SO SÁNH THUẬT TOÁN TĂNG CƯỜNG ĐỘ DỐC (XGBOOST) VỚI MỘT SỐ THUẬT TOÁN HỌC MÁY KHÁC

Trần Quý Nam
Học Viện Công Nghệ Bưu Chính Viễn Thông

Tóm tắt: Nghiên cứu này thử nghiệm một số thuật toán học máy gồm XGBoost, rừng ngẫu nhiên, cây quyết định, máy vectơ hỗ trợ, k láng giềng gần nhất cho bài toán dự đoán. Tập dữ liệu thị trường chứng khoán NASDAQ của Hoa Kỳ được lấy đến hết tháng 5/2021 làm dữ liệu thử nghiệm cho các thuật toán học máy. Giá trị dự đoán và huấn luyện là giá cổ phiếu trên thị trường chứng khoán dựa trên bộ dữ liệu của AAPL. Kết quả thử nghiệm cho thấy phương pháp XGBoost cho kết quả dự đoán tốt nhất so với các thuật toán học máy khác.

Từ khóa: Học máy, AI, XGBoost, rừng ngẫu nhiên, cây quyết định, máy vectơ hỗ trợ, k láng giềng gần nhất, NASDAQ, giá cổ phiếu.

I. GIỚI THIỆU

Ngành công nghệ tài chính (FinTech) hiện nay đang phát triển mạnh mẽ và có tác động tích cực đến một số hoạt động của ngành tài chính hiện đại. Các ứng dụng công nghệ phân tích dữ liệu để tìm quy luật thị trường đang ngày càng được sử dụng rộng rãi, cho kết quả nhanh và xử lý được nguồn dữ liệu lớn. Sự kết hợp giữa công nghệ thông tin và phân tích tài chính đã mở ra không gian phát triển mới, cả về học thuật và ứng dụng thực tế. Cùng với sự phát triển các thuật toán trí tuệ nhân tạo (AI) về tốc độ, độ chính xác và khả năng xử lý, các ứng dụng AI ngày càng giúp hỗ trợ con người nâng cao khả năng phân tích và xử lý các nguồn dữ liệu tài chính. Trên thực tế, các thuật toán trí tuệ nhân tạo có ứng dụng phổ biến trong nhiều lĩnh vực tài chính như: dự đoán thị trường chứng khoán, phân tích thị trường đầu tư cho doanh nghiệp, phân tích hành vi người tiêu dùng, gợi ý danh mục ngành nghề đầu tư có tiềm năng lãi suất cao, phân tích các yếu tố tác động và dự báo sự phát triển của các ngành tài chính, ngân hàng, bảo hiểm,...

Thị trường chứng khoán toàn cầu có giá trị vốn hóa rất lớn với một lưu lượng tiền khổng lồ. Năm 2019, tổng giá trị cổ phiếu toàn cầu của thị trường chứng khoán đạt khoảng 85 nghìn tỷ đô la Mỹ [1]. Trong những năm qua, bài toán dự đoán thị trường chứng khoán luôn thu hút được nhiều sự quan tâm của các nhà đầu tư và các nhà nghiên cứu để hỗ trợ ra quyết định, cung cấp thông tin tham khảo để ra quyết định đầu tư với mong muốn thu được lợi nhuận cao hơn. Trên thực tế, giá cổ phiếu không chỉ phụ thuộc vào biến động lịch sử của giá quá khứ trước kia mà còn phụ thuộc nhiều vào nhiều yếu tố khó lường, chẳng hạn như đầu tư của chính phủ, chỉ số chứng khoán của các quốc gia khác, chỉ số giá hàng hóa liên quan, thậm chí cả những tin đồn thất thiệt, các tin tức liên quan,… Tuy nhiên, việc dự đoán giá cổ phiếu dựa trên dữ liệu lịch sử và sự biến động của chúng trong quá khứ cũng giúp cung cấp thông tin tham khảo có giá trị cho các nhà đầu tư trên thị trường để có thể ra quyết định đầu tư tốt nhất cho họ.

II. CÁC NGHIÊN CỨU LIÊN QUAN

Trên thực tế, các học giả quốc tế đã có rất nhiều nghiên cứu về dự đoán giá cổ phiếu trên thế giới. Một trong số đó là nghiên cứu [2] đã cho thấy mối liên hệ chặt chẽ giữa mức độ dự đoán đúng giá chứng khoán với việc sử dụng các phương pháp học máy khác nhau. Qua rà soát các công trình nghiên cứu, bài toán dự đoán giá cổ phiếu thường được giải quyết bằng mô hình LSTM (Long Short Term Memory). Nghiên cứu [3] đã thử nghiệm mô hình LSTM áp dụng cho giá mở cửa (open price) hàng ngày của hai cổ phiếu NKE (Nike Inc) và GOOGL (Alphabet Inc) được niêm yết trên sàn giao dịch chứng khoán New York (NYSE). Kết quả cho thấy nếu dữ liệu đào tạo ít nhưng có nhiều lớp mạng hơn có thể cải thiện kết quả thử nghiệm và cung cấp các giá trị dự đoán tốt hơn. Công trình nghiên cứu [4] đã cung cấp đánh giá tổng thể 86 bài báo đã xuất bản từ năm 2015 đến đầu năm 2021 về giải quyết bài toán dự đoán giá chứng khoán. Bài báo của họ đã đánh giá các chỉ số hiệu năng chính của mô hình gồm RMSE, MAPE, MAE, MSE, độ chính xác, tỷ lệ Sharpe và tỷ lệ hoàn vốn. Bài báo này đã chỉ ra rằng phần lớn các nghiên cứu bài toán dự báo giá chứng khoán sử dụng mô hình LSTM, kết hợp với sử dụng các phương pháp khác (ví dụ DNN). Bài báo cũng chỉ ra các phương pháp học củng cố và các phương pháp học sâu khác cũng mang lại hiệu quả tốt.

