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

LUẬN VĂN THẠC SỸ KỸ THUẬT

(Theo định hướng ứng dụng)

TP. HCM – NĂM 2022

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

LUẬN VĂN THẠC SỸ KỸ THUẬT

(Theo định hướng ứng dụng)

NGƯỜI HƯỚNG DẪN KHOA HỌC:

TS. NGUYỄN XUÂN SÂM

TP. HCM – NĂM 2022

LỜI CAM ĐOAN

Tôi cam đoan rằng luận văn ***"Nghiên cứu mô hình học máy cho dự báo lưu lượng trong mạng di động"*** là công trình nghiên cứu của chính tôi.

Tôi cam đoan các số liệu, kết quả nêu trong luận văn là trung thực và chưa từng được ai công bố trong bất kỳ công trình nào 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.

TP. Hồ Chí Minh, ngày 25 tháng 01 năm 2022

Học viên thực hiện luận văn

Nguyễn Xuân Quốc

LỜI CẢM ƠN

Trong suốt quá trình học tập và nghiên cứu thực hiện luận văn, ngoài nỗ lực của bản thân, tôi đã nhận được sự hướng dẫn nhiệt tình quý báu của quý Thầy Cô, cùng với sự động viên và ủng hộ của gia đình, bạn bè và đồng nghiệp. Với lòng kính trọng và biết ơn sâu sắc, tôi xin gửi lời cảm ơn chân thành tới:

Ban Giám Đốc, Phòng đào tạo sau đại học và quý Thầy Cô đã tạo mọi điều kiện thuận lợi giúp tôi hoàn thành luận văn.

Tôi xin chân thành cảm ơn Thầy **TS. Nguyễn Xuân Sâm**, người thầy kính yêu đã hết lòng giúp đỡ, hướng dẫn, động viên, tạo điều kiện cho tôi trong suốt quá trình thực hiện và hoàn thành luận văn.

Tôi xin chân thành cảm ơn gia đình, bạn bè, đồng nghiệp trong cơ quan đã động viên, hỗ trợ tôi trong lúc khó khăn để tôi có thể học tập và hoàn thành luận văn.

Mặc dù đã có nhiều cố gắng, nỗ lực, nhưng 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 tôi ngày một hoàn thiện hơn.

Xin chân thành cảm ơn!

TP. Hồ Chí Minh, ngày 25 tháng 01 năm 2022

Học viên thực hiện luận văn

Nguyễn Xuân Quốc

DANH SÁCH HÌNH VẼ

DANH SÁCH BẢNG

DANH MỤC CÁC THUẬT NGỮ, CHỮ VIẾT TẮT

Từ viết tắt	Tiếng Anh
ML	Machine Learning
AI	Artificial Intelligence
RNN	Recurrent Neural Network
LTE	Long Term Evolution
CDMA	Code-division multiple access
TDMA	Time-division multiple access
GSM	The Global System for Mobile Communications
MLP	Multilayer perceptron
TDNN	Time delay neural network
LSTM	Long Short Term Memory
CEC	Consumer Electronics Control

MỤC LỤC

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.

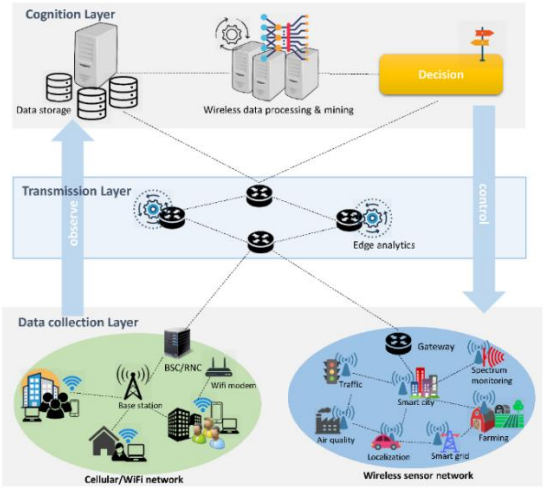
1.1.4 Thời gian giữ kênh

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 [1]

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.

Cách hoạt động của thuật toán học có giám sát

Cho một tập hợp tập dữ liệu huấn luyện N theo mẫu $\{(x_1, y_1),\ldots(x_N, y_N)\}$ sao cho x_i là vectơ đặc trưng của mẫu i-th và y_i là nhãn của nó (tức là lớp), một thuật toán học tìm kiếm một hàm $g : X \to Y$, trong đó X là không gian đầu vào và Y là không gian đầu ra. Hàm **g** là một phần tử của một số không gian của các hàm khả thi G, thường được gọi là không gian giả thuyết. Đôi khi sẽ thuận tiện khi biểu diễn **g** bằng hàm tính điểm $f : X \times Y \to \mathbb{R}$ sao cho **g** được xác định là trả về giá trị **y** cho điểm cao nhất: $g(x) = \arg_y \max f(x, y)$. Gọi F biểu thị không gian của các hàm tính điểm.

Mặc dù G và F có thể là bất kỳ không gian hàm nào, nhưng nhiều thuật toán học là mô hình xác suất trong đó **g** có dạng mô hình xác suất có điều kiện $g(x) = P(y \mid x)$, hoặc f có dạng mô hình xác suất chung $f(x, y) = P(x, y)$. Ví dụ, Naïve Bayes và phân tích phân biệt tuyến tính là mô hình xác suất chung, trong khi hồi quy logistic là mô hình xác suất có điều kiện.

Có hai cách tiếp cận cơ bản để chọn f hoặc g: giảm thiểu rủi ro theo kinh nghiệm và giảm thiểu rủi ro cấu trúc. Giảm thiểu rủi ro theo kinh nghiệm tìm kiếm chức năng phù hợp nhất với dữ liệu đào tạo. Giảm thiểu rủi ro cấu trúc bao gồm một chức năng phạt kiểm soát sự cân bằng độ lệch/phương sai.

Trong cả hai trường hợp, giả định rằng tập huấn luyện bao gồm một mẫu các cặp độc lập và được phân phối giống nhau, (x_i, y_i). Để đo lường mức độ phù hợp của một hàm với dữ liệu huấn luyện, hàm mất mát $L: Y \times Y \to \mathbb{R} \geq 0$ được xác định. Đối với ví dụ đào tạo (x_i, y_i), việc mất dự đoán giá trị $\hat{y}$ là $L(y_i, \hat{y})$.

Rủi ro R(g) của hàm g được xác định là tổn thất dự kiến của g. Điều này có thể được ước tính từ dữ liệu đào tạo như

$$R_{emp}(g) = \frac{1}{N}\sum_i L(y_i, g(x_i)) \qquad (2.1)$$

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:

2.1.2.1 Hồi quy (Linear Regression)

Hồi quy tuyến tính là một kỹ thuật học có giám sát được sử dụng để mô hình hóa mối quan hệ giữa một tập hợp các biến đầu vào độc lập là x và một biến đầu ra phụ thuộc là y, sao cho đầu ra là sự kết hợp tuyến tính của các biến đầu vào:

$$y = f(\mathbf{x}) := \theta_0 + \theta_1 x_1 + \ldots + \theta_n x_n + \epsilon = \theta_0 + \sum_{j=1}^{n} \theta_j x_j \,, \tag{2.2}$$

Trong đó:

$x = [x_1,\ldots x_n]^T$ và $\theta = [\theta_0,\theta_1,\ldots\theta_n]^T$ là vectơ tham số ước tính từ một tập huấn luyện nhất định (y_i,x_i), $j = 1,2,\ldots m$

a. Mô hình hồi quy tuyến tính đơn giản

Mối quan hệ giữa biến trả lời *Y* và biến dự đoán *X* được quy định là mô hình tuyến tính

$$\boldsymbol{Y = \beta_0 + \beta_1 X + \varepsilon}, \tag{2.3}$$

trong đó β_0 và β_1 là các hằng số được gọi là hệ số hồi quy mô hình hoặc tham số và ε là một lỗi hoặc nhiễu ngẫu nhiên. Giả định rằng trong phạm vi của các quan sát được nghiên cứu, phương trình tuyến tính (2.1) cung cấp một xấp xỉ chấp nhận được cho mối quan hệ thực sự giữa *Y* và *X*. Nói cách khác, *Y* xấp xỉ một hàm tuyến tính của *X* và ε đo lường sự khác biệt trong phép tính gần đúng đó. Cụ thể, ε không chứa thông tin có hệ thống để xác định *Y* chưa được ghi trong *X*. Hệ số β_1, được gọi là độ dốc, có thể được hiểu là sự thay đổi của *Y* đối với thay đổi đơn vị trong *X*. Hệ số β_0, được gọi là hệ số không đổi hoặc đánh chặn, là giá trị dự đoán của *Y* khi $X = 0$. Phương trình (2.1), có thể được viết như:

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i, \qquad i = 1, 2, \ldots, n, \tag{2.4}$$

Trong đó y_i đại diện cho giá trị thứ *i* của biến trả lời *Y*, x_i đại diện cho giá trị thứ *i* của biến dự đoán *X* và ε_i đại diện cho lỗi trong xấp xỉ của y_i.

Phân tích hồi quy khác với một cách quan trọng từ phân tích tương quan. Hệ số tương quan là đối xứng theo nghĩa Cor (*Y*, *X*) giống với Cor (*X*, *Y*). Các biến *X* và *Y* có tầm quan trọng như nhau. Trong phân tích hồi quy, biến trả lời *Y* có tầm quan trọng chính. Tầm quan trọng của yếu tố dự đoán *X* nằm ở khả năng tính đến sự biến thiên của biến trả lời *Y* và không phải là chính nó. Do đó *Y* có tầm quan trọng hàng đầu.

b. Ước tính tham số

Dựa trên dữ liệu có sẵn, chúng tôi muốn ước tính các tham số β_0 và β_1. Điều này tương đương với việc tìm đường thẳng cho điểm phù hợp nhất (đại diện) của các điểm trong biểu đồ phân tán của trả lời so với biến dự đoán. Chúng tôi ước tính các

tham số bằng phương pháp bình phương tối thiểu, đưa ra đường thẳng tối thiểu hóa tổng bình phương của khoảng cách dọc từ mỗi điểm đến đường thẳng. Khoảng cách dọc biểu thị các lỗi trong biến trả lời. có thể thu được bằng cách viết lại (2.2) như

$$\varepsilon_i = y_i - \beta_0 - \beta_1 x_i, \qquad i = 1, 2, \dots, n. \tag{2.5}$$

Tổng bình phương của các khoảng cách này sau đó có thể được viết là

$$S(\beta_0,\beta_1) = \sum_{i=1}^{n} \varepsilon_i^2 = \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_i)_i^2. \tag{2.6}$$

Các giá trị $\hat{\beta}_0$ và $\hat{\beta}_1$ tối thiểu hóa $S(\beta_0,\beta_1)$ được đưa ra bởi

$$\hat{\beta}_1 = \frac{\sum(y_i - \bar{y})(x_i - \bar{x})}{\sum(x_i - \bar{x})^2} \tag{2.7}$$

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}. \tag{2.8}$$

Chúng tôi cung cấp công thức cho $\hat{\beta}_1$ trước công thức cho $\hat{\beta}_0$ bởi vì $\hat{\beta}_0$ sử dụng $\hat{\beta}_1$. Các ước tính, $\hat{\beta}_0$ và $\hat{\beta}_1$ được gọi là ước lượng bình phương nhỏ nhất của β_0 và β_1 vì chúng là giải pháp cho phương pháp bình phương nhỏ nhất, đánh chặn và độ dốc của đường có tổng bình phương nhỏ nhất có thể có của khoảng cách dọc từ mỗi điểm đến đường. Vì lý do này, đường được gọi là đường hồi quy bình phương nhỏ nhất. Đường hồi quy bình phương nhỏ nhất được cho bởi

$$\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 X. \tag{2.9}$$

Lưu ý rằng một dòng bình phương tối thiểu luôn tồn tại bởi vì chúng ta luôn có thể tìm thấy một dòng cho tổng bình phương tối thiểu của khoảng cách dọc. Trong thực tế, trong một số trường hợp, một đường bình phương nhỏ nhất có thể không phải là duy nhất. Đối với mỗi quan sát trong dữ liệu của chúng tôi, chúng tôi có thể tính toán

$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i, \qquad i = 1, 2, \dots, n. \tag{2.10}$$

Chúng được gọi là các giá trị phù hợp. Do đó, giá trị phù hợp thứ *i*, $\hat{y}_i$, là điểm trên đường hồi quy bình phương nhỏ nhất (2.7) tương ứng với x_i. Khoảng cách dọc tương ứng với quan sát thứ *i* là

$$e_i = y_i - \hat{y}_i, \qquad i = 1, 2, \dots, n. \tag{2.11}$$

Những khoảng cách dọc này được gọi là phần dư bình phương nhỏ nhất thông thường. Một thuộc tính của phần dư trong (2.11) là tổng của chúng bằng 0. Điều này có nghĩa là tổng khoảng cách trên đường bằng tổng khoảng cách bên dưới đường.

c. Thử nghiệm các giả thuyết

Như đã nêu trước đó, tính hữu ích của *X* như một yếu tố dự đoán của *Y* có thể được đo lường một cách không chính thức bằng cách kiểm tra hệ số tương quan và biểu đồ phân tán tương ứng của *Y* so với *X*. Một cách chính thức hơn để đo tính hữu dụng của *X* như một yếu tố dự đoán của *Y* là tiến hành kiểm tra giả thuyết về tham số hồi quy β_1. Lưu ý rằng giả thuyết $\beta_1 = 0$ có nghĩa là không có mối quan hệ tuyến tính giữa *Y* và *X*. Một thử nghiệm của giả thuyết này đòi hỏi giả định sau đây. Đối với mỗi giá trị cố định của *X*, giả sử ε là các đại lượng ngẫu nhiên độc lập thường được phân phối chuẩn với giá trị trung bình bằng 0 và phương sai chung σ^2. Với các giả định này, các đại lượng, $\hat{\beta}_0$ và $\hat{\beta}_1$ là các ước tính không thiên vị của β_0 và β_1, tương ứng. Phương sai của chúng là

$$Var(\hat{\beta}_0) = \sigma^2 \left[\frac{1}{n} + \frac{\bar{x}^2}{\sum(x_i - \bar{x})^2}\right], \qquad (2.12)$$

và

$$Var(\hat{\beta}_1) = \frac{\sigma^2}{\sum(x_i - \bar{x})^2}. \qquad (2.13)$$

Hơn nữa, các phân phối lấy mẫu của các ước lượng bình phương nhỏ nhất $\hat{\beta}_0$ và $\hat{\beta}_1$ là chuẩn với các trung bình β_0 và β_1 và phương sai như được đưa ra trong (2.10) và (2 .11), tương ứng.

