ĐẠI HỌC QUỐC GIA TP. HỒ CHÍ MINH

ĐẠI HỌC BÁCH KHOA

KHOA KHOA HỌC & KĨ THUẬT MÁY TÍNH

BÀI DỰ THI OLYMPIC KINH TẾ LƯỢNG

Xây Dựng Công Cụ Hỗ Trợ Dự Đoán Giá Trị Bitcoin Bằng Học Máy

Người thực hiện:

Phan Sơn Tự 51204436

Giáo viên hướng dẫn:

TS. Nguyễn Đức Thái

Giáo viên cố vấn:

TS. Nguyễn An Khương

TP. Hồ Chí Minh, ngày 3 tháng 4 năm 2017

Lời cam kết

Tôi tên Phan Sơn Tự - 51204436, hiện đang là sinh viên khoa Khoa học và Kĩ thuật máy tính, Đại học Bách Khoa TP.HCM. Tôi xin cam kết bài dự thi với đề tài "Xây dựng công cụ hỗ trợ dự đoán giá trị bitcoin bằng Học máy" là công trình nghiên cứu độc lập, tự tìm hiểu của bản thân, không sao chép bất kì công trình nghiên cứu nào.

Đề tài được thực hiện cho mục đích tìm hiểu và nghiên cứu ở bậc đại học.

Tất cả những tài liệu tham khảo được ghi trong báo cáo đều được trích dẫn rõ ràng từ các nguồn đáng tin cậy và từ một số bài báo khoa học.

Tất cả số liệu trong bài báo cáo đều được thực hiện một cách trung thực, không gian dối, không sao chép từ bất kì nguồn nào.

Các công cụ hỗ trợ cho việc thực hiện giải thuật, đo đạt số liệu đều là mã nguồn mở và tập dữ liệu được cung cấp hoàn toàn công khai của chủ nhân, tổ chức sở hữu.

Hình ảnh trong bài báo cáo đều được trích dẫn nguồn gốc rõ ràng.

Lời giới thiệu

"Cơ hội càng lớn, rủi ro càng cao" đó là một trong những câu nói thường được nghe trong môi trường kinh tế và tài chính, điều này đa số đúng với các hành vi đầu tư kinh tế. Người đầu tư giỏi là người đầu tư có khả năng đoán biết rủi ro từ đó giảm thiểu rủi ro nhưng vẫn gia tăng lợi nhuận, để làm được điều này người đầu tư cần có kiến thức chuyên sâu về kinh tế và kinh nhgiệm, trong đó kinh nghiệm chiếm một vị trí rất quan trọng.

Về một lĩnh vực khác, ngành Công nghệ thông tin đang trở thành một ngành không thể thiếu đối mọi lĩnh vực, nó làm thay đổi phương thức lao động, tạo ra các giá trị hoàn toàn mới, thúc đẩy các lĩnh vực khác cực kỳ mạnh mẽ. Và Kinh tế cũng không nằm ngoài tác động đó. Cũng vì vậy mà lĩnh vực Business Intelligence được sinh ra, đây là một sản phẩm của quá trình sử dụng chất xám Công nghệ thông tin để giải quyết thông minh các vấn đề kinh tế.

Business Intelligence là một bức tranh rộng lớn, riêng trong phạm vi bài dự thi này, tác giả xin trình bày một đề tài cụ thể, đó là sử dụng Học máy để giải quyết bài toán giảm thiểu rủi ro trong đầu tư Bitcoin.

Mục lục

Danh sách hình vẽ

Danh sách bảng

Danh mục chữ viết tắt

MNN	Multilayer Neural Network
KNN	K-Nearest Neighbors
LR	Logistic Regression
SVM	Support Vector Machine
RF	Random Forest
RNN	Recurrent Neural Network
LSTM	Long Short Term Memory
RBF	Radial Basis Function
GDA	Gaussian Discriminant Analysis
QDA	Quadratic Discriminant Analysis
BTC	Bitcoin
USD	US Dollar
ROC	Rate of Change
SO	Stochastic Oscillator
RDP	Relative Difference Percentage
API	Application Programming Interface
JSON	JavaScript Object Notation
UI	User Interface

Chương 1

Giới thiệu đề tài

1.1 Tính cấp thiết của đề tài

Bitcoin - một hệ thống tiền mã hóa (hay tiền điện tử) được xuất hiện lần đầu tiên vào năm 2009 bởi Satoshi Nakamoto [5], với những đặc tính ưu việt hơn cả tiền tệ truyền thống hiện nay đã khiến cho sự tăng lên nhanh chóng về giá trị. Nhận thấy được sức mạnh của tiền mã hóa có thể sẽ là tương lai của kinh tế và chính trị nên việc hiểu rõ cũng như đầu tư vào Bitcoin là việc đáng để quan tâm.

