HỌC VIỆN CÔNG NGHỆ BƯU CHÍNH VIỄN THÔNG

NGUYỄN NGỌC THƠ

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

LUẬN VĂN THẠC SĨ KỸ THUẬT

(Theo định hướng ứng dụng)

THÀNH PHỐ HỒ CHÍ MINH - NĂM 2022

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

LUẬN VĂN THẠC SĨ KỸ THUẬT

(Theo định hướng ứng dụng)

NGƯỜI HƯỚNG DẪN KHOA HỌC

TS. NGUYỄN XUÂN SÂM

THÀNH PHỐ HỒ CHÍ MINH - NĂM 2022

LỜI CAM ĐOAN

Tôi tên là Nguyễn Ngọc Thơ, cam đoan rằng luận văn "***Hệ hỗ trợ quyết định phân nhóm các trạm BTS theo lưu lượng***" là bài nghiên cứu của chính tôi dưới sự hướng dẫn của **TS. Nguyễn Xuân Sâm.**

Ngoại trừ những tài liệu tham khảo được trích dẫn trong luận văn này, tôi cam đoan rằng toàn phần hay những phần nhỏ của luận văn này chưa từng được công bố hay được sử dụng để nhận bằng cấp ở những nơi khác.

Không có sản phẩm nghiên cứu nào của người khác được sử dụng trong luận văn này mà không được trích dẫn theo đúng quy định.

Luận văn này chưa bao giờ được nộp để nhận bất kỳ bằng cấp nào tại các trường đại học hoặc cơ sở đào tạo khác.

Tp.HCM, ngày 25 tháng 01 năm 2022

Học viên thực hiện luận văn

Nguyễn Ngọc Thơ

LỜI CẢM ƠN

Trong suốt quá trình học tập, nghiên cứu và thực hiện đề tài luận văn thạc sĩ, ngoài những cố gắng và nỗ lực của bản thân, tôi đã nhận được sự hướng dẫn, giúp đỡ quý báu của quý thầy cô, cùng với sự động viên, khích lệ và ủng hộ của các đồng nghiệp, bạn bè và gia đình. Với lòng kính trọng và biết ơn sâu sắc tôi xin được bày tỏ lời cảm ơn chân thành tới:

Ban Giám hiệu, Phòng đào tạo sau đại học và quý Thầy Cô Khoa công nghệ thông tin, trường Học viện Công nghệ Bưu chính viễn thông đã tạo mọi điều kiện thuận lợi giúp tôi hoàn thành luận văn.

Tôi xin chân thành cảm ơn **TS. NGUYỄN XUÂN SÂM**, người thầy kính mến đã hết lòng giúp đỡ, dạy bảo, động viên và tạo mọi điều kiện thuận lợi cho tôi trong suốt quá trình học tập và hoàn thành luận văn tốt nghiệp. Xin chân thành cảm ơn quý Thầy Cô trong hội đồng chấm luận văn đã cho tôi những đóng góp quý báu để hoàn chỉnh luận văn này.

Tôi xin chân thành cảm ơn mọi người trong gia đình tôi, đã tạo điều kiện, động viên khích lệ để tôi học tập và hoàn thành luận văn này.

Mặc dù đã cố gắng hết sức, song do thời gian và kinh nghiệm nghiên cứu khoa học còn hạn chế nên không thể tránh khỏi những thiếu sót. Tôi rất mong nhận được sự góp ý của quý thầy cô cùng bạn bè đồng nghiệp để kiến thức của mình ngày một hoàn thiện hơn.

Tp.HCM, ngày 25 tháng 01 năm 2022

Học viên thực hiện luận văn

Nguyễn Ngọc Thơ

DANH SÁCH HÌNH VẼ

DANH SÁCH BẢNG

MỤC LỤC

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 CÁC YÊU CẦU BÀI TOÁN LƯU LƯỢNG CÁC TRẠM BTS

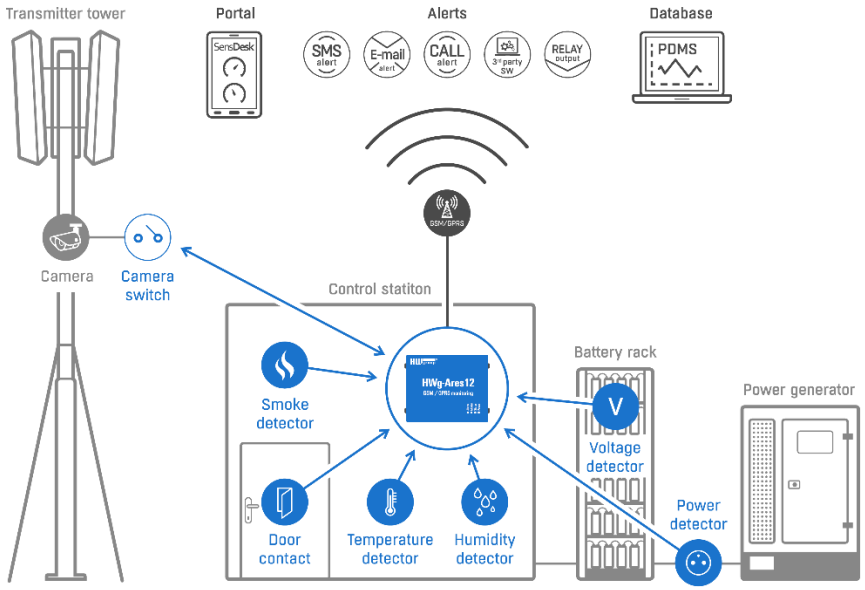
Chương 2: CƠ SỞ LÝ THUYẾT VÀ CÁC NGHIÊN CỨU LIÊN QUAN

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

Chương 4: THỰC NGHIỆM VÀ ĐÁNH GIÁ KẾT QUẢ

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

Trong hình 1.1, trạm thu phát gốc (BTS) là một thiết bị hỗ trợ giao tiếp không dây giữa thiết bị người dùng (UE) và mạng dùng để chuyển tiếp lưu lượng thoại và dữ liệu. 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. Lưu lượng mạng di động cung cấp một số tính năng như:

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 Mạng điện thoại di động

Ví dụ phổ biến nhất của mạng di động là mạng điện thoại di động (điện thoại di động). Điện thoại di động là điện thoại di động nhận hoặc thực hiện các cuộc gọi thông qua một điểm di động (trạm gốc) hoặc tháp truyền. Sóng vô tuyến được sử dụng để chuyển tín hiệu đến và đi từ điện thoại di động. Các mạng điện thoại di động hiện đại sử dụng tế bào vì tần số vô tuyến là một nguồn tài nguyên có giới hạn, được chia sẻ. Trang web di động và thiết bị cầm tay thay đổi tần số dưới sự điều khiển của máy tính và sử dụng bộ phát công suất thấp để nhiều người gọi có thể sử dụng đồng thời số lượng tần số vô tuyến giới hạn mà ít bị nhiễu hơn.

Nhà khai thác điện thoại di động sử dụng mạng di động để đạt được cả vùng phủ sóng và dung lượng cho thuê bao của họ. Các khu vực địa lý lớn được chia thành các ô nhỏ hơn để tránh mất tín hiệu đường ngắm và hỗ trợ một số lượng lớn điện thoại hoạt động trong khu vực đó. Tất cả các điểm di động đều được kết nối với tổng đài điện thoại (hoặc bộ chuyển mạch), đến lượt nó lại kết nối với mạng điện thoại công cộng. Ở các thành phố, mỗi điểm di động có thể có phạm vi lên đến khoảng 1⁄2 dặm (0,80 km), trong khi ở các vùng nông thôn, phạm vi có thể lên tới 5 dặm (8,0 km). Có thể là ở những khu vực thoáng đãng, người dùng có thể nhận được tín hiệu từ một điểm di động cách đó 25 dặm (40 km).

Vì hầu hết tất cả điện thoại di động đều sử dụng công nghệ di động, bao gồm GSM, CDMA và AMPS (tương tự), thuật ngữ "điện thoại di động" ở một số vùng, đặc biệt là Hoa Kỳ, được sử dụng thay thế cho "điện thoại di động". Tuy nhiên, điện

thoại vệ tinh là điện thoại di động không liên lạc trực tiếp với tháp di động trên mặt đất mà có thể hoạt động gián tiếp qua vệ tinh. Có một số công nghệ di động kỹ thuật số khác nhau, bao gồm: Hệ thống toàn cầu cho truyền thông di động (GSM), Dịch vụ vô tuyến gói chung (GPRS), cdmaOne, CDMA2000, Evolution-Data Optimized (EV-DO), Tốc độ dữ liệu nâng cao cho GSM Evolution (EDGE), Hệ thống viễn thông di động toàn cầu (UMTS), Viễn thông không dây nâng cao kỹ thuật số (DECT), AMPS kỹ thuật số (IS-136 / TDMA) và Mạng kỹ thuật số nâng cao tích hợp (iDEN). Việc chuyển đổi từ tiêu chuẩn tương tự sang tiêu chuẩn kỹ thuật số hiện có theo một con đường rất khác ở Châu Âu và Hoa Kỳ. [19] Kết quả là, nhiều tiêu chuẩn kỹ thuật số đã xuất hiện ở Mỹ, trong khi Châu Âu và nhiều quốc gia lại hướng tới tiêu chuẩn GSM.

Cấu trúc mạng di động của mạng vô tuyến bao gồm các yếu tố sau:

- Một mạng lưới các trạm gốc vô tuyến tạo thành hệ thống con của trạm gốc.
- Mạng chuyển mạch lõi để xử lý các cuộc gọi thoại và văn bản.
- Một mạng chuyển mạch gói để xử lý dữ liệu di động.
- Mạng điện thoại chuyển mạch công cộng để kết nối các thuê bao với mạng điện thoại rộng hơn.

Mạng này là nền tảng của mạng hệ thống GSM. Có nhiều chức năng được thực hiện bởi mạng này để đảm bảo khách hàng có được dịch vụ mong muốn bao gồm quản lý di động, đăng ký, thiết lập cuộc gọi và bàn giao. Bất kỳ điện thoại nào kết nối với mạng thông qua RBS (Trạm gốc vô tuyến) ở một góc của ô tương ứng, đến lượt nó sẽ kết nối với Trung tâm chuyển mạch di động (MSC). MSC cung cấp kết nối đến mạng điện thoại chuyển mạch công cộng (PSTN). Liên kết từ điện thoại đến RBS được gọi là đường lên trong khi cách khác được gọi là đường xuống. Các kênh vô tuyến sử dụng hiệu quả phương tiện truyền dẫn thông qua việc sử dụng các sơ đồ truy cập và ghép kênh sau: đa truy cập phân chia theo tần số (FDMA), đa truy cập phân

chia theo thời gian (TDMA), đa truy cập phân chia theo mã (CDMA) và đa truy cập phân chia theo không gian truy cập (SDMA).