Phương sai của $\hat{\beta}_0$ và $\hat{\beta}_1$ phụ thuộc vào tham số chưa biết σ^2. Vì vậy, chúng ta cần ước tính σ^2 từ dữ liệu. Một ước tính không thiên vị của σ^2 được đưa ra bởi

$$\hat{\sigma}^2 = \frac{\sum e_i^2}{n-2} = \frac{\sum(y_i - \hat{y}_i)^2}{n-2} = \frac{SSE}{n-2}, \qquad (2.14)$$

Trong đó SSE là tổng bình phương của phần dư (lỗi). Số n - 2 trong mẫu số của (2.14) được gọi là bậc tự do (df). Nó bằng số lượng quan sát trừ đi số lượng hệ số hồi quy ước tính.

Thay thế σ^2 trong (2.12) và (2.13) bằng $\hat{\sigma}^2$ trong (2.14), chúng tôi nhận được các ước tính không thiên vị về phương sai của $\hat{\beta}_0$ và $\hat{\beta}_1$. Ước tính độ lệch chuẩn được gọi là lỗi tiêu chuẩn (s.e.) của ước tính. Do đó, các lỗi tiêu chuẩn của $\hat{\beta}_0$ và $\hat{\beta}_1$ là

$$s.e.(\hat{\beta}_0) = \hat{\sigma}\sqrt{\frac{1}{n} + \frac{\bar{x}^2}{\sum(x_i - \bar{x})^2}} \qquad (2.15)$$

và $$s.e.(\hat{\beta}_1) = \frac{\hat{\sigma}}{\sqrt{\sum(x_i - \bar{x})^2}}, \qquad (2.16)$$

tương ứng, trong đó $\hat{\sigma}$ là căn bậc hai của $\hat{\sigma}^2$ trong (2.14). Lỗi tiêu chuẩn của $\hat{\beta}_1$ là số đo độ chính xác của độ dốc đã được ước tính. Lỗi tiêu chuẩn càng nhỏ thì công cụ ước tính càng chính xác.

d. Dự đoán

Phương trình hồi quy được điều chỉnh có thể được sử dụng để dự đoán. Chúng tôi phân biệt giữa hai loại dự đoán:

Dự đoán giá trị của biến trả lời *Y* tương ứng với bất kỳ giá trị được chọn nào, x_0, của biến dự đoán.

Ước tính của trả lời trung bình μ_0, khi $X = x_0$.

Trong trường hợp đầu tiên, giá trị dự đoán y_0 là

$$\hat{y}_0 = \hat{\beta}_0 + \hat{\beta}_1 x_0. \tag{2.17}$$

Lỗi tiêu chuẩn của dự đoán này là

$$s.e.(\hat{y}_0) = \hat{\sigma}\sqrt{1 + \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{\sum(x_i - \bar{x})^2}}. \tag{2.18}$$

Do đó, giới hạn tin cậy cho giá trị dự đoán với hệ số tin cậy $(1 - \alpha)$ được đưa ra bởi

$$\hat{y}_0 \pm t_{\left(n-2,\frac{\alpha}{2}\right)}\, s.e.(\hat{y}_0). \tag{2.19}$$

Đối với trường hợp thứ hai, trả lời trung bình μ_0 được ước tính bởi

$$\hat{\mu}_0 = \hat{\beta}_0 + \hat{\beta}_1 x_0. \tag{2.20}$$

Lỗi tiêu chuẩn của ước tính này là

$$s.e.(\hat{\mu}_0) = \hat{\sigma}\sqrt{\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{\sum(x_i - \bar{x})^2}}. \tag{2.21}$$

từ đó, theo đó các giới hạn tin cậy cho μ_0 với hệ số tin cậy $(1 - \alpha)$ được đưa ra bởi

$$\hat{\mu}_0 \pm t_{\left(n-2,\frac{\alpha}{2}\right)}\, s.e.(\hat{\mu}_0). \tag{2.22}$$

Lưu ý rằng ước tính điểm của μ_0 giống hệt với trả lời dự đoán $\hat{y}_0$. Điều này có thể được nhìn thấy bằng cách so sánh (2.17) với (2.20). Tuy nhiên, lỗi tiêu chuẩn của $\hat{\mu}_0$ là nhỏ hơn lỗi tiêu chuẩn của $\hat{y}_0$ và có thể được nhìn thấy bằng cách so sánh (2.18) với (2.21). Theo trực giác, điều này có ý nghĩa. Có sự không chắc chắn (tính biến thiên) lớn hơn trong việc dự đoán một quan sát (quan sát tiếp theo) so với ước tính

đáp ứng trung bình khi $X = x_0$. Tính trung bình được ngụ ý trong trả lời trung bình làm giảm tính biến thiên và độ không đảm bảo liên quan đến ước tính.

Để phân biệt giữa các giới hạn trong (2.19) và (2.22), các giới hạn trong (2.19) đôi khi được gọi là giới hạn dự đoán hoặc dự báo, trong khi các giới hạn được đưa ra trong (2.22) được gọi là giới hạn tin cậy.

e. Chất lượng đo lường của sự điều chỉnh

Kiểm tra biểu đồ phân tán của *Y* so với $\hat{Y}$. Tập hợp các điểm với đường thẳng càng gần, mối quan hệ tuyến tính giữa *Y* và *X* càng mạnh. Người ta có thể đo cường độ của mối quan hệ tuyến tính trong biểu đồ này bằng cách tính hệ số tương quan giữa *Y* và $\hat{Y}$, được đưa ra bởi

$$Cor(Y, \hat{Y}) = \frac{\Sigma(y_i - \bar{y})(\hat{y}_i - \bar{\hat{y}})}{\sqrt{\Sigma(y_i - \bar{y})^2 \Sigma(\hat{y}_i - \bar{\hat{y}})^2}}, \qquad (2.23)$$

Trong đó $\bar{y}$ là giá trị trung bình của biến trả lời *Y* và $\bar{\hat{y}}$ là giá trị trung bình của các giá trị phù hợp. Trong thực tế, biểu đồ phân tán của *Y* so với *X* và biểu đồ phân tán của *Y* so với $\hat{Y}$ là dư thừa vì các mẫu của các điểm trong hai biểu đồ là giống như nhau. Hai giá trị tương ứng của hệ số tương quan có liên quan theo phương trình sau:

$$Cor(Y, \hat{Y}) = |Cor(Y, X)|. \qquad (2.24)$$

Mặc dù các biểu đồ phân tán của *Y* so với $\hat{Y}$ và $Cor(Y, \hat{Y})$ là dư thừa trong hồi quy tuyến tính đơn giản, chúng cho chúng ta một dấu hiệu về chất lượng của sự phù hợp trong cả hồi quy đơn giản và đa biến. Hơn nữa, trong cả hai hồi quy đơn giản và đa biến, $Cor(Y, \hat{Y})$ có liên quan đến một thước đo hữu ích khác về chất lượng của sự phù hợp của mô hình tuyến tính với dữ liệu được quan sát. Biện pháp này được phát triển như sau. Sau khi chúng ta tính toán các ước lượng bình phương nhỏ nhất của các tham số của mô hình tuyến tính, chúng ta hãy tính các đại lượng sau:

$$\begin{aligned} SST &= \sum(y_i - \bar{y})^2, \\ SSR &= \sum(\hat{y}_i - \bar{y})^2, \\ SSE &= \sum(y_i - \hat{y}_i)^2, \end{aligned} \qquad (2.25)$$

Trong đó *SST* là tổng của độ lệch bình phương trong *Y* từ trung bình $\bar{y}$ của nó, *SSR* biểu thị tổng bình phương do hồi quy và *SSE* đại diện cho tổng số dư bình phương (lỗi). Các đại lượng $(\hat{y}_i - \bar{y})$, $(y_i - \bar{y})$ và $(y_i - \hat{y}_i)$ được mô tả trong Hình 2.1 cho một điểm điển hình (x_i, y_i). Đường $\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$ là đường hồi quy phù hợp dựa

trên tất cả các điểm dữ liệu (không hiển thị trên biểu đồ) và đường ngang được vẽ tại $Y = \bar{y}$. Lưu ý rằng với mỗi điểm (x_i, y_i), có hai điểm, $(x_i, \hat{y}_i)$, nằm trên đường phù hợp và $(x_i, \bar{y})$ nằm trên đường thẳng $Y = \bar{y}$.

Một đẳng thức cơ bản, trong cả hai hồi quy đơn giản và đa biến, được đưa ra bởi

$$SST = SSR + SSE. \quad (2.26)$$

Theo đó, tổng số độ lệch bình phương trong *Y* có thể được phân tách thành tổng của hai đại lượng, *SSR* thứ nhất, đo lường chất lượng của *X* như một công cụ dự đoán của *Y* và thứ hai, *SSE* đo lường sai số trong dự đoán này. Do đó, tỷ lệ $R^2 = SSR/SST$ có thể được hiểu là tỷ lệ của tổng biến thể trong *Y* được tính bởi biến dự đoán *X*. Sử dụng (2.24), chúng ta có thể viết lại R^2 như

$$R^2 = \frac{SSR}{SST} = 1 - \frac{SSE}{SST}. \quad (2.27)$$

Ngoài ra, nó có thể được hiển thị rằng

$$[Cor(Y, X)]^2 = \left[Cor\left(Y, \hat{Y}\right)\right]^2 = R^2. \quad (2.28)$$

Trong hồi quy tuyến tính đơn giản, R^2 bằng bình phương của hệ số tương quan giữa biến trả lời *Y* và yếu tố dự đoán *X* hoặc bình phương của hệ số tương quan giữa biến trả lời *Y* và giá trị phù hợp $\hat{Y}$. Định nghĩa được đưa ra trong (2.25) cung cấp cho chúng tôi một cách giải thích khác về các hệ số tương quan bình phương. Chỉ số mức độ phù hợp, R^2, có thể được hiểu là tỷ lệ của tổng biến thiên trong biến trả lời *Y* được tính bởi biến dự đoán *X*. Lưu ý rằng $0 \leq R^2 \leq 1$ bởi vì $SSE \leq SST$. Nếu R^2 ở gần 1, thì *X* giải thích một phần lớn của biến thể trong *Y*. Vì lý do này, R^2 được gọi là hệ số xác định vì nó cho chúng ta biết về cách biến dự đoán *X* đánh giá (xác định) biến trả lời *Y*.

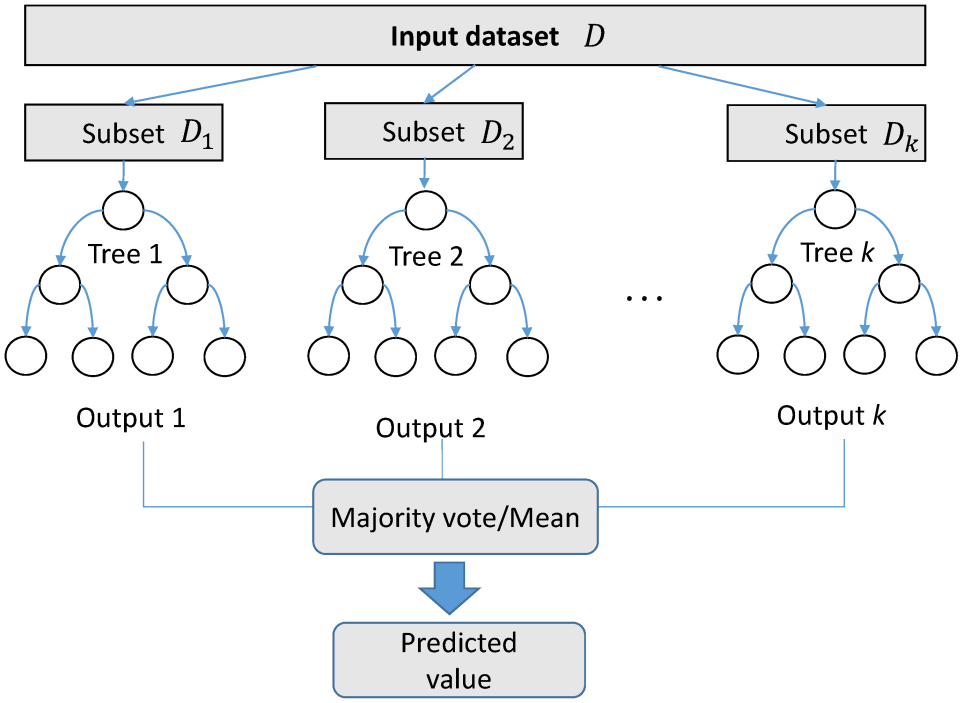
2.1.2.2 Cây quyết định (Decision Tree)

DT một thuật toán học có giám sát nhằm tạo ra một đồ thị hoặc mô hình dạng cây thể hiện các kết quả hoặc hệ quả có thể có của việc sử dụng các giá trị đầu vào nhất định. Cây bao gồm một nút gốc, các nút bên trong được gọi là nút quyết định kiểm tra đầu vào của nó dựa trên một biểu thức đã học và các nút lá tương ứng với một lớp hoặc quyết định cuối cùng. Cây học tập có thể được sử dụng để rút ra các quy tắc quyết định đơn giản có thể được sử dụng cho các vấn đề quyết định hoặc để phân loại các trường hợp trong tương lai bằng cách bắt đầu từ nút gốc và di chuyển

qua cây cho đến khi đạt đến nút lá nơi gán nhãn lớp. Tuy nhiên, cây quyết định chỉ có thể đạt được độ chính xác cao nếu dữ liệu có thể phân tách tuyến tính, tức là nếu tồn tại một siêu phẳng tuyến tính giữa các lớp.