Trong giai đoạn hiện nay, đối với nước ta, Bitcoin là một khái niệm mới vì thế mà việc đầu tư khi chưa có nền tảng kiến thức hoặc kinh nghiệm đầu tư là hết sức rủi ro. Nhận thấy vấn đề này, bản thân đã đặt ra vấn đề "Tại sao không tạo ra một công cụ để cho nhà đầu tư có thể dựa vào như một yếu tố tham khảo tin cậy?".

Đồng thời, trong lĩnh vực công nghệ thông tin nói riêng, Học máy đang là nền tảng cho hàng loạt các sản phẩm công nghệ mang tính dự đoán thông minh, ngoài ra còn ứng dụng trong các lĩnh vực về trí thông minh nhân tạo, xử lý ngôn ngữ tự nhiên... và điều đó đang đi đúng với mục tiêu của vấn đề được đưa ra trong phạm vi bài dự thi này.

1.2 Đặc tả đề tài

Trên một sàn giao dịch tiền mã hóa điển hình, quá trình mua bán BTC được chia ra thành các giai đoạn thời gian và được gọi là phiên giao dịch. Một phiên giao dịch được diễn tả bởi các giá trị điển hình như sau:

- Giá mở phiên: giá bán (mua) BTC của (các) giao dịch ngay tại thời điểm mở phiên.

- Giá đóng phiên: giá bán (mua) BTC của (các) giao dịch tại thời điểm kết thúc phiên.
- Giá cao nhất: giá bán (mua) BTC cao nhất của giao dịch trong khoảng thời gian mở phiên đến kết thúc phiên.
- Giá thấp nhất: giá bán (mua) BTC thấp nhất của giao dịch trong khoảng thời gian mở phiên đến kết thúc phiên.

Thời gian của một phiên giao dịch thường được chọn là 5 phút, 30 phút, 1 tiếng, 2 tiếng, 4 tiếng hoặc 1 ngày, ... Trong phạm vi đề tài chúng ta chọn thời gian một phiên giao dịch là 30 phút.

Vậy, bài toán cần giải quyết là đi dự đoán giá trị BTC trong phiên tiếp theo sẽ tăng hay giảm so với phiên hiện tại. Cụ thể, gọi n là phiên hiện tại và n_{close} là giá đóng phiên hiện tại, $(n+1)$ là phiên tiếp theo và $(n+1)_{close}$ là giá đóng phiên tiếp theo. Nếu $(n+1)_{close} > n_{close}$ thì giá tăng - Up, ngược lại thì, $(n+1)_{close} \leq n_{close}$ thì giá giảm - $Down$.

Sau khi cụ thể được yêu cầu bài toán, ta sẽ đi đặc tả hướng tiếp cận giải quyết vấn đề. Học máy là lựa chọn của đề tài này, cụ thể phương pháp giải quyết sẽ sử dụng giải thuật phân lớp để dự đoán nhãn của phiên giao dịch sẽ là Up hay $Down$.

1.3 Mục tiêu của đề tài

Vấn đề cơ bản của việc đầu tư là lợi nhuận, bám sát với mục tiêu này phương hướng đề ra sẽ đi giải quyết bài toán cụ thể như sau.

Sử dụng USD để mua/bán BTC, với mỗi phiên giao dịch là 30 phút, chúng ta sẽ đi dự đoán giá trị BTC trong phiên tiếp theo sẽ tăng hay giảm - bài toán phân lớp trong Học máy.

Để thực hiện được điều đó chúng ta cần vạch ra những bước đi cụ thể để hiện thực mục tiêu:

- Thu thập, xử lí dữ liệu BTC.
- Áp dụng các giải thuật phân lớp vào tập dữ liệu có được.
- Đánh giá trên lý thuyết hệ thống.
- Vận hành, khảo sát và đánh giá hệ thống trên thực tế.
- Xây dựng, hoàn thiện sản phẩm.

Sản phẩm hoàn thiện mà người dùng được sử dụng sẽ là một ứng dụng nền web cung cấp các thông tin về dự đoán và các thông số thống kê dùng để tham khảo cho việc đầu tư.

1.4 Phương pháp thực hiện đề tài

Các công trình hoặc bài báo liên quan đến đề tài dự đoán BTC bằng Học máy hầu như mới chỉ xuất hiện một vài năm gần đây và chỉ một số ít trong các nghiên cứu này được công bố một cách công khai. Vì những giới hạn này, quá trình nghiên cứu và phân tích đề tài của bài dự thi này ngoài tham khảo những công trình có liên quan trực tiếp đến đề tài, sẽ còn tham khảo các công trình nghiên cứu khác có mức độ liên quan một các tương đối. Cụ thể như các công trình liên quan đến sử dụng Học máy trong dự đoán giá trị của vàng hoặc cổ phiếu.