1.3.4 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. Có một số yếu tố phổ biến ảnh hưởng đến hiệu suất mạng vô tuyến mà hầu hết mọi người sẽ dễ dàng xác định, nhưng điều đó không làm giảm bớt tầm quan trọng của chúng khi xem xét quy hoạch mạng lưới, đó là:

- **Cản trở vật lý**: Tín hiệu vô tuyến có thể gặp khó khăn khi thâm nhập qua các vật thể rắn, có thể bất kỳ như là đồi núi, các tòa nhà, bức tường hoặc thậm chí con người. Càng nhiều vật cản giữa trạm phát và trạm nhận thì càng nhiều khả năng cường độ tín hiệu bị ảnh hưởng hơn, do đó bạn nên cố gắng duy trì thông thoáng các đường liên kết trong site (line-of-site) tốt nhất có thể. Điều này rõ ràng là không thực tế vì gần như luôn có thứ gì đó trên tuyến truyền dẫn, nhưng bạn có thể giảm thiểu ảnh hưởng của nó bằng cách sử dụng tần số cụ thể, khả dụng cho bạn. Như một quy luật, tần số càng thấp thì sóng có đặc điểm thâm nhập càng tốt hơn. Tuy nhiên cũng cần phải nói, với tần số càng cao thì khả năng phản xạ của sóng càng tốt, vì vậy trong một số trường hợp có thể lợi dụng yếu tố phản xạ của một tín hiệu để thực hiện gửi nó tới trạm thu mà không cần phải truyền xuyên qua vật cản.
- **Phạm vi mạng và khoảng cách giữa các thiết bị:** Các thiết bị đang hoạt động trong mạng luôn cố gắng kết nối truyền nhận với nhau nhiều hơn và điều đó gây ra giảm rớt cường độ tín hiệu rất nhiều. Điều này là do cách thức lan truyền các tín hiệu vô tuyến, phủ một vùng rộng lớn hơn khi chúng đi xa hơn và vì lý do này nên khi tín hiệu trãi rộng hơn,

nó sẽ trở nên yếu hơn. Cường độ tín hiệu giảm theo quan hệ nghịch đảo bậc ba với khoảng cách giữa hai thiết bị. Do đó, khi khoảng cách tăng gấp đôi tín hiệu sẽ suy yếu đi 8 lần.

- **Nhiễu trong mạng vô tuyến:** Mạng vô tuyến đang trở nên ngày càng thông dụng và do đó ngày càng nhiều truyền dẫn vô tuyến thực hiện truyền nhận qua môi trường không khí. Những tín hiệu hoạt động ở các tần số tương tự nhau có thể gây nhiễu với nhau và có tác động tiêu cực đáng kể đến hiệu suất của mạng. Điều này có nghĩa là những băng tần được sử dụng phổ biến như các băng không cần cấp phép 2.4GHz có thể bị ảnh hưởng nghiêm trọng bởi tình trạng dày đặc của các tín hiệu vô tuyến, do đó tai một điểm, chẳng hạn một thiết bị thu sẽ không thể hoạt động dù đạt mức công suất chấp nhận được. Những kỹ thuật vô tuyến khác có thể gây nhiễu giống nhau như điện thoại DECT và lò vi sóng. Nhiều dãi băng tần đang trở nên khả dụng cho hoạt động mạng vô tuyến để tránh vấn đề xuyên nhiễu này chẳng hạn như băng tần không cần cấp phép 5GHz, đây là một băng tần đang trở nên khá phổ biến. Khi hoạt động trong vùng mật độ mạng vô tuyến cao, băng tần 5GHz được khuyên dùng để bạn thiết lập hoạt động cho các mạng doanh nghiệp, các nhà điều hành, vv… xung quanh bạn để tránh các vấn đề về nhiễu trong tương lai.
- **Việc chia sẻ tín hiệu:** Mạng vô tuyến cho phép nhiều hơn một người để truyền thông giao tiếp với một nguồn mạng khác tại một thời điểm bất kỳ. Việc chia sẻ kết nối này có nghĩa là có nhiều thuê bao hơn sử dụng mạng, nhiều thiết bị hơn cố gắng kết nối truyền thông với một điểm truy cập trong một thời điểm. Các điểm truy cập phải ủy thác những tài nguyên của nó tới mỗi thuê bao riêng lẻ với mỗi lượng vô tuyến truyền dẫn để cho thuê bao hoạt động. Thiết bị có khả năng truyền dẫn song công (Full-Duplex) có thể truyền và nhận dữ liệu đồng thời,

trong khi thiết bị truyền dẫn bán song công (Half-Duplex) chỉ có thể gửi hoặc nhận vào tại một thời điểm bất kỳ.

- **Tải trọng và cách sử dụng mạng:** Bạn sẽ thấy rằng càng nhiều khách hàng thiết bị đang sử dụng mạng băng thông, thì càng ít người chia sẻ băng thông giữa chúng ta. Vì những đoạn băng yêu cầu tăng lên trong mạng của bạn (ví dụ video streaming là một ứng dụng chuyên sâu hỏi nhiều băng thông), bạn có thể muốn đầu tư nâng cấp thiết bị để đạt được mức độ máy chủ high quality information (or own the data transfer rate), an toàn hiệu suất và mạng tin cậy ở mức cao.
- **Bản chất của môi trường nội vùng:** Hầu hết những ảnh hưởng là dễ nhận thấy trong các mạng indoor, bản chất kết cấu tường có thể là một trong những cản trở lớn nhất của tín hiệu vô tuyến. Các vật liệu được sử dụng có mức độ ảnh hưởng khác nhau, bê tông là một mối ngờ thường trực trong việc liệu có ảnh hưởng xấu tới hiệu suất khi thiết lập mạng indoor. Nó gần như truyền đi mà không cho rằng các bức tường dày hơn thì khả năng thành công thấp hơn, tín hiệu sẽ xuyên qua nó trong khi duy trì một cường độ cao.
- **Giới hạn kênh phổ:** Điều này thường chỉ ảnh hưởng đến mạng vô tuyến hoạt động chỉ ở những dãi tần số phổ biến như 2.4GHz nhưng có thể bắt đầu ảnh hưởng đến dãi tần 5GHz trong tương lai nếu con người chuyển sang dùng ồ ạt trên dãi băng tần này. Mạng vô tuyến hoạt động trên những băng con (sub-band) còn gọi là kênh mà có băng thông nhỏ hơn so với băng thông trong toàn bộ những tần số hoạt động có thể của chúng. Băng tần 2.4GHz được chia thành 11 kênh, mỗi kênh hoạt động trên độ rộng kênh 25MHz và khoảng cách giữa các đỉnh kênh kế cận là 5MHz trong một dãi tổng thể từ 2412MHz đến 2462MHz. Rõ ràng không mất nhiều tính toán để nhận ra rằng các kênh phải chồng lấn lên nhau để có thể phù hợp trong tổng phạm vi hoạt động. Những vùng chồng lấn này gây ra nhiễu nếu các thiết bị vô tuyến đang sử dụng các

kênh lân cận, do đó các kênh được đề nghị sử dụng là chỉ các kênh 1, 6 và 11, đó là những kênh không chồng lấn. Tuy nhiên, điều này có nghĩa là chỉ có 3 thiết bị vô tuyến có thể sử dụng trong cùng khu vực trừ khi các kênh chồng lấn nhau được sử dụng.

- **Sự phản xạ của tín hiệu:** Phản xạ tín hiệu được biết một cách chính xác hơn với ảnh hưởng đa đường (Multi-Path Fade), thường xảy ra trong các tòa nhà có bố trí cấu trúc rắc rối và phức tạp. Các tuyến đường khác nhau mà các tín hiệu thực hiện truyền trên đó có thể bị phản xạ môi trường xung quanh gây ra sự khác biệt trong các chiều dài khoảng cách tuyến đường mà chúng đi tới trạm thu. Khi những tín hiệu khác nhau này đến trạm thu chúng có thể lệch pha nhau và điều này có thể gây ra chồng lấn sóng, đặt ra nghi vấn hoặc là mở rộng khả năng khuếch đại tín hiệu hoặc là có thể loại bỏ tín hiệu của nhau hoàn toàn. Thời gian mà các tín hiệu phản xạ đi tới trạm thu là khác nhau do khoảng cách khác nhau trong các tuyến đường RF mà chúng đi. Sự trãi rộng trễ giữa các tín hiệu tạo ra nhiễu liên ký tự ISI (Intersymbol Interference) là một trường hợp trong đó các tín hiệu bị trễ bắt đầu gây lỗi ký tự đi trên một tuyến đường RF ngắn hơn. Những vấn đề này có thể được khắc phục bằng cách sử dụng những anten phân tập (cài đặt nhiều hơn 1 anten trên máy phát với khoảng cách cụ thể phù hợp) đảm bảo rằng nếu một anten hoạt động kém, những anten khác có thể sẽ ổn định, hoặc bằng cách khác, sử dụng công nghệ như OFDM có thể khắc phục vấn đề bằng cách đưa ra các kênh sóng mang con được chia ra từ độ rộng mỗi băng kênh chính. Những kênh sóng mang con này gửi và nhận dữ liệu đồng thời, song song nhau. Việc phân chia nhiều kênh nhỏ hơn đảm bảo nhiều dữ liệu hơn có thể được truyền với mức suy hao thấp hơn do nhiễu tín hiệu.
- **Hạn chế của tín hiệu vô tuyến:** Vì lý do an ninh bạn có thể muốn hạn chế việc truyền tín hiệu vô tuyến của bạn để chỉ các khu vực mà bạn

muốn cung cấp được truy cập mạng. Điều này có thể khó khăn bởi vì việc kiểm soát sự lan truyền tín hiệu thì không phải là dễ dàng vì nó có thể đi xuyên qua tường vào các tòa nhà khác hoặc các khu vực bên ngoài site phủ sóng thiết lập nơi mà ai đó có thể cố gắng kết nối dù không được phép. Việc giới hạn mức công suất máy phát để chỉ phủ sóng các khu vực được yêu cầu là một trong những phương pháp giải quyết vấn đề này nhưng nó sẽ làm giảm bớt hiệu quả, cường độ tín hiệu sẽ bị yếu đi khi tới tất cả các thiết bị thu. Việc sử dụng anten định hướng để hạn chế sự trãi rộng tín hiệu cũng có thể có hiệu quả trong việc hạn chế vùng phủ sóng tín hiệu.

