

Ứng dụng mô hình mạng học sâu dự báo lượng khách hàng

Nguyễn Thu Hằng
Phạm Thị Nguyệt
Trường Đại học Mở - Địa chất

Bài báo được thực hiện trong khuôn khổ đề tài KHCN cấp cơ sở mã số T22 - 05, quyết định số 65/QĐ-MDC.

Trong cuộc cách mạng công nghệ 4.0, trí tuệ nhân tạo (AI) đang trở thành xu hướng toàn cầu và được ứng dụng phổ biến trong mọi ngành nghề, lĩnh vực, tổ chức, doanh nghiệp. Tại Việt Nam trong những năm gần đây, các giải pháp sử dụng trí tuệ nhân tạo trong phân tích dự báo thu hút sự quan tâm của các doanh nghiệp, giúp tối ưu hóa quy trình và tài nguyên, đồng thời giảm chi phí và nâng cao hiệu quả hoạt động của doanh nghiệp. Do đó ngày càng có nhiều nghiên cứu về các giải pháp phân tích dự đoán hỗ trợ bởi AI được triển khai và áp dụng trong kinh tế, tài chính.

1. Mở đầu

Phân tích dữ liệu và dự báo là một phần của khoa học dữ liệu, đây là bài toán rất có ý nghĩa, giúp hỗ trợ các nhà quản lý đưa ra quyết định đúng đắn dựa vào các thông tin hữu ích (được gọi là tri thức) được trích rút từ dữ liệu thu thập được. Trong các lớp bài toán dự báo, phân tích và dự báo dữ liệu dạng chuỗi thời gian là bài toán có thể được ứng dụng rộng rãi trong nhiều hoạt động khác nhau của bất kỳ tổ chức kinh doanh nào. Các doanh nghiệp và ngành công nghiệp hiện nay đều có số lượng lớn dữ liệu chuỗi thời gian, việc khai thác nguồn dữ liệu lớn đó một cách hiệu quả để rút ra những tri thức đắt giá giúp hỗ trợ ra quyết định, điều hành hoạt động sản xuất kinh doanh là nhu cầu cấp thiết. Điều đó đã làm cho dự báo chuỗi thời gian trở thành một lĩnh vực nghiên cứu quan trọng, và là công cụ hữu ích có thể giúp doanh nghiệp hiểu dữ liệu lịch sử ảnh hưởng như thế nào đến tương lai. Điều này được thực hiện bằng cách xem xét dữ liệu trong quá khứ, xác định các mẫu và đưa ra các dự đoán ngắn hạn hoặc dài hạn.

2. Những nghiên cứu liên quan

Tác giả Giang Thị Thu Huyền đã sử dụng các phương pháp học máy để chấm điểm tín dụng trong các tổ chức tài chính và ngân hàng hiện nay. Nghiên cứu thực hiện thực nghiệm trên bộ dữ liệu Kaggle, là bộ dữ liệu được sử dụng rộng rãi trong cộng đồng nghiên cứu.

Ngoài ra, nhóm nghiên cứu của Nguyễn Huy Hòa đã xây dựng thuật toán để định giá bất động sản. Nghiên cứu này đã áp dụng phương pháp LASSO (Nguyễn Huy Hòa và nnk 2016) để tinh chỉnh các hệ số hồi quy phi tuyến nhằm giảm phương sai của các ước lượng, từ đó giúp tăng độ chính xác của kết quả

định giá. Nhóm tác giả Huỳnh Quyết Thắng và cộng sự giới thiệu phương pháp dự đoán xu thế chỉ số chứng khoán Việt Nam (VN-Index) gồm bốn bước, trong đó dữ liệu đầu vào là chuỗi thời gian chứa lịch sử chỉ số giá của VN-Index (Huỳnh Quyết Thắng và nnk 2018). Nhóm áp dụng mô hình kết hợp GPR-ARMA, gồm mô hình tự hồi quy trung bình động (ARMA: AutoRegressive Moving Average) và mô hình phân tích hồi quy quá trình Gauss (GPR: Gaussian Process Regression) để dự đoán theo từng bước.

Những nghiên cứu nêu trên đã ứng dụng các phương pháp học máy, học sâu trong một số bài toán khác nhau. Tuy nhiên thực tế cho thấy hiện nay, các công trình nghiên cứu về kỹ thuật học máy, học sâu ứng dụng trong các ngành nghề chung và trong lĩnh vực kinh tế, tài chính nói riêng tại Việt Nam vẫn còn rất hạn chế.

Trong khi đó, trên thế giới các phương pháp học máy đã và đang được ứng dụng khá rộng rãi trong kinh tế, tài chính nói chung cũng như trong lớp bài toán phân tích dữ liệu dạng chuỗi thời gian nói riêng. Hai thuật toán học máy nổi tiếng trong nhóm này là Random Forest và Gradient Boosting.

