Khái lược nghiên cứu về ứng dụng các kỹ thuật mô hình học máy trong dự báo giá vàng

Trương Thị Thùy Dương
Phan Anh
Học viện Ngân hàng

Dự báo xu hướng của giá vàng chính xác và hiệu quả có ý nghĩa lý thuyết và thực tiễn rất lớn, nó không chỉ giúp ích cho các nhà đầu tư, các nhà quản lý tiền tệ trong việc lựa chọn thời điểm cung ứng mặt hàng này mà còn sử dụng vàng như một phương tiện để quản lý lạm phát và củng cố vị thế kinh tế quốc gia. Đã có nhiều phương pháp dự đoán được ứng dụng trong dự báo, từ các mô hình thống kê truyền thống đến các mô hình học máy. Do tính đa yếu tố và phi tuyến của thị trường vàng cùng với sự tiến bộ như vũ bão của công nghệ, các kỹ thuật mô hình học máy trong những năm gần đây đã đóng vai trò quan trọng trong việc dự đoán giá vàng.

1. Mở đầu

Các nghiên cứu về giá vàng và các yếu tố ảnh hưởng đến sự thay đổi của chúng đã được nhiều nhà nghiên cứu xem xét trong những thập kỷ qua và nó vẫn là một trong những chủ đề nóng trong các nghiên cứu kinh tế và tài chính toàn cầu. Các nghiên cứu về yếu tố quyết định giá vàng có thể được phân loại theo ba cách tiếp cận: (i) Lập mô hình biến động giá vàng theo giá lịch sử để dự đoán giá trong tương lai; (ii) Lập mô hình biến động giá vàng theo sự thay đổi của các biến kinh tế vĩ mô chính, được phân loại là phân tích hai biến và đa biến; (iii) Lập mô hình biến động giá vàng theo sự thay đổi của các biến số kinh tế vĩ mô và tài chính, chẳng hạn như đầu cơ trong biến động giá vàng và cả các chỉ số tài chính.

Machine Learning (ML) là một loại trí tuệ nhân tạo (AI) cho phép các ứng dụng của phần mềm trở nên chính xác hơn trong việc dự đoán kết quả, các thuật toán của ML sử dụng thông tin lịch sử làm đầu vào để dự báo tốc độ đầu ra mới nhất. Để dự đoán định tính tỷ giá vàng, nhiều kỹ thuật dự báo đã được sử dụng như mô hình xám, phân tích chuỗi thời gian và mô hình hồi quy. Nhiều kỹ thuật khác nhau như máy vectơ hỗ trợ (SVM), rừng ngẫu nhiên (RF), cây quyết định (DT), hồi quy logistic (LR) và nhiều kỹ thuật khác đã được sử dụng trong các nghiên cứu trước đây để dự đoán và cho kết quả chính xác. Giá vàng được xác định bởi các yếu tố như đồng bảng Anh, đồng đô la Mỹ thường xuyên suy yếu, khủng hoảng tín dụng trên toàn thế giới, nhu cầu tăng cao từ các nhà đầu tư tổ chức và Ngân hàng Trung ương, bất ổn chính trị, tấn công khủng bố, v.v... Điều bắt buộc là phải xác định yếu tố nào ảnh hưởng đến giá vàng, vì một trong những yêu cầu để sử dụng mô hình học máy là tính hợp lý của các biến đầu vào. Có hai cách để chọn biến đầu vào: (i) Cách tiếp cận đơn biến, chỉ sử dụng dữ liệu giá trước đó dựa trên mối quan hệ giữa giá cả và thời gian, phương pháp này đơn giản và dễ thực hiện (Zietz và Traian, 2014); (ii) Cách tiếp cận đa biến, xem xét các biến khác ảnh hưởng đến giá, tính toán tương đối phức tạp, nhưng độ chính xác dự đoán cao (Dedinec và cộng sự, 2016).

