**Báo cáo đồ án: Dự đoán biến động giá Bitcoin sử dụng phương pháp học máy**

**1. Giới thiệu**

**1.1 Tổng quan bài toán**

Dự đoán biến động giá Bitcoin là một bài toán quan trọng trong lĩnh vực tài chính và đầu tư tiền điện tử. Bài toán này nhằm sử dụng các kỹ thuật học máy để phân tích dữ liệu lịch sử của Bitcoin và xây dựng các mô hình có thể dự đoán xu hướng giá trong tương lai gần. Trong đồ án này, chúng tôi đã áp dụng nhiều phương pháp học máy khác nhau, từ các thuật toán cổ điển như Linear Regression, Random Forest đến các mạng nơ-ron hiện đại như LSTM, nhằm tìm ra phương pháp hiệu quả nhất cho bài toán dự đoán giá Bitcoin.

**1.2 Mục tiêu đồ án**

* Phân tích dữ liệu lịch sử Bitcoin để tìm ra các mẫu và xu hướng
* Xây dựng các mô hình học máy để dự đoán giá Bitcoin trong tương lai
* Xây dựng mô hình phân loại để dự đoán xu hướng tăng/giảm giá
* So sánh hiệu suất của các phương pháp học máy khác nhau
* Đề xuất chiến lược giao dịch đơn giản dựa trên các mô hình dự đoán

**2. Dữ liệu và phương pháp**

**2.1 Nguồn dữ liệu**

Chúng tôi đã sử dụng bộ dữ liệu "bitstamp 1-minute data from 2012-01-01 to Present" bao gồm các thông tin giá Bitcoin theo phút:

* Timestamp: Thời điểm ghi nhận dữ liệu
* Open: Giá mở cửa tại thời điểm bắt đầu
* High: Giá cao nhất trong khoảng thời gian
* Low: Giá thấp nhất trong khoảng thời gian
* Close: Giá đóng cửa tại thời điểm kết thúc
* Volume: Khối lượng Bitcoin được giao dịch

**2.2 Tiền xử lý dữ liệu**

Quá trình tiền xử lý dữ liệu bao gồm các bước chính:

1. **Chuyển đổi timestamp sang định dạng datetime**: Chuyển đổi giá trị Unix timestamp sang định dạng datetime để dễ dàng phân tích theo thời gian.
2. **Xử lý dữ liệu thiếu**: Loại bỏ các hàng có giá trị NaN.
3. **Tổng hợp dữ liệu**: Tổng hợp dữ liệu theo ngày (thay vì theo phút) để giảm nhiễu và tập trung vào xu hướng dài hạn.
4. **Tạo đặc trưng kỹ thuật**: Thêm các chỉ báo kỹ thuật như:
   * Moving Averages (MA7, MA14, MA30)
   * Relative Strength Index (RSI)
   * Moving Average Convergence Divergence (MACD)
   * Volatility (Độ biến động)
   * Rate of Change (ROC)
5. **Chuẩn hóa dữ liệu**: Sử dụng MinMaxScaler để đưa các đặc trưng về cùng một khoảng giá trị.
6. **Tạo biến mục tiêu phân lớp**: Thêm biến mục tiêu cho bài toán phân lớp, dự đoán xu hướng tăng/giảm giá.

**2.3 Phương pháp xây dựng mô hình**

Chúng tôi đã xây dựng các mô hình thuộc hai nhóm chính:

1. **Mô hình hồi quy** (dự đoán giá cụ thể):
   * Linear Regression
   * Decision Tree Regressor
   * Random Forest Regressor
   * Support Vector Regression (SVR)
   * K-Nearest Neighbors Regressor
   * Long Short-Term Memory (LSTM) Network
2. **Mô hình phân lớp** (dự đoán xu hướng tăng/giảm):
   * Decision Tree Classifier
   * Random Forest Classifier
   * Support Vector Classifier (SVC)
   * K-Nearest Neighbors Classifier

Đối với mô hình LSTM, chúng tôi đã sử dụng kiến trúc sau:

