HỌC VIỆN CÔNG NGHỆ BƯU CHÍNH VIỄN THÔNG

Nguyễn Xuân Quốc

NGHIÊN CỨU MÔ HÌNH HỌC MÁY CHO DỰ BÁO LƯU LƯỢNG TRONG MẠNG DI ĐỘNG

Chuyên ngành: HỆ THỐNG THÔNG TIN

Mã số: 8.48.01.04

TÓM TẮT LUẬN VĂN THẠC SĨ

(Theo định hướng ứng dụng)

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. Tính cấp thiết của đề tài

Tên đề tài: Nghiên cứu mô hình học máy cho dự báo lưu lượng trong mạng di động.

Việt Nam đã và đang nỗ lực hết sức để hiện đại hóa và mở rộng mạng lưới viễn thông. Trong nước, việc liên lạc giữa các tỉnh thành đều được số hóa và kết nối với 63/63 tỉnh thành, 705/705 quận/huyện/thị xã, 10.599/10.599 xã/phường/thị trấn thông qua mạng cáp quang hoặc sóng vô tuyến chuyển tiếp. Các đường dây chính được tăng lên đáng kể và việc sử dụng điện thoại di động đang phát triển nhanh chóng. Tính đến tháng 6 năm 2020, Việt Nam có 126,95 triệu thuê bao điện thoại di động, xếp hạng 6 trên toàn thế giới.

Tại Tây Ninh, 3 nhà cung cấp dịch vụ viễn thông lớn là Viettel, mobifone, vinaphone đã phát sóng trên 1154 trạm LTE, phủ sóng đến 9/9 thành phố/thị xã/huyện, 95/95 xã/phường/thị trấn góp phần thúc đẩy kết nối và chia sẻ dữ liệu, phát triển xã hội số.

Hiện tại dịch bệnh covid-19 rất nguy hiểm, một số thời điểm giãn cách xã hội, làm thúc đẩy tăng trưởng lưu lượng (traffic) dữ liệu di động.

Với sự phát triển dịch vụ di động nhanh, các nhà cung cấp viễn thông cần áp dụng công cụ khoa học kỹ thuật như mô hình máy học để thống kê và dự đoán tương đối chính xác sự tăng trưởng, dự đoán dung lượng của nhà cung cấp viễn thông đáp ứng để có kế hoạch phát triển mạng lưới di động phù hợp để vừa đảm bảo chất lượng, không để nghẽn cục bộ, đầu tư hạ tầng được hiệu quả và đáp ứng được chất lượng dịch vụ cho khách hàng với chi phí thấp nhất và hiệu quả nhất.

2. Tổng quan về vấn đề nghiên cứu

Máy học là một lĩnh vực rộng lớn, do đó không có một ngôn ngữ lập trình nào có thể một mình thực hiện mọi việc, do vậy nghiên cứu chủ yếu mô hình LSTM trên nền tảng sử dụng Python để ứng dụng trong dịch vụ mạng di động.

Nghiên cứu mô hình LSTM cho việc phân loại chuỗi dữ liệu theo thời gian ứng dụng trong phân tích dữ liệu mạng di động LTE của một nhà cung cấp dịch vụ trên địa bàn tỉnh Tây Ninh.

3. Mục đích nghiên cứu

Xây dựng, phát triển hệ thống phân tích, quản lý, giám sát hệ thống mạng access LTE dựa trên mô hình LSTM dự đoán sự tăng trưởng lưu lượng của mạng di động để đưa ra Phương án hành động đảm bảo tiến độ và hiệu quả đầu tư cao, chi phí phù hợp.

4. Đối tượng và phạm vi nghiên cứu

Đối tượng nghiên cứu: Mô hình LSTM, các công cụ thu thập, phân tích log và cảnh báo.

Phạm vi nghiên cứu: Xây dựng các rule tăng trưởng của mạng di động, công cụ hỗ trợ phân tích log và cảnh báo hiệu quả cho mạng di động LTE.

5. Phương pháp nghiên cứu

Phương pháp luận: Dựa trên cơ sở lý thuyết về mô hình máy học để xây dựng mối quan hệ mô hình LSTM.

Phương pháp đánh giá dựa trên cơ sở toán học: Trên cơ sở các lý thuyết về mô hình học máy, đề xuất ra thuật toán để dự báo lưu lượng trong mạng di động. Chứng minh thuật toán và đánh giá hiệu quả của thuật toán.

Phương pháp đánh giá bằng mô phỏng thực nghiệm: Xây dựng mô hình mô phỏng và thực nghiệm thuật toán đã đề xuất.

6. Bố cục luận văn

Ngoài phần mở đầu, mục lục, kết luận và tài liệu tham khảo, nội dung chính của luận án được chia thành 4 chương, cụ thể như sau:

Chương 1 trình bày tổng quan về mạng di động.

Chương 2 trình bày cơ sở lý thuyết và các công trình liên quan tới đề tài luận văn.

Chương 3 trình bày đề xuất, nghiên cứu mô hình học sâu cho dự báo lưu lượng trong mạng di động.

Chương 4 trình bày mô phỏng chương trình và đánh giá kết quả thực nghiệm.

CHƯƠNG 1. TỔNG QUAN VỀ ỨNG DỤNG HỌC MÁY PHÂN TÍCH LƯU LƯỢNG MẠNG DI ĐỘNG

1.1 Lưu lượng mạng di động

Mạng điện thoại di động được tạo thành từ một số lượng lớn các khu vực địa lý được gọi là cell (tạm dịch là tế bào). Các cell này được sắp xếp để cung cấp các vùng phủ sóng di động rộng lớn. Trong các cell này là các trạm gốc di động gửi và nhận các tín hiệu vô tuyến đến và từ các thiết bị cầm tay di động được đặt trong các cell đó để cho phép người dùng của họ kết nối với internet và thực hiện cuộc gọi.

Tất cả các trạm gốc này đều được liên kết thông qua mạng truyền dẫn trở lại mạng lõi của nhà cung cấp dịch vụ di động, mạng này quản lý các kết nối giữa khách hàng của mình và những người dùng di động khác cũng như giữa khách hàng của nó với internet.

Các yếu tố quan trọng của lưu lượng di động bao gồm: chất lượng dịch vụ, dung lượng lưu lượng và kích thước cell, hiệu suất phổ và phân vùng, dung lượng lưu lượng so với vùng phủ sóng và phân tích thời gian giữ kênh.