2. Ứng dụng mô hình học máy trong dự báo giá vàng

2.1. Support Vector Machine (SVM)

SVM là một kỹ thuật phân loại, mục dữ liệu có thể được vẽ dưới dạng một điểm trong không gian n chiều với giá trị của mọi thuộc tính là giá trị của một tọa độ chính xác. SVM là một công cụ tính toán hiệu quả trong không gian chiều cao, trong đó đặc biệt áp dụng cho các bài toán phân loại văn bản và phân tích quan điểm nơi chiều có thể cực kỳ lớn. Tiết kiệm bộ nhớ, do chỉ có một tập hợp con của các điểm được sử dụng trong quá trình huấn luyện và ra quyết định thực tế cho các điểm dữ liệu mới nên chỉ có những điểm cần thiết mới được lưu trữ trong bộ nhớ khi ra quyết định. Tính linh hoạt - phân lớp thường là phi tuyến tính. Khả năng áp dụng cho phép linh động giữa các phương pháp tuyến tính và phi tuyến tính từ đó khiến cho hiệu suất phân loại lớn hơn. Tuy nhiên, trong trường hợp số lượng thuộc tính của tập dữ liệu lớn hơn rất nhiều so với số lượng dữ liệu thì SVM cho kết quả không được tốt. Chưa thể hiện rõ tính xác suất, việc phân lớp của SVM chỉ là việc cố gắng tách các đối tượng vào hai lớp được phân tách bởi siêu phẳng SVM, chưa giải thích được xác suất xuất hiện của một thành viên trong một nhóm là như thế nào. Onsumran và cộng sự (2015) tập trung vào việc khai thác văn bản trước của kỹ thuật biến động chi phí vàng, kỹ thuật này được nâng cao để đánh giá xem các bài báo đã thao túng giá vàng như

thế nào, các tác giả đã sử dụng SVM cũng như các phương pháp thống kê chi bình phương và hai kỹ thuật phân loại khác như KNN và Bayes để dự đoán độ chính xác của các phương pháp trọng số. SVM được phát hiện là một phương pháp tốt hơn trong số tất cả các phương pháp được so sánh với tỷ lệ chính xác là 87,52% và được coi là một kỹ thuật vượt trội theo quan điểm của cả trình phân loại và trọng số thuộc tính.

2.2. Linear Regression (LR)

LR được sử dụng để đánh giá các giá trị thực dựa trên các biến liên tục, LR được sử dụng để thiết lập mối quan hệ giữa các biến phụ thuộc cũng như độc lập bằng cách đưa ra một giải pháp thích hợp. LR là một lựa chọn ưa thích để dự đoán giá vàng, Bingol và cộng sự đã giới thiệu một cách tiếp cận để khảo sát mối quan hệ của tỷ giá vàng với một số biến mô tả có xu hướng được đo lường như là dấu hiệu của thảm họa địa chính trị và kinh tế. Nghiên cứu đã khảo sát xác suất dự báo tỷ giá vàng. Các tác giả đã sử dụng bốn thuật toán ML khác nhau, chẳng hạn như LR, SVM, mô hình tự hồi quy véc tơ (VAR) và trung bình di chuyển tích hợp tự hồi quy (ARIMA) và nhận thấy rằng LR có thuật toán cho điểm cao nhất và ARIMA là thuật toán cho điểm thấp nhất. Sekar và cộng sự (2017) đã đề xuất một mô hình hồi quy tuyến tính đa biến để dự đoán hàng hóa vàng với việc loại bỏ sự không chắc chắn. Nó rất hữu ích đối với tầm quan trọng của việc hiểu các khoản đầu tư vào vàng (đặc biệt là trong thời gian biến động). Bằng cách xem xét dữ liệu của 5 năm, họ đã mô phỏng mô hình đã phát triển và chứng minh rằng nó là một công cụ dự đoán giá vàng hiệu quả.

2.3. Support Vector Regression (SVR)

SVR là một thuật toán nằm trong bộ thuật toán SVM dùng để giải quyết các vấn đề hồi quy. Thay vì giảm thiểu lỗi trong quá trình huấn luyện, SVR cố gắng giảm thiểu lỗi tổng quát bị ràng buộc để đạt được hiệu xuất tổng thể. Ý tưởng về SVR dựa trên tính toán của hàm hồi quy tuyến tính, trong không gian đặc trưng chiều cao nơi dữ liệu đầu vào bằng hàm phi. SVR đã được áp dụng trong các lĩnh vực khác nhau như phân tích và dự đoán theo chuỗi thời gian và tài chính (lọc nhiễu và rủi ro), xấp xỉ các phân tích kỹ thuật phức tạp, lập trình và lựa chọn các hàm mất mát... SVR sử dụng các nguyên tắc tương tự cho phân loại và sử dụng thêm loại mới của hàm mất mát. Với một tập dữ liệu huấn luyện nhất định, được biểu thị trong một không gian vectơ, trong đó mỗi dữ liệu của mẫu là một điểm. Phương thức này là tốt nhất, tại đó có thể chia các điểm trong không gian thành hai lớp riêng biệt, tương ứng với (lớp) + và (lớp) - (phân loại nhị phân). Đặc trưng của siêu phẳng này được xác định bởi khoảng cách (được gọi là ranh giới) của điểm dữ liệu gần nhất của mỗi lớp với mặt phẳng này. Do đó, ranh giới càng rộng cho thấy mặt phẳng phân chia và phân loại càng chính xác. Mục tiêu của phương pháp SVR là tìm ra khoảng cách ranh giới tối đa, xác định các giá trị cho các tham số SVR thông qua quá trình thử - lỗi. Ý tưởng cơ bản của SVR là ánh xạ không gian đầu vào sang một không gian đặc trưng nhiều chiều mà ở đó ta có thể áp dụng được hồi quy tuyến tính. SVR có đặc điểm là xây dựng được hàm hồi quy mà không cần phải sử dụng hết toàn bộ tất cả các điểm dữ liệu trong bộ huấn luyện, những điểm ở biên đóng góp vào việc xây dựng hàm hồi quy, việc phân lớp cho tập dữ liệu mới sẽ chỉ phụ thuộc vào các hàm này.