Từ những kinh nghiệm của các công trình này, bản thân sẽ đúc kết một phương pháp tổng quát, từ đó áp dụng trở lại cho vấn đề dự đoán xu hướng giá trị BTC.

Đồng thời, ngoài việc tham khảo các công trình liên quan, bản thân còn vận dụng những kinh nghiệm cá nhân về khai phá dữ liệu và kiến thức Học máy, để áp dụng vào quá trình nghiên cứu nhằm đem lại kết quả tốt nhất. Việc tìm ra lời giải tốt nhất sẽ tiến hành theo phương pháp so sánh kết quả giữa các giải thuật, chúng ta sẽ đi chạy các giải thuật phân lớp khác nhau từ đó đánh giá xem giải thuật nào là tốt hơn và sẽ tập trung tối ưu cho giải thuật đó.

Sản phẩm hoàn thiện là sản phẩm đã được chạy và khảo nghiệm trên thực tế, vì vậy sau khi xây dựng hoàn chỉnh, hệ thống sẽ được chạy thực tế và đánh giá kết quả trong một khoảng thời gian.

1.5 Bố cục đề tài

Để phục vụ tốt cho việc phát triển sau này, bố cục đề tài sẽ được trình bày theo hướng diễn dịch và được chia thành các phần nhỏ để người đọc có thể nắm bắt nội dung.

Trước hết, chúng ta sẽ đi tìm hiểu qua các công trình liên quan nhằm hiểu được công việc chúng ta sẽ làm là gì, và những hướng giải quyết tổng quát đã được sử dụng ra sao.

Sau đó, phần nền tảng lý thuyết sẽ đề cập đến các kiến thức liên quan

đến Bitcoin, một số khái niệm về tài chính, cũng như lý thuyết giải thuật MNN dùng cho phân lớp để phục vụ cho việc đọc hiểu nội dung các chương sau, đặc biệt là phục vụ cho quá trình phân tích giải thuật phân lớp trong Học máy.

Cuối cùng, thu thập dữ liệu và khai phá dữ liệu cho phù hợp với giải thuật, chạy giải thuật, đánh giá giải thuật và hiện thực sản phẩm.

Chương 2

Những công trình liên quan

Hai công trình được công bố công khai có đề tài tương đồng với đề tài của bài dự thi này đó là "Automated Bitcoin Trading via Machine Learning Algorithms" [6] và "Predicting the price of Bitcoin using Machine Learning" [7]. Dây là hai công trình sử dụng trực tiếp Học máy trong bài toán dự đoán xu hướng giá trị BTC, mỗi công trình đều có một hướng tiếp cận riêng biệt và đồng thời cũng cho ra các kết quả khác nhau.

Trong bài báo [6], nhóm tác giả sử dụng một tập dữ liệu về giá trị BTC, tập dữ liệu này chứa khoảng 120.000 mẫu được thu thập thông qua API của Coinbase và OKCoin (đây là những loại ví điện tử được dùng để lưu trữ và trao đổi BTC, Coinbase xây dựng ở San Francisco và OKCoin xây dựng ở Bắc Kinh) - gọi đây là tập dữ liệu thứ nhất. Ngoài ra, nhóm tác giả còn thu thập thêm tập dữ liệu giá thường nhật kèm theo các thông tin về mạng Bitcoin, tập dữ liệu này bao gồm 26 đặc trưng nhưng chỉ được dùng 16 đặc trưng cho quá trình phân tích và chạy giải thuật - gọi đây là tập dữ liệu thứ hai.

Với tập dữ liệu thứ nhất, nhóm tác giả tiếp cận vấn đề bằng phương pháp chuỗi thời gian (Time Series) với 2 khoảng thang thời gian là 10 giây và 10 phút. Còn lại, tập dữ liệu thứ hai được sử dụng như một tập dữ liệu phân lớp đơn thuần. Kết quả chi tiết của các giải thuật LR, SVM và RF:

Tham số đánh giá	10 giây - LR	10 phút - SVM	10 phút - RF
Recall	54.29%	52.40%	54.00%
Specificity	57.70%	57.60%	61.90%
Precision	57.40%	55.10%	58.10%
Accuracy	8.50%	53.90%	57.40%

Bảng 2.1: Bảng đánh giá với tập dữ liệu thứ nhất

Lưu ý, nhóm tác giả ký hiệu giải thuật LR là Binomial GLM và tham số

đánh giá Recall là Sensitivity.