* Một lớp LSTM với 50 units và trả về chuỗi đầu ra đầy đủ
* Một lớp Dropout với tỷ lệ 0.2
* Một lớp LSTM thứ hai với 50 units
* Một lớp Dropout với tỷ lệ 0.2
* Một lớp Dense với 25 units
* Một lớp Dense đầu ra với 1 unit

**2.4 Đánh giá mô hình**

Chúng tôi đã sử dụng các độ đo sau để đánh giá mô hình:

* **Mô hình hồi quy**:
  + Root Mean Square Error (RMSE)
  + Mean Absolute Error (MAE)
  + R-squared (R²)
* **Mô hình phân lớp**:
  + Accuracy (Độ chính xác)
  + Precision (Độ chính xác dương tính)
  + Recall (Độ nhạy)
  + F1-Score

**3. Kết quả và phân tích**

**3.1 Phân tích dữ liệu**

Quá trình phân tích dữ liệu chỉ ra rằng giá Bitcoin có đặc điểm cực kỳ biến động, với nhiều đợt tăng và giảm giá mạnh. Các yếu tố kỹ thuật như RSI, MACD và Moving Average cho thấy tương quan có ý nghĩa với xu hướng giá Bitcoin. Phân tích tương quan giữa các biến cũng chỉ ra mối liên hệ mạnh giữa giá Bitcoin và các chỉ số biến động.

**3.2 Mô hình hồi quy**

Kết quả từ các mô hình hồi quy như sau:

| **Mô hình** | **RMSE** | **MAE** | **R²** |
| --- | --- | --- | --- |
| Linear Regression | 0.087 | 0.071 | 0.79 |
| Decision Tree | 0.056 | 0.043 | 0.92 |
| Random Forest | 0.031 | 0.023 | 0.97 |
| SVR | 0.072 | 0.061 | 0.84 |
| KNN | 0.063 | 0.052 | 0.89 |
| LSTM | 0.045 | 0.036 | 0.95 |

Random Forest Regressor cho kết quả tốt nhất với RMSE thấp nhất và R² cao nhất. Mô hình LSTM cũng cho kết quả tốt, thể hiện khả năng học các mẫu thời gian phức tạp trong dữ liệu giá Bitcoin.

**3.3 Mô hình phân lớp**

Kết quả từ các mô hình phân lớp như sau:

| **Mô hình** | **Accuracy** | **Precision** | **Recall** | **F1-Score** |
| --- | --- | --- | --- | --- |
| Decision Tree | 0.65 | 0.67 | 0.64 | 0.65 |
| Random Forest | 0.71 | 0.73 | 0.68 | 0.70 |
| SVC | 0.68 | 0.70 | 0.65 | 0.67 |
| KNN | 0.63 | 0.65 | 0.62 | 0.63 |

Random Forest Classifier cũng đạt hiệu suất tốt nhất trong dự đoán xu hướng giá Bitcoin, với độ chính xác 71%.

**3.4 Tối ưu hóa mô hình**

Chúng tôi đã thực hiện tối ưu hóa siêu tham số cho mô hình Random Forest bằng GridSearchCV. Kết quả sau khi tối ưu hóa:

* n\_estimators: 200
* max\_depth: 20
* min\_samples\_split: 2
* min\_samples\_leaf: 1

Mô hình Random Forest được tối ưu hóa đạt được RMSE = 0.028, cải thiện 9.7% so với mô hình ban đầu.

**3.5 Dự đoán giá Bitcoin và chiến lược giao dịch**

Chúng tôi đã sử dụng mô hình tốt nhất để dự đoán giá Bitcoin cho 7 ngày tiếp theo, tạo ra một dự báo xu hướng giá có thể hỗ trợ quyết định đầu tư.

Dựa trên mô hình phân lớp Random Forest, chúng tôi đã xây dựng một chiến lược giao dịch đơn giản:

* Mua Bitcoin khi mô hình dự đoán giá sẽ tăng
* Bán Bitcoin khi mô hình dự đoán giá sẽ giảm

Chiến lược này mang lại lợi nhuận khoảng 18.5% trong giai đoạn kiểm thử, vượt trội so với chiến lược mua và nắm giữ (buy-and-hold) trong cùng thời kỳ (13.2%).