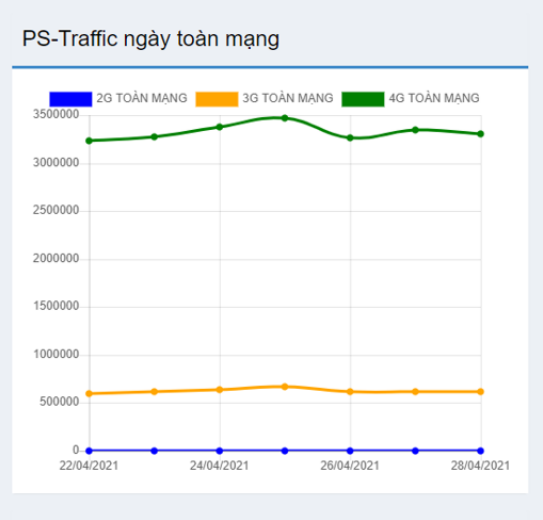
- **Hạn chế công suất trạm phát:** Những quy định được thiết lập bởi OFCOM chỉ ra rằng các thiết bị hoạt động trên dãi tần số nhất định nào đó phải tuân thủ mức công suất ngõ ra máy phát tối đa. Những giá trị này thay đổi phụ thuộc vào tần số thiết bị của bạn đang hoạt động trên, ví dụ ở 2.4GHz là100mW EIRP và ở 5GHz thì nằm trong khoảng giữa 200mW - 4W EIRP, vì thế bạn có thể thấy rằng đối với các dãi băng tần khác nhau, bạn có thể hoạt động trên những mức công suất khác nhau. Điều này có ảnh hưởng lớn đến cường độ tín hiệu và trong trường hợp tốt nhất, các mạng được hưởng lợi từ việc các thiết bị hoạt động trên mức năng lượng cao hơn theo quy định cho phép, tăng cường cường độ tín hiệu trên một khoảng cách dài hơn. Nếu định hướng, khuếch đại anten được sử dụng thì những mức công suất hoạt động có thể phải được giảm xuống bởi vì OFCOM khai báo giới hạn tối đa cho mức cường độ tín hiệu liên quan bởi cách kết hợp anten và card vô tuyến phát. Nếu sử dụng một anten búp sóng nhỏ cho định hướng cao, mức công suất có thể phải được giảm đáng kể để giữ cho truyền dẫn vô tuyến theo những quy định cường độ tín hiệu tối đa.
- **Suy giảm tốc độ do các mào đầu quản lý (overhead) vô tuyến:** Do mã hóa, biên dịch gói và sử dụng một phần băng thông kênh cho dữ

liệu người dùng, tốc độ truyền dẫn dữ liệu tối đa không phải là tốc độ truyền dẫn thực, tốc độ truyền dẫn dữ liệu hữu ích ngoài thực tế được trải nghiệm bởi người dùng cuối. Các mào đầu quản lý (overhead) của giao thức mạng vô tuyến thông thường sẽ dẫn đến tốc độ truyền dẫn dữ liệu thực tế chỉ bằng khoảng một nửa tốc độ truyền dẫn dữ liệu tối đa được quảng bá và điều này sau đó có thể được giảm thêm bởi các yếu tố khác liên quan đến việc trang bị các gói dữ liệu. Về cơ bản, tốc độ truyền dẫn dữ liệu tối đa được quảng bá nhìn chung là cao hơn nhiều so với những gì bạn có thể trải nghiệm, nhưng một số nhà sản xuất vẫn liệt kê ra như là tốc độ thực trên các sản phẩm của họ để cung cấp cho khách hàng một ý tưởng tốt hơn những gì họ mong đợi để đạt được.

- **Làm giảm hiệu suất để duy trì kết nối:** Một số thiết bị mạng vô tuyến có thể làm giảm tốc độ thông lượng hoạt động tới các thiết bị vì để duy trì kết nối trong những vùng tín hiệu thấp do khoảng cách tăng lên giữa các thiết bị hoặc do nhiễu … Điều này ảnh hưởng đến toàn bộ mạng lưới và những người dùng khác mà được kết nối với nó vì thời gian truyền dữ liệu tăng lên giữa các thiết bị chậm hơn. Ngoài ra, việc truyền lại (retransmission) dữ liệu do rớt gói tin cũng hạn chế sự khả dụng của các điểm truy cập cho giao tiếp với những client khác.

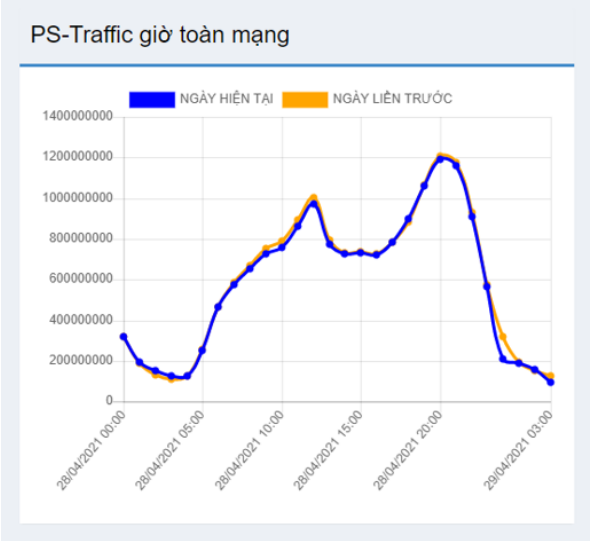
1.4 Mô tả tập dữ liệu

Trong những năm gần đây, các thiết bị di động (UE) kết nối dữ liệu qua mạng di động gia tăng nhanh chóng qua mạng 3G và 4G trên địa bàn tỉnh Tây Ninh. Hình 1.2 mô tả lưu lượng mạng di động 2G, 3G, và 4G đo được trên địa bạn tỉnh Tây Ninh từ ngày 22 – 28/04/2022, trong đó lưu lượng qua mạng 4G cao hơn rất nhiều và biến thiên theo thời gian so với mạng 3G. Việc lưu lượng qua mạng di động gia tăng và biến thiên gây áp lực đến công tác quản lý, vận hành mạng viễn thông. Do vậy, việc khảo sát và phân tích các số liệu liên quan đến lưu lượng mạng di động hiện nay là công việc cấp bách.



Hình 1.2: Thống kê lưu lượng theo ngày

Do đặc trưng lưu lượng mạng di động là chuỗi dữ liệu được sắp xếp có trật tự theo thời gian (time series), việc sử dụng các đặc trưng theo thời gian này có ý nghĩa trong việc phát hiện các thay đổi lưu lượng bất thường tại các trạm BTS. Hình 1.3 mô tả lưu lượng mạng được thống kê trong ngày từ 0h ngày 28/04/2021 đến 3h ngày 29/04/2022, theo đó, lưu lượng mạng được thống kê giữa 2 ngày gần tương đương nhau, riêng mức lưu lượng vào khoảng 23h ngày 28/04/2022 tăng nhiều hơn so với ngày 29/04/2022.



Hình 1.3: Thống kê lưu lượng theo giờ

1.4.1 Chuẩn bị dữ liệu

Trong nghiên cứu này, chúng tôi kiểm thử và phân tích chuỗi lưu lượng thời gian thực của mạng di động tỉnh Tây Ninh (với số mẫu và số trạm hạn chế) tại mạng di động tỉnh Tây Ninh. Bộ dữ liệu về lưu lượng mạng chúng tôi khảo sát có tổng cộng 24 trường và 1000 dòng được dùng trong thực nghiệm để đánh giá và so sánh hiệu năng các mô hình học máy tiêu biểu. Trong đó, các trường dữ liệu liên quan đến lưu lượng như Traffic_Volume_UL_GB, Traffic_Volumn_DL_GB, vv…

Dữ liệu thô này được sử dụng làm đầu vào cho mô hình huấn luyện sau khi đã được (tiền) xử lý. Dữ liệu sau khi được phân tích này có vai trò gắn liền với việc ra quyết định trong việc quản trị và tối ưu các hoạt động khai thác trong mạng viễn thông. Trong thời gian đầu, một trong các mong muốn trong các hoạt động khai thác này là cung cấp thông tin hỗ trợ giám sát và ra quyết định phân nhóm các trạm theo lưu lượng.

1.4.2 Nhu cầu về ra quyết định

Trên cơ sở phân tích lưu lượng tại các trạm BTS, người quản lý xác định được các vấn đề bất thường về lưu lượng và các vấn đề kỹ thuật liên quan đến việc quản lý, giám sát các trạm. Việc ra quyết định quản lý tốt có thể tối ưu các hoạt động liên quan quan đến lưu lượng qua trạm. Do vậy, việc hoàn thiện hệ thống hỗ trợ quyết

định phân nhóm lưu lượng (trong tương lai) dựa trên các mô hình học máy phù hợp sẽ mang lại thành công trong việc ra quyết định cho nhà quản lý.

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 [1]

Ngày nay, 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.

Để áp dụng thành công học máy trong bối cảnh kinh doanh, điều quan trọng là phải hiểu sự khác biệt của các phương pháp tiếp cận, bất kể nền tảng chuyên môn hay kinh nghiệm trước đây với Học máy. Biết khi nào nên sử dụng phương pháp nào và cách bắt đầu, có thể giúp phát hiện và nhận ra tiềm năng phát triển. Khi đối mặt với các loại thách thức khác nhau, có thể tìm ra kỹ thuật phù hợp cho một trường hợp sử dụng cụ thể là yếu tố quyết định để thành công. Từ việc phân loại hình ảnh hoặc dự đoán tác động của chiến dịch tiếp thị, kiến thức về các cách tiếp cận Học máy khác nhau sẽ giúp bạn đi đúng hướng.

Học máy có thể được sử dụng để giải quyết nhiều loại vấn đề khác nhau. Sau khi cung cấp cho nó dữ liệu đào tạo, bao gồm các ví dụ với các tính năng đầu vào và kết quả liên quan, nó sẽ tính toán một bộ quy tắc để đưa ra dự đoán về tương lai. Bộ quy tắc được tạo ra bởi thuật toán được sử dụng thông qua kỹ thuật đã chọn. Đây là một điểm khác biệt chính đối với các phương pháp phân tích dữ liệu truyền thống thường xây dựng một logic rõ ràng theo cách thủ công.

Trong một số điều kiện nhất định, các kỹ thuật học máy khác nhau có thể được áp dụng để phù hợp với các bài toán ứng dụng trong thực tế. Lợi ích của học máy có thể được minh họa bằng ví dụ phân loại thư rác email. Bằng cách sử dụng các dữ liệu cho trước có gắn nhãn, học máy có thể xây dựng phân qui tắc phân nhóm để lọc thư

rác tự động và tránh được việc triển khai thủ công danh sách dài các quy tắc cho các cụm từ thư rác trong email.

Bài toán phân loại hiện là một chủ đề được cộng đồng khoa học trong và ngoài nước quan tâm. Dựa vào các cách tiếp cận và thiết lập logic gán nhãn và huấn luyện, có nhiều phương pháp học máy khác nhau được sử dụng để phân biệt cho các cách tiếp cận khác nhau. Trong đó, có thể phân chia thành các cách tiệp cận như:

- Học có giám sát
- Học không giám sát
- Học bán giám sát

2.1.1 Học tập có giám sát

Tính năng chính của học có giám sát là các nhãn của đầu ra được xác định trước. Con người quyết định cấu trúc của kết quả và do đó trực tiếp giám sát các nỗ lực đào tạo của mô hình. Trong học tập có giám sát, các loại nhiệm vụ chính là phân loại và hồi quy.