2.1.2.3 Rừng ngẫu nhiên (Random Forest)

RF cây quyết định có đóng bao. Đóng bao là một kỹ thuật liên quan đến việc đào tạo nhiều nhóm phân loại và xem xét sản lượng trung bình của tổng thể. Bằng cách này, phương sai của bộ phân loại tập hợp tổng thể có thể được giảm đáng kể. Tính năng đóng gói thường được sử dụng với các DT vì chúng không chắc chắn lắm đối với các lỗi do sự khác biệt trong dữ liệu đầu vào



Hình 2.1: Sơ đồ biểu diễn thuật toán RF

2.1.2.4 Support Vector Machine (SVM)

SVM một thuật toán học giải quyết các vấn đề phân loại bằng cách ánh xạ dữ liệu đầu vào đầu tiên vào một không gian đặc trưng có chiều cao hơn, trong đó nó trở nên có thể phân tách tuyến tính bằng một siêu phẳng, được sử dụng để phân loại. Trong hồi quy vectơ Hỗ trợ, siêu phẳng này được sử dụng để dự đoán đầu ra giá trị liên tục. Ánh xạ từ không gian đầu vào đến không gian đặc trưng chiều cao là phi tuyến tính, đạt được bằng cách sử dụng các hàm nhân. Các chức năng nhân khác nhau tuân thủ tốt nhất cho các miền ứng dụng khác nhau. Các hàm nhân phổ biến nhất

được sử dụng trong SVM là: nhân tuyến tính, nhân đa thức và hàm nhân cơ sở (RBF), công thức được biểu diễn như sau:

$$k(x_i, x_j) = x_i^T x_j k(x_i, x_j) = \left(x_i^T x_j + 1\right)^d k(x_i, x_j) = e^{-\frac{\left(x_i - x_j\right)^2}{\sigma^2}} \quad (2.29)$$

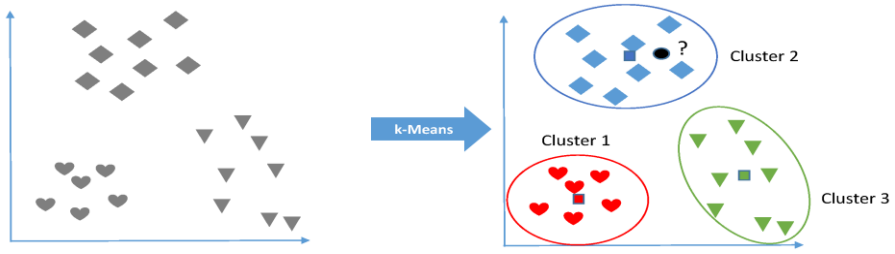
Trong đó: σ là một tham số do người dùng xác định

2.1.2.5 KNN (k nearest neighbors)

KNN một thuật toán học tập có thể giải quyết các vấn đề phân loại và hồi quy bằng cách xem xét khoảng cách (độ gần) giữa các cá thể đầu vào. Nó được gọi là thuật toán học không tham số bởi vì, không giống như các thuật toán học có giám sát khác, nó không học một chức năng mô hình rõ ràng từ dữ liệu huấn luyện. Thay vào đó, thuật toán chỉ cần ghi nhớ tất cả các trường hợp trước đó và sau đó dự đoán đầu ra bằng cách tìm kiếm tập huấn luyện đầu tiên cho k trường hợp gần nhất và sau đó: (1) để phân loại - dự đoán lớp đa số trong số k hàng xóm gần nhất đó, trong khi (2) để hồi quy -dự đoán giá trị đầu ra là giá trị trung bình của các giá trị của k lân cận gần nhất của nó. Do cách tiếp cận này, k-NN được coi là một hình thức học tập dựa trên cá thể hoặc dựa trên bộ nhớ. K-NN được sử dụng rộng rãi vì nó là một trong những hình thức học đơn giản nhất. Nó cũng được coi là lười học vì người học thụ động cho đến khi phải thực hiện một dự đoán, do đó không cần tính toán cho đến khi thực hiện nhiệm vụ dự đoán.

2.1.2.6 K-Means

K-Means một thuật toán học tập không giám sát được sử dụng cho các bài toán phân cụm. Mục đích là gán một số điểm, x1,… xm thành K nhóm hoặc cụm, sao cho độ tương đồng nội bộ kết quả là cao, trong khi độ tương tự giữa các cụm thấp. Sự tương tự được đo lường đối với giá trị trung bình của các điểm dữ liệu trong một cụm. Hình 2.2 minh họa một ví dụ về phân cụm k-mean, trong đó K = 3 và tập dữ liệu đầu vào bao gồm hai đặc điểm với các điểm dữ liệu được vẽ dọc theo trục x và y.



Hình 2.2: Sơ đồ biểu diễn ý tưởng thuật toán K-means

Ở phía bên trái của Hình 8 là các điểm dữ liệu trước khi áp dụng phương tiện k, trong khi ở phía bên phải là 3 cụm đã được xác định và các trọng tâm của chúng được biểu diễn bằng các hình vuông.

2.1.2.7 Mạng thần kinh nhân tạo (Neural Networks)

Neural Networks hay mạng thần kinh nhân tạo (ANN) là một thuật toán học tập có giám sát lấy cảm hứng từ hoạt động của bộ não, thường được sử dụng để lấy ra các ranh giới quyết định phức tạp, phi tuyến tính để xây dựng mô hình phân loại, nhưng cũng thích hợp cho các mô hình hồi quy huấn luyện khi mục tiêu là dự đoán các đầu ra có giá trị thực. Mạng nơron được biết đến với khả năng xác định các xu hướng phức tạp và phát hiện các mối quan hệ phi tuyến tính phức tạp giữa các biến đầu vào với chi phí là gánh nặng tính toán cao hơn. Một mô hình mạng nơron bao gồm một đầu vào, một số lớp ẩn và một lớp đầu ra, như thể hiện trên

Công thức tổng quát cho một lớp như sau:

$$y = g(w^T x + b), \qquad (2.30)$$

Trong đó x là đầu vào huấn luyện và y là đầu ra của lớp, w là trọng số của lớp, trong khi b là số hạng thiên vị.

Lớp đầu vào tương ứng với các biến dữ liệu đầu vào. Mỗi lớp ẩn bao gồm một số phần tử xử lý được gọi là tế bào thần kinh xử lý đầu vào của nó (dữ liệu từ lớp trước) bằng cách sử dụng một hàm kích hoạt hoặc truyền để chuyển tín hiệu đầu vào thành tín hiệu đầu ra, g (). Các hàm kích hoạt thường được sử dụng là: hàm bước đơn vị, hàm tuyến tính, hàm sigmoid và hàm tiếp tuyến hypebol. Các phần tử giữa mỗi lớp được kết nối cao bằng các kết nối có trọng số bằng số được thuật toán học. Lớp đầu ra đưa ra dự đoán (tức là lớp) cho các đầu vào đã cho và theo trọng số kết nối được xác định thông qua lớp ẩn. Thuật toán đang trở lại phổ biến trong những năm gần đây do các kỹ thuật mới và phần cứng mạnh mẽ hơn cho phép đào tạo các mô hình phức tạp để giải quyết các tác vụ phức tạp. Nói chung, mạng nơ-ron được cho là có thể xấp xỉ với bất kỳ hàm nào được quan tâm khi được điều chỉnh tốt, đó là lý do tại sao chúng được coi là bộ xấp xỉ phổ quát.

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.

2.2.1 Phân loại các loại chuỗi thời gian

Có nhiều cách phân loại chuỗi thời gian khác nhau dựa trên các tiêu chí cụ thể. Các yếu tố phụ thuộc quan trọng nhất là: độ dài của bước thời gian, trí nhớ và tính ổn định. Tùy thuộc vào khoảng cách giữa các giá trị được ghi lại, dữ liệu chuỗi thời gian được phân loại thành: Chuỗi thời gian cách đều và chuỗi thời gian không đều nhau.

Chuỗi thời gian lỏng được hình thành, khi các giá trị của nó được ghi lại định kỳ với độ dài chu kỳ không đổi. Rất nhiều quá trình vật lý hoặc môi trường được mô tả bằng loại chuỗi thời gian này. Chuỗi thời gian không cách đều là những chuỗi thời gian không giữ khoảng cách không đổi giữa các lần quan sát. Các chỉ số kinh tế lượng, chẳng hạn như giá cổ phiếu không cần thiết được thực hiện trong những khoảng thời gian đều đặn, chúng được điều chỉnh bởi tỷ lệ cung và cầu cụ thể trên thị trường cụ thể. Do đó, loại chuỗi này thể hiện một cách phù hợp ví dụ chuỗi thời gian không đều nhau.

Theo tỷ lệ phụ thuộc giữa các giá trị mới được quan sát và các giá trị trước đó, chuỗi thời gian được chia thành: chuỗi thời gian nhớ dài, chuỗi thời gian nhớ ngắn.

Chuỗi thời gian có bộ nhớ dài là những chuỗi mà hàm tự tương quan giảm chậm. Loại chuỗi thời gian này thường mô tả các quy trình không có vòng quay

nhanh. Tắc nghẽn giao thông, tiêu thụ năng lượng điện, các chỉ số vật lý hoặc khí tượng khác nhau, như đo nhiệt độ không khí, tất cả các quá trình này thường được mô tả bằng chuỗi thời gian bộ nhớ dài. Chuỗi thời gian bộ nhớ ngắn là những chuỗi mà hàm tự tương quan giảm nhanh hơn. Ví dụ điển hình chứa các quy trình từ lĩnh vực kinh tế lượng. Một cách phân loại khác của chuỗi thời gian dựa trên tính ổn định của chúng đó là chuỗi thời gian tĩnh và chuỗi thời gian không cố định.

Chuỗi thời gian tĩnh là chuỗi thời gian, trong đó các thuộc tính thống kê như giá trị trung bình hoặc phương sai, không đổi theo thời gian. Các chuỗi thời gian này luôn ở trạng thái cân bằng tương đối so với các giá trị trung bình tương ứng của nó. Các chuỗi thời gian khác thuộc chuỗi thời gian không cố định. Trong ngành công nghiệp, thương mại hoặc kinh tế, chuỗi thời gian thường xuyên hơn thuộc về loại không cố định. Để xử lý công việc dự báo, các chuỗi thời gian không cố định thường được chuyển đổi thành các chuỗi thời gian tĩnh, bằng các phương pháp tiền xử lý thích hợp.

2.2.2 Mục tiêu của phân tích chuỗi thời gian

Phân tích chuỗi thời gian hợp nhất một nhóm các phương pháp làm việc với dữ liệu chuỗi thời gian, để trích xuất thông tin hữu ích tiềm năng. Có hai mục tiêu chính của phân tích chuỗi thời gian:

- Xác định hành vi của chuỗi thời gian - Xác định các tham số và đặc tính quan trọng, mô tả đầy đủ hành vi của chuỗi thời gian.
- Dự báo chuỗi thời gian - Dự báo giá trị tương lai của chuỗi thời gian, tùy thuộc vào giá trị thực tế và quá khứ của nó.

Cả hai mục tiêu này đều yêu cầu xác định mô hình chuỗi thời gian. Ngay sau khi mô hình được xác định, nó có thể được khai thác để diễn giải hành vi của chuỗi thời gian, ví dụ, để hiểu những thay đổi theo mùa của giá cả hàng hóa. Mô hình cũng có thể được sử dụng để ngoại suy chuỗi thời gian, tức là để dự báo các giá trị trong tương lai của nó.

2.2.3 Các thành phần chuỗi thời gian

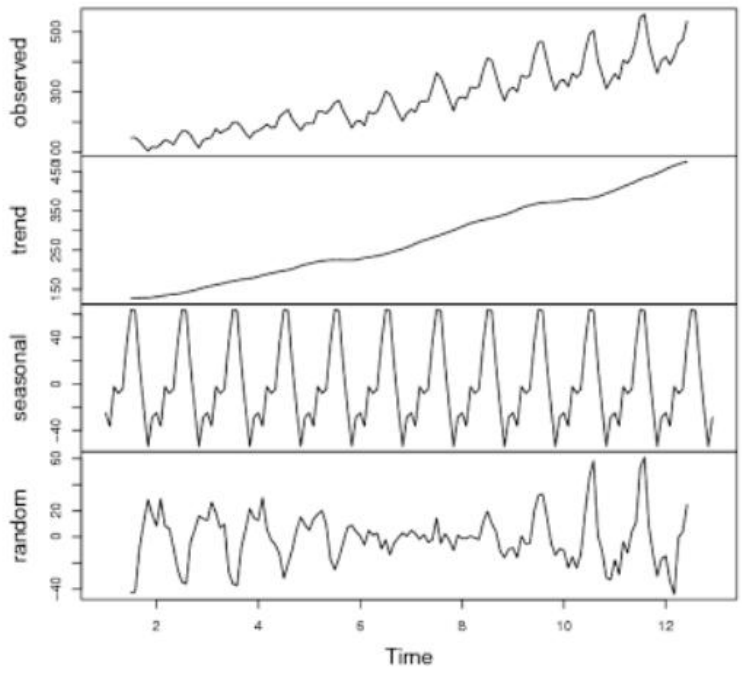
Thông thường, hầu hết các phương pháp phân tích đều giả định rằng dữ liệu chuỗi thời gian chứa thành phần hệ thống (thường bao gồm một số thành phần) và nhiễu ngẫu nhiên (lỗi), làm phức tạp việc phát hiện các thành phần thông thường. Do

đó, phần lớn các phương pháp, bao gồm các phương pháp lọc nhiễu khác nhau, để phát hiện các thành phần thông thường, hoặc nó phải thực hiện trong quá trình tiền xử lý dữ liệu.

Hầu hết các thành phần thông thường thuộc về hai lớp chính. Chúng thuộc về xu hướng hoặc thành phần theo mùa. Xu hướng là một thành phần tuyến tính hoặc phi tuyến tính có hệ thống chung, có thể thay đổi theo thời gian. Thành phần thời vụ là thành phần lặp lại định kỳ. Cả hai loại thành phần thông thường này thường được trình bày đồng thời trong chuỗi thời gian. Ví dụ: doanh số bán hàng có thể tăng từ năm này sang năm khác, nhưng có một thành phần theo mùa, điều này phản ánh sự tăng trưởng đáng kể của doanh số bán hàng vào tháng 12 và giảm xuống trong tháng 8.

Mô hình này có thể được chứng minh trên chuỗi đại diện cho lượng hành khách hàng tháng của các hãng hàng không quốc tế từ năm 1949 đến năm 1960. Biểu đồ số lượng hành khách hàng tháng thể hiện rõ xu hướng gần như tuyến tính, tức là tăng ổn định từ năm này sang năm khác (số lượng hành khách vận chuyển năm 1960 là bốn lần lớn hơn năm 1949). Đồng thời, diễn biến của giá cước hàng tháng trong vòng một năm được lặp lại và tương tự từ năm này sang năm khác (ví dụ: tỷ lệ hành khách cao hơn trong các kỳ nghỉ lễ).