1.1.1 Chất lượng dịch vụ (Quality of Service – QoS)

Tại thời điểm mà các ô của một hệ thống con vô tuyến được thiết kế, các mục tiêu Chất lượng Dịch vụ (QoS) được đặt ra, cho: tắc nghẽn và chặn giao thông, vùng phủ sóng chi phối, C / I, xác suất ngừng hoạt động, tỷ lệ chuyển giao thất bại, tỷ lệ cuộc gọi thành công tổng thể, tốc độ dữ liệu, độ trễ.

1.1.2 Dung lượng lưu lượng và kích thước cell

Càng tạo ra nhiều lưu lượng, càng cần nhiều trạm gốc để phục vụ khách hàng. Số lượng trạm gốc của một mạng di động đơn giản bằng số lượng cell. Kỹ sư giao thông có thể đạt được mục tiêu đáp ứng số lượng khách hàng ngày càng tăng bằng cách tăng số lượng cell trong khu vực liên quan, do đó, điều này cũng sẽ làm tăng số lượng trạm cơ sở. Phương pháp này được gọi là tách tế bào (và kết hợp với

sectorization) là cách duy nhất để cung cấp dịch vụ cho dân số đang phát triển. Điều này chỉ đơn giản hoạt động bằng cách chia các cell đã có sẵn thành các kích thước nhỏ hơn do đó tăng dung lượng lưu lượng. Việc giảm bán kính cell cho phép cell chứa thêm lưu lượng truy cập. Chi phí thiết bị cũng có thể được cắt giảm bằng cách giảm số lượng trạm gốc thông qua việc thiết lập ba cell lân cận, với các cell phục vụ ba cung 120 ° với các nhóm kênh khác nhau.

Mạng vô tuyến di động được vận hành với tài nguyên hữu hạn, hạn chế (phổ tần số có sẵn). Các tài nguyên này phải được sử dụng một cách hiệu quả để đảm bảo rằng tất cả người dùng đều nhận được dịch vụ, tức là chất lượng dịch vụ được duy trì một cách nhất quán. Điều này cần phải sử dụng một cách cẩn thận phổ tần hạn chế, mang lại sự phát triển của các tế bào trong mạng di động, cho phép tái sử dụng tần số bởi các cụm tế bào liên tiếp. Các hệ thống sử dụng hiệu quả phổ có sẵn đã được phát triển, ví dụ: hệ thống GSM. Bernhard Walke định nghĩa hiệu suất phổ là đơn vị dung lượng lưu lượng chia cho tích của phần tử băng thông và diện tích bề mặt, và phụ thuộc vào số kênh vô tuyến trên mỗi cell và kích thước cụm (số cell trong một nhóm cell)

1.1.3 Dung lượng lưu lượng so với vùng phủ sóng

Hệ thống di động sử dụng một hoặc nhiều trong bốn kỹ thuật truy cập khác nhau (TDMA, FDMA, CDMA, SDMA). Xem các khái niệm về Di động. Giả sử một trường hợp Đa truy nhập phân chia theo mã được xem xét cho mối quan hệ giữa dung lượng lưu lượng và vùng phủ sóng (khu vực được bao phủ bởi các ô). Hệ thống di động CDMA có thể cho phép tăng dung lượng lưu lượng với chi phí chất lượng dịch vụ.

1.1.4 Thời gian giữ kênh

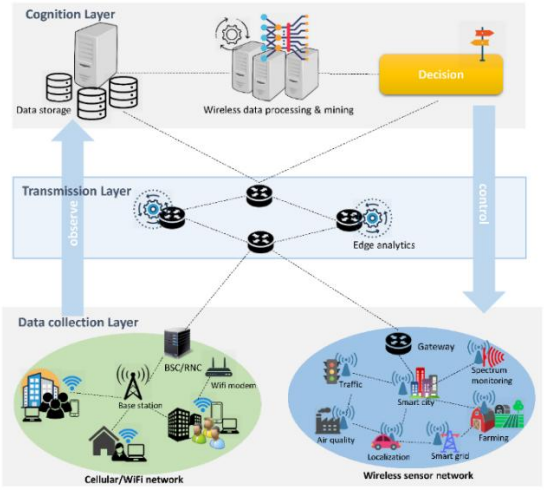
Các thông số quan trọng như tỷ lệ sóng trên nhiễu (C / I), hiệu suất phổ và khoảng cách tái sử dụng xác định chất lượng dịch vụ của mạng di động. Thời gian giữ kênh là một tham số khác có thể ảnh hưởng đến chất lượng dịch vụ trong mạng di động, do đó nó được xem xét khi lập kế hoạch mạng. Tuy nhiên, việc tính toán thời

gian giữ kênh không phải là điều dễ dàng. (Đây là thời gian một Trạm di động (MS) vẫn ở trong cùng một ô trong khi gọi). Do đó, thời gian giữ kênh sẽ nhỏ hơn thời gian giữ cuộc gọi nếu MS di chuyển nhiều hơn một ô vì quá trình chuyển giao sẽ diễn ra và MS từ bỏ kênh. Trên thực tế, không thể xác định chính xác thời gian giữ kênh. Do đó, tồn tại các mô hình khác nhau cho phân phối thời gian giữ kênh. Trong ngành công nghiệp, một ước lượng tốt về thời gian giữ kênh thường đủ để xác định khả năng lưu lượng mạng.

1.2 Ứng dụng học máy trong phân tích lưu lượng

Lưu lượng mạng di động được tạo ở các trạm ngày càng trở nên phức tạp hơn và khó hiểu hơn. Ví dụ: mạng không dây mang lại nhiều chỉ số hiệu suất mạng (ví dụ: tỷ lệ tín hiệu trên nhiễu (SNR), tốc độ truy cập liên kết / tỷ lệ xung đột, tỷ lệ mất gói, tỷ lệ lỗi bit (BER), độ trễ, chỉ báo chất lượng liên kết, thông lượng, năng lượng tiêu thụ, v.v.) và các thông số hoạt động ở các lớp khác nhau của ngăn xếp giao thức mạng (ví dụ: ở lớp PHY: kênh tần số, sơ đồ điều chế, công suất máy phát; ở lớp MAC: lựa chọn giao thức MAC và các tham số của các giao thức MAC cụ thể như CSMA: kích thước cửa sổ tranh chấp, số lượng dự phòng tối đa, số mũ dự phòng; TSCH: trình tự nhảy kênh, v.v.) có tác động đáng kể đến hiệu suất truyền thông.