Tham số đánh giá	LR	SVM	RF
Recall	97.90%	3.48%	100%
Specificity	99.39%	55.14%	93.92%
Precision	97.90%	8.39%	77.62%
Accuracy	98.79%	27.16%	94.98%

Bảng 2.2: Bảng đánh giá với tập dữ liệu thứ hai

Với công trình nghiên cứu [7], tác giả sử dụng một hướng tiếp cận khác hơn so với bài báo [6]. Với tập dữ liệu giá BTC được thu thập thông qua CoinDesk từ ngày 19/8/2013 đến ngày 19/07/2016, tác giả sử dụng các kiến thức về Học sâu (Deep Learning) như là một công cụ chính để phân tích và giải quyết bài toán. Ngoài ra, trong công trình này còn nhắc đến việc sử dụng mô hình ARIMA - một dạng mô hình phân tích chuỗi thời gian - như là một giải thuật dùng để so sánh với giải thuật chính. Giải thích quyết định này, tác giả đã đưa ra lý luận phân tích rằng, ARIMA mặc dù là một mô hình dữ đoán chuỗi thời gian phổ biến, tuy nhiên, mô hình này lại phụ thuộc vào giả định tập dữ liệu là tuyến tính về thời gian và điều này không phù hợp với tập dữ liệu đang được sử dụng - tập dữ liệu được sử dụng là phi tuyến tính.

Đi vào chi tiết, giải thuật học sâu được sử dụng đó là RNN một dạng của mạng neural (xem phần 3.3.3), RNN có khả năng sử dụng cả hai giải thuật lan truyền là lan truyền thuận và lan truyền ngược. Ngoài ra, tác giả còn sử dụng một biến thể khác của RNN, đó là LSTM, ở RNN các tham số mạng được tìm bằng giải thuật di truyền (Genetic Algorithm), khác với LSTM sử dụng giải thuật tối ưu Bayesian (Bayesian Optimisation) để tìm tham số mạng. Kết quả của quá trình chạy giải thuật:

Tham số đánh giá	LSTM	RNN	ARIMA
Recall	37.00%	40.40%	14.7%
Specificity	61.30%	56.65%	100%
Precision	35.50%	39.08%	100%
Accuracy	52.78%	50.25%	50.05%

Bảng 2.3: Bảng đánh giá các giải thuật học sâu và ARIMA

Ngoài hai công trình trên, chúng ta sẽ tiếp tục tham khảo các vấn đề liên quan khác như là dự đoán xu hướng giá trị vàng và dự đoán xu hướng giá trị cổ phiếu. Mặc dù các công trình này không đi giải quyết vấn đề dự đoán giá trị BTC, nhưng với cách nhìn tổng quát, vấn đề chung vẫn là sử dụng Học máy để dự đoán giá trị các tài sản giao dịch. Vàng, cổ phiếu là hai tài sản giao dịch đặc trưng và lâu đời, vì thế các vấn đề về dự đoán của hai tài sản

giao dịch này cũng đã được khai thác và phát triển từ lâu. Hai công trình cụ thể được tham khảo trong đề tài là: "Predicting Gold Prices" [8] và "Machine Learning in Stock Price Trend Forecasting" [9]

Bài báo [8] liên quan đến ứng dụng Học máy cho việc dự đoán giá vàng, tác giả đã chọn hướng tiếp cận học có giám sát và cụ thể là bài toán phân lớp. Tập dữ liệu mà tác giả sử dụng là tập dữ liệu giá vàng từ đầu năm 2007 đến cuối năm 2013 với khoảng 1700 mẫu và được lấy từ trang web của USA Gold, đồng thời sử dụng hai giải thuật phân lớp là SVM và LR. Trong đó, khi gặp phải vấn đề mất cân đối trong tập dữ liệu (nhãn *positive* lớn hơn rất nhiều sao với nhãn *negative*) tác giả đã sử dụng giải thuật SVM với nhiều lần điều chỉnh mô hình như: sử dụng kernel RBF, sử dụng kernel tuyến tính với L1,... nhưng đều cho ra kết quả thấp (Accuracy nằm trong khoảng 50% - 51%). Với kết quả như vậy, tác giả đã quyết định sử dụng SVM như một giải thuật dùng để so sánh và đánh giá với giải thuật còn lại. Với một hướng tiếp cận khác, LR được sử dụng để giải quyết bài toán, kết quả được ghi nhận như bảng sau.