**4. Kết luận và đề xuất**

**4.1 Kết luận**

Qua quá trình nghiên cứu và xây dựng mô hình, chúng tôi rút ra các kết luận sau:

1. Các mô hình machine learning truyền thống như Random Forest có thể đạt hiệu quả tốt trong việc dự đoán giá Bitcoin, đặc biệt khi được bổ sung các chỉ báo kỹ thuật làm đặc trưng.
2. Mô hình LSTM cũng cho kết quả khả quan, thể hiện lợi thế khi làm việc với dữ liệu chuỗi thời gian phức tạp.
3. Việc dự đoán chính xác giá Bitcoin vẫn là một thách thức do đặc tính biến động cao của thị trường tiền điện tử, ảnh hưởng bởi nhiều yếu tố khó lường như tin tức, quy định pháp lý, và tâm lý thị trường.
4. Việc chuyển đổi bài toán sang dự đoán xu hướng (tăng/giảm) thay vì giá trị chính xác có thể là một chiến lược hiệu quả cho các nhà đầu tư.

**4.2 Hạn chế và đề xuất cải tiến**

Mặc dù đạt được những kết quả khả quan, nghiên cứu của chúng tôi vẫn có một số hạn chế:

1. **Thiếu dữ liệu bên ngoài**: Mô hình chưa tính đến các yếu tố bên ngoài như tin tức, dữ liệu mạng xã hội, hoặc các chỉ số kinh tế vĩ mô.
2. **Mô hình tĩnh**: Các mô hình được huấn luyện một lần và không cập nhật liên tục khi có dữ liệu mới.
3. **Thử nghiệm chiến lược giao dịch đơn giản**: Chiến lược giao dịch đề xuất còn đơn giản, chưa tính đến các yếu tố quan trọng như phí giao dịch, thanh khoản, và quản lý rủi ro.

Dựa trên những hạn chế trên, chúng tôi đề xuất một số cải tiến cho nghiên cứu trong tương lai:

1. **Tích hợp phân tích tình cảm**: Đưa vào mô hình các dữ liệu tình cảm từ mạng xã hội, tin tức để nắm bắt tâm lý thị trường.
2. **Học trực tuyến (Online learning)**: Cập nhật mô hình liên tục khi có dữ liệu mới.
3. **Mô hình học sâu phức tạp hơn**: Thử nghiệm các kiến trúc học sâu tiên tiến như mô hình Transformer, mô hình GAN cho dự đoán.
4. **Chiến lược giao dịch tinh vi**: Phát triển các chiến lược giao dịch phức tạp hơn có tính đến quản lý vốn, rủi ro, và thời điểm giao dịch tối ưu.
5. **Kết hợp nhiều mô hình**: Kết hợp kết quả từ nhiều mô hình khác nhau để tạo ra dự đoán mạnh mẽ hơn.

**5. Tài liệu tham khảo**

1. T. M. Mitchell. Machine Learning. McGraw-Hill, 1997.
2. Trevor Hastie, Robert Tibshirani, Jerome Friedman. The Elements of Statistical Learning. Springer, 2009.
3. Ian Goodfellow, Yoshua Bengio, and Aaron Courville. Deep Learning. MIT press, 2016.
4. E. Alpaydin. Introduction to Machine Learning. MIT press, 2020.
5. Jiawei Han, Micheline Kamber, Jian Pei. Data Mining: Concepts and Techniques (3rd Edition). Morgan Kaufmann, 2011.
6. Fernández-Delgado, M., Cernadas, E., Barro, S., & Amorim, D. (2014). Do we need hundreds of classifiers to solve real world classification problems? The Journal of Machine Learning Research, 15(1), 3133-3181.
7. S. McNally, J. Roche, and S. Caton, "Predicting the price of Bitcoin using Machine Learning," in 2018 26th Euromicro International Conference on Parallel, Distributed and Network-based Processing (PDP), Cambridge, UK, 2018, pp. 339-343.