Công trình [5] đã áp dụng mô hình mạng nơ-ron tích chập có cải tiến tính năng học nâng cao để áp dụng dự đoán biến động giá cổ phiếu. Kết quả của họ cho thấy độ chính xác dự đoán cao nhất của mô hình FA-CNN được đề xuất là 64,81% và cao hơn 7,38% so với mô hình LSTM truyền thống. Công trình [6] đã sử dụng các thuật toán học máy để dự đoán giá thị trường chứng khoán S&P 500. Các tác giả đã áp dụng các kỹ thuật Linear Regression,

Tác giả liên hệ: Trần Quý Nam
Email: namtq@ptit.edu.vn
Đến tòa soạn: 6/2021, chỉnh sửa: 7/2021, chấp nhận đăng: 7/2021

Multivariate Regression, Support Vector Regression, Decision Tree, Random Forest Regressor and Extra Tree Regressor. Kết quả nghiên cứu đã chứng minh Random Forest Regressor and Extra Tree Regressor là các thuật toán hồi quy tốt nhất cho bài toán dự đoán chứng khoán. Nghiên cứu [7] đã thử nghiệm thuật toán LSTM và Random Forest cho bài toán dự đoán giá cổ phiếu và cho kết quả dự đoán phụ thuộc vào kích thước bộ dữ liệu. Thuật toán LSTM và Random Forest chưa cung cấp được giá trị dự đoán chính xác cao đối với bộ dữ liệu nhỏ, nhưng nếu áp dụng cho bộ dữ liệu dài hạn và đủ lớn, kết quả dự đoán đã khá chính xác. Nghiên cứu [7] cũng chứng minh rằng thuật toán Random Forest có bổ sung một số đặc trưng kết xuất từ bộ dữ liệu huấn luyện cho kết quả dự đoán giá cổ phiếu chính xác hơn.

Công trình [8] đã thực hiện nghiên cứu áp dụng một số thuật toán học máy như Averaging, Linear Regression và các kỹ thuật học sâu như LSTM áp dụng vào tập dữ liệu của thị trường chứng khoán Ấn Độ. Bộ dữ liệu thử nghiệm bao gồm lợi nhuận, tỷ lệ phần trăm thay đổi giá đóng cửa (close price) của giá cổ phiếu các công ty khác nhau. Công trình [8] đã áp dụng một số kỹ thuật khác nhau như hồi quy tuyến tính, K-Means Clustering, K láng giềng gần nhất, LSTM,... Nghiên cứu [8] đã cho kết quả rằng việc sử dụng các thuật toán khác nhau áp dụng trên cùng tập dữ liệu sẽ cho kết quả khác nhau khi giải quyết bài toán dự đoán giá tương lai dựa trên dữ liệu giá cổ phiếu trong quá khứ. Điều này cho thấy kết quả dự đoán không chỉ phụ thuộc thuật toán áp dụng mà còn cả tập dữ liệu thử nghiệm.

Công trình nghiên cứu [9] đã thực hiện thử nghiệm một số thuật toán với các tham số khác nhau để dự đoán thị trường chứng khoán. Các tập dữ liệu thử nghiệm được lấy từ Kaggle.com. Các tác giả đã triển khai các thuật toán học có giám sát với tỷ lệ phân chia khác nhau đối với bộ dữ liệu huấn luyện và bộ dữ liệu dự đoán. Các tác giả thử nghiệm thuật toán áp dụng cho kích thước phân chia bộ dữ liệu huấn luyện và dự đoán theo các tỷ lệ 70:30, 50:50 và 30:70. Kết quả nghiên cứu chỉ ra thuật toán cho hiệu quả tốt nhất là KNN (K-Nearest Neighbor) với tỷ lệ phân chia dữ liệu có kich thước là 70:30.

Nghiên cứu [10] thử nghiệm giải quyết bài toán dự báo thị trường chứng khoán Trung Quốc với bộ dữ liệu trong 8 năm liên tiếp. Nội dung nghiên cứu đã áp dụng có cải tiến ba mô hình tổng hợp để dự đoán giá cổ phiếu, đó là XGBoost, LightGBM và CatBoost. Kết quả nghiên cứu đã chứng minh 3 mô hình cải tiến nói trên cho độ chính xác tốt hơn so với các mô hình truyền thống. Nghiên cứu cũng đã thử nghiệm 3 mô hình XGBoost, LightGBM và CatBoost với điều chỉnh thông số Bayes trên tập dữ liệu. Kết quả cho thấy 3 mô hình này hoạt động tốt hơn và được cải thiện cao hơn sau khi điều chỉnh tham số Bayes. Ngoài ra, các tác giả cũng nghiên cứu việc tích hợp 3 mô hình này và áp dụng vào cùng bộ dữ liệu. Kết quả thử nghiệm tích hợp 3 mô hình cho giá trị dự đoán tốt hơn với các chỉ số đo: độ chính xác, giá trị F1, giá trị AUC so với áp dụng trên từng mô hình đơn lẻ.

III. PHƯƠNG PHÁP VÀ DỮ LIỆU ÁP DỤNG

Trong nghiên cứu này, phương pháp thực hiện dựa trên các kỹ thuật hồi quy (regression) để dự đoán biến động giá cổ phiếu tương lai dựa trên các số liệu lịch sử của bộ dữ liệu. Nghiên cứu đã triển khai 5 kỹ thuật để kiểm tra mô hình, đó là: Hồi quy máy vectơ hỗ trợ (SVM- Support Vector Machine), Cây quyết định (Decision Tree), Rừng ngẫu nhiên (Random Forest), K láng giềng gần nhất (K-Nearest Neighbor) và tăng cường độ dốc XGBoost. Trong các nội dung tiếp theo, bài báo sẽ tóm tắt ngắn gọn các khái niệm và ý tưởng của 5 thuật toán này. Bài báo này thực hiện áp dụng 5 phương pháp này vào bài toán dự đoán giá cổ phiếu trên thị trường chứng khoán.