Gần đây, các mô hình học sâu DNNs (Deep Neural Networks) được tập trung nghiên cứu phát triển và ứng dụng ngày càng rộng rãi trong kinh tế tài chính cũng như trong vấn đề dự báo chuỗi thời. Trong số đó, có thể kể đến mạng RNN (Recurrent Neural Networks) và LSTM (Long Short-Term Memory) được dùng khá phổ biến và được tích hợp trong thư viện Keras. Các mạng này tìm ra sự phụ thuộc thứ tự giữa các mẫu trong một chuỗi dữ liệu để phân tích dự báo. N-BEATS là một thuật toán học sâu tùy chỉnh dựa trên các liên kết lan truyền ngược và liên kết chuyển tiếp. Ngoài ra, các mô hình DNNs khác

như DeepAR, TFT (Temporal Fusion Transformer) và các mô hình dự báo đa biến sử dụng kết hợp chuỗi dữ liệu lịch sử với các thông tin phụ trợ như giá cả, chiết khấu, điều kiện thời tiết, dân số,...) cũng được đánh cao trong dự báo kinh tế tài chính.

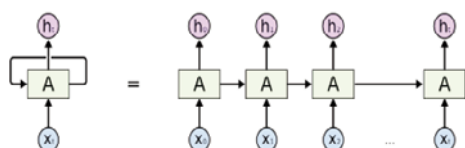
Có thể thấy, mặc dù các lĩnh vực kinh tế, tài chính trên thế giới đã áp dụng phương pháp học máy và học sâu khá là rộng rãi, nhưng các hoạt động nghiên cứu về ứng dụng này tại Việt Nam vẫn còn hạn chế. Thúc đẩy bởi xu hướng này, bài báo của chúng tôi tập trung vào việc nghiên cứu, ứng dụng mô hình mạng học sâu vào bài toán dự báo với dữ liệu chuỗi thời gian, một dạng dữ liệu phổ biến trong kinh tế, tài chính. Với những phân tích và nghiên cứu thu thập được trong thời gian qua, bài báo đề xuất sử dụng hai mô hình mạng học sâu là RNN và LSTM cho bài toán dự đoán tổng lượng khách hàng tới các nhà hàng tại Mỹ, có tính đến sự ảnh hưởng của các ngày nghỉ lễ trong năm.

3. Các mô hình mạng học sâu dự báo

3.1. Recurrent Neural Network (RNN)

Mạng nơ-ron hồi quy (RNN) là loại mạng có cấu trúc được thiết kế để xử lý chuỗi dữ liệu tuần tự (hình 1). RNN không chỉ sử dụng dữ liệu đầu vào mà còn sử dụng các đầu ra trước đó để dự đoán đầu ra hiện tại. Sự liên kết giữa các nút mạng trong mô hình RNN tạo nên một sơ đồ có hướng và bản thân bộ nhớ trong được sử dụng để xử lý các chuỗi đầu vào liên tục. Trạng thái của mỗi nút thay đổi theo thời gian bởi hàm kích hoạt có giá trị thực. Mô hình huấn luyện trong RNN được xác định bằng cách chuyển đổi giữa các trạng thái nên nó luôn có cùng kích thước đầu vào. Mặt khác, cùng một hàm chuyển đổi sẽ có cùng hệ số ở mỗi bước đã được sử dụng trong hệ thống.

Hình 1: Mô hình RNN



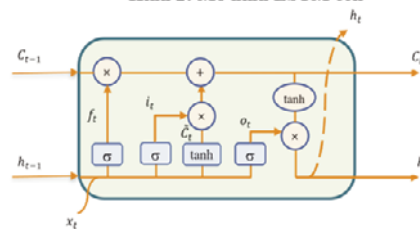
Mạng RNN có cấu trúc tuần hoàn, có thể ghi nhớ các đầu vào trước đó, ghi nhớ từng thông tin theo thời gian, do đó rất hữu ích trong dự đoán cho dữ liệu chuỗi thời gian. Thậm chí, mạng RNN còn được sử dụng với các lớp mạng tích chập để mở rộng các giá trị trong ma trận pixel một cách hiệu quả. Ngoài ra, mối tương quan giữa các mẫu cũng được giải thích thông qua các mô hình hóa chuỗi dữ liệu trong mạng RNN.

3.2. Long Short - Term Memory (LSTM)

Mạng LSTM là một dạng đặc biệt của RNN, với cấu trúc gồm các cổng có khả năng ghi nhớ và học được các thông tin và ràng buộc trong khoảng thời

gian dài mà không cần bất kỳ sự can thiệp nào. Các mạng hồi quy đều có dạng là một chuỗi các mô-đun lặp đi lặp lại của mạng nơ-ron. Đối với mạng RNN chuẩn, các mô-đun có cấu trúc rất đơn giản, thường là một tầng tanh. Mặc dù LSTM cũng có kiến trúc dạng chuỗi như mạng RNN chuẩn, nhưng các mô-đun bên trong sẽ có cấu trúc phức tạp hơn. Thay vì chỉ có một tầng mạng nơ-ron, chúng có tới 4 tầng tương tác với nhau một cách rất đặc biệt.