Nó đã được đề cập, mô hình chung của chuỗi thời gian thường chứa một số thành phần: thành phần xu hướng T(t), thành phần theo mùa S(t), thành phần nhiễu ngẫu nhiên R(t), và đôi khi có đề cập đến thành phần chu kỳ C(t). Sự khác biệt giữa các thành phần theo chu kỳ và theo mùa là, các thành phần theo mùa thể hiện tính chu kỳ theo mùa thường xuyên, trong khi thành phần chu kỳ có ảnh hưởng lâu dài hơn và có thể thay đổi theo từng chu kỳ. Thông thường, thành phần chu kỳ được tích hợp vào một thành phần xu hướng T(t). Hình 2.3 minh họa một ví dụ về phân rã chuỗi thời gian.



Hình 2.3: Các thành phần chuỗi thời gian

Điều quan trọng là phải mô tả, cách các thành phần này tương tác với nhau về mặt toán học, để tạo ra một chuỗi thời gian. Mối quan hệ chức năng cụ thể giữa các thành phần có thể khác nhau đối với các loạt sản phẩm khác nhau. Tuy nhiên, có hai mô hình chính, cách chúng tương tác với nhau:

- Mô hình cộng

$$Z(t) = T(t) + C(t) + S(t) + R(t) \quad (2.31)$$

- Mô hình nhân

$$Z(t) = T(t) \times C(t) \times S(t) \times R(t) \quad (2.32)$$

Sự khác biệt chính giữa hai mô hình này có thể được quan sát thấy ở tốc độ tăng trưởng. Ví dụ đã đề cập trước đây về số lượng hành khách hàng không hàng tháng, thể hiện một mô hình nhân điển hình, trong đó biên độ thay đổi theo mùa tăng theo xu hướng. Sự tăng trưởng của xu hướng hoặc các thành phần theo mùa có thể được biểu thị bằng phần trăm (mô hình số nhân) hoặc bằng giá trị tuyệt đối (mô hình cộng).

2.2.4 Dự báo chuỗi thời gian

Dự báo chuỗi thời gian thuộc về hầu hết các phương pháp phân tích quan trọng, được thực hiện trên dữ liệu chuỗi thời gian. Ý tưởng chung là dựa trên thực tế, rằng thông tin về các sự kiện trong quá khứ có thể được khai thác một cách hiệu quả để

tạo ra các dự đoán về các sự kiện trong tương lai. Từ quan điểm của dữ liệu chuỗi thời gian, điều này có nghĩa là các mô hình dự báo sử dụng các giá trị đã được đo lường để dự đoán các giá trị trong tương lai trước khi chúng được quan sát.

Khi nói về dự báo chuỗi thời gian, cần nhấn mạnh tầm quan trọng của sự phân biệt giữa hai thuật ngữ, "phương pháp dự báo" và "mô hình dự báo". Mặc dù thực tế là cả hai thuật ngữ này đều có nghĩa được chỉ định chính xác, nhưng trong thực tế, chúng thường bị sử dụng nhầm lẫn với các nghĩa hỗn hợp.

Phương pháp dự báo - Biểu thị một chuỗi các hành động theo thuật toán, cần thiết để thực hiện, để có được mô hình dự báo chuỗi thời gian. Ngoài ra, các phương pháp dự báo xác định cách thức đo lường đánh giá chất lượng.

Mô hình dự báo - Biểu thị một biểu diễn chức năng, mô tả đầy đủ một chuỗi thời gian. Trên cơ sở mô hình dự báo này, các giá trị tương lai của chuỗi thời gian được dự báo.

Có hai cách chính, cách xác định các nhiệm vụ dự báo chuỗi thời gian. Tùy chọn đầu tiên dựa trên các phép tính, chỉ sử dụng các giá trị trong quá khứ của cùng một chuỗi thời gian, để dự đoán các giá trị trong tương lai. Tùy chọn thứ hai cho phép không chỉ sử dụng các giá trị trong quá khứ của cùng một chuỗi thời gian mà còn sử dụng các yếu tố bên ngoài khác, có thể hữu ích cho việc dự báo. Trong những trường hợp này, các yếu tố bên ngoài thường được trình bày dưới dạng một chuỗi thời gian khác. Chuỗi thời gian của các yếu tố bên ngoài không bắt buộc phải có cùng khoảng thời gian bước như dữ liệu chuỗi thời gian gốc. Do đó, các bước bổ sung phải được thực hiện để đối phó với vấn đề này. Người ta cũng mong đợi rằng các yếu tố bên ngoài sẽ có một số ảnh hưởng đến tiến trình của chuỗi thời gian ban đầu. Ví dụ, một yếu tố bên ngoài trực quan của mức tiêu thụ năng lượng có thể là các chỉ số khí tượng khác nhau, như nhiệt độ không khí hoặc độ ẩm không khí.

Dự báo không có yếu tố bên ngoài

Dự báo chuỗi thời gian không có yếu tố bên ngoài. Nếu các quan sát của một số quá trình ngẫu nhiên có sẵn tại các đơn vị thời gian rời rạc t = (1, 2, …, T) thì dãy giá trị $Z(t) = \{Z(i) \mid i \in T\} = \{Z(1), Z(2), \dots, Z(T)\}$ được ký hiệu là một chuỗi thời gian.

Giả sử rằng tại thời điểm đơn vị thời gian $-T$, cần phải đưa ra dự báo $-l$ về các giá trị trong tương lai quá trình đã cho Z (t). Nói cách khác, cần xác định các giá trị có thể xảy ra nhất trong tương lai cho mỗi đơn vị thời gian {T + 1,…T + l }. Đơn vị thời gian $-T$ là thời điểm khi dự báo được thực hiện, nó thường được đặt tên theo thuật ngữ "điểm gốc". Tham số $-l$ được biểu thị là "thời gian dẫn đầu", nó đại diện cho số lượng giá trị trong tương lai sẽ được dự đoán.

Để tính toán các giá trị của chuỗi thời gian tại các đơn vị thời gian trong tương lai, cần phải xác định phụ thuộc hàm mô tả mối quan hệ giữa các giá trị trong quá khứ và tương lai của chuỗi thời gian đã cho. Dự báo dựa trên - k giá trị trong quá khứ, được biểu thị là một vectơ đầu vào Z_T. Kết quả là sẽ thu được vectơ của $-l$ dự đoán trong tương lai, được ký hiệu là vectơ đầu ra $\widehat{Z_T}$. Tất cả các giá trị dự đoán $\widehat{Z_{(l)}}$sẽ được đánh dấu bằng dấu ^ để gắn nhãn chúng là dự đoán, không phải giá trị thực.

$$Z_T = \begin{pmatrix} Z(T) \\ Z(T-1) \\ Z(T-2) \\ \vdots \\ Z(T-k) \end{pmatrix} \qquad \hat{Z}_T = \begin{pmatrix} \hat{Z}(T+1) \\ \hat{Z}(T+2) \\ \vdots \\ \hat{Z}(T+l) \end{pmatrix} \tag{2.33}$$

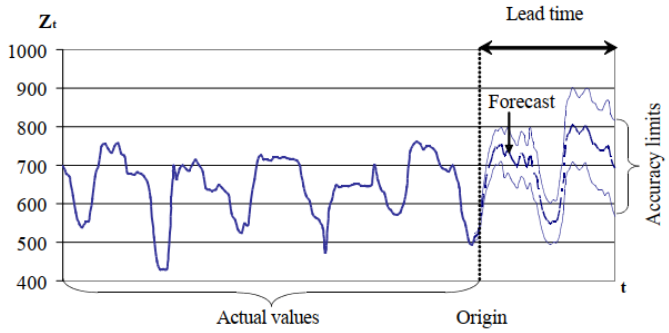
$$\mathrm{F}(Z_T) = \widehat{Z_T} \tag{2.34}$$

Phụ thuộc hàm (2.32) thường được ký hiệu là hàm dự báo và nó đại diện cho mô hình dự báo. Mục đích trực quan là tìm ra hàm dự báo sao cho độ lệch giữa giá trị dự đoán và giá trị thực tế, sẽ được quan sát sau này trong tương lai, càng nhỏ càng tốt.

$$\varepsilon_T = \begin{pmatrix} Z(T+1) \\ Z(T+2) \\ \vdots \\ Z(T+l) \end{pmatrix} - \begin{pmatrix} \hat{Z}(T+1) \\ \hat{Z}(T+2) \\ \vdots \\ \hat{Z}(T+l) \end{pmatrix} \tag{2.35}$$

Phân tích vectơ độ lệch (2.3) đại diện cho một cơ sở của cái gọi là hàm mất mát hoặc hàm lỗi. Chức năng này đo lường chất lượng của dự báo, dựa trên độ lệch đo được. Có nhiều lựa chọn hơn, cách tính tỷ lệ chất lượng từ véc tơ độ lệch, thường là sai số trung bình bình phương căn bậc hai hoặc độ lệch tuyệt đối trung bình được tính. Chi tiết hơn về các hàm lỗi sẽ được thảo luận trong phần 2.2. Mục tiêu chính thức của dự báo chuỗi thời gian sau đó được xây dựng dưới dạng hàm giảm thiểu của hàm mất mát.

Ngoài việc tính toán các giá trị trong tương lai, đôi khi yêu cầu xác định giới hạn độ chính xác. Độ chính xác của các dự báo có thể được thể hiện bằng cách tính toán các giới hạn xác suất ở hai bên của mỗi dự báo. Các giới hạn này có thể được tính toán cho bất kỳ tập hợp xác suất thuận tiện nào. Chúng sao cho giá trị thực của chuỗi thời gian, khi nó xảy ra cuối cùng, sẽ được bao gồm trong các giới hạn này với xác suất đã nêu.



Hình 2.4: Dự báo chuỗi thời gian không có yếu tố bên ngoài

Dự báo với các yếu tố bên ngoài

Quá trình chuỗi thời gian Z (t) được xác định tại các đơn vị thời gian rời rạc t = (1, 2, …, T). Theo giả thiết, chuỗi thời gian này bị ảnh hưởng bởi tập hợp các yếu tố bên ngoài $\{X_1(t_1),X_2(t_2),\ldots,X_m(t_m)\}$. Mỗi yếu tố bên ngoài được biểu diễn như một quá trình chuỗi thời gian độc lập. Ví dụ, một yếu tố bên ngoài $X_1(t_1)$ được xác định tại các đơn vị thời gian rời rạc tương ứng $t_1 = \{1, 2, \ldots, T_1\}$.

Chuỗi thời gian gốc Z(t) và các yếu tố bên ngoài $X_i(t_i)$ không bắt buộc phải xác định cùng một đơn vị thời gian. Nếu các đơn vị thời gian t, t_1, t_2 ,…, t_m không bằng nhau, khi đó cần phải tính toán lại các giá trị của yếu tố bên ngoài thành một thang đo duy nhất t.

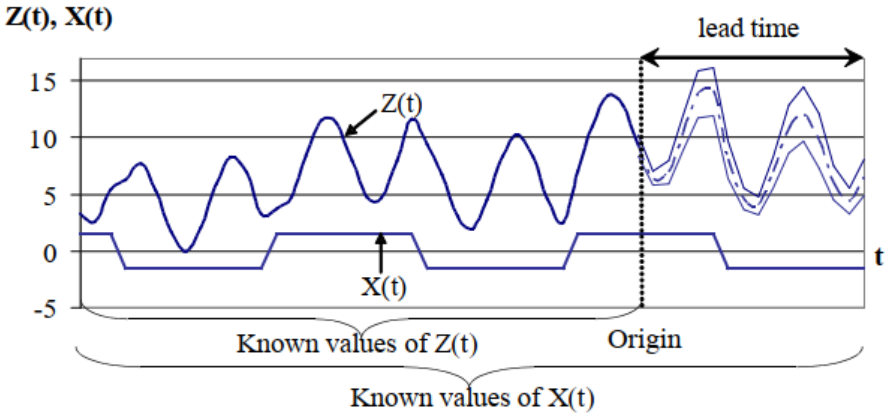
Giả sử rằng tại thời điểm của đơn vị thời gian T, cần phải đưa ra dự báo về $-\ l$ giá trị trong tương lai của quá trình đã cho Z(t). Để tính toán các dự đoán, cần phải xác định phụ thuộc hàm, mô tả mối quan hệ giữa các giá trị trong quá khứ và tương lai, đồng thời xem xét tác động của các yếu tố bên ngoài.

$$Z_T = \begin{pmatrix} Z(T) \\ Z(T-1) \\ Z(T-2) \\ \vdots \\ Z(T-k) \end{pmatrix} \quad X_{i,\,T} = \begin{pmatrix} X_i(T+l) \\ \vdots \\ X_i(T+1) \\ X_i(T) \\ X_i(T-1) \\ \vdots \\ X_i(T-k) \end{pmatrix} \quad \hat{Z}_T = \begin{pmatrix} \hat{Z}(T+l) \\ \vdots \\ \hat{Z}(T+2) \\ \hat{Z}(T+1) \end{pmatrix} \tag{2.36}$$

$$f(Z_T,\ X_{1,T},\ X_{2,T},\ \ldots,\ X_{m,T}) = \hat{Z}_T \tag{2.37}$$

Phụ thuộc hàm (2.35) là một hàm dự báo và nó thể hiện mô hình dự báo với các yếu tố bên ngoài. Các nhiệm vụ còn lại được thực hiện giống như trong trường hợp dự báo mà không có các yếu tố bên ngoài. Mục tiêu chính là tìm ra hàm dự báo sao cho độ lệch giữa giá trị dự đoán và giá trị thực tế, sẽ được quan sát sau này trong tương lai, càng nhỏ càng tốt. Mục tiêu này hình thành nhiệm vụ giảm thiểu của cái gọi là "chức năng mất mát" hoặc "chức năng lỗi".

Các giới hạn độ chính xác có thể được tính toán cho bất kỳ tập hợp xác suất thuận tiện nào. Giới hạn độ chính xác sao cho giá trị thực của chuỗi thời gian, khi nó xảy ra cuối cùng, sẽ được đưa vào các giới hạn này với xác suất đã nêu.