Dữ liệu sử dụng trong bài báo này bao gồm giá mở cửa hàng ngày (daily open price) của mã cổ phiếu AAPL (Apple Inc) trên sàn giao dịch chứng khoán NASDAQ thông qua cơ sở dữ liệu của Yahoo Finance. Bộ dữ liệu được trích xuất từ Yahoo Finance cho chuỗi dữ liệu giá mã chứng khoán AAPL bao gồm khoảng thời gian từ ngày 01 tháng 6 năm 2001 đến ngày 01 tháng 6 năm 2021. Chi tiết về phương pháp và tập dữ liệu được giải thích ngắn gọn trong các đoạn sau.

1. Support Vector Machine (SVM)

SVM là một thuật toán học máy phổ biến thường được sử dụng trong cả 2 loại bài toán hồi quy và phân loại. Xét trên một số góc nhìn, hồi quy vectơ hỗ trợ (SVR) là tìm siêu phẳng để phù hợp với tập dữ liệu. Hồi quy SVM được coi là một kỹ thuật phi tham số vì nó dựa vào các chức năng của hàm hạt nhân.

Theo tham khảo bài báo [11], mô hình SVR dựa trên lý thuyết học thống kê, có thể cải thiện khả năng tổng quát hóa của máy học bằng cách tìm kiếm cấu trúc có sai biệt tối thiểu.

Giả sử rằng:

SV = {(x_1, y_1), (x_2, y_2), …, (x_m, y_m)} là một tập hợp gồm m số lượng mẫu đào tạo, mỗi mẫu x_m là một biến đầu vào chứa các thông tin ảnh hưởng đến biến động giá cổ phiếu trong tập dữ liệu. Trong khi đó, y_m là giá trị phần trăm thay đổi của giá cổ phiếu tương ứng với x_m. Các mẫu huấn luyện này được điều chỉnh theo hàm: $f(x) = \omega^T x+b$ và các kết quả được điều chỉnh phải thỏa mãn với độ chính xác sai số ε, tức là:

$$||\omega^T x_i+b-y_i|| \le \varepsilon, \text{ với } i=1, 2, \ldots, m \quad (1)$$

Theo nguyên tắc tìm kiếm cấu trúc có sai biệt tối thiểu, f(x) cần tạo ra giá trị $\frac{1}{2}||\omega||^2$ là nhỏ nhất.

Tiếp theo, thuật toán SVR áp dụng các nguyên tắc kép, thiết lập các bước khác nhau, chẳng hạn như giải thuật Lagrange, áp dụng một số hàm hạt nhân phổ biến như hàm polynomial kernel function, radial basis function, sigmoid kernel function và tính toán trong không gian Euclide... để tìm ra mô hình hồi quy tốt nhất với kết quả được chỉ ra bởi các hệ số ω. Trong mô hình SVR, các hàm kernel function khác nhau có thể được cấu trúc hóa các bề mặt hồi quy khác nhau dẫn đến đưa ra các kết quả khác nhau. Trong thực tế, cần phải chọn hàm kernel function phù hợp với các tham số hạt nhân tối ưu trong mô hình SVR.

2. Decision Trees

Cây quyết định cũng là một thuật toán học máy rất phổ biến, thường được áp dụng cho cả bài toán hồi quy và phân loại. Phương pháp này sử dụng giải quyết nhiều vấn đề bằng cách xử lý cả tập dữ liệu số và tập dữ liệu danh nghĩa. Phương pháp này không yêu cầu thông tin rõ ràng về phân bố dữ liệu và các giới hạn khác trên tập dữ liệu. Có nhiều thuật toán khác nhau trong phương pháp cây quyết định và một trong những thuật toán phổ biến nhất là CART (Classification And Regression Trees).

Theo mô tả ngắn gọn được nêu trong bài báo [12] thì CART phân bố các mẫu dữ liệu dưới dạng liên kết một tập thông số các đặc trưng X tương ứng cho một biến đầu ra Y. Thuật toán chính được sử dụng là một cấu trúc cây nhị phân với các mẫu con không giao nhau được gọi là các nút, xử lý theo một số các quy tắc xác định. Quy tắc cho mỗi lần tách của nút cha Ƭ như sau: "Nếu $X_i \leq X_{th}$, thì nếu $Y_i \in$ Ƭ sẽ cho Y di chuyển sang nút con bên trái của Ƭ, nếu ngược lại Y được chuyển sang nút con bên phải của Ƭ". Trong đó, X_i là giá trị thứ i của yếu tố dự đoán X và X_{th} là giá trị ngưỡng (threshold) của X. Thực hiện lặp lại từng bước ở mỗi nút quyết định cho bước di chuyển tiếp theo, giá trị mô hình dự đoán của đầu ra là giá trị trung bình hoặc một giá trị tương đối của các trường hợp đã được xác định trong một nút nhất định. Thuật toán tìm kiếm tất cả các yếu tố dự đoán X và tất cả các giá trị ngưỡng X_{th} để xác định những yếu tố giảm thiểu giá trị lỗi nhất định cho mô hình ở mỗi bước. Trên thực tế, thuật toán cây quyết định cũng phải đối mặt với vấn đề cây phát triển quá sâu (over-deeping) và mô hình quá khớp (over-fitting). Để giải quyết những điểm yếu này, thuật toán thiết lập và kiểm soát một số tham số, chẳng hạn như số trường hợp tối thiểu cho nút cha, số trường hợp tối thiểu cho nút con, xác thực chéo được áp dụng trên tập dữ liệu, giới hạn độ sâu cây, xác định trước mức độ lỗi chấp nhận được,... Phương pháp hồi quy CART hoạt động dựa trên các đặc điểm của tập dữ liệu để đạt được biến phụ thuộc có độ chính xác cao và đưa ra các mô hình được chấp nhận, phù hợp với tập dữ liệu cho bài toán hồi quy.