2.4. Random Forests (RF)

RF là thuật toán học có giám sát, có thể được sử dụng cho cả phân lớp và hồi quy. Đây là thuật toán linh hoạt và dễ sử dụng nhất, người ta nói rằng càng có nhiều cây thì rừng càng mạnh. RF tạo ra cây quyết định trên các mẫu dữ liệu được chọn ngẫu nhiên, được dự đoán từ mỗi cây và chọn giải pháp tốt nhất bằng cách bỏ phiếu. Về mặt kỹ thuật, nó là một phương pháp tổng hợp (dựa trên cách tiếp cận phân chia và chinh phục) của các cây quyết định được tạo ra trên một tập dữ liệu được chia ngẫu nhiên. Bộ sưu tập phân loại cây quyết định này còn được gọi là rừng, cây quyết định riêng lẻ được tạo ra bằng cách sử dụng chỉ báo chọn thuộc tính như tăng thông tin, tỷ lệ tăng và chỉ số Gini cho từng thuộc tính. Mỗi cây phụ thuộc vào một mẫu ngẫu nhiên độc lập. Trong bài toán phân loại, mỗi phiếu bầu chọn và lớp phổ biến nhất được chọn là kết quả cuối cùng. Trong trường hợp hồi quy, mức trung bình của tất cả các kết quả đầu ra của cây được coi là kết quả cuối cùng. Nó đơn giản và mạnh mẽ hơn so với các thuật toán phân loại phi tuyến tính khác, hoạt động theo bốn bước: Chọn các mẫu ngẫu nhiên từ tập dữ liệu đã cho; Thiết lập cây quyết định cho từng mẫu và nhận kết quả dự đoán từ mỗi quyết định cây; Bỏ phiếu cho mỗi kết quả dự đoán; Chọn kết quả được dự đoán nhiều nhất là dự đoán cuối cùng. RF được coi là một phương pháp chính xác và mạnh mẽ vì số cây quyết định tham gia vào quá trình này, nó mất trung bình của tất cả các dự đoán, trong đó hủy bỏ những thành kiến. Thuật toán có thể được sử dụng trong cả hai vấn đề phân loại và hồi quy. RF cũng có thể xử lý các giá trị còn thiếu, bằng hai cách là sử dụng các giá trị trung bình để thay thế các biến liên tục và tính toán mức trung bình gần kề của các giá trị bị thiếu. Tuy nhiên, RF chậm tạo dự đoán bởi vì nó có nhiều cây quyết định. bất cứ khi nào nó đưa ra dự đoán, tất cả các cây trong rừng phải đưa ra dự

đoán cho cùng một đầu vào cho trước và sau đó thực hiện bỏ phiếu trên đó. Pierdzioch và Risse (2017) đã đề xuất một mô hình dự đoán dựa trên rừng ngẫu nhiên đa biến để dự đoán tỷ giá của vàng, bạc cùng với hai kim loại quý khác