Phân loại dự đoán khả năng các tập dữ liệu thuộc về hoặc dẫn đến một lớp đầu ra được xác định trước. Trong quá trình đào tạo phân loại, dữ liệu, bao gồm các tính năng (còn được gọi là dự đoán) và nhãn (còn được gọi là lớp), được đưa vào thuật toán và khả năng xảy ra mối quan hệ của mỗi tập dữ liệu với các lớp và gán lớp.

Phương pháp chính khác, **hồi quy**, nhằm dự đoán một giá trị số cụ thể dựa trên các tính năng của dữ liệu đầu vào. Như trong phân loại, việc đào tạo được thực hiện dựa trên các đặc điểm đầu vào cũng như kết quả tương ứng. Khi dữ liệu mới được đưa vào mô hình, giá trị đầu ra có thể được dự đoán. Các trường hợp sử dụng cho nhiệm vụ hồi quy trong học tập có giám sát bao gồm dự đoán giá cổ phiếu, thống kê kinh tế và điểm hài lòng của khách hàng.

Đối với hai phương pháp, phân loại và hồi quy, các thuật toán sau thường được sử dụng: K-NN (K Nearest Neighbor), Hồi quy tuyến tính (Linear Regression), Hồi quy logistic, Máy Véc tơ hỗ trợ (Support Vector Machine), Cây quyết định (Decision

Tree) và rừng ngẫu nhiên (Random Forest), Mạng thần kinh (Convolutional Neural Network).

2.1.2 Học không giám sát

So với học có giám sát, trong đó con người quyết định cấu trúc của đầu ra bằng cách chỉ định các nhãn của đầu ra, học không giám sát hoàn toàn không quan tâm đến việc sử dụng các nhãn được xác định trước - hệ thống học độc lập và đề xuất một cấu trúc đầu ra. Các loại vấn đề học tập không giám sát chính là phân cụm, phát hiện bất thường & mới lạ, trực quan hóa & giảm kích thước và học quy tắc kết hợp.

Bằng cách sử dụng thuật toán **phân nhóm**, bạn có thể phân chia tập dữ liệu thành các nhóm, ví dụ: phân nhóm khách hàng thành các phân khúc khách hàng khác nhau có cùng đặc điểm hoặc hành vi. Trên hết, phân nhóm phân cấp cho phép phân đoạn bổ sung trong các cụm đã được tìm thấy, do đó, phân khúc khách hàng có thể được chia thành các đơn vị nhỏ hơn.

2.1.3 Học bán giám sát

Đằng sau thuật ngữ học bán giám sát là sự kết hợp của các phương pháp từ cả hai lĩnh vực, học có giám sát và không giám sát. Phương pháp kết hợp này hoạt động với một phần dữ liệu đào tạo được gắn nhãn - hầu hết nó thực sự không được gắn nhãn và một phần nhỏ được gắn nhãn. Ví dụ nổi bật cho việc học bán giám sát là ứng dụng ảnh của các công ty công nghệ (tức là Google, Apple), nơi ứng dụng có thể xác định những người nào thường xuyên xuất hiện trong ảnh, nhóm ảnh cho phù hợp (= không giám sát) và cho phép người dùng cung cấp nhãn cho các cụm đó, đó là tên của những người được xác định (= được giám sát).

Các thuật toán phổ biến cho việc học bán giám sát là **mạng niềm tin sâu** (DBNs), hoạt động với **các máy Boltzmann bị hạn chế** (RBM), một loại mạng nơ-ron nhân tạo cụ thể, cho phần không được giám sát. Chúng cũng có thể được xếp chồng lên nhau. Sau khi đào tạo các RBM (không giám sát), họ có thể được nâng cao bằng cách sử dụng các loại phương pháp học có giám sát khác nhau.

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

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

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 ∈ {0,1} và ước lượng xác suất p = Pr(y = 1), công thức logloss là:

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

2.3 Các công trình liên quan

Vào năm 2016, nhóm tác giả gồm Fengli Xu [1] và những người cộng sự của mình hiểu được mô hình lưu lượng di động của các tháp di động quy mô lớn trong môi trường đô thị là vô cùng quý giá đối với các nhà cung cấp dịch vụ Internet, người dùng di động và các nhà quản lý chính phủ của các đô thị hiện đại. Bài báo này nhằm mục đích trích xuất và mô hình hóa các mô hình giao thông của các tòa tháp quy mô lớn được triển khai trong một thành phố đô thị.

Để đạt được mục tiêu này cần giải quyết một số thách thức, bao gồm thiếu các công cụ thích hợp để xử lý dữ liệu đo lường giao thông quy mô lớn, các mô hình giao thông chưa xác định, cũng như xử lý các yếu tố phức tạp của sinh thái đô thị và hành vi của con người ảnh hưởng đến mô hình giao thông. Đóng góp cốt lõi của nhóm nghiên cứu là một mô hình mạnh mẽ kết hợp thông tin ba chiều (thời gian, vị trí của các tháp và phổ tần số giao thông) để trích xuất và lập mô hình lưu lượng của hàng nghìn tháp di động. Phân tích thực nghiệm của nhóm nghiên cứu cho thấy những quan sát quan trọng sau đây. Đầu tiên, chỉ có năm mẫu giao thông cơ bản theo thời gian tồn tại trong số 9600 tháp di động. Thứ hai, mỗi mô hình giao thông được trích xuất lập bản đồ cho một loại vị trí địa lý liên quan đến sinh thái đô thị, bao gồm khu dân cư, khu kinh doanh, giao thông, giải trí và khu vực toàn diện. Thứ ba, phân tích phổ lưu lượng miền tần số của nhóm nghiên cứu cho thấy rằng lưu lượng của bất kỳ tháp

nào trong số 9600 có thể được xây dựng bằng cách sử dụng kết hợp tuyến tính của bốn thành phần chính tương ứng với các hành vi hoạt động của con người. Nhóm tác giả tin rằng phương pháp trích xuất và mô hình hóa các mô hình giao thông được đề xuất, kết hợp với phân tích thực nghiệm về lưu lượng di động, mở đường cho việc hiểu sâu sắc về các mô hình giao thông của các tháp di động quy mô lớn trong đô thị hiện đại.

Việc phân loại tự động các ứng dụng và dịch vụ là một tính năng vô giá đối với các mạng di động thế hệ mới. Ở đây, nhóm tác giả đề xuất và xác nhận các thuật toán để thực hiện tác vụ này, trong thời gian chạy, từ kênh điều khiển vật lý thô của mạng di động đang hoạt động, mà không cần phải giải mã và / hoặc giải mã các luồng đã truyền [2]. Hướng tới điều này, nhóm tác giả giải mã các bản tin Thông tin điều khiển đường xuống (DCI) được mang trong Kênh điều khiển đường xuống vật lý LTE (PDCCH). Các bản tin DCI được gửi bởi tế bào vô tuyến dưới dạng văn bản rõ ràng và trong bài báo này, được sử dụng để phân loại các ứng dụng và dịch vụ được thực thi tại các thiết bị đầu cuối di động được kết nối. Hai bộ dữ liệu được thu thập thông qua một chiến dịch đo lường lớn: một bộ được gắn nhãn, được sử dụng để huấn luyện các thuật toán phân loại và một bộ không được gắn nhãn, được thu thập từ bốn ô vô tuyến ở khu vực đô thị Barcelona, ở Tây Ban Nha. Trong số các phương pháp tiếp cận khác, bộ phân loại Mạng thần kinh hợp pháp (CNN) cung cấp độ chính xác phân loại cao nhất là 98%. Sau đó, bộ phân loại CNN được tăng cường với khả năng từ chối các phiên có các mẫu không phù hợp với những gì đã học trong giai đoạn đào tạo và sau đó được sử dụng để đạt được sự phân rã chi tiết lưu lượng cho bốn ô vô tuyến được giám sát, trong một trực tuyến và không bị giám sát thời trang.

Biến đổi khí hậu đặt các thành phố vào nguy cơ ngày càng tăng và đặt ra thách thức nghiêm trọng cho việc thích ứng [3]. Như một phản ứng, các nguồn dữ liệu mới kết hợp với lôgic hướng dữ liệu và các kỹ thuật mô hình không gian tiên tiến có tiềm năng tạo ra sự thay đổi đáng kể về vai trò của thông tin trong quy hoạch đô thị. Tuy nhiên, có rất ít hướng dẫn thực tế về các cơ hội tiềm năng do dữ liệu điện thoại di

động mang lại để nâng cao năng lực thích ứng ở các khu vực thành thị. Dựa trên việc xem xét các nghiên cứu không gian huy động dữ liệu điện thoại di động, bài báo này khám phá các cơ hội được cung cấp bởi thông tin kỹ thuật số như vậy để cung cấp các đánh giá rõ ràng về không gian về mức độ dễ bị tổn thương của đô thị và chỉ ra những cách mà chúng có thể giúp phát triển các chiến lược và công cụ năng động hơn cho quy hoạch đô thị và thảm họa quản lý rủi ro. Cuối cùng, dựa trên những hạn chế của phân tích dữ liệu điện thoại di động, nó thảo luận về những thách thức quản lý đô thị chính cần được giải quyết để hỗ trợ sự xuất hiện của sự thay đổi mang tính chuyển đổi trong các khuôn khổ quy hoạch hiện tại.

Những tiến bộ mới nhất trong công nghệ không dây đã dẫn đến sự gia tăng của các thiết bị và dịch vụ di động dữ liệu. Kết quả là, các mạng di động đã có sự gia tăng đáng kể về lưu lượng dữ liệu, trong khi lưu lượng thoại gần như không tăng trưởng. Do đó, điều cần thiết là các nhà khai thác phải hiểu hành vi lưu lượng dữ liệu ở cấp độ người dùng để đảm bảo trải nghiệm khách hàng tốt. Trong mạng truy cập vô tuyến (RAN), các giải pháp truyền thống dựa trên các phép đo mức tế bào không đủ để phân tích hiệu suất của từng người dùng. Thay vào đó, các lựa chọn thay thế mới như sử dụng dấu vết cuộc gọi và định nghĩa các chỉ báo lấy người dùng làm trung tâm mới sẽ cung cấp thông tin chi tiết và có giá trị cho mỗi kết nối. Một trong những phép đo quan trọng liên quan đến dịch vụ dữ liệu là thông lượng của người dùng. Trong công việc này, thông lượng người dùng được sử dụng làm thuộc tính chính để tiến hành chẩn đoán trong RAN, vốn thường là điểm nghẽn cho các dịch vụ dữ liệu. Vì vậy, một cây phân loại nhị phân được đề xuất để xác định nguyên nhân gốc rễ của thông lượng kém trong các phiên dữ liệu cấp người dùng. Sau đó, thông tin này được tổng hợp ở cấp độ tế bào để đưa ra chẩn đoán hiệu quả về các tế bào bị suy thoái. Đặc biệt, một phân tích dựa trên mối tương quan về tình trạng tế bào được đề xuất để xác định các hành vi bất thường của tế bào một cách tự động. Đánh giá đã được thực hiện với bộ dữ liệu từ các mạng di động trực tiếp. Kết quả cho thấy rằng phương pháp chẩn đoán được đề xuất [4] là một phương tiện hữu hiệu để xác định các yếu tố chính hạn chế thông lượng người dùng trong các tế bào mạng.