3. Random Forest

Rừng ngẫu nhiên là một phương pháp ensemble (kết hợp) được sử dụng để giải quyết cả bài toán phân loại và hồi quy, được phát triển bởi Breiman [13]. Thuật toán thực hiện bằng cách kết hợp các thuật toán cây quyết định trên tập dữ liệu với sự thay đổi có kiểm soát.

Theo mô tả của bài báo [14], thuật toán rừng ngẫu nhiên trong hồi quy được mô tả ngắn gọn theo ngôn ngữ thuật toán như sau:

*) Giai đoạn đào tạo (Training Phase):

Cho:

- D: tập huấn luyện với n giá trị quan sát, có p đặc trưng và biến đầu ra.

- B: số hồi quy trong tập hợp.

Thủ tục thực hiện:

Cho b = 1 đến B

1. Tạo một tập mẫu bootstrapped D_b^* từ tập huấn luyện D.

2. Tạo cây hồi quy bằng cách sử dụng mẫu bootstrapped D_b^*.

Đối với một nút t cho trước,

(i) Lấy mẫu ngẫu nhiên một số đặc trưng (features) từ bộ đặc trưng đầy đủ.

(ii) Tìm quy tắc tách tốt nhất bằng cách sử dụng tập hợp con đặc trưng trong mẫu ngẫu nhiên.

(iii) Tách nút t thành hai nút con bằng cách sử dụng quy tắc tách tốt nhất.

Lặp lại bước (i)-(iii) cho đến khi đáp ứng các quy tắc dừng.

3. Lấy một bộ hồi quy đã được huấn luyện R_b.

*) Giai đoạn thử nghiệm (Test Phase):

Đối với trường hợp áp dụng bộ thử nghiệm x, giá trị dự đoán được ước tính bởi các bộ hồi quy B và được đưa ra dưới dạng công thức:

$$R(x) = \frac{1}{B}\sum_{b=1}^{B} R_b(x) \quad (2)$$

4. K-Nearest Neibourgh (k-NN)

Trong phần này, bài báo đề cập đến thuật toán K-NN được mô tả ngắn gọn bởi Guo-Feng và cộng sự trong công bố [15]. Thuật toán K-NN làm việc theo hướng tìm ra k mẫu huấn luyện gần nhất với đối tượng đích (biến phản hồi hoặc biến đầu ra) trong tập huấn luyện. Hơn nữa, phương pháp K-NN còn tìm ra đặc điểm nổi bật nhất từ k mẫu huấn luyện và sau đó áp dụng tính năng này cho đối tượng đích (với điều kiện k là số mẫu đào tạo).

Theo nghiên cứu [15], ý tưởng của thuật toán K-NN được mô tả đơn giản và ngắn gọn như trong các nội dung sau. Thuật toán K-NN nhằm mục đích tính toán khoảng cách giữa điểm dữ liệu dự báo và điểm dữ liệu đã biết, để chọn dữ liệu có nhãn k gần nhất, $\{y_1, y_2, \ldots, y_k\}$, trong đó y_1 đại diện cho điểm dữ liệu đã biết gần nhất với điểm dự báo; y_2 đại diện cho điểm dữ liệu đã biết gần thứ hai với điểm dự báo,… Do đó, chuỗi điểm dữ liệu dự báo có thể được thực hiện bằng hồi quy thuật toán K-NN như phương trình dưới đây:

$$S_i = \frac{1}{k} * \sum_{j=1}^{k} S_{yj} \quad (3)$$

trong đó S_i đại diện cho giá trị dự báo thứ i, là giá trị trung bình của S_{yj} (j = 1, 2,…, k); S_{yj} đại diện cho giá trị dự báo của điểm dữ liệu đã biết gần nhất thứ j (y_j).

5. XGBoost

XGBoost viết tắt của cụm từ Extreme Gradient Boosting, một thuật toán học máy hiệu quả cao dựa trên sự kết hợp các kỹ thuật để điều chỉnh các trọng số lỗi trên các mô hình yếu hơn để tạo ra một mô hình mạnh hơn. Nguyên tắc thuật toán XGBoost dựa trên cây quyết định và kỹ thuật tăng cường độ dốc để đưa ra mô hình tối ưu. Các cây mới sinh ra tuần tự được giảm thiểu lỗi từ cây trước đó bằng cách học lại lỗi của cây trước đó, thực hiện sửa lỗi để được cây tốt hơn. XGBoost ban đầu được Chen và Guestrin (2016) giới thiệu để cải thiện hiệu suất và tốc độ của cây

quyết định theo nguyên tắc tăng cường độ dốc (gradient-boosted) [16].

Theo mô tả thuật toán được đưa ra bởi các tác giả Chen và Guestrin [16], XGBoost hoạt động như sau:

Đối với một tập dữ liệu đã cho có n mẫu dữ liệu và m đặc trưng $D = \{(x_i, y_i)\}$ ($|D| = n$, $x_i \in R^m$, $y_i \in R$), áp dụng một mô hình kết hợp các cây sử dụng K hàm tăng cường để dự đoán đầu ra.

$$\hat{y}_i = \emptyset(x_i) = \sum_{k=1}^{K} f_k(x_i), f_k \in F \quad (4)$$

trong đó $F = \{f(x) = w_q(x)\}$ ($q : R^m \rightarrow T$, $w \in R^T$) là không gian của cây hồi quy (còn được gọi là CART). Ở đây q là đại lượng biểu diễn cho cấu trúc của mỗi cây, ánh xạ một mẫu dữ liệu cho chỉ số lá tương ứng. T là số lá trên cây. Mỗi f_k tương ứng với một cấu trúc cây độc lập q và trọng lượng lá w.