Precision	69.90%
Recall	72.31%
Accuracy	69.30%

Bảng 2.4: Bảng đánh giá giải thuật LR

Với kết quả trên, các tham số đánh giá cho kết quả trong khoảng gần bằng 70% và tác giả xem đây là một kết quả có ý nghĩa.

Bài báo [9] liên quan đến ứng dụng Học máy cho việc dự đoán giá trị cổ phiếu (cụ thể là công ty 3M), nguồn dữ liệu được sử dụng là Bloomberg Data Terminal với khoảng 1400 mẫu. Nhóm tác giả đã sử dụng bốn giải thuật trong quá trình phân tích và giải quyết bài toán đó là: GDA, LR, SVM, QDA.

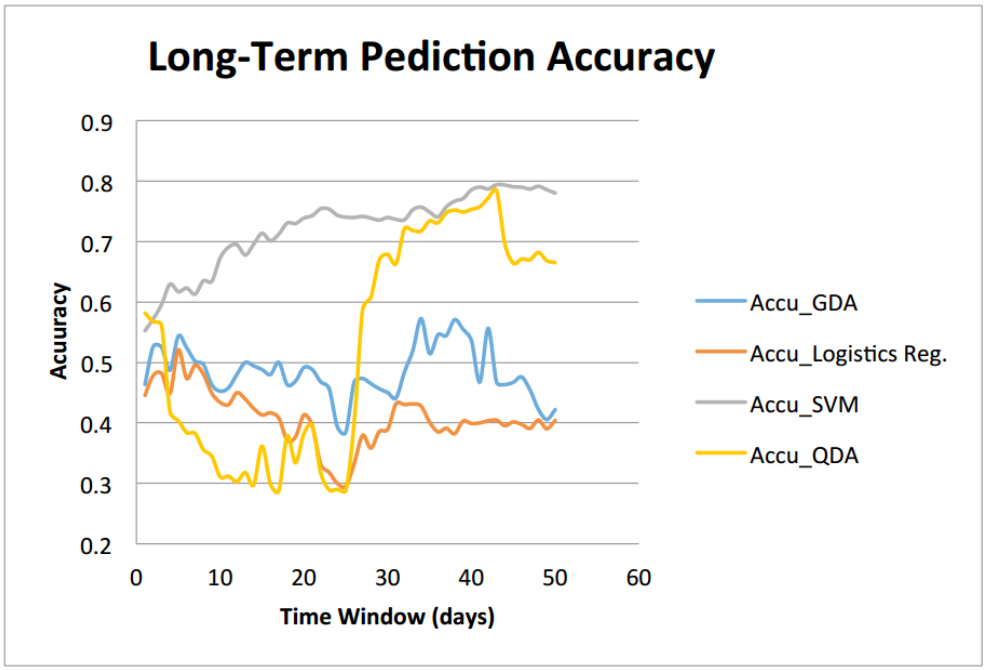
Ngoài ra, để có thể tìm ra một hướng giải quyết tối ưu, nhóm tác giả tiếp cận vấn đề dựa trên hai mô hình khác nhau. Thứ nhất, mô hình Ngày tiếp theo (Next-Day Model) với mục tiêu đi dự đoán xu hướng giá trị của cổ phiếu trong ngày tiếp theo. Và thứ hai, mô hình Dài hạn (Long-Term Model) với mục tiêu đi dự đoán xu hướng giá trị của n ngày tiếp theo.

Kết quả đánh giá của 4 giải thuật cho mô hình Ngày tiếp theo:

Model	LR	GDA	QDA	SVM
Accuracy	44.5%	46.4%	58.2%	55.2%

Bảng 2.5: Bảng đánh giá mô hình Ngày tiếp theo

Đồ thị đánh giá của 4 giải thuật cho mô hình Dài hạn:



Hình 2.1: Đồ thị đánh giá mô hình Dài hạn

Hai mô hình với hai kết quả, ta thấy ở mô hình Dài hạn hai giải thuật SVM và QDA cho kết quả Accuracy gần bằng 80% với $n = 44$ ngày, so sánh với mô hình Ngày tiếp theo Accuracy gần bằng 60% ta có thể thấy mô hình dài hạn cho kết quả có ý nghĩa hơn. Nhưng một hạn chế rất lớn của công trình này, nhóm tác giả chỉ sử dụng một tham số đánh giá duy nhất là Accuracy mà bỏ qua Precision và Recall. Việc đánh giá chỉ dựa trên duy nhất một tham số sẽ không thể tổng quát được kết quả đầu ra, vì vậy mà dẫn đến nguy cơ không phát hiện các vấn đề về lệch dữ liệu hoặc overfitting.