Việc phân tích các dấu vết lưu lượng di động thực rất hữu ích để hiểu các kiểu sử dụng của mạng di động. Cụ thể, dữ liệu di động có thể được sử dụng để tối ưu hóa và quản lý mạng về tài nguyên vô tuyến, quy hoạch mạng, tiết kiệm năng lượng, chẳng hạn. Tuy nhiên, dữ liệu mạng thực từ các nhà khai thác thường khó được truy cập do các vấn đề pháp lý và quyền riêng tư. Trong bài báo này [5], nhóm tác giả khắc phục tình trạng thiếu thông tin mạng bằng cách sử dụng bộ dò tìm LTE có khả năng giải mã kênh điều khiển LTE không được mã hóa và nhóm nghiên cứu trình bày phân tích không gian và thời gian của các dấu vết được ghi lại. Hơn nữa, nhóm nghiên cứu trình bày một phương pháp luận để rút ra đặc tính ngẫu nhiên cho sự biến đổi hàng ngày của lưu lượng LTE. Mô hình được đề xuất dựa trên chuỗi Markov thời gian rời rạc và được so sánh với các dấu vết thực. Kết quả cho thấy rằng, với một số trạng thái hạn chế, mô hình của nhóm nghiên cứu thể hiện mức độ chính xác cao về thống kê đơn hàng thứ nhất và thứ hai.

Dữ liệu sẵn có trong những năm gần đây đã tăng lên theo cấp số nhân, cho phép các nhà nghiên cứu trong lĩnh vực giao thông vận tải khai thác thông tin có giá trị liên quan đến các luồng giao thông. Theo nghĩa đó, dữ liệu mạng di động đại diện cho thông tin lưu lượng có giá trị khi xử lý các khu vực rộng lớn về mặt không gian do thuộc tính thu thập dữ liệu tuyến đường bằng cách sử dụng các trạm gốc di động ở xa. Thuộc tính này cho phép thu thập tự động dữ liệu điểm đến, được thu thập theo cách truyền thống bằng cách sử dụng bảng câu hỏi trực tuyến hoặc thực địa. Bài báo này nhằm mục đích trình bày khả năng sử dụng dữ liệu điểm đến được trích xuất từ tập dữ liệu mạng di động để phân loại các phương thức di chuyển. Một nghiên cứu điển hình [6] đã được thực hiện trên tập dữ liệu được thu thập tại Thành phố Rijeka, Croatia. Dataset được đánh giá dựa trên năm thuật toán học máy, dẫn đến Random forest là thuật toán có hiệu suất cao nhất với điểm chính xác là 99,93%.

Sự phổ biến của công nghệ di động có ý nghĩa quan trọng trong nghiên cứu vận tải. Mặc dù ngày càng có nhiều ứng dụng sử dụng điện thoại thông minh và mối quan tâm trong suy luận về tính di động, người ta đã có rất ít nỗ lực thảo luận về các lý

thuyết, mô hình và chủ đề nghiên cứu dựa trên một nghiên cứu có hệ thống về các nguồn học thuật bắt nguồn từ lĩnh vực liên ngành của công nghệ di động và giao thông [7]. Vì vậy, cần phải có sự tổng hợp kịp thời và toàn diện về tình trạng nghiên cứu hiện tại. Một phân tích tài liệu, tuân theo các hướng dẫn của PRISMA, nhằm xác định sự phát triển và triển khai thành công của công nghệ di động có thể được sử dụng cho các nghiên cứu hành vi trong vận tải. Việc xem xét các Bộ sưu tập cốt lõi của Web of Science, cơ sở dữ liệu JSTOR và SAGE được thực hiện.

Một quy trình sàng lọc nghiêm ngặt được sử dụng để thu thập các bài báo chính nhằm xây dựng hình ảnh chung về kiến thức hiện có. Ngoài ra, nghiên cứu này đề xuất một mô hình nghiên cứu tích hợp để tóm tắt cách các nghiên cứu trước đây đạt được kết quả hành vi và chương trình nghiên cứu để xác định các câu hỏi nghiên cứu chưa được giải đáp mà nghiên cứu trong tương lai có thể giải quyết. Hai trăm bốn mươi tám bài báo đáp ứng các tiêu chí bao gồm. 'là nghiên cứu chứng minh rằng công nghệ di động rất hữu ích để hiểu rõ hơn về các loại hành vi giao thông khác nhau. 'ey có thể được phân loại theo thiết kế hệ thống và chủ đề nghiên cứu của chúng: (1) Các ứng dụng điện thoại thông minh trong quy hoạch giao thông và du lịch bền vững đã được nghiên cứu trong một bộ sưu tập các bài báo đáng chú ý. (2) Vì tính di động của mỗi cá nhân đang được đặt ra, nên dữ liệu tín hiệu di động rất nổi bật để xây dựng các mô hình phân tích. (3) Dữ liệu CDR, WiFi và GPS ngày càng được sử dụng nhiều hơn, nhưng tỷ trọng của các kỹ thuật mô hình hóa cho tất cả các hệ thống thông tin di động vẫn còn thấp. Nó cho thấy rằng các nhà thiết kế hệ thống có thể cung cấp các tính năng hấp dẫn và mong muốn hơn trong hầu hết các lĩnh vực. Tuy nhiên, các ứng dụng cho việc di chuyển hàng hóa còn hạn chế, mặc dù việc vận chuyển hàng hóa đã tiến tới số hóa.

Việc phát hiện những điểm bất thường ở đô thị là điều quan trọng hàng đầu đối với công tác quản lý trật tự công cộng, vì chúng có thể gây ra những rủi ro nghiêm trọng đối với an toàn công cộng nếu không được xử lý kịp thời. Tuy nhiên, việc giám sát các khu vực đô thị lớn đòi hỏi các hệ thống phức tạp có thể dẫn đến chi phí tăng

cao. Trong bài báo này [8], nhóm tác giả thảo luận về cơ hội khai thác mạng di động như một nền tảng cảm biến bổ sung để phát hiện các bất thường ở đô thị. Để hỗ trợ khả năng nhận dạng bất thường đáng tin cậy và độ trễ thấp, nhóm nghiên cứu dựa trên kiến trúc Điện toán biên đa truy cập (MEC), cho phép mô tả đặc tính lưu lượng di động chi tiết và sâu gần như trong thời gian thực và cho phép cung cấp dịch vụ đáp ứng hiệu suất, điều này rất quan trọng trong vấn đề của nhóm nghiên cứu.

Với sự xuất hiện của mạng 5G, các mạng di động đang di chuyển theo hướng đa dạng, băng thông rộng, tích hợp và mạng thông minh. Đồng thời, sự phổ biến của các thiết bị đầu cuối thông minh khác nhau đã dẫn đến sự phát triển bùng nổ trong lưu lượng di động. Dự đoán lưu lượng mạng chính xác đã trở thành một phần quan trọng của trí tuệ mạng di động. Trong bối cảnh đó, bài báo này đề xuất một phương pháp học sâu để lập mô hình không-thời gian và dự đoán lưu lượng truyền thông mạng di động. Đầu tiên, nhóm tác giả phân tích các đặc điểm thời gian và không gian của lưu lượng mạng di động từ Telecom Italia. Trên cơ sở này, nhóm tác giả đề xuất một mạng không gian hỗn hợp (HSTNet) [9], là một phương pháp học sâu sử dụng mạng nơ-ron tích tụ để nắm bắt các đặc điểm không gian của lưu lượng truyền thông. Công việc này bổ sung tích chập có thể biến dạng vào mô hình tích chập để cải thiện hiệu suất dự đoán. Thuộc tính thời gian được giới thiệu dưới dạng thông tin bổ trợ. Một cơ chế chú ý dựa trên dữ liệu lịch sử để điều chỉnh trọng lượng được đề xuất để cải thiện độ chắc chắn của mô-đun. Nhóm nghiên cứu sử dụng tập dữ liệu của Telecom Italia để đánh giá hiệu suất của mô hình được đề xuất. Kết quả thử nghiệm cho thấy so với các phương pháp thống kê và thuật toán máy học hiện có, HSTNet đã cải thiện đáng kể độ chính xác của dự đoán dựa trên MAE và RMSE.

Dự báo lưu lượng di động cho phép các nhà khai thác thích ứng với nhu cầu lưu lượng trong thời gian thực để cải thiện việc sử dụng tài nguyên mạng và trải nghiệm người dùng [10]. Để dự đoán lưu lượng di động, các nghiên cứu trước đây đã áp dụng Mạng thần kinh định kỳ (RNN) tại các trạm gốc riêng lẻ hoặc Mạng thần kinh kết hợp (CNN) đã điều chỉnh để hoạt động tại các ô lưới trong một lưới được xác định

theo địa lý. Các giải pháp này không xem xét rõ ràng ảnh hưởng của chuyển giao đối với các đặc điểm không gian của giao thông, điều này có thể dẫn đến độ chính xác của dự đoán thấp hơn. Hơn nữa, các giải pháp RNN chậm được đào tạo, và các giải pháp CNN-lưới không hoạt động cho các tế bào và khó áp dụng cho các trạm gốc. Bài báo này đề xuất một mô hình dự đoán mới, STGCN-HO, sử dụng ma trận xác suất chuyển đổi của biểu đồ chuyển giao để cải thiện dự đoán lưu lượng. STGCN-HO xây dựng cấu trúc mạng nơ-ron dư xếp chồng lên nhau kết hợp các chập đồ thị và các đơn vị tuyến tính có kiểm soát để nắm bắt cả khía cạnh không gian và thời gian của lưu lượng truy cập. Không giống như RNN, STGCN-HO đào tạo nhanh và đồng thời dự đoán nhu cầu giao thông cho tất cả các trạm gốc dựa trên thông tin thu thập được từ toàn bộ biểu đồ. Không giống như CNN-grid, STGCNHO có thể đưa ra dự đoán không chỉ cho các trạm gốc mà còn cho các ô bên trong các trạm gốc. Các thử nghiệm sử dụng dữ liệu từ một nhà khai thác mạng di động lớn chứng minh rằng mô hình của nhóm nghiên cứu vượt trội hơn các giải pháp hiện có về độ chính xác của dự đoán.