Để tìm hiểu tập hợp các hàm được sử dụng trong mô hình, thuật toán tối thiểu hóa hàm mục tiêu quy chuẩn sau:

$$\mathcal{L}(\emptyset) = \sum_{i=1}^{n} l(\hat{y}_i, y_i) + \sum_{k}^{K} \Omega(f_k) \quad (5)$$

trong đó $\Omega(f) = \gamma T + \frac{1}{2}\lambda||w||^2$

Trong đó, l là một hàm mất mát lồi khả vi được dùng để đo lường sự khác biệt giữa giá trị dự đoán $\hat{y}_i$ và giá trị thực tế y_i. Thành phần thứ hai Ω là mức độ phạt cho độ phức tạp của mô hình (ví dụ: hàm của cây hồi quy). Thành phần quy chuẩn bổ sung giúp làm trơn các trọng số cuối cùng đã học được để tránh hiện tượng quá khớp. Về mặt trực quan, mục tiêu quy chuẩn có xu hướng chọn một mô hình sử dụng các hàm đơn giản nhưng có tính dự đoán cao.

Thuật toán tăng cường độ dốc trên cây (Gradient Tree Boosting) được thực hiện khi mô hình liên tục được đào tạo theo cách bổ sung đặc tính. Về mặt hình thức, đặt $\hat{y}_i^{(t)}$ là giá trị dự đoán thứ i tại vòng lặp thứ t, thuật toán sẽ cần bổ sung thành phần f_t để thu nhỏ hàm mục tiêu như sau:

$$\mathcal{L}^{(t)} = \sum_{i=1}^{n} l\left(y_i, \hat{y}_i^{(t-1)} + f_t(x_i)\right) + \Omega(f_t) \quad (6)$$

Tối ưu xấp xỉ bậc 2 được sử dụng để tối ưu hóa nhanh hơn hàm mục tiêu trong cài đặt thuật toán.

$$\mathcal{L}^{(t)} \cong \sum_{i=1}^{n} \left[l\left(y_i, \hat{y}^{(t-1)} + g_i f_t(x_i)\right) + \frac{1}{2} h_i f_t^2(x_i) \right] + \Omega(f_t) \quad (7)$$

trong đó $g_i = \partial_{\hat{y}^{(t-1)}} l\left(y_i, \hat{y}^{(t-1)}\right)$ và $h_i = \partial^2_{\hat{y}^{(t-1)}} l\left(y_i, \hat{y}^{(t-1)}\right)$ là gradient bậc nhất và bậc hai trên hàm mất mát. Chúng ta có thể loại bỏ các hằng số để thu được hàm mục tiêu đơn giản hơn như sau ở bước t.

$$\tilde{\mathcal{L}}^{(t)} = \sum_{i=1}^{n} [g_i f_t(x_i) +] l\left(, \hat{y}_i^{(t-1)} + \frac{1}{2} h_i f_t^2(x_i)\right) + \Omega(f_t) \quad (8)$$

Định nghĩa $I_j = \{i|q(x_i) = j\}$ là tập biểu diễn thành phần của lá j. Chúng ta có thể tính trọng số tối ưu w_j^* của lá j bằng cách:

$$w_j^* = -\frac{\sum_{i \in I_j} g_i}{\sum_{i \in I_j} h_i + \lambda} \quad (9)$$

tính giá trị tối ưu tương ứng bằng cách:

$$\tilde{\mathcal{L}}^{(t)} = -\frac{1}{2} \sum_{j=1}^{T} \frac{\left(\sum_{i \in I_j} g_i\right)^2}{\sum_{i \in I_j} h_i + \lambda} + \gamma T \quad (10)$$

Công thức (10) có thể được sử dụng như một hàm chấm điểm để đo chất lượng của một cấu trúc cây q. Điểm này giống như điểm phân loại để đánh giá cây quyết định, ngoại trừ việc nó được tính cho một phạm vi rộng hơn của các hàm mục tiêu.

Về bản chất, thuật toán XGBoost sử dụng sử dụng kỹ thuật tăng cường độ dốc (gradient boosting) để xác định các cây mới được sinh ra trên cơ sở giảm thiểu lỗi từ cây trước đó, điều chỉnh trọng số lỗi để có cây tốt hơn. Do đó, những điểm bị lỗi ở cây trước đó sẽ có cơ hội được điều chỉnh chính xác hơn ở cây tiếp theo. Thuật toán XGBoost đã được chứng minh tối ưu hóa về tốc độ và hiệu năng cho việc xây dựng các mô hình dự đoán. Đồng thời, thuật toán XGBoost sử dụng đa dạng các định dạng dữ liệu, kể cả dữ liệu dạng bảng với kích thước khác nhau và cả các dạng dữ liệu phân lớp.

6. Mô tả dữ liệu sử dụng cho nghiên cứu

Bài báo này sử dụng dữ liệu của sàn giao dịch chứng khoán NASDAQ thông qua cơ sở dữ liệu của Yahoo Finance. Dữ liệu trích xuất sau khi lọc bỏ các bản ghi có số liệu rỗng và chuẩn hóa dữ liệu, ta thu được 5029 bản ghi của mã cổ phiếu AAPL (Apple Inc) trên sàn giao dịch NASDAQ. Dữ liệu được thu thập trong khoảng 20 năm từ ngày 01/6/2001 đến ngày 01/6/2021. Bộ dữ liệu được trích xuất trực tiếp từ Yahoo Finance, là các số liệu thực tế giao dịch cho chuỗi dữ liệu giá mã chứng khoán AAPL bao gồm các thuộc tính của các giao dịch hàng ngày.