2.5. K- Nearest Neighbors (KNN)

KNN là một thuật toán học máy có giám sát, có thể được sử dụng trong cả phân loại và hồi quy. Giá trị của một điểm dữ liệu được xác định bởi các điểm dữ liệu xung quanh nó. Ví dụ: Nếu bạn có một người bạn rất thân và dành phần lớn thời gian cho anh ấy / cô ấy, bạn sẽ có chung sở thích và tận hưởng những điều giống nhau. Đó là KNN với k = 1. Nếu bạn luôn đi chơi với một nhóm 5 người, mỗi người trong nhóm có ảnh hưởng đến hành vi của bạn và bạn sẽ là trung bình của 5. Đó là KNN với k = 5. Bộ phân loại KNN xác định lớp (class) của một điểm dữ liệu theo nguyên tắc biểu quyết đa số. Nếu k được đặt là 5, các lớp của 5 điểm gần nhất sẽ được kiểm tra. Dự đoán đưa ra kết quả lớp của điểm dữ liệu dựa vào lớp nào chiếm đa số trong 5 điểm gần nhất. Tương tự, hồi quy KNN lấy giá trị trung bình của 5 điểm gần nhất. Làm thế nào các điểm dữ liệu được xác định là gần nhau? Trước hết cần đo khoảng cách giữa các điểm dữ liệu. Có nhiều phương pháp để đo khoảng cách, phép đo khoảng cách Euclid là một trong những phép đo khoảng cách được sử dụng phổ biến nhất. Ưu điểm: Đơn giản và dễ giải thích; Không dựa trên bất kỳ giả định nào, vì thế nó có thể được sử dụng trong các bài toán phi tuyến tính; Hoạt động tốt trong trường hợp phân loại với nhiều lớp; Sử dụng được trong cả phân loại và hồi quy. Nhược điểm: Trở nên rất chậm khi số lượng điểm dữ liệu tăng lên vì mô hình cần lưu trữ tất cả các điểm dữ liệu; Tốn bộ nhớ; Nhạy cảm với các dữ liệu bất thường (nhiễu). Al-Dhuraibi và Ali (2018) đã sử dụng các phương pháp phân loại để dự báo giá vàng. Dự đoán tỷ giá vàng rất quan trọng trong các lĩnh vực chính như môi trường chính trị, kinh tế, thương mại và đầu tư. Đánh giá đầu tư tốt hơn có thể được hoàn thành khi giá trị của tỷ giá vàng được dự báo chính xác. Mục đích chính của phương pháp đề xuất của tác giả là dự đoán xem vàng sẽ tăng hay giảm trong tương lai sắp tới. KNN trong số tất cả các thuật toán tìm thấy hiệu suất tốt hơn.

3. Kết quả và thảo luận

Bảng 1: Hiệu suất dự báo giá vàng từ các kỹ thuật mô hình học máy

Tác giả	Kỹ thuật	Chính xác	Tác giả	Kỹ thuật	Chính xác
1.Risse (2019)	SVR	93%	5. Al-Dhuraibi and Ali (2018)	SVM	73.49%
2.Alameer và cộng sự (2019)	WOA-NN	99%	6. Dubey (2016)	SVR	99%
3.Raghuram (2020)	SVM	60%	7. Makala and Li (2020)	LR	73%
4.Onsumran và cộng sự (2015)	SVM. KNN	87.52% 85.57%			

Nguồn: Nhóm tác giả tự tổng hợp

Nhiều thuật toán ML khác nhau như SVM, RF, LR, SVR, DT đã được xem xét để dự đoán giá vàng và cho thấy hiệu quả của chúng trong việc dự báo, trong đó SVR chiếm phần lớn tỷ lệ chính xác và được xếp vào loại kỹ thuật ML cao nhất trong việc dự đoán giá vàng, tiếp theo là SVM. Bài báo này chủ yếu tập trung vào các loại kỹ thuật ML khác nhau được sử dụng để dự báo giá vàng, bằng chứng cho thấy ML rất phổ biến và nó đã nhận được sự quan tâm từ nhiều nhà nghiên cứu khác nhau. Các mô hình học máy đã được chỉ ra có hiệu quả tốt hơn so với các mô hình thống kê truyền thống trong nhiều lĩnh vực khác nhau, ưu điểm của mô hình học máy không đòi hỏi chặt chẽ về điều kiện của dữ liệu và mối quan hệ giữa biến độc lập với biến phụ thuộc, do đó khả năng ứng dụng rộng hơn và linh hoạt hơn. Tuy nhiên, ngoài các kỹ thuật ML chung như SVM, DT, RF, v.v., thực tế còn có một số phương pháp khác như kỹ thuật lai, mô hình ML nâng cao và nhiều phương pháp khác cho thấy hiệu quả của chúng trong dự đoán giá vàng chưa được đề cập trong nghiên cứu này./.

Tài liệu tham khảo

Hu, Y., Ni, J., & Wen, L. (2020). A hybrid deep learning approach by integrating LSTM-ANN networks with GARCH model for copper price volatility prediction. Physica A: Statistical Mechanics and its Applications, 557, 124907.

Jianwei, E., Ye, J., & Jin, H. (2019). A novel hybrid model on the prediction of time series and its application for the gold price analysis and forecasting. Physica A: Statistical Mechanics and its Applications, 527, 121454.

Parisi, A., Parisi, F., & Díaz, D. (2008). Forecasting gold price changes: Rolling and recursive neural network models. Journal of Multinational financial management, 18(5), 477-487.

Sarangi, P., & Dublish, S. (2013). Prediction of gold bullion return using GARCH family and artificial neural network models. Asian Journal of Research in Business Economics and Management, 3(10), 217-230.

Wen, F., Yang, X., Gong, X., & Lai, K. K. (2017). Multi-scale volatility feature analysis and prediction of gold price. International Journal of Information Technology & Decision Making, 16(01), 205-223.

S. Das et al. (2020). Gold Price Forecasting Using Machine Learning Techniques: Review of a Decade. 679-695