Mạng thành phố thông minh bao gồm nhiều ứng dụng đặt ra các yêu cầu về Chất lượng Dịch vụ (QoS) cụ thể, do đó đại diện cho một kịch bản đầy thách thức đối với việc quản lý mạng . Các giải pháp nhằm đảm bảo hỗ trợ QoS chưa được triển khai trong các mạng quy mô lớn. Phân loại lưu lượng là một cơ chế được sử dụng để quản lý các khía cạnh khác nhau, bao gồm cả các yêu cầu QoS. Tuy nhiên, các phương pháp phân loại lưu lượng thông thường, chẳng hạn như phương pháp dựa trên cổng, không hiệu quả vì chúng không có khả năng xử lý phân bổ và mã hóa cổng động. Phân loại lưu lượng bằng cách sử dụng máy học đã thu hút được sự quan tâm nghiên cứu như một phương pháp thay thế để đạt được hiệu suất cao [11]. Trên thực tế, học máy nhúng thông minh vào các chức năng mạng, do đó cải thiện khả năng quản lý mạng. Trong nghiên cứu này, nhóm tác giả áp dụng các thuật toán học máy để dự đoán phân loại lưu lượng mạng. Nhóm tác giả áp dụng bốn thuật toán học có giám sát: SVM, rừng ngẫu nhiên, k-láng giềng gần nhất và cây quyết định. Nhóm tác giả cũng áp dụng phương pháp phân loại lưu lượng dựa trên cổng dựa trên số cổng được

chỉ định phổ biến của ứng dụng. Sau đó, nhóm tác giả so sánh kết quả của phương pháp này với kết quả thu được từ các thuật toán học máy. Kết quả đánh giá chỉ ra rằng thuật toán cây quyết định cung cấp độ chính xác trung bình cao nhất trong số các thuật toán được đánh giá, ở mức 99,18%. Hơn nữa, phân loại lưu lượng mạng bằng cách sử dụng máy học cung cấp kết quả chính xác hơn và hiệu suất cao hơn so với phương pháp dựa trên cổng.

Dự báo lưu lượng trong thành phố là rất quan trọng đối với an toàn công cộng, quản lý giao thông và quy hoạch đô thị. Các nhà nghiên cứu đề xuất một số phương pháp để dự báo luồng đám đông. Tuy nhiên, họ đã bỏ qua việc thu thập dữ liệu luồng toàn diện. Dữ liệu quỹ đạo GPS của taxi và dữ liệu hệ thống chia sẻ xe đạp thường được sử dụng trong các công trình này làm dữ liệu luồng. Nhưng chúng không thể phản ánh toàn diện luồng đám đông trong một thành phố, vì chúng chỉ chứa những chuyển đổi của một phương thức giao thông cụ thể. Trong bài báo này [12], nhóm tác giả đề xuất trích xuất các luồng đám đông toàn diện từ các bản ghi luồng di động (MFR), một dữ liệu di động được tinh chỉnh. Nhóm nghiên cứu cũng sử dụng phương pháp dựa trên Convolution Neural Network (CNN) để dự báo luồng đám đông và so sánh nó với các mô hình hồi quy chuỗi thời gian truyền thống. Các thử nghiệm trên tập dữ liệu di động quy mô lớn cho thấy rằng phương pháp dựa trên CNN có thể giảm sai số từ 28% đến 77%.

Dự đoán lưu lượng di động không dây là một vấn đề quan trọng đối với các nhà nghiên cứu và thực hành trong lĩnh vực 5G / B5G [13]. Tuy nhiên, nó rất khó khăn vì lưu lượng di động không dây thường hiển thị độ phi tuyến tính cao và các mẫu phức tạp. Hầu hết các phương pháp dự đoán lưu lượng di động không dây hiện có, thiếu khả năng mô hình hóa các mối tương quan không gian-thời gian động của dữ liệu lưu lượng di động không dây, do đó không thể mang lại kết quả dự đoán thỏa đáng. Để cải thiện độ chính xác của dự đoán lưu lượng mạng di động 5G / B5G, nhiều dữ liệu miền chéo hơn đã được xem xét, một chiến lược học tập chuyển giao hợp nhất giữa các dịch vụ và khu vực (Fusion-transfer) dựa trên mô hình mạng nơ-ron xuyên

miền theo không gian-thời gian (STC -N) đã được đề xuất. Nhiều bộ dữ liệu tên miền chéo đã được tích hợp. Độ chính xác đào tạo của miền dịch vụ mục tiêu dựa trên các đặc tính dữ liệu của miền dịch vụ nguồn của nó theo sự giống nhau giữa các dịch vụ và sự giống nhau giữa các vùng khác nhau đã được cải thiện, do đó hiệu suất dự đoán của mô hình được nâng cao. Kết quả thử nghiệm cho thấy độ chính xác dự đoán của mô hình dự báo lưu lượng được cải thiện đáng kể sau khi tích hợp nhiều bộ dữ liệu tên miền chéo, hiệu suất RMSE của dịch vụ SMS, Call và Internet có thể được cải thiện lần lượt khoảng 8,39%, 13,76% và 5,7%. Ngoài ra, so với chiến lược chuyển giao hiện tại, RMSE của ba dịch vụ có thể được cải thiện khoảng 2,48% ~13,19%.

Thiếu nghiên cứu về phân tích lưu lượng truy cập của người dùng trong mạng di động, nhóm tác giả gồm Amin Azari và cộng sự quyết định thực hiện một nghiên cứu nhằm mục đích tìm ra và tuân theo việc quản lý mạng để nhận biết lưu lượng [14]. Trên thực tế, phương pháp thiết kế kế thừa, trong đó việc cung cấp tài nguyên và kiểm soát hoạt động được thực hiện dựa trên các kịch bản lưu lượng tổng hợp theo tế bào, không quá tiết kiệm năng lượng và chi phí và cần được thay thế bằng phân tích dự đoán lấy người dùng làm trung tâm về lưu lượng mạng di động và quản lý tài nguyên mạng chủ động. Ở đây, nhóm nghiên cứu làm sáng tỏ vấn đề này bằng cách thiết kế các công cụ dự đoán lưu lượng sử dụng các công cụ máy học tiêu chuẩn (ML), bao gồm bộ nhớ ngắn hạn dài (LSTM) và đường trung bình động tích hợp tự động phục hồi (ARIMA) trên dữ liệu của mỗi người dùng. Nhóm tác giả trình bày một đánh giá thực nghiệm mở rộng về các giải pháp được thiết kế trên tập dữ liệu lưu lượng mạng thực. Trong phân tích này, tác động của các tham số khác nhau, chẳng hạn như mức độ chi tiết về thời gian, độ dài của các dự đoán trong tương lai và lựa chọn đối tượng địa lý được điều tra. Là một ứng dụng tiềm năng của các giải pháp này, nhóm nghiên cứu trình bày sơ đồ Tiếp nhận không liên tục (DRX) được hỗ trợ bởi ML để tiết kiệm năng lượng. Hướng tới mục tiêu này, nhóm nghiên cứu tận dụng các mô hình ML bắt nguồn để điều chỉnh thông số DRX động với lưu lượng truy cập của người dùng. Kết quả đánh giá hiệu suất cho thấy sự vượt trội của LSTM so với ARIMA nói chung, đặc biệt khi độ dài của chuỗi thời gian đào tạo đủ cao và nó được

tăng cường bởi một bộ tính năng được lựa chọn một cách khôn ngoan. Hơn nữa, kết quả cho thấy rằng việc điều chỉnh các thông số DRX bằng cách dự đoán trực tuyến lưu lượng truy cập trong tương lai giúp tiết kiệm năng lượng hơn nhiều với chi phí độ trễ thấp so với việc điều chỉnh thông số DRX trên toàn di động cũ.

Phân loại lưu lượng sẽ là một khía cạnh quan trọng trong hoạt động của mạng di động 5G trong tương lai [15], nơi các dịch vụ có bản chất rất khác nhau sẽ cùng tồn tại. Thật không may, mã hóa dữ liệu làm cho nhiệm vụ này rất khó khăn. Để khắc phục vấn đề này, các sơ đồ dựa trên luồng đã được đề xuất dựa trên các tính năng không phụ thuộc vào trọng tải được trích xuất từ luồng lưu lượng Giao thức Internet (IP). Tuy nhiên, cách tiếp cận như vậy dựa vào việc sử dụng các đầu dò lưu lượng đắt tiền trong mạng lõi. Ngoài ra, trong bài báo này, một phương pháp ngoại tuyến để phân loại lưu lượng được mã hóa trong giao diện vô tuyến được trình bày. Phương pháp này phân chia các kết nối cho mỗi lớp dịch vụ bằng cách chỉ phân tích các tính năng trong các dấu vết kết nối vô tuyến do các trạm gốc thu thập. Với mục đích này, nó dựa vào việc học tập không được giám sát, cụ thể là phân nhóm phân cấp tích tụ. Do đó, nó có thể được áp dụng trong trường hợp không có dữ liệu được dán nhãn (hiếm khi có sẵn trong các mạng di động đang hoạt động). Tương tự như vậy, nó cũng có thể xác định các dịch vụ mới được khởi chạy trong mạng. Đánh giá phương pháp được thực hiện trên một tập dữ liệu theo dõi thực lấy từ mạng Tiến hóa dài hạn (LTE) trực tiếp. Kết quả cho thấy rằng chia sẻ lưu lượng truy cập trên mỗi lớp ứng dụng được ước tính theo phương pháp đề xuất tương tự như những chia sẻ được cung cấp bởi báo cáo của nhà cung cấp.

Dự đoán lưu lượng người dùng trong mạng di động đã thu hút sự chú ý sâu sắc để cải thiện việc sử dụng tài nguyên. Trong bài báo này [17], nhóm tác giả nghiên cứu vấn đề dự đoán và phân loại lưu lượng mạng bằng cách sử dụng phương pháp dự đoán chuỗi thời gian học thống kê và máy học tiêu chuẩn, bao gồm bộ nhớ ngắn hạn dài (LSTM) và đường trung bình động tích hợp tự động phục hồi (ARIMA), tương ứng. Nhóm nghiên cứu trình bày một đánh giá thử nghiệm mở rộng về các công cụ

được thiết kế trên tập dữ liệu lưu lượng mạng thực. Trong phân tích này, nhóm nghiên cứu khám phá tác động của các thông số khác nhau đến hiệu quả của các dự đoán. Nhóm nghiên cứu mở rộng thêm phân tích của mình đến vấn đề phân loại lưu lượng mạng và dự đoán các bùng nổ lưu lượng. Một mặt, kết quả cho thấy hiệu suất vượt trội của LSTM so với ARIMA nói chung, đặc biệt khi độ dài của chuỗi thời gian đào tạo đủ cao và nó được tăng cường bởi một bộ tính năng được lựa chọn một cách khôn ngoan. Mặt khác, kết quả làm sáng tỏ các trường hợp mà ARIMA thực hiện gần với mức tối ưu với độ phức tạp thấp hơn.

2.4 Kết luận chương

Chương 2 đã trình bày một số cơ sở lý thuyết và công trình nghiên cứu liên quan đến đề tài thực hiện. Các nghiên cứu về ứng dụng học máy và các thuật toán đề xuất vào việc khai phá dữ liệu, xây dựng các mô hình huấn luyện để dự báo và phân loại lưu lượng mạng di động, giúp hiểu rõ hơn về quá trình khai phá dữ liệu nói chung cũng như các mô hình học máy và thuật toán nói riêng. Hiểu được các ưu nhược điểm của từng mô hình cũng như thuật toán đã được nêu trong một số công trình nghiên cứu, từ đó luận văn có thể lựa chọn và ứng dụng các thuật toán cũng như mô hình phù hợp vào đề tài.