Việc điều chỉnh các thông số vận hành này và đạt được tối ưu hóa nhiều lớp để tối đa hóa hiệu suất đầu cuối là một nhiệm vụ đầy thách thức. Điều này đặc biệt phức tạp do nhu cầu lưu lượng lớn và tính không đồng nhất của các công nghệ không dây được triển khai. Để giải quyết những thách thức này, học máy (ML) ngày càng được sử dụng nhiều hơn để phát triển các phương pháp tiếp cận nâng cao có thể tự động trích xuất các mẫu và dự đoán xu hướng (ví dụ: ở lớp PHY: nhận dạng giao thoa, ở lớp MAC: dự đoán chất lượng liên kết, ở lớp mạng: ước tính nhu cầu giao thông) dựa trên các phép đo môi trường và các chỉ số hiệu suất làm đầu vào. Các mẫu như vậy có thể được sử dụng để tối ưu hóa cài đặt tham số ở các lớp giao thức khác nhau, ví dụ: PHY, MAC hoặc lớp mạng.



Hình 1.1: Kiến trúc mô hình phân tích dữ liệu lớn của mạng vô tuyến [5]

Với những tiến bộ về phần cứng và sức mạnh tính toán cũng như khả năng thu thập, lưu trữ và xử lý một lượng lớn dữ liệu, học máy (ML) đã dần tiếp cận vào nhiều lĩnh vực khoa học khác nhau. Những thách thức mà mạng không dây và tương lai phải đối mặt cũng thúc đẩy lĩnh vực mạng không dây tìm kiếm các giải pháp sáng tạo để đảm bảo hiệu suất mạng như mong đợi. Để giải quyết những thách thức này, ML ngày càng được sử dụng rộng rãi trong các mạng không dây.

Trong luận văn này sẽ sử dụng thuật toán học máy có giám sát là LSTM (Long short term memory) và phương pháp time series để tiến hành dự báo lưu lượng mang di động dựa vào chuỗi thời gian, hỗ trợ cho việc phát hiện những trạm có lưu lượng quá cao hoặc quá thấp để có những kế hoạch cũng như chiến lược xử lý phù hợp.

1.3 Kết luận chương

Chương một đã giới thiệu và trình bày sơ lược về mạng di động, lưu lượng mạng cũng như các trạm thu phát và quản lý mạng di động. Ngoài ra, các khái niệm liên quan đến học máy và sự ảnh hưởng của học máy đến nhiều lĩnh vực khác nhau trong đó mạng di động là một trong những lĩnh vực có tiềm năng để có thể áp dụng các kĩ thuật liên quan đến học máy, nhằm cải thiện chất lượng và nâng cao dịch vụ.

CHƯƠNG 2. CƠ SỞ LÝ THUYẾT VÀ CÁC CÔNG TRÌNH LIÊN QUAN

2.1 Cơ sở lý thuyết về học máy

2.1.1 Giới thiệu học máy

Học máy (ML) là một loại trí tuệ nhân tạo (AI) cho phép các ứng dụng phần mềm trở nên chính xác hơn trong việc dự đoán kết quả mà không cần được lập trình rõ ràng để làm như vậy. Các thuật toán học máy sử dụng dữ liệu lịch sử làm đầu vào để dự đoán các giá trị đầu ra mới.

Học máy thường được phân loại theo cách một thuật toán học để trở nên chính xác hơn trong các dự đoán của nó. Có bốn cách tiếp cận cơ bản: học có giám sát, học không giám sát, học bán giám sát và học tăng cường.

2.1.1.1 Học có giám sát (Supervised learning)

Trong loại học máy này, các nhà khoa học dữ liệu cung cấp các thuật toán với dữ liệu huấn luyện được gắn nhãn và xác định các biến mà họ muốn thuật toán đánh giá về các mối tương quan. Cả đầu vào và đầu ra của thuật toán đều được chỉ định.

Để giải quyết một vấn đề nhất định về học có giám sát, người ta phải thực hiện các bước sau:

Bước 1: Xác định loại ví dụ đào tạo. Trước khi làm bất cứ điều gì khác, người dùng nên quyết định loại dữ liệu nào sẽ được sử dụng làm tập huấn luyện. Ví dụ, trong trường hợp phân tích chữ viết tay, đây có thể là một ký tự viết tay đơn lẻ, toàn bộ từ viết tay, toàn bộ câu chữ viết tay hoặc có thể là một đoạn văn viết tay đầy đủ.

Bước 2: Tập hợp một tập hợp đào tạo. Tập huấn luyện cần phải đại diện cho việc sử dụng hàm trong thế giới thực. Do đó, một tập hợp các đối tượng đầu vào được tập hợp và các đầu ra tương ứng cũng được thu thập, từ các chuyên gia con người hoặc từ các phép đo.

Bước 3: Xác định biểu diễn đặc điểm đầu vào của hàm đã học. Độ chính xác của hàm đã học phụ thuộc nhiều vào cách biểu diễn đối tượng đầu vào. Thông thường, đối tượng đầu vào được chuyển đổi thành một vectơ đặc trưng, chứa một số đặc điểm mô tả đối tượng. Số lượng các đối tượng địa lý không được quá lớn, vì điều này có thể xảy ra; nhưng phải chứa đủ thông tin để dự đoán chính xác kết quả đầu ra.

Bước 4: Xác định cấu trúc của hàm đã học và thuật toán học tương ứng. Ví dụ, kỹ sư có thể chọn sử dụng máy vectơ hỗ trợ hoặc cây quyết định.

Bước 5: Hoàn thiện thiết kế. Chạy thuật toán học tập trên tập huấn luyện đã tập hợp. Một số thuật toán học có giám sát yêu cầu người dùng xác định các thông số điều khiển nhất định. Các tham số này có thể được điều chỉnh bằng cách tối ưu hóa hiệu suất trên một tập hợp con (được gọi là tập xác nhận) của tập huấn luyện hoặc thông qua xác nhận chéo.

Bước 6: Đánh giá độ chính xác của hàm đã học. Sau khi điều chỉnh tham số và học hỏi, hiệu suất của chức năng kết quả phải được đo trên một bộ thử nghiệm tách biệt với bộ huấn luyện

2.1.1.2 Học không giám sát (Unsupervised learning)

Loại học máy này liên quan đến các thuật toán đào tạo trên dữ liệu không được gắn nhãn. Thuật toán quét qua các tập dữ liệu để tìm kiếm bất kỳ kết nối có ý nghĩa nào. Dữ liệu mà các thuật toán đào tạo cũng như các dự đoán hoặc khuyến nghị mà chúng xuất ra được xác định trước.