Hình 2.5: Dự báo chuỗi thời gian với các yếu tố bên ngoài

2.2.5 Các trường hợp sử dụng phân tích chuỗi thời gian

Phân tích chuỗi thời gian được sử dụng cho dữ liệu không cố định - những thứ liên tục biến động theo thời gian hoặc bị ảnh hưởng bởi thời gian. Các ngành như tài chính, bán lẻ và kinh tế thường sử dụng phân tích chuỗi thời gian vì tiền tệ và doanh số luôn thay đổi. Phân tích thị trường chứng khoán là một ví dụ tuyệt vời về phân tích chuỗi thời gian trong thực tế, đặc biệt là với các thuật toán giao dịch tự động. Tương tự như vậy, phân tích chuỗi thời gian là lý tưởng để dự báo những thay đổi thời tiết, giúp các nhà khí tượng học dự đoán mọi thứ từ báo cáo thời tiết ngày mai đến những năm biến đổi khí hậu trong tương lai. Ví dụ về phân tích chuỗi thời gian trong thực tế bao gồm:

- Phân tích dữ liệu thời tiết
- Đo lượng mưa
- Đọc nhiệt độ
- Theo dõi nhịp tim (EKG)
- Theo dõi hoạt động não bộ (EEG)
- Phân tích doanh số theo quý
- Phân tích giá cổ phiếu
- Phân tích giao dịch chứng khoán tự động
- Các ngành liên quan đến dự báo
- Phân tích lãi suất

Bởi vì phân tích chuỗi thời gian bao gồm nhiều danh mục hoặc các biến thể của dữ liệu, các nhà phân tích đôi khi phải đưa ra các mô hình phức tạp. Tuy nhiên,

các nhà phân tích không thể giải thích tất cả các phương sai và họ không thể tổng quát hóa một mô hình cụ thể cho mọi mẫu. Mô hình quá phức tạp hoặc cố gắng làm quá nhiều việc có thể dẫn đến thiếu phù hợp. Các mô hình thiếu phù hợp hoặc trang bị quá mức dẫn đến các mô hình đó không phân biệt được mối quan hệ giữa sai số ngẫu nhiên và đúng, làm cho phân tích bị sai lệch và dự báo không chính xác.

Các loại phân tích chuỗi thời gian

Trong phân tích chuỗi thời gian, có nhiều loại và mô hình phân tích khác nhau cho những kết quả đạt được khác nhau.

- **Phân loại (Classification)**: Xác định và gán các danh mục cho dữ liệu. Điều chỉnh đường cong (Curve fitting): Vẽ đồ thị dữ liệu dọc theo đường cong để nghiên cứu mối quan hệ của các biến trong dữ liệu.
- **Phân tích mô tả (Descriptive analysis)**: Xác định các mẫu trong dữ liệu chuỗi thời gian, như xu hướng, chu kỳ hoặc biến đổi theo mùa.
  Phân tích giải thích: Cố gắng hiểu dữ liệu và các mối quan hệ bên trong nó, cũng như nguyên nhân và kết quả.
- **Phân tích thăm dò (Exploratory analysis)**: Làm nổi bật các đặc điểm chính của dữ liệu chuỗi thời gian, thường ở định dạng trực quan.
- **Dự báo (Forecasting)**: Dự đoán dữ liệu trong tương lai. Loại này dựa trên xu hướng lịch sử. Nó sử dụng dữ liệu lịch sử làm mô hình cho dữ liệu trong tương lai, dự đoán các tình huống có thể xảy ra dọc theo các điểm cốt truyện trong tương lai.
- **Phân tích can thiệp (Intervention analysis)**: Nghiên cứu cách một sự kiện có thể thay đổi dữ liệu.
- **Phân đoạn (Segmentation)**: Tách dữ liệu thành các phân đoạn để hiển thị các thuộc tính cơ bản của thông tin nguồn.

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 (MSLE) 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

- Mean absolute percentage error (MSLE)

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

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.

- Root Mean squared error (RMSE)

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

Trong thống kê, sai số bình phương trung bình (MSE) hoặc độ lệch bình phương trung bình (MSD) của một công cụ ước lượng (của một thủ tục ước tính một đại lượng không được quan sát) đo giá trị trung bình của các bình phương của các lỗi - nghĩa là, sự khác biệt bình phương trung bình giữa các giá trị ước tính và giá trị thực tế. MSE là một hàm rủi ro, tương ứng với giá trị kỳ vọng của tổn thất sai số bình phương. Thực tế là MSE hầu như luôn luôn dương (và không phải bằng 0) là do ngẫu nhiên hoặc do công cụ ước lượng không tính đến thông tin có thể đưa ra ước tính chính xác hơn

- Mean squared error (MSE)

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

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:

- Mean absolute error (MAE)

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

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.

- 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 công trình nghiên cứu liên quan

Năm 2021, Xun Xu; Shuo Zeng; Yuanjie He [2], đã công bố nghiên cứu về, thông qua các bằng chứng thực nghiệm từ Airbnb từ tám thành phố lớn ở Hoa Kỳ, nhóm tác giả xem xét vai trò của thông tin tiết lộ trong việc tác động đến hành vi mua hàng của người tiêu dùng trên nền tảng kinh tế chia sẻ này.

Chúng tôi phân tích việc công bố thông tin từ bốn khía cạnh - đó là thông tin gì (tức là nội dung thông tin), từ đâu (tức là, nguồn thông tin), ở định dạng nào (hình thức trình bày thông tin), và bao nhiêu (số lượng thông tin). Chúng tôi tìm thấy cả ba nguồn thông tin - nhà cung cấp, nền tảng và người tiêu dùng ngang hàng - ảnh hưởng đến người tiêu dùng hành vi mua hàng. Liên quan đến thông tin do các nhà cung cấp đăng tải, chúng tôi nhận thấy mối quan hệ lõm giữa thông tin (tức là số lượng ảnh và độ dài của mô tả) và hành vi mua hàng của người tiêu dùng. Tuy nhiên, không có mối quan hệ đáng kể nào giữa phần tự mô tả của nhà cung cấp (văn bản và ảnh) và hành vi mua hàng của người tiêu dùng hành vi được tìm thấy.

Về thông tin được đăng bởi nền tảng, cả khuyến nghị từ nền tảng và thông tin xác minh nhà cung cấp ảnh hưởng tích cực đến hành vi mua hàng của người tiêu dùng. Đối với thông tin về tương tác giữa nhà cung cấp và người tiêu dùng, tỷ lệ phản hồi cao và tốc độ phản hồi nhanh của nhà cung cấp sẽ nâng cao khả năng mua hàng

của người tiêu dùng hành vi. Tuy nhiên, việc cung cấp kết nối với hồ sơ mạng xã hội của nhà cung cấp ảnh hưởng tiêu cực đến người tiêu dùng hành vi mua hàng. Về thông tin từ người tiêu dùng ngang hàng, chúng tôi nhận thấy mặc dù về tổng thể, người tiêu dùng xếp hạng ảnh hưởng tích cực đến hành vi mua hàng của người tiêu dùng, ảnh hưởng đó giảm đi khi xếp hạng vượt quá mức nhất định các ngưỡng. Nghiên cứu của chúng tôi cung cấp các gợi ý cho chủ sở hữu nền tảng để tối ưu hóa bố cục trình bày thông tin trực tiếp thông qua thiết kế nền tảng hoặc gián tiếp thông qua hướng dẫn tiết lộ thông tin của nhà cung cấp để tạo điều kiện thuận lợi cho việc tìm kiếm và thu thập thông tin của người tiêu dùng nhằm giảm rủi ro được nhận thức, nâng cao lòng tin đối với các nhà cung cấp và nền tảng, đồng thời nâng cao ý định và hành vi mua hàng của họ.

Vào năm 2019, Byoungsuk Ji cùng Ellen J. Hong [3] đã đề xuất một phương pháp dự đoán giao thông đường bộ theo thời gian thực dựa trên nghiên cứu sâu sử dụng dữ liệu truy cập tiến hóa dài hạn (LTE). Hệ thống được đề xuất tạo ra một mô hình học tốc độ giao thông đường bộ dựa trên dữ liệu tốc độ đường bộ và dữ liệu LTE lịch sử được thu thập từ nhiều trạm cơ sở nằm trong bán kính xác định trước từ đường. Dữ liệu LTE thời gian thực là đầu vào cho mô hình học tập được tạo để dự đoán tốc độ thời gian thực của lưu lượng. Vì hệ thống được phát triển bằng cách sử dụng mô hình học tốc độ giao thông đường bộ theo chuỗi thời gian dựa trên dữ liệu LTE từ quá khứ, nên nó có thể được sử dụng cho những con đường mà môi trường đã thay đổi. Hơn nữa, ngay cả trên những con đường mà việc thu thập dữ liệu giao thông không hợp lệ, chẳng hạn như vùng bóng vô tuyến, có thể nhập trực tiếp dữ liệu truyền thông không dây theo thời gian thực vào mô hình học tốc độ giao thông để dự đoán tốc độ giao thông trên đường trong thời gian thực và do đó, nâng cao độ chính xác của các dự đoán về giao thông đường bộ trong thời gian thực.

Sự tiến hóa của sự phát triển công nghệ tế bào đã dẫn đến sự phát triển bùng nổ trong lưu lượng mạng di động. Các mô hình chuỗi thời gian chính xác để dự đoán lưu lượng di động di động đã trở nên rất quan trọng để tăng chất lượng dịch vụ (QoS) với mạng. Việc mô hình hóa và dự báo tải mạng di động đóng một vai trò quan trọng trong việc đạt được phân bổ tài nguyên thuận lợi nhất bằng cách cung cấp băng thông thuận tiện và đồng thời duy trì mức sử dụng mạng cao nhất. Tính mới của nghiên cứu

được đề xuất là phát triển một mô hình có thể giúp dự đoán lưu lượng tải trong mạng di động một cách thông minh. Trong bài báo này [4], một mô hình kết hợp làm mịn theo cấp số nhân với bộ nhớ ngắn hạn dài (SES-LSTM) được đề xuất để dự đoán lưu lượng di động. Mô hình chuẩn hóa tối thiểu-tối đa đã được sử dụng để chia tỷ lệ tải mạng. Phương pháp làm trơn đơn lẻ theo cấp số nhân được áp dụng để điều chỉnh lưu lượng mạng do lưu lượng mạng rất phức tạp và có nhiều dạng khác nhau. Đầu ra e từ mô hình hàm mũ đơn được xử lý bằng cách sử dụng mô hình LSTM để dự đoán tải mạng. Hệ thống thông minh điện tử được đánh giá bằng cách sử dụng lưu lượng mạng di động thực đã được thu thập trong tập dữ liệu kaggle. Kết quả của thử nghiệm cho thấy rằng phương pháp được đề xuất có độ chính xác vượt trội, đạt được các giá trị thước đo bình phương R lần lượt là 88,21%, 92,20% và 89,81% trong ba khoảng thời gian một tháng. Người ta quan sát thấy rằng các giá trị dự đoán rất gần với các quan sát. Một so sánh của kết quả dự đoán giữa mô hình LSTM hiện có và hệ thống đề xuất của chúng tôi được trình bày. Hệ thống đề xuất đạt được hiệu suất vượt trội để dự đoán lưu lượng mạng di động.

Trong một bài báo nghiên cứu vào năm 2019, Diogo Clemente và các cộng sự [5] đã đề xuất một phương pháp luận để cải thiện độ chính xác của dự báo lưu lượng di động bằng cách tiếp cận máy học. Để phát triển phương pháp luận này, trước tiên, chúng tôi thực hiện một phân tích có hệ thống để giảm độ lệch bằng cách chọn các ô có ít dữ liệu bị thiếu hơn. Sau đó, chúng tôi đã chọn các tính năng và đào tạo một bộ phân loại để phân bổ các ô giữa có thể dự đoán và không thể dự đoán, có tính đến lỗi dự báo lưu lượng truy cập trước đó. Phương pháp phân loại Naive Bayes và Holt-Winters đã được chọn để thực hiện phương pháp luận được đề xuất trong thời gian thực. Hệ thống được áp dụng cho một tập hợp 786 ô trong một mạng thực. Bộ phân loại đưa ra độ chính xác 91%, dẫn đến các ô có thể dự đoán, sử dụng Holt-Winters, đưa ra RMSE trung bình là 2,74%. Điều này có nghĩa là bây giờ có thể triển khai các thuật toán tối ưu hóa có độ nhạy cao với dự đoán lưu lượng truy cập. Dự đoán lưu lượng thời gian thực chính xác được yêu cầu trong nhiều ứng dụng mạng như phân bổ tài nguyên động và quản lý nguồn., là bài báo khám phá một số công cụ dự đoán và tìm kiếm một công cụ dự đoán có độ chính xác cao, độ phức tạp tính toán thấp và mức tiêu thụ điện năng. Nhiều bộ dự đoán từ ba lớp khác nhau, bao gồm chuỗi thời

gian cổ điển, mạng nơ-ron nhân tạo và bộ dự đoán dựa trên biến đổi wavelet, được so sánh. , các dự báo ese được đánh giá bằng cách sử dụng các dấu vết mạng thực. So sánh độ chính xác và chi phí, cả về độ phức tạp của tính toán và mức tiêu thụ điện năng, được trình bày. Người ta quan sát thấy rằng một công cụ dự báo làm mịn theo cấp số nhân kép cung cấp sự cân bằng hợp lý giữa hiệu suất và chi phí chung.