Tổng quan qua các công trình trên và tham khảo một số công trình khác, nhận thấy đa số các hướng tiếp cận đều đi theo một phương pháp tổng quát chung, nó bao gồm các bước cơ bản như:

1. Xây dựng không gian vector đặc trưng phù hợp với tính chất bài toán
2. Lựa chọn mô hình phân tích
3. Sử dụng các giải thuật phân lớp điển hình trong Học máy như là SVM, LR ...
4. Đánh giá giải thuật bằng các tham số Accuracy, Recall, Precision.

Từ những đục kết trên, bản thân nhận thấy các bước trên cũng chính là phương pháp nên dùng để tiếp cận đề tài. Ngoài ra, nhận thấy ở hai công

trình trên chưa đề cập đến việc sử dụng một giải thuật rất được phổ biến hiện nay, nó nổi lên như một đại diện của Học sâu đó là MNN. Do đó mà đề tài này sẽ sử dụng MNN như là một giải thuật bổ sung trong quá trình so sánh và đánh giá so với các giải thuật phân lớp khác.

Chương 3

Nền tảng lý thuyết

3.1 Bitcoin

Các hình thức thương mại trên Internet ngày nay hầu như đều dựa vào một tổ chức bên thứ ba đáng tin cậy để xử lý các hoạt động thanh toán điện tử. Tuy rằng sau nhiều năm phát triển, các tổ chức bên thứ ba này đều đã nâng cao mức độ tin cậy, an toàn nhưng đa số vẫn còn tồn tại những điểm yếu: không thể tránh khỏi những tranh chấp, phí trung gian, đòi hỏi phải cung cấp các thông tin cá nhân... Và Bitcoin - hệ thống tiền điện tử ngang hàng (A Peer-to-Peer Electronic Cash System) được sinh ra để giải quyết các vấn đề trên [5].

3.1.1 Sơ lược thông tin về Bitcoin

Khi so sánh với tiền tệ truyền thống, Bitcoin có hình thức và cách thức hoạt động khác biệt. Bitcoin không hề có bất kỳ một tổ chức tập trung nào để quản lý, thay vào đó, Bitcoin sử dụng mạng ngang hàng để hoạt động [1].

Trong cách viết, Bitcoin được hiểu như một hệ thống, giao thức hoặc một cộng đồng, cụ thể trong phạm vi bài dự thi này, Bitcoin được hiểu là một hệ thống. Còn BTC được hiểu là một tài sản hoặc một đơn vị tiền tệ.

BTC được sinh ra bằng cách sử dụng các tài nguyên phần cứng và năng lượng (CPU, GPU, phần cứng ASIC, điện năng...) để đi giải một "câu đố mã hóa", phần thưởng cho người chiến thắng chính là BTC. Trong thời gian đầu (210.000 block đầu tiên) (tham khảo về block ở mục 3.1.5), phần thưởng là 50 BTC và các giai đoạn sau sẽ giảm dần đi một nửa (25 BTC, 12.5 BTC ...) cứ sau khoảng 210.000 block. Đặc biệt, số lượng BTC là hữu hạn và chính xác là 21 triệu BTC, ước tính đến năm 2140 lượng BTC khai thác sẽ cạn kiệt [2]. Cũng chính tính chất này là một trong những yếu tố tạo nên giá trị của BTC, vì bị giới hạn số lượng nên BTC có tính khan hiếm và không bị

lạm phát, các tính chất này khác biệt so với tiền tệ truyền thống và được ví như vàng 2.0 - vàng của mạng Internet.

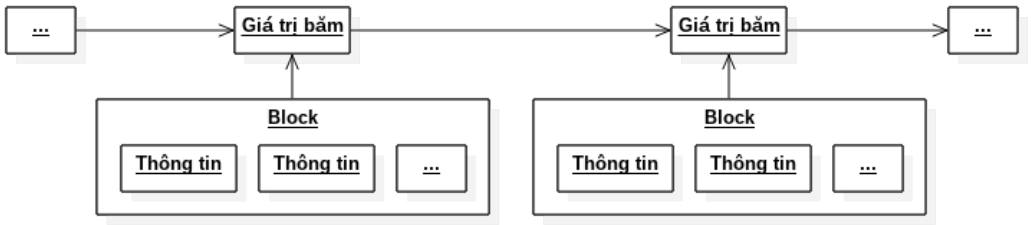
Ngoài các yếu tố vượt trội trên, Bitcoin còn được xây dựng là một hệ thống có tính chất ẩn danh, các địa chỉ chứa BTC đều là những dãy ký tự trừu tượng, không có ý nghĩa về mặt xác minh cá nhân và rất khó để biết ai là chủ nhân thật sự của một địa chỉ. Mặc dù, Bitcoin có tính ẩn danh nhưng tất cả những giao dịch trên hệ thống đều được công khai, điều đó có nghĩa là một giao dịch phải được sự xác minh về tính hợp lệ của đa số các thành viên trong mạng. Việc xác minh dựa vào các quan hệ, cấu trúc liên quan đến toán học, mật mã... [1, 5]

Tính công khai của Bitcoin được thể hiện ở quá trình xác minh các giao dịch, ngoài ra nó còn được thể hiện ở phương diện kỹ thuật, các mã nguồn lập trình và các đoạn lập trình đều được công bố công khai.