2.1.1.3 Học bán giám sát (Semi-supervised learning)

Cách tiếp cận này đối với học máy liên quan đến sự kết hợp của hai loại trước đó. Các nhà khoa học dữ liệu có thể cung cấp một thuật toán chủ yếu là dữ liệu đào tạo được gắn nhãn, nhưng mô hình có thể tự do khám phá dữ liệu và phát triển sự hiểu biết của riêng mình về tập dữ liệu.

2.1.1.4 Học tăng cường (Reinforcement learning)

Các nhà khoa học dữ liệu thường sử dụng học tăng cường để dạy máy hoàn thành một quy trình gồm nhiều bước trong đó có các quy tắc được xác định rõ ràng.

Các nhà khoa học dữ liệu lập trình một thuật toán để hoàn thành một nhiệm vụ và cung cấp cho nó các tín hiệu tích cực hoặc tiêu cực khi nó tìm ra cách hoàn thành một nhiệm vụ. Nhưng phần lớn, thuật toán tự quyết định những bước cần thực hiện trong quá trình thực hiện.

2.1.2 Các thuật toán học máy

Có rất nhiều thuật toán được sử dụng trong học máy, tuy nhiên ở phạm vi của đề tài nghiên cứu cũng như lĩnh vực liên quan đến mạng di động, một số thuật toán thường được sử dụng trong lĩnh vực này được bài báo [1] liệt kê như sau: Hồi quy (Linear Regression), Cây quyết định (Decision Tree), Rừng ngẫu nhiên (RF), Support Vector Machine (SVM), KNN (k nearest neighbors), K-Means, Mạng thần kinh (Neural Networks)

2.2 Kỹ thuật phân tích và dự báo theo chuỗi thời gian

Phân tích chuỗi thời gian là một cách cụ thể để phân tích một chuỗi các điểm dữ liệu được thu thập trong một khoảng thời gian. Trong phân tích chuỗi thời gian, các nhà phân tích ghi lại các điểm dữ liệu theo các khoảng thời gian nhất quán trong một khoảng thời gian nhất định thay vì chỉ ghi các điểm dữ liệu một cách gián đoạn hoặc ngẫu nhiên. Tuy nhiên, loại phân tích này không chỉ đơn thuần là hành động thu thập dữ liệu theo thời gian. Điều làm cho dữ liệu chuỗi thời gian khác biệt với các dữ liệu khác là phân tích có thể cho thấy các biến thay đổi như thế nào theo thời gian.

Nói cách khác, thời gian là một biến quan trọng vì nó cho thấy cách dữ liệu điều chỉnh trong quá trình của các điểm dữ liệu cũng như kết quả cuối cùng. Nó cung cấp một nguồn thông tin bổ sung và một thứ tự phụ thuộc giữa các dữ liệu. Phân tích chuỗi thời gian thường yêu cầu một số lượng lớn các điểm dữ liệu để đảm bảo tính nhất quán và độ tin cậy. Tập dữ liệu mở rộng đảm bảo bạn có cỡ mẫu đại diện và phân tích có thể cắt bỏ dữ liệu nhiễu. Nó cũng đảm bảo rằng bất kỳ xu hướng hoặc kiểu mẫu nào được phát hiện không phải là ngoại lệ và có thể giải thích cho phương

sai theo mùa. Ngoài ra, dữ liệu chuỗi thời gian có thể được sử dụng để dự báo — dự đoán dữ liệu trong tương lai dựa trên dữ liệu lịch sử.

2.3 Các tiêu chuẩn đánh giá

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)

$$\text{MAPE} = \frac{1}{N}\sum_{t=1}^{N}\frac{|Z(t)-\hat{Z}(t)|}{Z(t)}.100\% \qquad (2.38)$$

- Root Mean squared error (RMSE)

$$\text{RMSE} = \sqrt{\frac{1}{N}\sum_{t=1}^{N}(Z(t)-\hat{Z}(t))^2} \qquad (2.39)$$

- Mean squared error (MSE)

$$\text{MSE} = \frac{1}{N}\sum_{t=1}^{N}(Z(t)-\hat{Z}(t))^2 \qquad (2.40)$$

- Mean absolute error (MAE)

$$\text{MAE} = \frac{1}{N}\sum_{t=1}^{N}\left|Z(t)-\hat{Z}(t)\right| \qquad (2.41)$$

- Sum of squared errors (SSE)

$$\text{SSE} = \sum_{t=1}^{N}(Z(t)-\hat{Z}(t))^2 \qquad (2.42)$$

Mức độ phù hợp của các độ đo MSE, RMSE, MAE và SSE là khá giống nhau. Chúng chỉ khác nhau một chút, ví dụ các lỗi sai số RMSE thì ít hơn các độ đo khác. MAE và RMSE đại diện cho một thước đo phụ thuộc vào quy mô, trong khi những độ đo khác không phụ thuộc vào quy mô. Tất cả các tiêu chuẩn đánh giá này phù hợp để so sánh các phương pháp dự báo khác nhau trên cùng một dữ liệu thử nghiệm.

2.4 Một số nghiên cứu liên quan

- Fabio Ricciato, Peter Widhalm, Francesco Pantisano, Massimo Craglia, "Beyond the “single-operator, CDR-only” paradigm: An interoperable framework for mobile phone network data analyses and population density estimation," *Pervasive and Mobile Computing*, 2017.

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

- 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.

- J. Brownlee, "A Gentle Introduction to Long Short-Term Memory Networks by the Experts," Machine Learning Mastery, 24 May 2017. [Online]. Available: https://machinelearningmastery.com/gentle-introduction-long-short-term-memory-networks-experts/.

- TensorFlow, "TensorFlow," [Online]. Available: https://www.tensorflow.org/.

- 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.

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

- Rahul Awati, "TechTarget," June 2021. [Online]. Available: https://www.techtarget.com/searchnetworking/definition/TDMA.

CHƯƠNG 3. NGHIÊN CỨU MÔ HÌNH HỌC MÁY CHO DỰ BÁO LƯU LƯỢNG TRONG MẠNG DI ĐỘNG

3.1 Phương pháp Time Series

Time Series (tạm dịch là Dự báo chuỗi thời gian) là quá trình phân tích dữ liệu chuỗi thời gian sử dụng số liệu thống kê và mô hình hóa để đưa ra dự đoán và thông báo cho việc ra quyết định chiến lược. Nó không phải lúc nào cũng là một dự đoán chính xác và khả năng dự báo có thể rất khác nhau - đặc biệt là khi xử lý các biến thường xuyên dao động trong dữ liệu chuỗi thời gian cũng như các yếu tố nằm ngoài tầm kiểm soát của chúng tôi. Tuy nhiên, dự báo cái nhìn sâu sắc về kết quả nào có nhiều khả năng - hoặc ít khả năng hơn - xảy ra hơn các kết quả tiềm năng khác.