Đo kích thước mạng là một nhiệm vụ quan trọng trong các mạng di động hiện tại, vì bất kỳ lỗi nào trong quá trình này đều dẫn đến trải nghiệm người dùng bị xuống cấp hoặc nâng cấp tài nguyên mạng không cần thiết. Với mục đích này [6], các công cụ lập kế hoạch vô tuyến thường dự đoán lưu lượng dữ liệu giờ bận hàng tháng để phát hiện trước các tắc nghẽn về dung lượng. Học có giám sát (SL) ra đời như một giải pháp đầy hứa hẹn để cải thiện các dự đoán thu được bằng các phương pháp tiếp cận kế thừa. Các công trình trước đây đã chỉ ra rằng học sâu vượt trội hơn phân tích chuỗi thời gian cổ điển khi dự đoán lưu lượng dữ liệu trong mạng di động trong ngắn hạn (giây / phút) và trung hạn (giờ / ngày) từ chuỗi dữ liệu lịch sử dài. Tuy nhiên, dự báo dài hạn (khoảng thời gian vài tháng) được thực hiện trong các công cụ lập kế hoạch vô tuyến dựa trên chuỗi thời gian ngắn và ồn ào, do đó đòi hỏi một phân tích riêng biệt. Trong công trình này, nhóm tác giả trình bày nghiên cứu đầu tiên so sánh phương pháp tiếp cận phân tích chuỗi thời gian và SL để dự đoán lưu lượng dữ liệu theo giờ bận hàng tháng trên cơ sở ô trong mạng LTE trực tiếp. Để đạt được mục tiêu này, một tập dữ liệu mở rộng được thu thập, bao gồm lưu lượng dữ liệu trên mỗi ô cho cả một quốc gia trong suốt 30 tháng. Các phương pháp được xem xét bao gồm rừng ngẫu nhiên, các mạng nơ ron khác nhau, hồi quy vectơ hỗ trợ, đường trung bình động tích hợp tự động hồi quy theo mùa và Holt-Winters cộng thêm. Kết quả cho thấy các mô hình SL hoạt động tốt hơn các phương pháp tiếp cận chuỗi thời gian, đồng thời giảm yêu cầu về dung lượng lưu trữ dữ liệu. Quan trọng hơn, không giống như trong dự báo lưu lượng ngắn hạn và trung hạn, các phương pháp tiếp cận SL không sâu cạnh tranh với học sâu đồng thời hiệu quả hơn về mặt tính toán.

Trong một nghiên cứu khác vào năm 2019, nhóm tác giả gồm Vincent Le Guen, và Nicolas Thome [7] đã giải quyết vấn đề dự báo chuỗi thời gian cho các tín hiệu không đứng yên và dự đoán nhiều bước trong tương lai. Để xử lý nhiệm vụ đầy thách thức này, chúng tôi giới thiệu DILATE (Mất phân số bao gồm shApe và TimE), một

chức năng mục tiêu mới để đào tạo mạng nơ-ron sâu. DILATE nhằm mục đích dự đoán chính xác những thay đổi đột ngột và kết hợp rõ ràng hai thuật ngữ hỗ trợ phát hiện hình dạng chính xác và thay đổi thời gian. Chúng tôi giới thiệu một chức năng mất mát có thể phân biệt phù hợp để đào tạo mạng lưới thần kinh sâu và cung cấp triển khai hỗ trợ tùy chỉnh để tăng tốc độ tối ưu hóa. Chúng tôi cũng giới thiệu một biến thể của DILATE, cung cấp sự tổng quát hóa mượt mà về Độ cong thời gian động (DTW) bị hạn chế về mặt thời gian. Các thử nghiệm được thực hiện trên các bộ dữ liệu không cố định khác nhau cho thấy hoạt động rất tốt của DILATE so với các mô hình được đào tạo với chức năng mất mát sai số trung bình bình phương (MSE) tiêu chuẩn, cũng như đối với DTW và các biến thể. DILATE cũng không tin vào việc lựa chọn mô hình và chúng tôi nhấn mạnh lợi ích của nó đối với việc đào tạo các mạng được kết nối đầy đủ cũng như các kiến trúc lặp lại chuyên biệt, cho thấy khả năng của nó để cải thiện các phương pháp tiếp cận dự báo quỹ đạo hiện đại.

Với sự phát triển của các thiết bị không dây và sự gia tăng của người dùng di động, trọng tâm của nhà khai thác đã chuyển từ việc xây dựng mạng truyền thông sang vận hành và bảo trì mạng. Các nhà khai thác rất muốn biết hành vi của mạng di động và trải nghiệm thời gian thực của người dùng, điều này yêu cầu sử dụng dữ liệu lịch sử để dự đoán chính xác các điều kiện mạng trong tương lai. Phân tích dữ liệu lớn và tính toán được áp dụng rộng rãi có thể được sử dụng như một giải pháp. Tuy nhiên, vẫn còn một số thách thức trong việc phân tích và dự đoán dữ liệu để tối ưu hóa mạng di động, chẳng hạn như tính kịp thời và chính xác của dự đoán. Bài báo này [8] đề xuất một hệ thống phân tích và dự đoán lưu lượng phù hợp với các mạng truyền thông không dây đô thị bằng cách kết hợp phân tích dữ liệu bản ghi chi tiết cuộc gọi (CDR) thực tế và các thuật toán dự đoán đa biến. Thứ nhất, mô hình không gian-thời gian được sử dụng để trích xuất dữ liệu lưu lượng truy cập lịch sử. Sau đó, phân tích quan hệ nhân quả được áp dụng cho phân tích dữ liệu truyền thông lần đầu tiên. Dựa trên phân tích nhân quả, các mô hình bộ nhớ ngắn hạn đa biến dài hạn được sử dụng để dự đoán dữ liệu trong tương lai cho dữ liệu CDR. Cuối cùng, thuật toán dự đoán được sử dụng để xử lý dữ liệu thực của các cảnh khác nhau trong thành phố nhằm xác minh hiệu suất của toàn bộ hệ thống.

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 [9], Lorenza Giupponi và các cộng sự đã 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 độ trễ bất thường đáng tin cậy và độ trễ thấp, chúng tôi 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 dịch vụ đáp ứng hiệu suất, điều này rất quan trọng trong vấn đề của chúng tôi. Chúng tôi tập trung vào việc phát hiện sự bất thường ở đô thị, bằng cách theo dõi các sự kiện đã biết tập trung nhiều người. Thông tin mạng di động được thu thập từ Kênh điều khiển đường xuống vật lý LTE (PDCCH), kênh này chứa thông tin lập lịch vô tuyến và có lợi ích là không được mã hóa và chi tiết, vì các tin nhắn được trao đổi sau mỗi khung con LTE 1 ms. Với mục đích này, chúng tôi thiết kế một hệ thống phát hiện bất thường dựa trên Mạng thần kinh bộ nhớ ngắn hạn dài (LSTM), để xử lý các đầu vào tuần tự và lặp lại. Chúng tôi chứng minh rằng kiến trúc LSTM xếp chồng lên nhau có thể xác định các bất thường về lưu lượng gây ra bởi sự gia tăng nhanh chóng về số lượng người dùng, khi một sự kiện đông đúc diễn ra gần khu vực được giám sát. Kết quả số cho thấy thuật toán được đề xuất đạt điểm F = 1 và vượt qua hiệu suất của các điểm chuẩn hiện đại khác.

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 nghiên

cứu này [10], 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 là một phương tiện hữu hiệu để xác định các yếu tố chính hạn chế thông lượng người dùng trong các ô 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 [11], Hoang Duy Trinh và các cộng sự đã 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 chúng tôi trình bày phân tích theo thời gian và không gian của các dấu vết được ghi lại. Hơn nữa, chúng tôi 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 chúng tôi thể hiện mức độ chính xác cao về thống kê đơn hàng thứ nhất và thứ hai.

2.5 Kết luận chương

Trong chương này thông qua việc nghiên cứu tìm hiểu được một số thuật toán và những công trình liên quan tới mô hình học máy, giải quyết các bài toán liên quan về dữ liệu mạng, từ đó hiểu được những ưu nhược điểm của các thuật toán, tạo tiền đề và cơ sở vững chắc cho nghiên cứu của đề tài luận văn này.

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.

Thông thường, dữ liệu càng toàn diện thì dự báo càng chính xác. Mặc dù dự báo và "dự đoán" thường có nghĩa giống nhau, nhưng có một điểm khác biệt đáng chú ý. Trong một số ngành, dự báo có thể đề cập đến dữ liệu tại một thời điểm cụ thể trong tương lai, trong khi dự đoán đề cập đến dữ liệu tương lai nói chung. Dự báo chuỗi thường được sử dụng cùng với phân tích chuỗi thời gian. Phân tích chuỗi thời gian liên quan đến việc phát triển các mô hình để có được sự hiểu biết về dữ liệu để hiểu được nguyên nhân cơ bản. Phân tích có thể cung cấp "lý do" đằng sau những kết quả mà bạn đang thấy. Sau đó, dự báo sẽ thực hiện bước tiếp theo về những việc cần làm với kiến thức đó và các phép ngoại suy có thể dự đoán được về những gì có thể xảy ra trong tương lai.

Có những hạn chế khi đối phó với những điều không thể đoán trước và những điều chưa biết. Dự báo chuỗi thời gian không sai và không phù hợp hoặc hữu ích cho mọi tình huống. Vì thực sự không có bộ quy tắc rõ ràng nào về thời điểm bạn nên hoặc không nên sử dụng dự báo, nên các nhà phân tích và nhóm dữ liệu phải biết các hạn chế của phân tích và những gì mô hình của họ có thể hỗ trợ. Không phải mọi mô hình sẽ phù hợp với mọi tập dữ liệu hoặc trả lời mọi câu hỏi. Nhóm dữ liệu nên sử dụng dự báo chuỗi thời gian khi họ hiểu câu hỏi kinh doanh và có dữ liệu và khả năng dự báo thích hợp để trả lời câu hỏi đó. Dự báo tốt hoạt động với dữ liệu rõ ràng, được đóng dấu thời gian và có thể xác định các xu hướng và mẫu chính xác trong dữ liệu lịch sử. Các nhà phân tích có thể cho biết sự khác biệt giữa biến động ngẫu nhiên hoặc ngoại lệ và có thể tách những thông tin chi tiết xác thực khỏi các biến thể theo

mùa. Phân tích chuỗi thời gian cho biết dữ liệu thay đổi như thế nào theo thời gian và dự báo tốt có thể xác định hướng dữ liệu đang thay đổi.

Ứng dụng của phương pháp dự báo Time Series

Dự báo có một loạt các ứng dụng trong các ngành công nghiệp khác nhau. Nó có rất nhiều ứng dụng thực tế bao gồm: dự báo thời tiết, dự báo khí hậu, dự báo kinh tế, dự báo kỹ thuật dự báo chăm sóc sức khỏe, dự báo tài chính, dự báo bán lẻ, dự báo kinh doanh, dự báo nghiên cứu môi trường, dự báo nghiên cứu xã hội và hơn thế nữa. Về cơ bản, bất kỳ ai có dữ liệu lịch sử nhất quán đều có thể phân tích dữ liệu đó bằng các phương pháp phân tích chuỗi thời gian và sau đó lập mô hình, dự báo và dự đoán. Đối với một số ngành, toàn bộ điểm của phân tích chuỗi thời gian là để tạo điều kiện thuận lợi cho việc dự báo. Một số công nghệ, chẳng hạn như phân tích tăng cường, thậm chí có thể tự động chọn dự báo trong số các thuật toán thống kê khác nếu nó mang lại sự chắc chắn nhất.

Một số ví dụ từ một loạt các ngành để làm cho các khái niệm về phân tích chuỗi thời gian và dự báo cụ thể hơn:

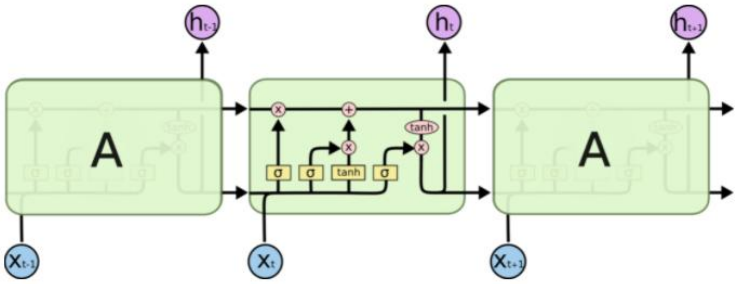
- Dự báo giá đóng cửa của cổ phiếu mỗi ngày.
- Dự báo doanh số bán sản phẩm theo đơn vị bán ra mỗi ngày cho một cửa hàng.
- Dự báo thất nghiệp cho một tiểu bang mỗi quý.
- Dự báo giá xăng trung bình mỗi ngày.

Những thứ ngẫu nhiên sẽ không bao giờ được dự báo chính xác, cho dù chúng ta thu thập bao nhiêu dữ liệu hay mức độ nhất quán. Ví dụ: chúng ta có thể quan sát dữ liệu hàng tuần về mọi người trúng xổ số, nhưng chúng ta không bao giờ có thể dự đoán ai sẽ thắng tiếp theo. Cuối cùng, tùy thuộc vào dữ liệu và phân tích dữ liệu chuỗi thời gian về thời điểm nên sử dụng dự báo, bởi vì dự báo rất khác nhau do các yếu tố khác nhau. Sử dụng phán đoán của bạn và biết dữ liệu của bạn.

3.2 Thuật toán LSTM

Long Short-Term Memory (LSTM) [12] 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.



Hình 3.1: Mô-đun lặp lại trong một LSTM chứa bốn lớp tương tác

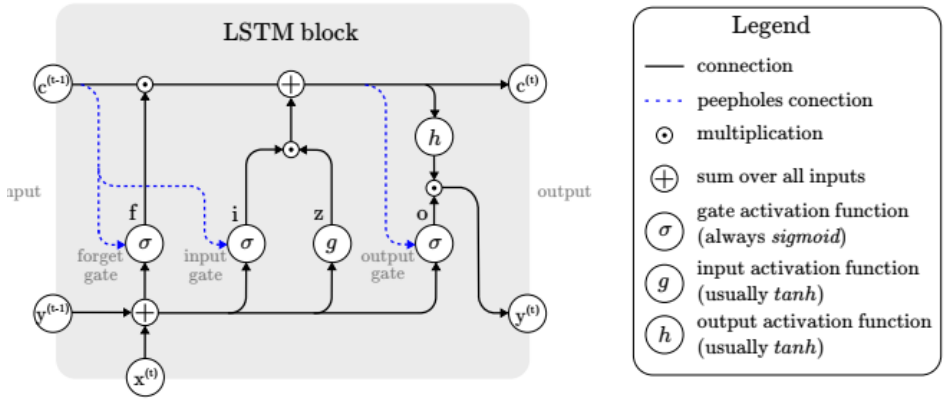
Ý tưởng của thuật toán

Mô hình LSTM (Long Short Term Memory) là một hệ thống thần kinh tái phát mạnh mẽ được thiết kế đặc biệt để khắc phục các vấn đề về độ dốc exploding /vanishing thường phát sinh khi học các phụ thuộc dài hạn, ngay cả khi thời gian trễ tối thiểu là rất dài. Nhìn chung, điều này có thể được ngăn chặn bằng cách sử dụng băng chuyền lỗi không đổi (CEC), duy trì tín hiệu lỗi trong từng ô của đơn vị. Trên thực tế, bản thân các tế bào như vậy là các mạng lặp lại, với một kiến trúc thú vị theo cách CEC được mở rộng với các tính năng bổ sung, cụ thể là cổng đầu vào và cổng đầu ra, tạo thành ô nhớ. Các kết nối tự lặp lại cho biết phản hồi với độ trễ là một bước thời gian.