Hình 2: Mô hình LSTM cell



Trong cấu trúc của mạng LSTM (hình 2), có thể nói trạng thái tế bào - cell state được coi là chìa khóa. Thông tin có thể chạy xuyên suốt tất cả các nút mạng, dễ dàng được truyền đi mà không bị thay đổi nhờ cơ chế hoạt động giống như băng chuyền của cell state. Cell state có thể nhận được các thông tin hữu ích thông qua điều chỉnh của các cổng (gate), nơi có thể bỏ hoặc nhận thêm các thông tin cần thiết. Thông tin đi qua các cổng sẽ được sàng lọc, sau đó được kết hợp thông qua một tầng mạng sigmoid và một phép nhân. Tầng sigmoid sẽ cho đầu ra là một số trong khoảng [0, 1], kết quả sẽ mô tả có bao nhiêu thông tin có thể được thông qua. Không có thông tin nào được cho qua khi kết quả là 0, ngược lại tất cả thông tin sẽ được đi qua nó khi đầu ra là 1. Mỗi một cell của LSTM gồm có 3 cổng: đầu vào (it), đầu ra (ot) và cổng quên (ft).

Mô hình LSTM cơ bản chỉ bao gồm một lớp ẩn. Mạng được mở rộng bằng cách thêm các lớp LSTM ẩn được gọi là mô hình LSTM xếp chồng (Stacked LSTM). Mỗi lớp ẩn của mạng này bao gồm nhiều cell xếp chồng lên nhau. Hơn nữa, ta sẽ nhận được một kết quả sau mỗi một bước thời gian đầu vào qua mỗi lớp. Do đó, mỗi lớp ẩn của LSTM có thể có một chuỗi đầu ra cho tất cả các bước thời gian đầu vào, thay vì một đầu ra như các mô hình khác.

4. Thử nghiệm và kết quả

4.1. Dữ liệu thử nghiệm

Để đánh giá và so sánh các mô hình đề xuất, chúng tôi thực hiện thử nghiệm với dữ liệu chuỗi thời gian dự báo lượng khách hàng với 2 biến, bộ dữ liệu lấy từ trang web Pierian Training, dữ liệu được xây dựng dựa trên cuộc thi Kaggle gần đây. Bộ dữ liệu thu thập số lượng khách, hàng ngày đến bốn nhà hàng tại Mỹ, và sự ảnh hưởng của các ngày lễ của Mỹ. Bộ dữ liệu gồm 478 ngày dữ liệu của nhà hàng, với 39 dữ liệu về ngày nghỉ lễ tại Mỹ trong khoảng thời gian từ 01/01/2016 đến 31/5/2017 (xem dữ liệu

trong bảng 1). Tập dữ liệu đều được chia thành hai phần để huấn luyện và thử nghiệm các mô hình đề xuất theo tỷ lệ: 80% cho dữ liệu huấn luyện, và 20% còn lại là tập dữ liệu thử nghiệm.

Bảng 1: Dữ liệu về lượng khách đến nhà hàng tại Mỹ

| TT | Ngày (date) | Ngày nghỉ (holiday) | Tổng lượng khách (total) |
|------|-------------|---------------------|--------------------------|
| 1 | 1/1/2016 | 1 | 296 |
| 2 | 1/2/2016 | 0 | 191 |
| 3 | 1/3/2016 | 0 | 202 |
| 4 | 1/4/2016 | 0 | 105 |
| 5 | 1/5/2016 | 0 | 98 |
| | | | |

4.2. Kiến trúc và các hệ số mô hình dự báo

Chúng tôi sử dụng thư viện Keras cho Python để cài đặt thực nghiệm mô hình mạng học sâu DNN và công cụ jupyter notebook để thực hiện phân tích. Khi huấn luyện dữ liệu, cả 2 mạng RNN và Stacked LSTM đều sử dụng lỗi bình phương trung bình (MSE) cho hàm mất mát (loss), và đều được tối ưu hóa (optimized) bằng thuật toán Adam. Số lượng epochs cho quá trình huấn luyện mô hình là 1000.

Mô hình RNN có hàm kích hoạt được sử dụng là ReLU, các lớp mạng và hệ số tương ứng được thiết lập như bảng 2.

