Chương 4 của luận văn trình bày mô hình thực nghiệm là TF-DF, bộ dữ liệu thực nghiệm cũng như quá trình xử lý và áp dụng bộ dữ liệu vào mô hình cũng đã được thực hiện và đánh giá. Việc thực nghiệm trên bộ dữ liệu cũng đã cho thấy được kết quả mô hình đạt độ chính xác cao, bên cạnh đó, độ đo mất mát cũng cho thấy sự cải thiện đáng kể qua các cây quyết định trong mô hình.

KẾT LUẬN

Trong khuôn khổ của luận văn, cơ sở lý thuyết về học máy và một số thuật toán áp dụng giải bài lựa chọn thuộc tính đã được tìm hiểu. Chúng tôi cũng đã tập trung nghiên cứu về mô hình Tensorflow-Decision Forest và thuật toán Random Forest. Từ những tìm hiểu này này chúng tôi đề xuất hướng cải tiến cách đánh nhãn cho các đặc trưng nhằm tăng hiệu quả của thuật toán phân loại đặc biệt với dữ liệu có số chiều cao.

Để chứng minh tính hiệu quả của mô hình cải tiến, thực nghiệm được tiến hành trên bộ dữ liệu về lưu lượng mạng. Từ những kết quả thực nghiệm đạt được trên bộ dữ liệu về lưu lượng mạng thấy rằng độ chính xác của mô hình Decision Forest sử dụng thuật toán Random Forest đạt hiệu năng cao. Qua đó, có thể đóng góp thêm một chọn lựa cho các nhà phát triển ứng dụng khi phát triển các ứng dụng liên quan đến phân loại dữ liệu. Với những đóng góp trong luận văn này, chúng tôi hi vọng đã góp phần giải quyết một phần nhỏ liên quan đến bài toán khai phá dữ liệu nói chung cũng như bài toán phân loại dữ liệu nói riêng.

Tôi cũng hi vọng từ các đóng góp của mình có thể xây dựng lên các hệ thống đánh giá và dự đoán áp dụng một cách thiết thực vào đời sống xã hội.