Một đơn vị vanilla LSTM bao gồm một tế bào, một cổng vào, một cổng ra và một cổng quên. Cổng quên này ban đầu không phải là một phần của mạng LSTM, nhưng được đề xuất bởi *Gers et al* để cho phép mạng thiết lập lại trạng thái của nó. Tế bào ghi nhớ các giá trị trong khoảng thời gian tùy ý và ba cổng điều chỉnh luồng thông tin liên kết với ô. Trong phần còn lại của phần này, LSTM sẽ đề cập đến phiên bản vanilla vì đây là kiến trúc LSTM phổ biến nhất.

Nhìn chung, kiến trúc LSTM bao gồm một tập hợp các mạng con được kết nối lặp lại, được gọi là các khối bộ nhớ. Ý tưởng đằng sau khối bộ nhớ là duy trì trạng thái của nó theo thời gian và điều chỉnh luồng thông tin nghĩ đến các đơn vị đo lường phi tuyến. Hình 1 hiển thị kiến trúc của một khối LSTM vani, bao gồm các cổng, tín

hiệu đầu vào $x^{(t)}$, đầu ra $y^{(t)}$, các chức năng kích hoạt và kết nối "lỗ nhìn trộm". Đầu ra của khối được kết nối liên tục trở lại đầu vào của khối và tất cả các cổng.



Hình 3.2: Kiến trúc của một khối LSTM vani điển hình

Nhằm mục đích làm rõ cách hoạt động của mô hình LSTM, chúng ta hãy giả sử một mạng bao gồm N khối xử lý và M đầu vào. Chuyển tiếp trong hệ thống thần kinh tái phát này được mô tả dưới đây.

Đầu vào khối. Bước này dành để cập nhật thành phần đầu vào khối, kết hợp đầu vào hiện tại $x^{(t)}$ và đầu ra của đơn vị LSTM $y^{(t-1)}$ trong lần lặp cuối cùng. Điều này có thể được thực hiện như mô tả bên dưới:

$$z^{(t)} = g(W_z x^{(t)} + R_z y^{(t-1)} + b_z)$$

W_z và R_z là các trọng số kết hợp với $x^{(t)}$ và $y^{(t-1)}$ tương ứng

b_z đại diện cho vector trọng số dự báo

Cổng đầu vào. Trong bước này, chúng tôi cập nhật cổng đầu vào kết hợp đầu vào hiện tại $x^{(t)}$, đầu ra của đơn vị LSTM đó $y^{(t-1)}$ và giá trị tế bào $c^{(t-1)}$ trong lần lặp cuối cùng. Phương trình sau đây cho thấy quy trình này:

$$i^{(t)} = \sigma(W_i x^{(t)} + R_i y^{(t-1)} + p_i \odot c^{(t-1)} + b_i)$$

$\odot$ biểu thị phép nhân theo từng điểm của hai vector

W_i, R_i và p_i là trọng số liên quan đến $x^{(t)}$, $y^{(t-1)}$ và $c^{(t-1)}$ tương ứng

b_i đại diện cho vector dự báo được liên kết với thành phần này

Trong các bước trước, lớp LSTM xác định thông tin nào sẽ được giữ lại trong các trạng thái tế bào của mạng $c^{(t)}$. Điều này bao gồm việc lựa chọn các giá trị

ứng cử viên $z^{(t)}$ có thể được thêm vào các trạng thái tế bào và các giá trị kích hoạt $i^{(t)}$ của các cổng đầu vào.

Cổng quên. Trong bước này, đơn vị LSTM xác định thông tin nào cần được xóa khỏi các trạng thái tế bào trước đó $c^{(t-1)}$. Do đó, các giá trị kích hoạt $f^{(t)}$ của các cổng quên ở bước thời gian t được tính dựa trên đầu vào hiện tại $x^{(t)}$, đầu ra $y^{(t-1)}$ và trạng thái $c^{(t-1)}$ của ô nhớ ở bước thời gian trước đó (t - 1), các kết nối lỗ nhìn trộm và các điều khoản thiên vị b_f của các cổng quên. Điều này có thể được thực hiện như sau:

$$f^{(t)} = \sigma(W_f x^{(t)} + R_f y^{(t-1)} + p_f \odot c^{(t-1)} + b_f)$$

W_f, R_f và p_f là trọng số liên quan đến $x^{(t)}$, $y^{(t-1)}$ và $c^{(t-1)}$ tương ứng

b_f đại diện cho vector trọng số dự báo

Xử lí tế bào. Bước này tính toán giá trị tế bào, kết hợp giá trị đầu vào khối $z^{(t)}$, cổng đầu vào $i^{(t)}$ và cổng quên $f^{(t)}$, với giá trị tế bào trước đó. Điều này có thể được thực hiện như mô tả bên dưới:

$$c^{(t)} = z^{(t)} \odot i^{(t)} + c^{(t-1)} \odot f^{(t)}.$$

Đầu ra cổng. Bước này tính toán cổng đầu ra, kết hợp đầu vào hiện tại $x^{(t)}$, đầu ra của đơn vị LSTM đó $y^{(t-1)}$ và giá trị ô $c^{(t-1)}$ trong lần lặp cuối cùng. Điều này có thể được thực hiện như mô tả bên dưới:

$$o^{(t)} = \sigma(W_o x^{(t)} + R_o y^{(t-1)} + p_o \odot c^{(t)} + b_o)$$

W_o, R_o và p_o là trọng số liên quan đến $x^{(t)}$, $y^{(t-1)}$ và $c^{(t-1)}$ tương ứng

b_o đại diện cho vector trọng số dự báo

Khối đầu ra. Cuối cùng, đầu ra khối được tính toán, kết hợp giá trị ô hiện tại c (t) với giá trị cổng đầu ra hiện tại như sau:

$$y^{(t)} = g(c^{(t)}) \odot o^{(t)}.$$

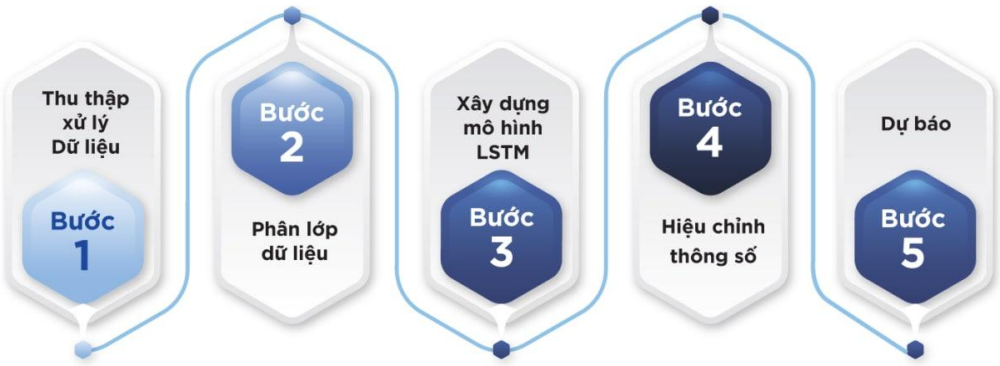
Trong các bước trên, σ, g và h biểu thị các hàm kích hoạt phi tuyến tính theo từng điểm. Hàm luận lý sigmoid $\sigma(x) = \frac{1}{1+e^{1-x}}$ được sử dụng làm hàm kích hoạt cổng,

trong khi tiếp tuyến hyperbol g(x) = h(x) = tanh(x) thường được sử dụng làm hàm kích hoạt đầu vào và đầu ra khối.

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.



Hình 3.3: Các bước thực nghiệm cho mô hình

3.4 Kết luận chương

Dự báo chuỗi thời gian xảy ra khi đưa ra các dự đoán khoa học dựa trên dữ liệu đóng dấu thời gian lịch sử. Nó liên quan đến việc xây dựng các mô hình thông qua phân tích lịch sử và sử dụng chúng để quan sát và thúc đẩy việc đưa ra quyết định

chiến lược trong tương lai. Cụ thể trong luận văn này, các mốc thời gian về lưu lượng mạng sẽ được quan sát và xử lí nhờ vào phương pháp Time Series và thuật toán học máy LSTM. Chương tiếp theo sẽ tiến hành thực nghiệm các phương pháp và thuật toán đề xuất trên bộ dữ liệu về lưu lượng mạng di động của một nhà mạng tại Việt Nam.

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 [13]: 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.

Tensorflow [14]: 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

Trong nghiên cứu này, chúng tôi đánh giá mô hình LSTM trên tập dữ liệu chuỗi lưu lượng thời gian thực của thu thập từ mạng di động tỉnh Tây Ninh (với số mẫu và số trạm hạn chế). Dữ liệu chuỗi thời gian này được thu thập theo giờ từ ngày 1/9/2021 đến ngày 10/12/2021 với 13 trường (biến) dữ liệu liên quan đến lưu lượng như là PS Traffic Total (GB), DL User Thput (Mbps), vv...

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.

Bộ dữ liệu trên được lấy trực tiếp trên hệ thống OSS của đơn vị cung cấp dịch vụ viễn thông tại Tây Ninh.

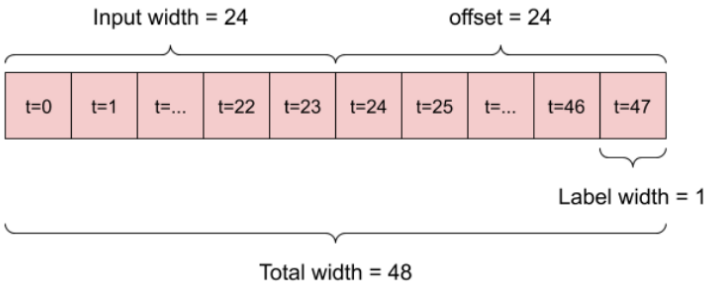
Bước tiếp theo cũng quan trọng không kém chính là chuẩn hóa dữ liệu. Điều quan trọng là phải mở rộng quy mô các đặc trưng trước khi huấn luyện mô hình mạng nơ-ron. Chuẩn hóa là một cách phổ biến để thực hiện việc chia tỷ lệ: trừ giá trị trung bình và chia cho độ lệch chuẩn của mỗi đặc trưng. Giá trị 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

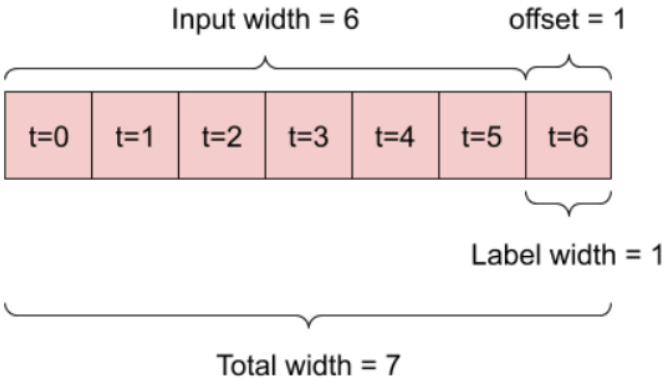
trưng chính cho khung là: chiều rộng (số lượng các bước nhảy trong thời gian) đầu vào và nhãn của khung, thời gian bù giữa các khung và các loại đặc trưng như dữ liệu đầu vào, nhãn hoặc cả hai. Ở bước này sẽ tập trung vào hoàn tất việc thiết lập khung thời gian, vì thế nó có thể được tái sử dụng nhiều 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:

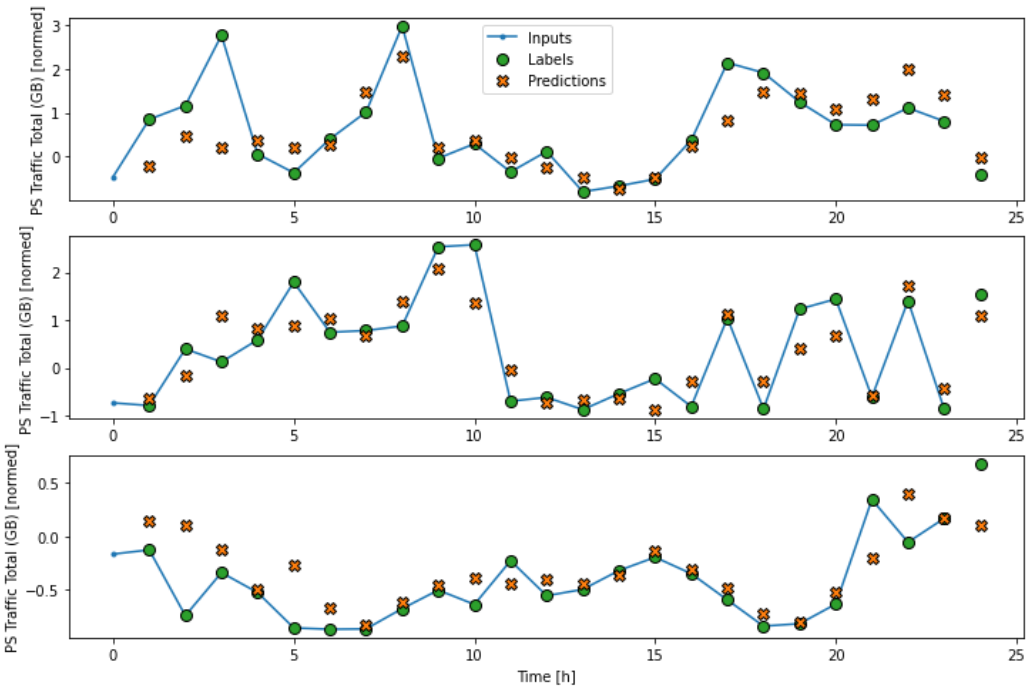
- **Tập dữ liệu có nhãn A**: có kích thước gồm 1075250 dòng dữ liệu với các trường dữ liệu khác nhau, trong đó trường dữ liệu "PS Traffic Total (GB)" có lưu lượng từ dưới 1.65 GB.

- **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 quá trình thực nghiệm với bộ dữ liệu, ngoài thuật toán LSTM, luận văn cũng đã áp dụng mô hình sử dụng thuật toán CNN, tuy nhiên kết quả thực nghiệm của mô hình LSTM cho thấy các độ đo mất mát như MAE, MSE, MSLE khả quan và tốt hơn, từ đó luận văn đã áp dụng mô hình sử dụng thuật toán LSTM trong suốt quá trình thực nghiệm.