3.2 Thuật toán LSTM

Long Short-Term Memory (LSTM) [19] là một mô hình được đề xuất vào năm 1997 và nó chính là một loại mạng đặc biệt của RNN. Đặc điểm chính là các mạng đó có thể lưu trữ thông tin có thể được sử dụng cho quá trình xử lý cell (tạm dịch là tế bào) trong tương lai. LSTM hoạt động rất tốt trên nhiều vấn đề và hiện đang được sử dụng rộng rãi. LSTM được thiết kế rõ ràng để tránh vấn đề phụ thuộc lâu dài. Ghi nhớ thông tin trong thời gian dài thực tế là hành vi mặc định của chúng, không phải là thứ mà chúng phải vật lộn để học! Tất cả các mạng RNN đều có dạng một chuỗi các môđun lặp lại của mạng nơron. Trong các RNN tiêu chuẩn, mô-đun lặp này sẽ có cấu trúc rất đơn giản, chẳng hạn như một lớp tanh.

3.3 Áp dụng LSTM vào dự báo lưu lượng mạng di động

Xây dựng đề xuất dự báo lưu lượng mạng di động sử dụng LSTM gồm các bước sau:

- Bước 1: Thu thập và xử lý dữ liệu mạng di động, phân lớp dữ liệu dựa trên tính chất lưu lượng của mạng di động. Dữ liệu thu được bao gồm 11 đặc trưng như PS Traffic Total (GB), DL User Thput (Mbps), AE-RAB attempt, … Sau đó trường dữ liệu về thời gian PERIOD_START_TIME sẽ được lấy ra và định dạng lại cho phù hợp với mô hình.

- Bước 2: Tiến hành tách bộ dữ liệu thành các tập training, testing và validation với tỉ lệ 70%, 20%, 10% tương ứng. Sau đó thiết lập các khung chuỗi thời gian và xây dựng mô hình dự báo với thuật toán LSTM kết hợp với bộ dữ liệu đã có.
- Bước 3: Xây dựng mô hình LSTM, chạy thực nghiệm với bộ dữ liệu.
- Bước 4: Hiệu chỉnh các thông số và các đặc trưng từ bộ dữ liệu để tối ưu hiệu năng của mô hình.
- Bước 5: Quan sát kết quả thu được và độ đo mất mát để đánh giá hiệu quả mô hình và khả năng áp dụng mô hình.

CHƯƠNG 4. MÔ PHỎNG CHƯƠNG TRÌNH VÀ ĐÁNH GIÁ KẾT QUẢ

4.1 Môi trường và bộ dữ liệu thực nghiệm

4.1.1 Môi trường thực nghiệm

Dựa vào dữ liệu về lưu lượng mạng di động mà ta có thể biết, ta sử dụng thuật toán LSTM để kết nối giữa các nút tạo thành một đồ thị có hướng hoặc vô hướng dọc theo một trình tự thời gian. Kết hợp với đánh giá số lần sai, và sai số, ta cải thiện thuật toán bằng cách áp dụng máy học vào, tuy nhiên, việc áp dụng này sẽ ít diễn ra vì có sai số cho phép.

Để xây dựng và thực nghiệm được mô hình, luận văn này đã kết hợp sử dụng các thư viện mã nguồn mở và các công cụ tự xây dựng để xử lý dữ liệu, huấn luyện mô hình và dự báo:

Google Colab [20]: Là một môi trường dành cho việc nghiên cứu từ google, cho phép viết code và thực thi bất kì đoạn code python nào trực tiếp trên trình duyệt web, đặc biệt là phù hợp với việc phân tích dữ liệu, xây dựng mô hình học máy và mục đích giáo dục. Về mặt kỹ thuật, Colab là một dịch vụ máy tính xách tay Jupyter được lưu trữ không yêu cầu thiết lập để sử dụng, đồng thời cung cấp quyền truy cập miễn phí vào các tài nguyên máy tính bao gồm GPU.

STensorflow [21]: Một khung làm việc mã nguồn mở, do Google phát hành, được sử dụng để xây dựng các mô hình học máy, tạo môi trường nghiên cứu, thực hiện các thử nghiệm một cách nhanh chóng và dễ dàng, đặc biệt là có khả năng chuyển đổi các bản thiết kế prototype tới các ứng dụng trong sản xuất.

4.1.2 Dữ liệu thực nghiệm

Bộ dữ liệu thực nghiệm với hơn hai triệu dòng và gồm nhiều trường dữ liệu liên quan đến lưu lượng mạng di động. Tuy nhiên, dựa vào các phạm vi của đề tài, bộ dữ liệu đã lọc và lựa chọn lại với các trường dữ liệu phù hợp.

4.2 Thực nghiệm và kết quả thực nghiệm của mô hình

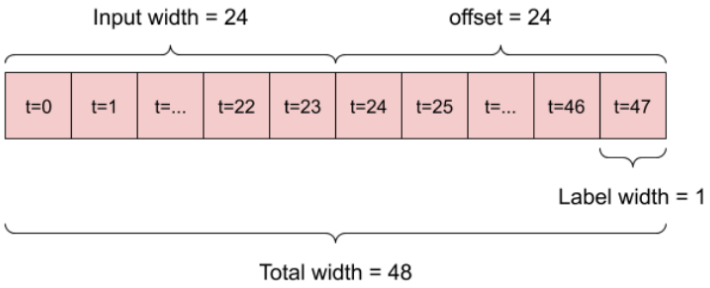
Trước khi bắt đầu xây dựng mô hình, điều quan trọng là phải hiểu dữ liệu và đảm bảo rằng việc đưa các trường và điều chỉnh số chiều dữ liệu được định dạng phù hợp cho mô hình, hay có thể nói cách khác đây chính là bước Feature Engineering. Trường dữ liệu về thời gian là một trong các đặc trưng quan trọng cho mô hình, tuy nhiên để phù hợp với mô hình thì dữ liệu thời gian dạng chuỗi cần được biến đổi thành đơn vị giây. Bên cạnh đó, để có kết quả đánh giá khách quan hơn, bộ dữ liệu sẽ được đánh nhãn theo trường "PS Traffic Total (GB)" (trường dữ liệu về lưu lượng của cell) và chia thành 3 tập dữ liệu con theo nhãn tương ứng lần lượt là A, B và C. Dựa trên thông tin tóm tắt dữ liệu của bảng 4.1, tiến hành đánh nhãn theo qui tắc sau: lưu lượng cell có dung lượng từ dưới 1.65 GB (<= 1.65) sẽ có nhãn là A, dung lượng từ dưới 3.81 GB (<= 3.81) có nhãn là B và còn lại là C.