Mỗi bản ghi dữ liệu giao dịch hàng ngày gồm có các thuộc tính sau:

- Date: ngày thực hiện giao dịch
- Open: giá mở cửa của cổ phiếu
- High: giá cổ phiếu cao nhất trong ngày
- Low: giá cổ phiếu thấp nhất trong ngày
- Close: giá đóng cửa của cổ phiếu
- Volume: khối lượng cổ phiếu giao dịch
- Dividends: lợi tức cổ phiếu
- Stock Splits: chia tách cổ phiếu

Chúng ta xem xét đặc điểm dữ liệu của 2 trường thông tin là Dividends và Stock Splits theo Bảng 1 bên dưới.

Bảng 1: Dữ liệu lợi tức và chia tách

Items	Dividends	Stock Splits
count	5029	5029
mean	0.001083	0.002584
std	0.013150	0.117106
min	0.000000	0.000000
25%	0.000000	0.000000
50%	0.000000	0.000000
75%	0.000000	0.000000
max	0.220000	7.000000

Như vậy, các giá trị của 2 trường thông tin Dividends và Stock Splits chủ yếu là 0. Hơn nữa bài toán dự đoán giá cổ phiếu không phụ thuộc nhiều vào các yếu tố lợi tức và chia tách cổ phiếu. Vì vậy, chúng ta loại bỏ 2 trường thông tin này.

Dữ liệu áp dụng cho bài toán còn lại 5 trường thông tin, bao gồm: ngày, giá mở cửa, giá đóng cửa, giá cao nhất, giá thấp nhất và khối lượng cổ phiếu giao dịch như bảng 2.

Bảng 2: Mô tả dữ liệu giá giao dịch

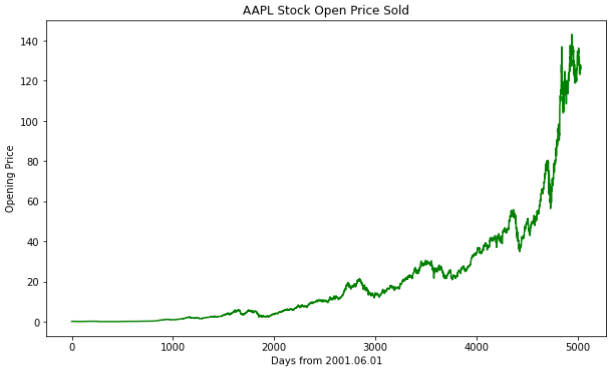
Date	Open	High	Low	Close	Volume
1-Jun-01	0.3091	0.3238	0.3068	0.3208	456075200
4-Jun-01	0.3237	0.3242	0.3142	0.3172	281920800
5-Jun-01	0.3194	0.3240	0.3125	0.3215	471794400
6-Jun-01	0.3214	0.3214	0.3122	0.3183	223176800
7-Jun-01	0.3180	0.3332	0.314	0.3326	325180800
…	…	…	…	…	…
24-May-21	126.01	127.94	125.94	127.10	63092900
25-May-21	127.82	128.32	126.32	126.90	72009500
26-May-21	126.96	127.39	126.42	126.85	56575900
27-May-21	126.44	127.64	125.08	125.28	94625600
28-May-21	125.57	125.80	124.55	124.61	71311100

Để có bộ dữ liệu phục vụ dự đoán giá cổ phiếu, ta tiếp tục lược bỏ trường dữ liệu ngày giao dịch. Thông tin mô tả 5 trường dữ liệu chính còn lại để áp dụng vào bài toán được mô tả trong Bảng 3 dưới đây.

Bảng 3: Mô tả dữ liệu cổ phiếu

Items	Open	High	Low	Close	Volume
count	5029	5029	5029	5029	5.02E+03
mean	21.116	21.34	20.883	21.119	4.36E+08
std	28.299	28.628	27.942	28.294	3.84E+08
min	0.1995	0.2025	0.1953	0.2015	3.93E+07
25%	2.1378	2.1789	2.0964	2.1372	1.62E+08
50%	10.753	10.856	10.665	10.764	3.18E+08
75%	27.381	27.568	27.195	27.418	5.90E+08
max	143.14	144.63	140.92	142.7	3.37E+09

Bài báo này sử dụng giá mở cửa (opening prices) để làm giá dự đoán (response) cho bài toán dự đoán giá cổ phiếu. Hình vẽ 1 bên dưới mô tả xu hướng biến động hàng ngày của giá mở cửa cổ phiếu AAPL trong 20 năm qua.



Hình 1 : Biến động giá mở cửa

IV. THỰC NGHIỆM VÀ KẾT QUẢ

Giá trị dự đoán là giá mở cửa (opening prices) của phiên giao dịch trong ngày. Bài toán đặt ra là xác định hướng biến động giá cổ phiếu. Vì vậy, các dữ liệu sẽ được xác định mức độ biến động giá cổ phiếu của ngày hiện tại (x_i) so với giá cổ phiếu ngày trước đó (x_{i-1}). Công thức cụ thể là:

$$x^* = \frac{x_i - x_{i-1}}{x_i} \quad (11)$$

Sau khi chuyển đổi tập dữ liệu để thu được các thông số thể hiện sự biến động giá cổ phiếu, xu hướng thay đổi giá theo thời gian (dương là tăng, âm là giảm), chúng ta thu được tập dữ liệu mô tả như trong Bảng 4 dưới đây.

Bảng 4: Mức độ biến động các chỉ số

Index	Open	Close	High	Low	Volume
…	…	…	…	…	…
5024	-1.4161	1.3314	-0.0469	0.5830	-20.4331
5025	1.4364	-0.1574	0.2970	0.3017	14.1325
5026	-0.6728	-0.0394	-0.7248	0.0792	-21.4327
5027	-0.4096	-1.2377	0.1962	-1.0600	67.2543
5028	-0.6881	-0.5348	-1.4416	-0.4237	-24.6387

Để thực hiện quá trình huấn luyện và đánh giá, tập 5029 bản ghi dữ liệu được lấy ngẫu nhiên và chia thành 2 tập: dữ liệu đào tạo (traning set) chiếm 90% và dữ liệu kiểm tra (testing set) chiếm 10% tổng số bản ghi.