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 (GBDT). 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.

Thuật toán cây quyết định dựa trên tiếp cận học có gám sát, được sử dụng rộng rãi cho các bài toán phân loại phù hợp với các loại đầu rời rạc, liên tục, chuỗi thời gian, vv… thuật toán này quyết định chia tách toàn bộ tập dữ liệu (root node) thành các tập con (leaf nodes). Độ chính xác của thuật toán cây quyết định dựa trên cấu trúc phân chia. Các quyết định này thường dựa trên một tiếp cận như chỉ số Gini, information entropy, vv… Nhược điểm của cách tiếp cận này là sai lệch và phương sai lớn.

Thuật toán rừng ngẫu nhiên (RF) cho phương sai thấp hơn do cách thức lấy mẫu ngẫu nhiên để chọn ra các tập tập con (leaf nodes) đa dạng và có mức độ ngẫu nhiên tốt hơn. Việc này cho phép cấu trúc của RF nông hơn trong việc ra quyết định. Trong thuật toán này, số lượng đối tượng được xem xét tại một lần tách nhất định xấp xỉ bằng căn bậc hai của tổng số đối tượng dung để phân loại.

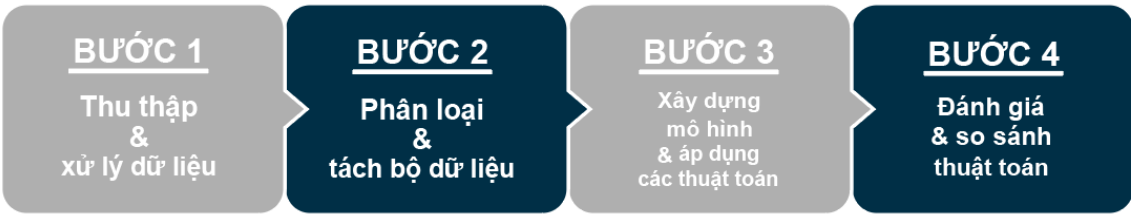
Giống như thuật toán rừng ngẫu nhiên (RF), GBDT là một tập hợp các cây quyết định nhưng khác khác biệt chính là cách các cây được xây dựng. Trong khi RF xây dựng từng cây một cách độc lập, GBDT xây dựng từng cây tại một thời điểm và

gộp các bước này lại. Kết quả là RF kết hợp các kết quả vào cuối quá trình (bằng cách lấy trung bình hoặc "quy tắc đa số") trong khi GBDT kết hợp các kết quả.

Trên thực tế, cả hai mô hình trên là những mô hình học máy được áp dụng rộng rãi trong nhiều lĩnh vực trong khoa học dữ liệu và phân tích dữ liệu. cả hai mô hình diều cho kết quả vượt trội so với các mô hình học máy khác khi sử dụng kiến trúc của thuật toán cây quyết định nhưng sử dụng qui tắc quyết định khác nhau. Thwo hiểu biết của chúng tôi, một phân tích và so sánh hoàn thiện hai kỹ thuật này chưa từng áp dụng với việc phân tích lưu lượng mạng viễn thông với chuỗi dữ liệu biến thiên theo thời gian.

Trong nghiên cứu này, chúng tôi đề xuất nghiên cứu bằng việc xem xét phan tích và so sánh hai kỹ thuật trên áp dụng cho bài toán phân loại lưu lương nhằm nâng cao độ chính xác và giảm thiểu sai lệch và phương sai. Đề xuất chúng tôi xây dựng bao gồm dựa trên lưu lượng tram BTS bao gồm các bước như sau:

- 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.

Bagging là một thuật toán tổng hợp hiệu quả vì mỗi cây quyết định phù hợp với một tập dữ liệu đào tạo hơi khác nhau và do đó có hiệu suất hơi khác nhau. Không giống như các mô hình cây quyết định thông thường, chẳng hạn như cây phân loại và cây hồi quy (CART), các cây được sử dụng trong tập hợp không được cắt tỉa, khiến chúng hơi quá phù hợp với tập dữ liệu đào tạo. Điều này là mong muốn vì nó giúp làm cho mỗi cây khác nhau hơn và có ít dự đoán tương quan hoặc sai số dự đoán. Các dự đoán từ các cây được tính trung bình trên tất cả các cây quyết định dẫn đến hiệu suất tốt hơn bất kỳ cây đơn lẻ nào trong mô hình.

Mỗi mô hình trong tập hợp sau đó được sử dụng để tạo dự đoán cho một mẫu mới và các mẫu dự đoán này được tính trung bình để đưa ra dự đoán của khu rừng. Dự đoán về một bài toán hồi quy là giá trị trung bình của dự đoán trên các cây trong nhóm. Dự đoán về một vấn đề phân loại là đa số phiếu bầu cho nhãn lớp trên các cây trong quần thể.

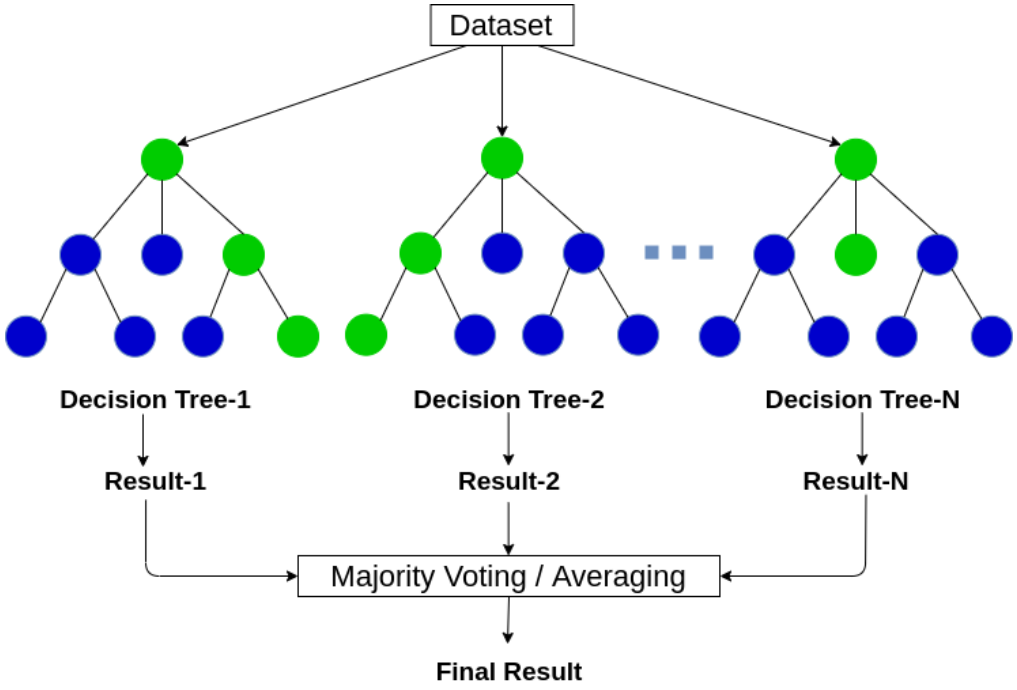
- Regression: Kết quả dự đoán sẽ là kết quả trung bình trên các cây quyết định.
- Classification: Kết quả dự đoán là đa số bình chọn của các lớp chọn làm nhãn được dự đoán trên các cây quyết định.

Giống như với bagging, mỗi cây trong rừng sẽ bỏ phiếu để phân loại một mẫu mới và tỷ lệ phiếu bầu ở mỗi lớp trên toàn bộ quần thể là vectơ xác suất dự đoán.

RF liên quan đến việc xây dựng một số lượng lớn cây quyết định từ các mẫu bootstrap từ tập dữ liệu đào tạo, giống như đóng bao. Không giống như đóng bao, rừng ngẫu nhiên cũng liên quan đến việc lựa chọn một tập hợp con các đặc điểm đầu vào (cột hoặc biến) tại mỗi điểm phân tách trong quá trình xây dựng cây. Thông thường, việc xây dựng cây quyết định bao gồm việc đánh giá giá trị cho từng biến đầu vào trong dữ liệu để chọn điểm phân tách. Bằng cách giảm các tính năng thành một tập hợp con ngẫu nhiên có thể được xem xét tại mỗi điểm phân tách, nó buộc mỗi cây quyết định trong tập hợp phải khác biệt hơn. RF cung cấp một sự cải tiến so với cây đóng bao bằng một chỉnh sửa nhỏ để trang trí tương quan với các cây. Nhưng khi xây dựng các cây quyết định này, mỗi lần phân chia trong cây được xem xét, một mẫu ngẫu nhiên gồm m yếu tố dự đoán được chọn làm ứng viên phân tách từ tập hợp đầy đủ các yếu tố dự đoán *p*.

Các kết quả dự đoán, lỗi dự đoán, được thực hiện bởi mỗi cây trong quần thể có tương quan ít nhiều khác nhau. Khi các dự đoán từ các cây ít tương quan hơn này được lấy trung bình để đưa ra dự đoán, nó thường mang lại hiệu suất tốt hơn so với các cây quyết định được đóng gói. Có lẽ siêu tham số quan trọng nhất để điều chỉnh cho khu rừng ngẫu nhiên là số lượng các đặc trưng ngẫu nhiên cần xem xét tại mỗi điểm phân tách.

RF thu thập các kết quả và tiến hành việc dự đoán dựa trên các cây quyết định để đưa ra lựa chọn tốt nhất cho dữ liệu đầu ra. RF sẽ tách tập dữ liệu thành hai phần: tập dữ liệu dùng để huấn luyện (training dataset) và tập dữ liệu dùng để thử nghiệm (testing dataset). Sau đó thuật toán RF sẽ bắt đầu chọn ngẫu nhiên một số dữ liệu mẫu từ tập train, sử dụng cây quyết định cho mỗi mẫu dữ liệu (chia mẫu dữ liệu thành hai tập con bằng phép chia tốt nhất). Tiếp sau đó, quá trình này sẽ lặp lại cho đến bước cuối cùng và kết quả dự đoán được biểu quyết nhiều nhất sẽ là kết quả cuối cùng.



Hình 3.2: Sơ đồ thuật toán Random Forest

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.

Cây quyết định được tăng cường độ dốc là một phương pháp phổ biến để giải quyết các vấn đề dự đoán trong cả lĩnh vực phân loại và hồi quy. Phương pháp này cải thiện quá trình học tập bằng cách đơn giản hóa mục tiêu và giảm số lần lặp lại để đạt được giải pháp tối ưu đủ. Các mô hình được tăng cường Gradient đã tự chứng minh hết lần này đến lần khác trong các cuộc thi khác nhau về độ chính xác và hiệu quả, khiến chúng trở thành một thành phần cơ bản trong bộ công cụ của nhà khoa học dữ liệu.