Bước tiếp theo cũng quan trọng không kém chính là chuẩn hóa dữ liệu. Điều quan trọng là phải mở rộng quy mô các đặc trưng trước khi huấn luyện mô hình mạng nơ-ron. Chuẩn hóa là một cách phổ biến để thực hiện việc chia tỷ lệ: trừ giá trị trung bình và chia cho độ lệch chuẩn của mỗi đặc trưng. Giá trị trung bình và độ lệch chuẩn chỉ nên được tính bằng cách sử dụng dữ liệu huấn luyện để các mô hình không có quyền truy cập vào các giá trị trong bộ kiểm tra và xác nhận. Cũng có thể cho rằng mô hình không nên có quyền truy cập vào các giá trị tương lai trong tập huấn luyện khi huấn luyện và việc chuẩn hóa này nên được thực hiện bằng cách sử dụng các đường trung bình động.

Sau khi chuẩn hóa dữ liệu, tiến hành tách bộ dữ liệu thành các tập huấn luyện (training set), tập thử nghiệm trong quá trình huấn luyện (validation set) và tập thử nghiệm sau quá trình huấn luyện (testing set) tương ứng với các mức 70%, 20% và 10%. Việc chia dữ liệu như vậy nhằm đảm bảo rằng kết quả kiểm thử trong và sau quá trình huấn luyện thực tế hơn, được đánh giá dựa trên dữ liệu thu thập được sau khi mô hình được huấn luyện.

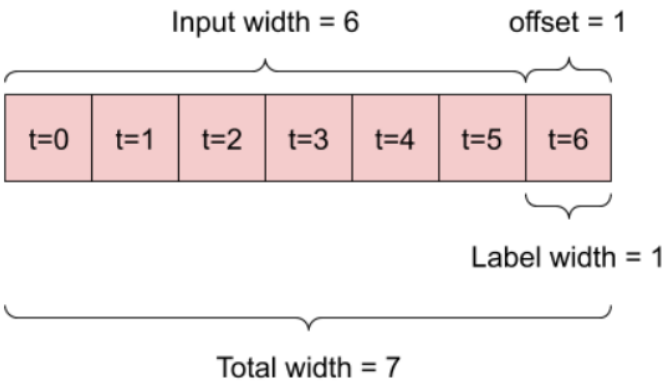
Một trong những bước xử lý dữ liệu và cài đặt cuối cùng đó chính là thiết lập khung dữ liệu dự đoán thời gian chuỗi (kiểu dữ liệu liên tục) cho mô hình. Các đặc trưng chính cho khung là: chiều rộng (số lượng các bước nhảy trong thời gian) đầu vào và nhãn của khung, thời gian bù giữa các khung và các loại đặc trưng như dữ liệu đầu vào, nhãn hoặc cả hai. Ở bước này sẽ tập trung vào hoàn tất việc thiết lập khung thời gian, vì thế nó có thể được tái sử dụng nhiều lần ở các mô hình khác nhau trong quá trình thực nghiệm. Dựa vào từng nhiệm vụ và loại mô hình, có thể tạo ra các đa dạng các khung thời gian.

- Ví dụ như để đưa ra một dự đoán duy nhất trong vòng 24h tới trong tương lai, giả sử đã ghi nhận 24h lịch sử:



Hình 4.1: Khung thời gian 48h với offset là 24

- Mô hình dự đoán 1h trong tương lai, giả sử đã ghi nhận 6h lịch sử:



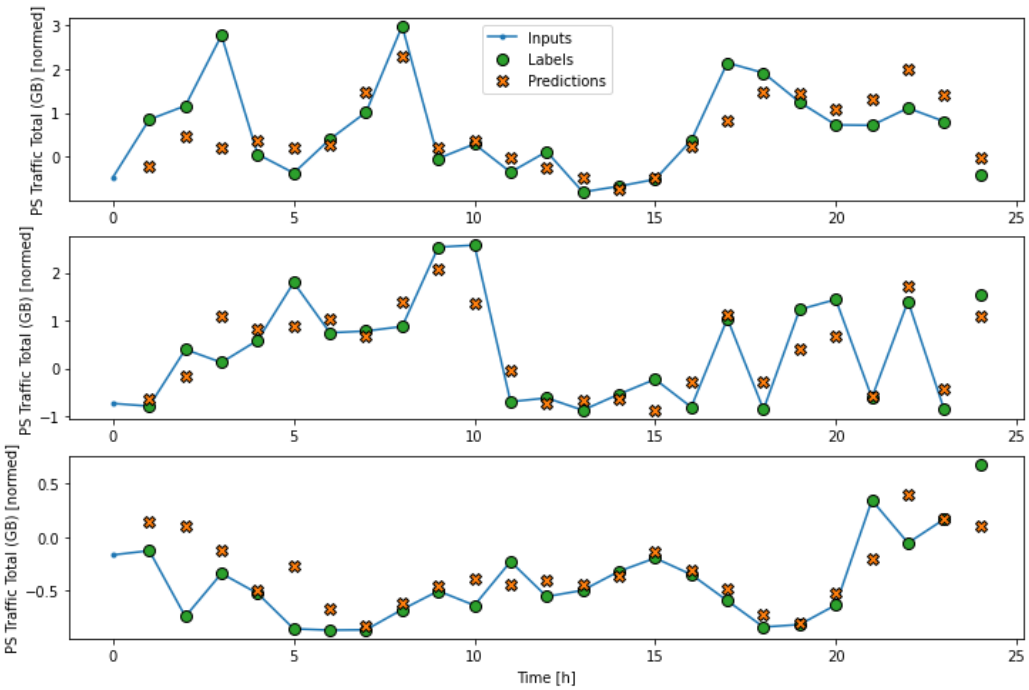
Hình 4.2: Khung thời gian 6h với offset là 1

Sau khi hoàn tất các bước thiết lập cho mô hình, tiến hành huấn luyện mô hình và thu được các kết quả như sau:

- **Tập dữ liệu có nhãn A**: có kích thước gồm 1075250 dòng dữ liệu với các trường dữ liệu khác nhau, trong đó trường dữ liệu "PS Traffic Total (GB)" có lưu lượng từ dưới 1.65 GB.
- **Tập dữ liệu có nhãn B**: có kích thước gồm 537589 dòng dữ liệu với các trường dữ liệu khác nhau, trong đó trường dữ liệu "PS Traffic Total (GB)" có lưu lượng từ dưới 3.81 GB.
- **Tập dữ liệu có nhãn C**: có kích thước gồm 535979 dòng dữ liệu với các trường dữ liệu khác nhau, trong đó trường dữ liệu "PS Traffic Total (GB)" có lưu lượng từ dưới 47.37 GB.