Để đánh giá hiệu năng (performance) của các mô hình, chúng ta sử dụng các độ đo thông thường áp dụng cho bài toán hồi quy, do áp dụng dữ liệu trên miền giá trị liên tục (khác với gán nhãn trong bài toán phân loại), gồm có: MAE (Mean Absolute Error), MSE (Mean Squared Error), R-Squared (Coefficient of determination). Công thức tính các độ đo này như sau:

$$MAE = \frac{1}{N}\sum_{i=1}^{N}|y_i - \hat{y}| \quad (12)$$

$$MSE = \frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{y})^2 \quad (13)$$

$$R^2 = 1 - \frac{\sum_{i=1}^{N}(y_i - \hat{y})^2}{\sum_{i=1}^{N}(y_i - \bar{y})^2} \quad (14)$$

Trong đó, N là tổng số bản ghi, y_i là giá trị thực tế, $\hat{y}$ là giá trị dự đoán của y và $\bar{y}$ là giá trị trung bình (mean value) của y.

Đối với R-Squared, giá trị càng cao thì mô hình càng tốt (lớn nhất bằng 1), đối với MAE và MSE thì giá trị càng thấp thể hiện sự sai khác giữa giá trị dự đoán và giá trị thực tế càng nhỏ, tức là mô hình càng tốt.

Như phần trên đã trình bày, chúng ta thử nghiệm tập dữ liệu với 5 thuật toán học máy gồm có: Hồi quy vectơ hỗ trợ (SVR), Cây quyết định, Rừng ngẫu nhiên, K-Nearest Neighbor và XGBoost. Bảng 5 dưới đây thể hiện các siêu tham số áp dụng cho mỗi thuật toán. Môi trường thử nghiệm sử dụng công cụ Google Colab có hỗ trợ GPU, với ngôn ngữ lập trình Python sử dụng thư viện TensorFlow.

Bảng 5: Siêu tham số cho mỗi thuật toán

Algorithm	Hyperparameter
SVR	C=1 epsilon=0.1 gamma='scale' kernel='rbf'
Decision Tree	max_depth=3 random_state=14 min_samples_leaf=1 min_samples_split=2 splitter='best'
Random Forest	max_depth=3 max_features=auto min_samples_leaf=1 min_samples_split=2 n_estimators=10 random_state=14
K-NN	max_depth=3 max_features='auto' n_estimators=10 random_state=14 verbose=0
XGBoost	base_score=0.5, booster='gbtree' gamma=0 learning_rate=0.1 max_depth=3 n_estimators=100 reg='linear' scale_pos_weight=1 max_features='auto' random_state=14

Các giá trị độ đo giữa các thuật toán học máy gồm R-Squared, MAE và MSE được trình bày trong bảng 6 dưới đây.

Bảng 6: Kết quả độ đo giữa các thuật toán

Độ đo	MAE	MSE	R-Squared
SVR	0.9421	1.7789	0.6610
Random Forest	0.9416	1.8199	0.6532
Decision Tree	0.9873	2.0945	0.6009
K-NN	1.0129	2.0553	0.6084
XGBoost	**0.8040**	**1.2992**	**0.7524**

Kết quả thực nghiệm cho thấy mô hình dự đoán sự biến động giá cổ phiếu với thuật toán XGBoost cho độ chính xác tốt nhất. Giá trị R^2 score đạt 0,75 là cao nhất khi so sánh với các mô hình hồi quy vectơ hỗ trợ (SVR), cây quyết định (Decision Tree), rừng ngẫu nhiên (Random Forest), k láng giềng gần nhất (K-Nearest Neighbor). Đồng thời giá trị dung sai giữa giá trị thực tế và giá trị dự đoán thông qua giá trị của các độ đo MAE và MSE cũng thấp nhất, chứng tỏ mô hình dự đoán dựa trên XGBoost cho kết quả phù hợp nhất với tập dữ liệu thử nghiệm trong bài toán.

V. KẾT LUẬN VÀ HƯỚNG PHÁT TRIỂN

Bài báo này đã thử nghiệm 5 thuật toán học máy gồm có: Hồi quy vectơ hỗ trợ, Cây quyết định, Rừng ngẫu nhiên, k láng giềng gần nhất và XGBoost cho tập dữ liệu giá chứng khoán của mã AAPL trên sàn giao dịch NASDAQ. Dữ liệu được cập nhật chính xác hàng ngày trong thời gian đủ dài (20 năm) để làm cơ sở dự đoán giá cổ phiếu. Kết quả thực nghiệm cho thấy thuật toán XGBoost cho kết quả tốt nhất, với độ đo chỉ số R^2 và các chỉ số MAE, MSE đều tốt hơn các thuật toán còn lại. Điều này cho thấy thuật toán XGBoost có hiệu năng tốt không chỉ trong các bài toán phân loại mà cả các bài toán hồi quy, với dữ liệu chuỗi thời gian (time-series).

Trên thực tế, giá cổ phiếu phụ thuộc rất nhiều yếu tố khác nhau, không chỉ dựa vào giá trị lịch sử trong quá khứ của nó. Ví dụ, các chỉ số kinh tế vĩ mô, vi mô, đầu tư chính phủ, chính sách thương mại, lãi suất ngân hàng, tỷ giá tiền tệ… Vì vậy, nghiên cứu này chỉ mang tính chất tham khảo. Trong tương lai, các nghiên cứu sẽ mở rộng dữ liệu gồm có các chỉ số ảnh hưởng khác đến giá cổ phiếu, cập nhật các dữ liệu mới nhất. Đồng thời, áp dụng các mô hình học sâu (deep learning) để thử nghiệm cho bài toán hồi quy với dữ liệu chuỗi thời gian. Đồng thời, trong tương lai có thể xem xét thử nghiệm phương án thực hiện bài toán với phương pháp phân loại (classification) với giá trị mục tiêu là âm hoặc dương, thể hiện mã chứng khoán tăng hoặc giảm so với một mốc thời gian trước đó.