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 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:** Thư viện giúp việc thao tác dữ liệu trở nên dễ dàng hơn đó là thư viện Pandas. Thư viện này có một cấu trúc dữ liệu riêng được gọi là Dataframe. Để làm việc trên cấu trúc này, Pandas đã tạo ra nhiều chức năng để xử lý dữ liệu. Pandas cho phép ta đọc/ ghi dữ liệu trên nhiều định dạng file khác nhau: text, excel, csv, … và cho phép ta thay đổi bố cục dữ liệu một cách dễ dàng.
- **Numpy:**
  - Numpy – Numerical Python - là một trong những thư viện của Python. Đây là một thư viện rất hữu dụng trong lĩnh vực Machine learning. Numpy giúp việc tính toán trên các mảng đa chiều trở nên dễ dàng. Không chỉ cung cấp các hàm để phục vụ việc tính toán trên mảng đa chiều mà Numpy còn cung cấp phương thức để tạo mảng đa chiều.
  - Numpy Array cũng tương tự như list có sẵn trong Python. Đối với list, ta có thể lưu nhiều giá trị với các kiểu dữ liệu khác nhau, nhưng với Numpy Array thì các giá trị lưu trữ phải có cùng kiểu dữ liệu. Điều này giúp chúng ta quản lý dữ liệu tốt hơn khi thao tác với các tập dữ liệu lớn. Không những thế, Numpy còn cung cấp nhiều hàm để việc thao tác trên mảng dễ dàng

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).

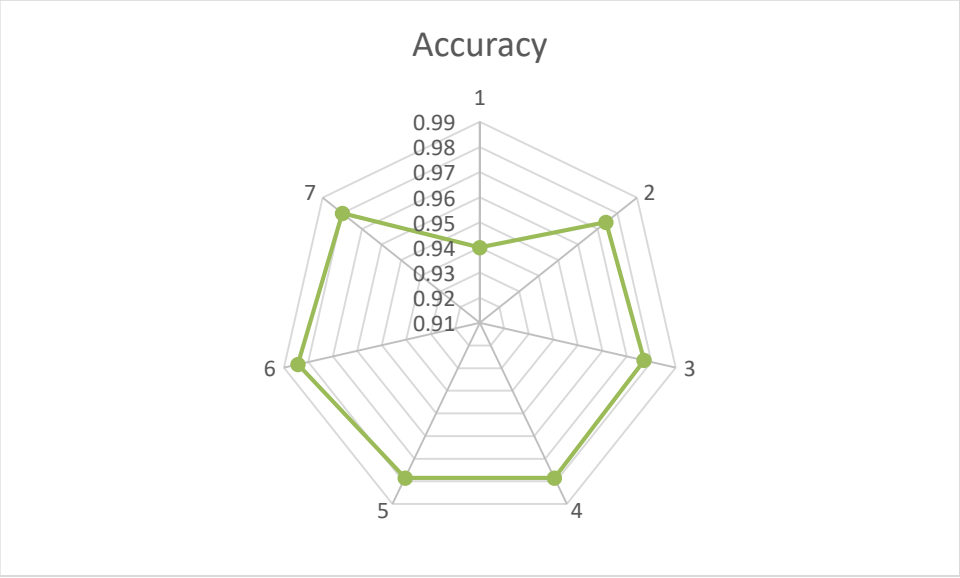
Mô hình DF sử dụng các đặc trưng dạng số, đặc trưng phân loại nguyên bản và các giá trị bị thiếu (missing-values). Các đặc trưng số không cần phải được chuẩn hóa. Các giá trị chuỗi phân loại không cần được mã hóa.

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

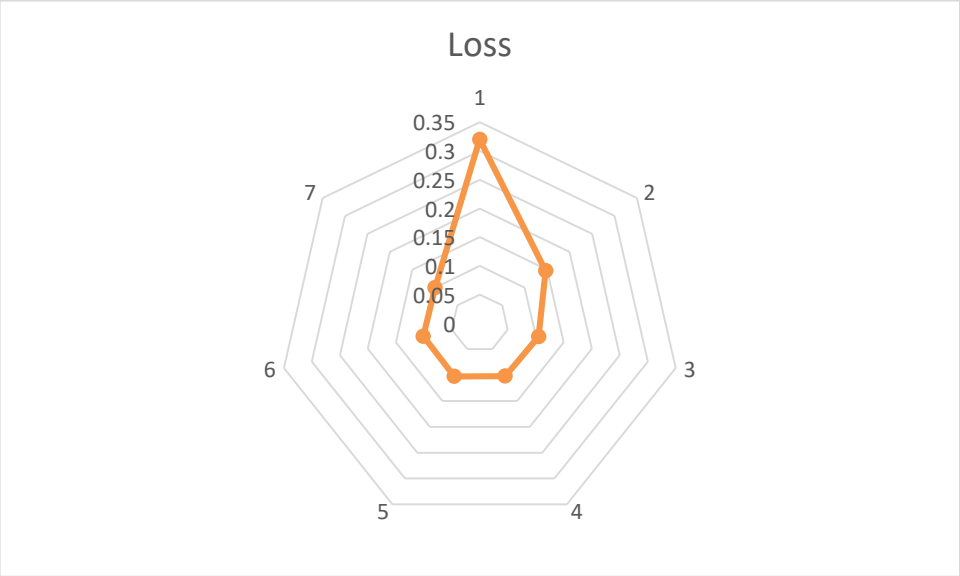
STT	Số cây	Độ chính xác (Accuracy)	Độ mất mát (Loss)
1	1	0.94	0.320204
2	51	0.974212	0.146839
3	101	0.977077	0.105082
4	151	0.97851	0.101884
5	201	0.97851	0.10259
6	251	0.984241	0.101035
7	300	0.979943	0.09969

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%).



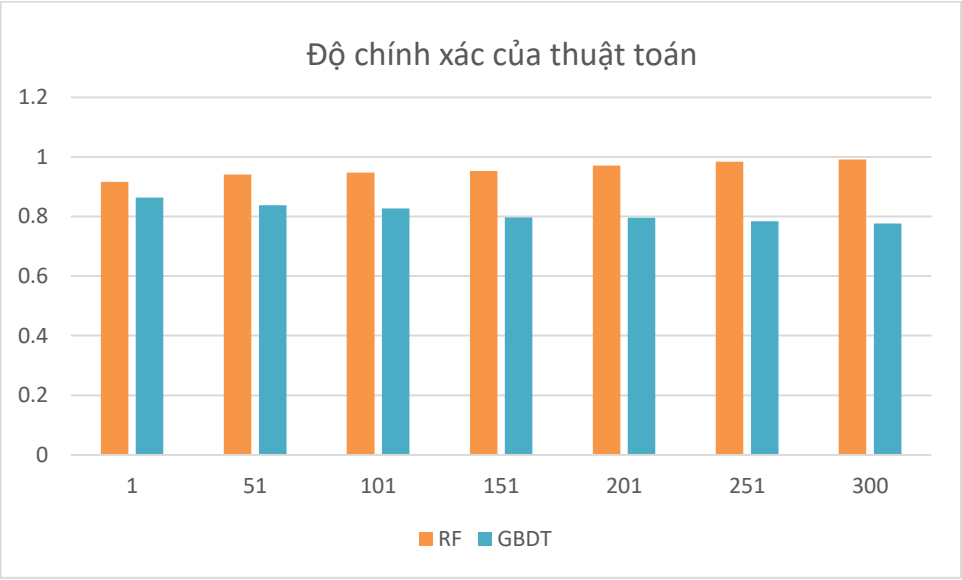
Hình 4.2: Độ mất mát của mô hình RF trong lần thực nghiệm đầu tiên

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. Luận văn cũng đã tập trung nghiên cứu về mô hình Tensorflow-Decision Forest và thuật toán Random Forest. Từ những tìm hiểu này này, luận văn 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, 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.

TÀI LIỆU THAM KHẢO

[1] 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.

[2] 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.

[3] 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.

[4] 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.

[5] 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.

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

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

[8] 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.

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

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

[11] 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.

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

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

[14] 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.

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

[16] 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.

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

[18] Fayez Tarsha Kurdi, Wijdan Amakhchan and Zahra Gharineiat, "Random Forest Machine Learning Technique for Automatic Vegetation Detection and Modelling in LiDAR Data," 04 June 2021.

[19] R K Priyadarshini, Bazila Banu, T Nagamani, "Gradient Boosted Decision Tree based Classification for Recognizing Human Behavior," in *2019 International*

Conference on Advances in Computing and Communication Engineering (ICACCE), 2019.

[20] Udit Narayana Kar, Debarshi Kumar Sanyal, "An overview of device-to-device communication in cellular networks," 9 October 2017.

BẢN CAM ĐOAN

Tôi cam đoan đã thực hiện việc kiểm tra mức độ tương đồng nội dung luận văn/luận án qua phần mềm *Kiểm tra tài liệu* một cách trung thực và đạt kết quả mức độ tương đồng 7% toàn bộ nội dung luận văn/luận án. Bản luận văn/luận án kiểm tra qua phần mềm là bản cứng luận văn đã nộp để bảo vệ trước hội đồng. Nếu sai tôi xin chịu các hình thức kỷ luật theo quy định hiện hành của Học viện.

Tp.HCM, ngày 25 tháng 01 năm 2022

HỌC VIÊN CAO HỌC

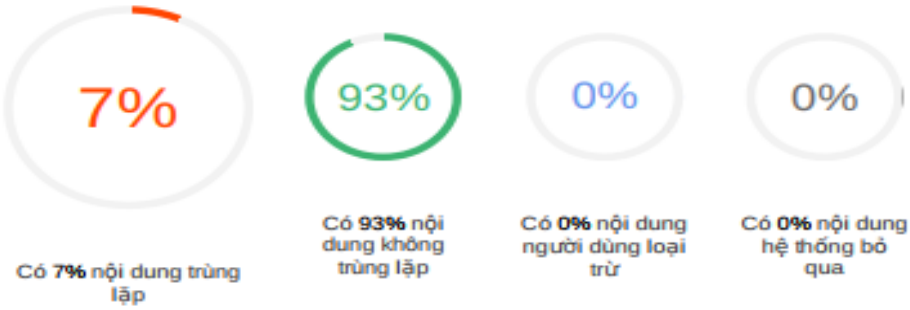
Nguyễn Ngọc Thơ

BÁO CÁO KIỂM TRA TRÙNG LẶP

Thông tin tài liệu

Tên tài liệu:	Hệ hỗ trợ quyết định phân nhóm các trạm BTS theo lưu lượng
Tác giả:	Nguyễn Ngọc Thơ
Điểm trùng lặp:	7
Thời gian tải lên:	16:19 25/01/2022
Thời gian sinh báo cáo:	16:22 25/01/2022
Các trang kiểm tra:	63/63 trang

Kết quả kiểm tra trùng lặp



Nguồn trùng lặp tiêu biểu

123doc.net tailieu.vn vi.wikipedia.org

Học viên	Người hướng dẫn khoa học
Nguyễn Ngọc Thơ	TS. Nguyễn Xuân Sâm