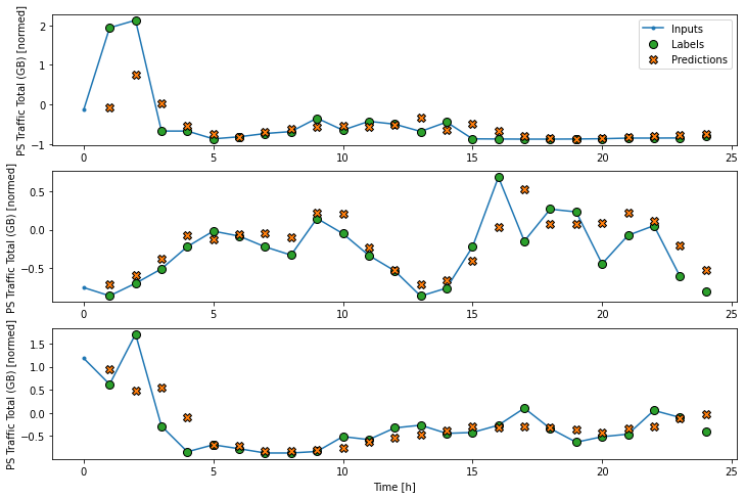
Trong các lần thử nghiệm, mô hình LSTM với thời gian huấn luyện là 24 giờ tại một thời điểm.

Kết quả khi sử dụng huấn luyện với tập dữ liệu có nhãn A



Hình 4.3: Mô hình tập dữ liệu nhãn A với độ đo MAE

Thử nghiệm với tập dữ liệu nhãn A cho kết quả dự báo tương đối, tuy nhiên vẫn còn nhiều điểm dữ liệu dự báo thưa thớt, chưa đạt kết quả cao. Kết quả được đánh giá dựa trên độ đo mất mát Mean Absolute Error – MAE, tuy nhiên vì hiệu quả chưa đạt được như mong muốn, nên độ đo được thay đổi thành Mean Squared Error - MSE và Mean Squared Logarithmic Error – MAPE, thu được kết quả như sau:



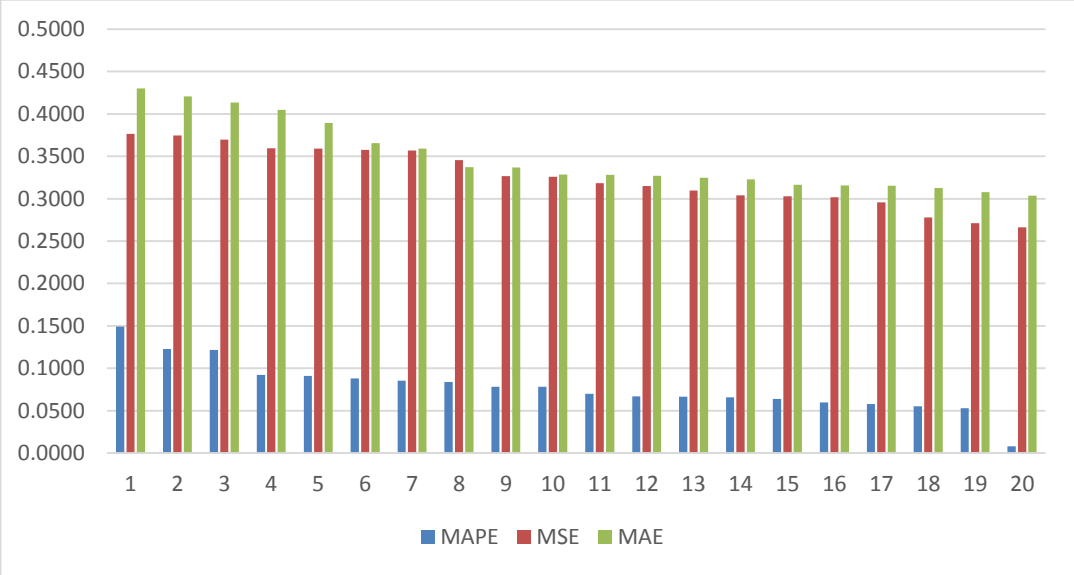
Hình 4.4: Mô hình tập dữ liệu nhãn A với độ đo MAPE

Bảng 4.1: So sánh các độ đo mất mát của tập A

Epoch	Độ đo mất mát (Loss)		
	MAPE	**MSE**	**MAE**
1	0.1491	0.3765	0.4301
2	0.1227	0.3745	0.4207
3	0.1216	0.3695	0.4133
4	0.0921	0.3596	0.4048
5	0.0910	0.3590	0.3893
6	0.0879	0.3576	0.3657
7	0.0852	0.3567	0.3591
8	0.0839	0.3455	0.3371