TÀI LIỆU THAM KHẢO

[1] Pound, J. (2019, December 24). Global stock markets gained $17 trillion in value in 2019. Retrieved from https://www.cnbc.com/2019/12/24/global-stock-marketsgained-17-trillion-in-value-in-2019.html.

[2] Strader, Troy J.; Rozycki, John J.; ROOT, THOMAS H.; and Huang, Yu-Hsiang (John) (2020) "Machine Learning Stock Market Prediction Studies: Review and Research Directions," *Journal of International Technology and Information Management*: Vol. 28 : Iss. 4 , Article 3. Available at: https://scholarworks.lib.csusb.edu/jitim/vol28/iss4/3

[3] Adil M., Mhamed. (2020) "Stock Market Prediction Using LSTM Recurrent Neural Network", International Workshop on Statistical Methods and Artificial Intelligence (IWSMAI 2020) April 6-9, 2020, Warsaw, Poland.

[4] Hu, Z.; Zhao, Y.; Khushi, M. A Survey of Forex and Stock Price Prediction Using Deep Learning. Appl. Syst. Innov. 2021, 4, 9. https://doi.org/10.3390/asi4010009

[5] Zhang,X.;Liu,S.;Zheng,X. (2021) "Stock Price Movement Prediction Based on a Deep Factorization Machine and the

Attention Mechanism". Mathematics 2021, 9, 800. https://doi.org/10.3390/math9080800
[6] Subba R. P., Srinivas K., Krishna M. A. (2019) "Stock Market Prices Prediction using Random Forest and Extra Tree Regression", International Journal of Recent Technology and Engineering (IJRTE), ISSN: 2277-3878, Volume-8 Issue-3, September 2019.
[7] Han, Shangxuan, "Stock Prediction with Random Forests and Long Short-term Memory" (2019). Creative Components. 393. https://lib.dr.iastate.edu/creativecomponents/393
[8] Aryendra S., Priyanshi G., Narina T. (2020) "An Empirical Research and Comprehensive Analysis of Stock Market Prediction using Machine Learning and Deep Learning techniques", IOP Conf. Series: Materials Science and Engineering 1022 (2021) 012098 IOP Publishing doi:10.1088/1757-899X/1022/1/012098
[9] Ranjeet K., Yogesh K. S., Devershi P. B. (2020) "Measuring Accuracy of Stock Price Prediction Using Machine Learning Based Classifiers", ASCI-2020 IOP Conf. Series: Materials Science and Engineering 1099 012049 IOP Publishing, doi:10.1088/1757-899X/1099/1/012049
[10] Jie Ni, Linghong Zhang, Jiaming Tao and Xiaorong Yang (2020) "Prediction of stocks with high transfer based on ensemble learning", ICAITA 2020 Journal of Physics: Conference Series 1651 (2020) 012124 IOP Publishing doi:10.1088/1742-6596/1651/1/012124
[11] Zhang J, Liao Y, Wang S, Han J. "Study on Driving Decision-Making Mechanism of Autonomous Vehicle Based on an Optimized Support Vector Machine Regression", *Applied Sciences*, 2018; 8(1):13. https://doi.org/10.3390/app8010013
[12] Ivanov A. (2020) "Decision Trees for Evaluation of Mathematical Competencies in the Higher Education: A Case Study", *Mathematics* 2020; 8(5):748, https://doi.org/10.3390/math8050748
[13] Breiman, L. (2001) "Random Forests", *Machine Learning*, 45, 5–32. https://doi.org/10.1023/A:1010933404324
[14] Han, Sunwoo; Kim, Hyunjoong (2021) "Optimal Feature Set Size in Random Forest Regression", *Appl. Sci.* 11, no. 8: 3428. https://doi.org/10.3390/app11083428
[15] Guo-Feng F., Yan-Hui G., Jia-Mei Z. & Wei-Chiang H. (2019) "Application of the Weighted K-Nearest Neighbor Algorithm for Short-Term Load Forecasting", Energies 2019, 12, 916; doi:10.3390/en12050916
[16] Chen and Guestrin (2016) "XGBoost: A Scalable Tree Boosting System", KDD'16: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, August 2016, Pages 785–794, https://doi.org/10.1145 /2939672.2939785.

EVALUATING XGBOOST WITH OTHER MACHINE LEARNING ALGORITHMS

Abstract: This study tests a number of machine learning algorithms such as XGBoost, random forest, decision tree, support vector machine, k nearest neighbors for predicting problem. The dataset on US NASDAQ stock market are collected until end of May 2021 is experimental data for implementation of machine learning algorithms. The predicting and training data are the stock prices on the stock market given by AAPL. The tested results show that the XGBoost method gives the best prediction results compared to other machine learning algorithms.

Keywords: Machine learning, AI, XGBoost, random forest, decision tree, support vector machine, k-nearest neighbors, NASDAQ, stock price.

Trần Quý Nam tốt nghiệp Kỹ sư (năm 1999) và Thạc sỹ (năm 2005) ngành công nghệ thông tin hệ chính quy tại Đại học Bách Khoa Hà Nội, nhận bằng Tiến sỹ quản lý công nghệ thông tin tại Đại học Quốc gia Xê-Un Hàn Quốc năm 2010. Lĩnh vực nghiên cứu: chính phủ điện tử, chuyển đổi số, hiệu năng các hệ thống thông tin, ứng dụng trí tuệ nhân tạo, phát triển ứng dụng đa phương tiện.