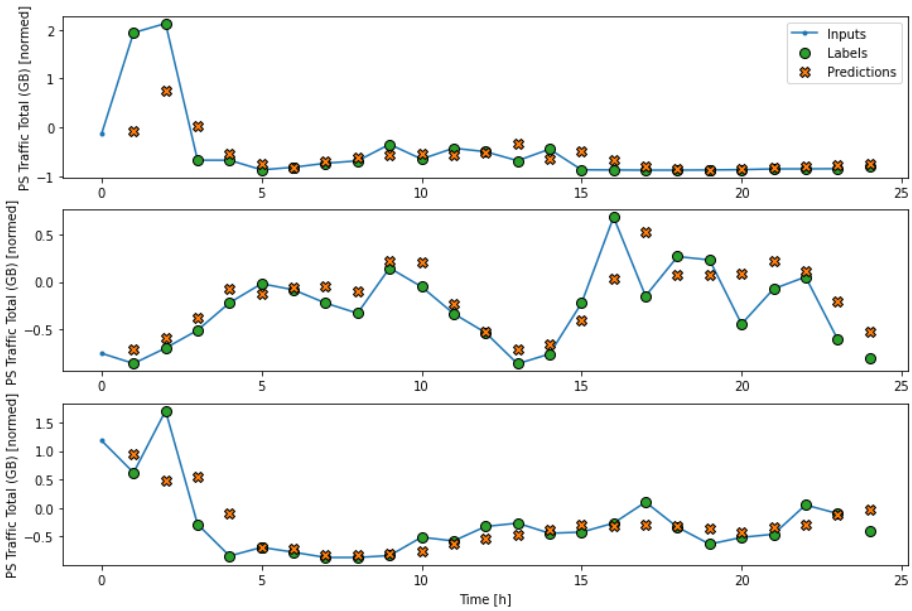
Trong các lần thử nghiệm, mô hình LSTM với thời gian huấn luyện là 24 giờ tại một thời điểm.

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 – MSLE, thu được kết quả như sau:

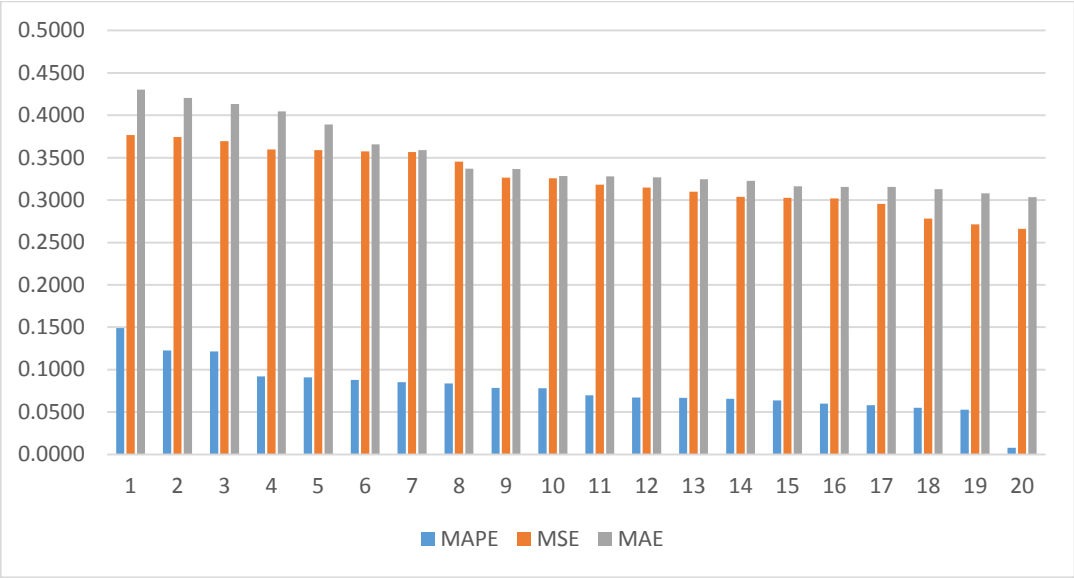


Hình 4.4: Mô hình tập dữ liệu nhãn A với độ đo MSLE

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

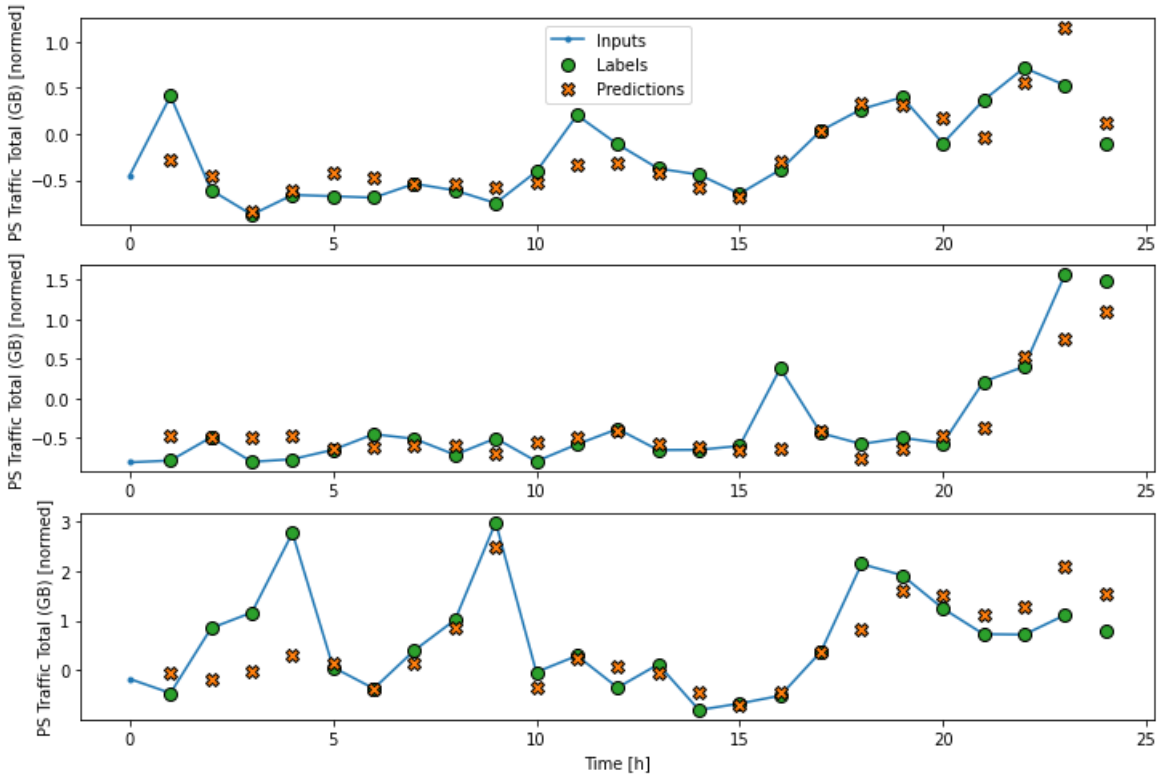
Epoch	Độ đo mất mát (Loss)		
	MSLE	**MSE**	**MAE**
1	0.1491	0.3765	0.4301
2	0.1227	0.3745	0.4207
3	0.1216	0.3695	0.4133
4	0.0921	0.3596	0.4048
5	0.0910	0.3590	0.3893
6	0.0879	0.3576	0.3657
7	0.0852	0.3567	0.3591
8	0.0839	0.3455	0.3371
9	0.0784	0.3265	0.3367
10	0.0783	0.3258	0.3284
11	0.0698	0.3182	0.3281
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 MSLE 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.

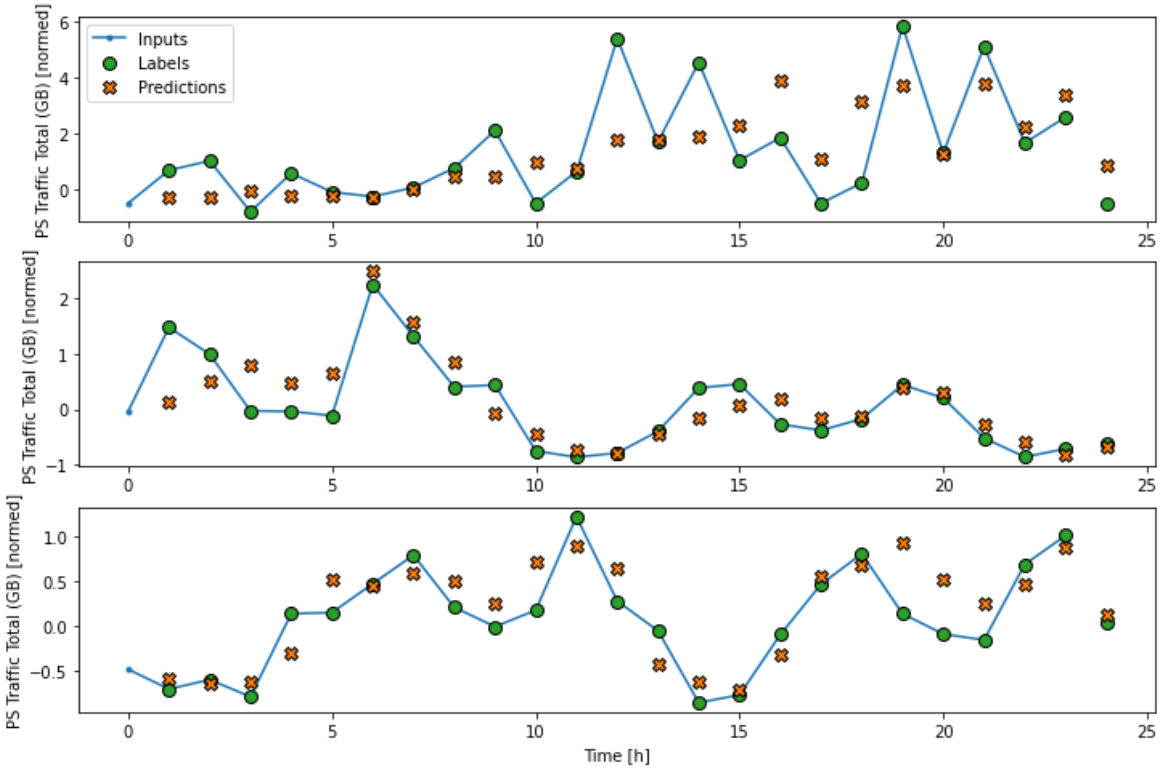


Hình 4.5: Biểu đồ so sánh độ đo mất mát tập dữ liệu A

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 MSLE



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

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.

Mô hình dự báo đề xuất sử dung LSTM với chuỗi thời gian có thể dự báo các giá trị trong tương lai dựa trên dữ liệu tuần tự, trước đó. Thêm vào đó, mô hình này cung cấp độ chính xác cao hơn so với các mô hình học máy đơn lẽ được công bố [11]

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 cho mô hình. 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.

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] Xun Xu, Shuo Zeng and Yuanjie He, "The impact of information disclosure on consumer purchase behavior on sharing economy platform Airbnb," *International Journal of Production Economics,* 2021.

[3] Ji, Byoungsuk, and Ellen J. Hong, "Deep-Learning-Based Real-Time Road Traffic Prediction Using Long-Term Evolution Access Data," *Sensors 19,* 2019.

[4] Fawaz Waselallah Alsaade, Mosleh Hmoud Al-Adhaileh, "Cellular Traffic Prediction Based on an Intelligent Model," *Mobile Information Systems,* 2021.

[5] D. Clemente, G. Soares, D. Fernandes, R. Cortesao, P. Sebastiao and L. S. Ferreira, "Traffic Forecast in Mobile Networks: Classification System Using Machine Learning," *2019 IEEE 90th Vehicular Technology Conference (VTC2019-Fall),* 2019.

[6] D. Clemente, D. Fernandes, R. Cortesão, G. Soares, P. Sebastião and L. S. Ferreira, "Assessment of Traffic Prediction Models for Mobile Communication Networks," *2019 22nd International Symposium on Wireless Personal Multimedia Communications (WPMC),* 2019.

[7] Guen, V.L. and Thome, N, "Shape and time distortion loss for training deep time series forecasting models," *arXiv preprint arXiv:1909.09020,* 2019.

[8] 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.

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

[10] 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.

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

[12] 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/.

[13] Google, "Google Colaboratory," [Online]. Available: https://colab.research.google.com/.

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

[15] 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.

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

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

[18] M. Schuster and K. K. Paliwal, "Bidirectional recurrent neural networks," pp. 2673-2681, Nov 1997.

[19] E. Alpaydin, Introduction to Machine Learning, Fourth Edition.

[20] Schmidt, J., Marques, M.R.G., Botti, S. et al, "Recent advances and applications of machine learning in solid-state materials science," *npj Comput Mater 5*, 2019.

[21] Wikipedia contributors, "Code-division multiple access," 9 December 2021. [Online]. Available: https://en.wikipedia.org/wiki/Code-division_multiple_access.

[22] S. Ndungu, "GSM (Global System for Mobile communication)," Search Mobile Computing, [Online]. Available: https://searchmobilecomputing.techtarget.com/definition/GSM.

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 DoIT một cách trung thực và đạt kết quả mức độ tương đồng 4% 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.

TPHCM, ngày 25 tháng 01 năm 2022

HỌC VIÊN CAO HỌC

Nguyễn Xuân Quốc

Hệ thống hỗ trợ nâng cao chất lượng tài liệu

KẾT QUẢ KIỂM TRA TRÙNG LẶP TÀI LIỆU

THÔNG TIN TÀI LIỆU

Tác giả	Nguyễn Xuân Quốc
Tên tài liệu	NGHIEN CUU MO HINH HOC MAY CHO DỰ BAO LƯU LƯỢNG TRONG MẠNG DI ĐỘNG
Thời gian kiểm tra	25-01-2022, 02:13:00
Thời gian tạo báo cáo	25-01-2022, 02:15:19

KẾT QUẢ KIỂM TRA TRÙNG LẶP

Điểm	4
Nguồn trùng lặp tiêu biểu	[text.123doc.org, 123doc.org, vi.wikipedia.org]

Học viên **Người hướng dẫn khoa học**

Nguyễn Xuân Quốc **TS. Nguyễn Xuân Sâm**