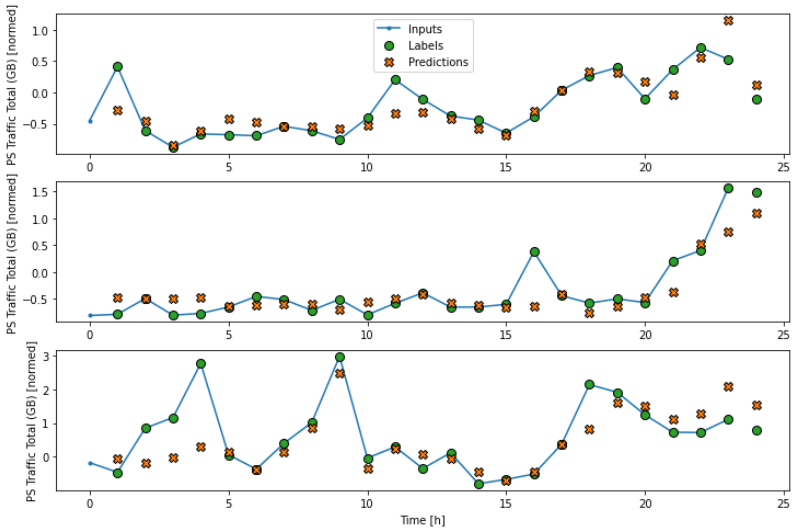
Epoch	Độ đo mất mát (Loss)		
	MAPE	**MSE**	**MAE**
9	0.0784	0.3265	0.3367
10	0.0783	0.3258	0.3284
11	0.0698	0.3182	0.3281
12	0.0671	0.3149	0.3270
13	0.0667	0.3097	0.3248
14	0.0657	0.3038	0.3227
15	0.0638	0.3029	0.3163
16	0.0599	0.3018	0.3156
17	0.0580	0.2956	0.3154
18	0.0552	0.2780	0.3128
19	0.0529	0.2713	0.3078
20	0.0080	0.2661	0.3036

Từ bảng 4.1, ta có thể thấy độ đo MAPE cho kết quả đánh giá mô hình tốt hơn hai độ đo còn lại, nhiều điểm dữ liệu dự báo gần với nhãn hơn, nhìn chung đạt được các mục tiêu đề ra.

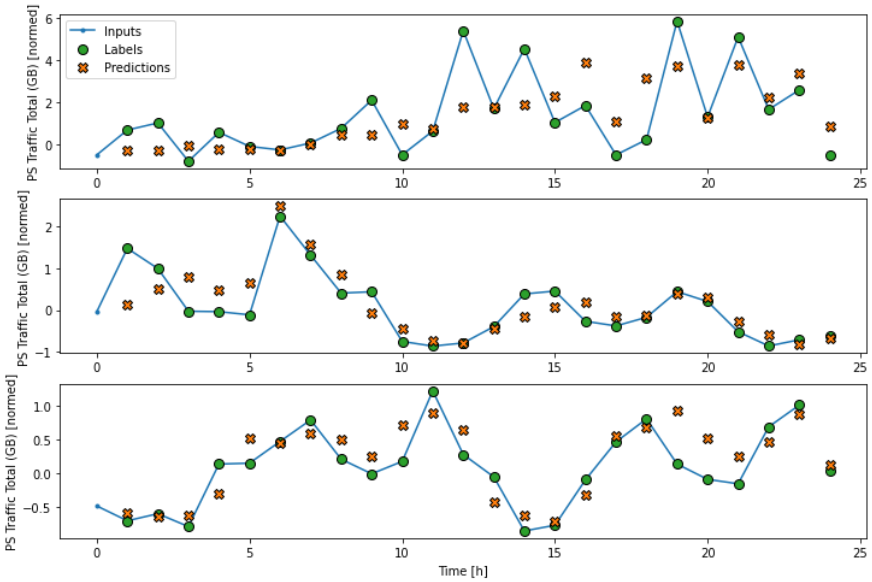


Hình 4.5: Biểu đồ so sánh độ đo mất mát tập dữ liệu A

Tương tự như vậy ta tiến hành thực nghiệm lại trên tập dữ liệu có nhãn B và C tương đương với mức lưu lượng lần lượt là <= 3.68 GB và <= 47.37 GB.



Hình 4.6: Mô hình tập dữ liệu nhãn B với độ đo MAPE



Hình 4.7: Mô hình tập dữ liệu nhãn C với độ đo MAPE

Qua hai lần thực nghiệm ở bộ dữ liệu B và C, các kết qua thu được tương đối ổn định ở một số điểm dữ liệu dự báo. Bên cạnh đó vẫn còn nhiều điểm dữ liệu dự báo chưa chính xác, cho thấy được rằng khi mức lưu lượng sử dụng mạng di động càng cao thì hiệu quả dự báo của mô hình chưa đạt được các kết quả như mong muốn.

KẾT LUẬN

1. Kết quả nghiên cứu của đề tài

Thông qua đề tài *"Nghiên cứu mô hình học máy cho dự báo lưu lượng trong mạng di động"*, luận văn đã đề xuất và thực nghiệm được mô hình dự đoán lưu lượng mạng di động dựa trên dữ liệu người dùng thực tế. Mô hình và kết quả nghiên cứu đã đạt được hiệu suất và khả năng dự báo tốt về lưu lượng sử dụng, từ đó giúp cho nhà mạng quản lý và kiểm soát tốt hạ tầng mạng viễn thông. Dựa vào mô hình dự báo của luận văn này, nhà mạng có thể áp dụng đưa ra khuyến nghị thời điểm nâng/hạ cấp mạng lưới để đảm bảo được tài nguyên được sử dụng tài nguyên một cách hiệu quả nhất, nhất là các thời điểm dị biệt của mạng di động LTE như mất điện, lễ hội sự kiện, ngày khuyến mãi.

2. Hạn chế của luận văn

Hầu hết các giải thuật khai phá dữ liệu chuỗi thời gian thường đòi hỏi phải xác định giá trị một số thông số đầu vào và việc xác định các thông số này thường không dễ dàng đối với người nghiên cứu. Việc xác định các thông số đầu vào thường đòi hỏi ở người nghiên cứu một quá trình thử nghiệm và kiểm tra kết quả (try-and-error) bằng thực nghiệm vô cùng tốn thời gian. Vì vậy mô hình đề xuất trong luận văn này chưa phải là tối ưu do các thông số đầu vào chưa đo đạc trong thời gian dài và tônge thể của bộ dữ liệu. Mô hình và giải thuật được đề xuất trong luận văn này cũng không tránh khỏi những hạn chế nêu trên, nghĩa là người nghiên cứu vẫn có thể xác định giá trị các thông số đầu vào giúp bài toán dự báo tốt hơn và hiệu quả hơn.

3. Hướng phát triển của luận văn

Từ những nghiên cứu và kết quả đạt được của luận văn này, người nghiên cứu đề nghị hướng nghiên cứu tiếp theo nhằm cải thiện và nâng cao hiệu suất dự báo, tăng độ chính xác và phù hợp theo. Để làm điều này, hướng phát triển của đề tài này có thể kết hợp với các mô hình học máy cải tiến (CNN kết hợp RNN...), hoặc là xây dựng bộ dữ liệu nhiều và đủ, đặc trưng cho lưu lượng mạng di động khu vực tỉnh Tây Ninh.