Đơn vị giao dịch, BTC có thể được chia thành nhiều đơn vị nhỏ hơn, tối thiểu là đơn vị *Satoshi*. Cụ thể $1\,Satoshi = 0.01\,\mu BTC = 0.00000001\,BTC$ hoặc $1\,BTC = 100.000.000\,Satoshi$ [3].

3.1.2 Máy chủ nhãn thời gian

Máy chủ nhãn thời gian (Timestamp Server) hoạt động bằng cách lấy giá trị băm (hash) của block liền trước và thông tin của block hiện tại, cho qua hàm băm để được một giá trị băm mới. Giá trị băm sau tính toán sẽ được công bố rộng rãi, nó mang ý nghĩa, giá trị băm này chứng minh rằng block thật sự tồn tại. Các block được nối với nhau thành một chuỗi xác định.

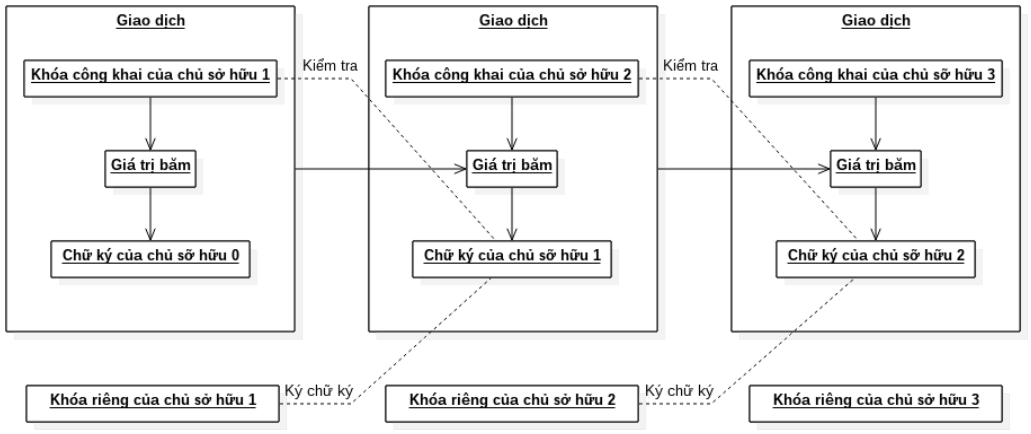


Hình 3.1: Máy chủ nhãn thời gian

3.1.3 Giao dịch (trên Blockchain)

Bitcoin tổ chức các giao dịch bằng cách xây dựng một chuỗi thông tin được đảm bảo bởi các chữ ký số. Một địa chỉ có chứa một lượng BTC được gọi là

một chủ sở hữu, một chủ sở hữu chuyển một lượng BTC cho một chủ sở hữu khác - người thụ hưởng - bằng cách ký lên giá trị băm, trong đó giá trị băm là kết quả sau khi đi qua hàm băm của tổ hợp giá trị băm giao dịch trước với địa chỉ người thụ hưởng.



Hình 3.2: Giao dịch

3.1.4 Proof-of-Work

Proof-of-Work được hiểu là bằng chứng để chứng minh quá trình lao động, nó dùng để kiểm tra quá trình tạo ra kết quả hợp lệ là một quá trình "lao động" có sử dụng và tiêu tốn tài nguyên.

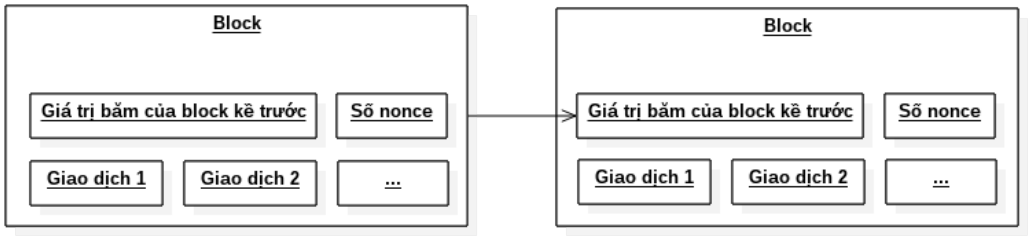
Proof-of-Work được sử dụng trong Bitcoin có cơ chế dựa trên hàm băm, ví dụ như SHA-256. Quá trình proof-of-work là quá trình đi tăng một con số - gọi là số *nonce* - sao cho giá trị băm của số *nonce* này cho kết quả đầu ra phải thỏa mãn tồn tại n bit 0 ở vị trí đầu, với n xác định và được gọi là số bit 0 yêu cầu.