Bảng 2: Cấu trúc chi tiết lớp của mạng RNN

| Lớp số (Layer number) | Tên lớp (Layer name) | Hệ số (Parameter) |
|-----------------------|----------------------|-------------------|
| 1 | LSTM | 100 |
| 2 | Dense | 1 |

Mô hình Stacked LSTM được thực nghiệm cho các trường hợp với 2 lớp LSTM, số lượng các nút trong các lớp tuân theo quy tắc: số lượng các nút trong các lớp ẩn bằng 2/3 kích thước của lớp đầu vào, đồng thời 1 lớp Dropout được đặt trước lớp Dense để tránh hiện tượng overfitting. Chi tiết cấu trúc của mạng Stacked LSTM được thể hiện trong bảng 3.

Bảng 3: Cấu trúc chi tiết lớp của mạng Stacked LSTM với 2 lớp LSTM

| Lớp số (Layer number) | Tên lớp (Layer name) | Hệ số (Parameter) |
|-----------------------|----------------------|-------------------|
| 1 | LSTM | 45 |
| 2 | LSTM | 30 |
| 3 | Dropout | Rate = 0.3 |
| 4 | Dense | 1 |

4.3. Ma trận đánh giá mô hình

Để đánh giá chất lượng của mô hình dự báo, chúng tôi sử dụng hai hệ số được sử dụng rộng rãi trong cộng đồng nghiên cứu là sai số toàn phương trung bình MSE (Mean squared error) và lỗi trung bình bình phương (RMSE - Root Mean Squared Error). Các hệ số được tính tương ứng trong công thức như sau:

$$MSE(x, \hat{x}) = \frac{1}{N} \sum_{t=1}^N (x_t - \hat{x}_t)^2$$

$$RMSE(x, \hat{x}) = \sqrt{\frac{1}{N} \sum_{t=1}^N (x_t - \hat{x}_t)^2}$$

Trong đó: N là số lượng mẫu trong tập dữ liệu; x_t là giá trị quan sát được (hay giá trị chính xác); \hat{x}_t là giá trị dự báo của mô hình.

4.4. Kết quả và thảo luận

Kết quả đánh giá sau khi chạy thực nghiệm hai mô hình mạng RNN và Stacked LSTM được thể hiện trong bảng 4. Mô hình dự báo của RNN có MSE và RMSE lần lượt là 1969 và 44.37. Với mô hình Stacked LSTM ta thu được kết quả dự báo tốt hơn với ma trận đánh giá MSE là 1738, RMSE là 41.69.

Bảng 4: Kết quả ma trận đánh giá mô hình

| Mô hình mạng | Hệ số MSE | Hệ số RMSE |
|--------------|-----------|------------|
| RNN | 1969 | 44.37 |
| Stacked LSTM | 1738 | 41.69 |

Như vậy có thể nói với cấu trúc của mạng LSTM với các “cổng tính toán” (“computational gates”), đã có sự cân bằng giữa bộ nhớ ngắn hạn của mạng (short-term memory), gồm dữ liệu được dùng để huấn luyện gần đây, với bộ nhớ dài hạn của mạng (long-term memory), gồm các dữ liệu đầu tiên được huấn luyện, đã cho kết quả tốt hơn khi dự báo với bộ dữ liệu chuỗi thời gian so với cấu trúc của mạng RNN./.

Tài liệu tham khảo

O. B. Sezer, M. U. Gudelek, and A. M. Ozbayoglu, “Financial time series forecasting with deep learning: A systematic literature review: 2005–2019,” *Applied Soft Computing*, vol. 90, p. 106181, May 2020,

A. Agapitos, A. Brabazon, and M. O’Neill, “Regularised gradient boosting for financial time-series modelling,” *Comput Manag Sci*, vol. 14, no. 3, pp. 367–391, Jul. 2017,

B. N. Oreshkin, D. Carпов, N. Chapados, and Y. Bengio, “N-BEATS: Neural basis expansion analysis for interpretable time series forecasting,” arXiv:1905.10437 [cs, stat], Feb. 2020, Accessed: Oct. 26, 2021. [Online]. Available: <http://arxiv.org/abs/1905.10437>

B. M. Pavlyshenko, “Machine-Learning Models for Sales Time Series Forecasting,” *Data*, vol. 4, no. 1, Art. no. 1, Mar. 2019, doi: 10.3390/data4010015.

R. P. Durbin, “Letter: Acid secretion by gastric mucous membrane,” *Am J Physiol*, vol. 229, no. 6, p. 1726, Dec. 1975, .

S. Hochreiter, “The Vanishing Gradient Problem During Learning Recurrent Neural Nets and Problem Solutions,” *Int. J. Unc. Fuzz. Knowl. Based Syst.*, vol. 06, no. 02, pp. 107–116, Apr. 1998,

N.-T. Nguyen, N.-N. Dao, Q.-D. Pham, and H. A. Le, Eds., *Intelligence of Things: Technologies and Applications: The First International Conference on Intelligence of Things (ICIoT 2022)*, Hanoi, Vietnam,