3.1.5 Blockchain

Blockchain được hình thành dựa trên sự kết hợp giữa máy chủ nhãn thời gian và proof-of-work. Blockchain là một chuỗi các block được kết nối một cách luận lý với nhau thông qua các mỗi quan hệ toán học và mật mã, các quan hệ này đảm bảo cho hệ thống luôn đúng đắn, không thể sửa và dễ dàng để kiểm tra. Block mới được sinh ra phải dựa trên quá trình proof-of-work.

Quá trình hình thành blockchain được bắt đầu bằng proof-of-work, các peer sẽ đi tìm số *nonce* sao cho sau khi cho qua hàm băm kết quả đạt được là một

giá trị băm thỏa mãn số bit 0 yêu cầu. Số *nonce* vừa được tìm ra sẽ được đưa và block mới cùng với các thông tin khác như: thông tin các giao dịch, giá trị băm của block kề trước... Tiếp tục như vậy, các block mới được sinh ra và được kết nối với block cuối cùng của chuỗi.



Hình 3.3: Blockchain

Vì hệ thống có tính phân tán nên trên toàn mạng sẽ có nhiều phiên bản của blockchain, cũng chính vì thế để giải quyết tính đồng nhất, chỉ blockchain có độ dài lớn nhất mới được xem là blockchain hợp lệ. Đồng thời để kiểm soát được tốc độ sinh block mới, hệ thống sẽ quy định một độ khó, nếu toàn mạng có tốc độ sinh block (số block được sinh ra trong một giờ đồng hồ) cao hơn mức quy định, độ khó sẽ tăng lên để điều chỉnh lại tốc độ của toàn mạng.

3.1.6 Mạng

Mỗi thành viên (máy tính, phần cứng ASIC, thiết bị di động...) khi tham gia vào quá trình tính toán của toàn mạng thì sẽ được xem như một node. Toàn mạng sẽ hiện thực hệ thống bằng cách thực hiện các bước như sau:

1. Một giao dịch mới được truyền đi cho tất cả các node (broadcast).
2. Mỗi node sẽ lựa chọn và thu thập các giao dịch để đưa vào block.
3. Mỗi node sẽ thực hiện proof-of-work, tìm ra số *nonce*.
4. Khi một node hoàn thành proof-of-work, node này sẽ đóng block và truyền thông tin của block này cho tất cả các node khác trên toàn mạng.
5. Các node khác sẽ kiểm tra thông tin của block nhận được (thông tin các giao dịch, thông tin proof-of-work...) và chấp nhận block này nếu tất cả các thông tin đều được kiểm tra chính xác.
6. Các node sẽ thể hiện sự chấp nhận của mình bằng cách thực hiện proof-of-work để sinh ra block mới và block này sẽ được gắn vào liền

sau block mà node đã chấp nhận (thêm giá trị băm của block trước mà node chấp nhận vào trong block mới sinh ra).

Lưu ý, một giao dịch mới không nhất thiết phải được truyền đến tất cả các node. Chỉ cần việc truyền đến số node đủ nhiều để đảm bảo việc sẽ được đưa vào một block và được đóng trong blockchain với độ dài lớn nhất. Cũng như vậy đối với block, block không nhất thiết phải được truyền đến tất cả các node, khi một node nhận được một block kế block bị thiếu, bằng quá trình kiểm tra node có thể biết được và yêu cầu các node khác trong mạng gửi cho node này block bị thiếu sót.

3.1.7 Phần thưởng khích lệ

Trong tập các giao dịch được đóng trong một block sẽ luôn tồn tại một giao dịch đặc biệt, giao dịch này khác với các giao dịch bình thường, nó không có người chủ sở hữu mà chỉ có người thụ hưởng. Điều này giải thích cách mà BTC mới được sinh ra, cứ mỗi block được tìm ra nhờ quá trình proof-of-work sẽ có một lượng BTC được sinh ra và chính là phần thưởng cho người tạo ra block - người tạo ra block này cần được toàn mạng xác minh proof-of-work và công nhận là hợp lệ, địa chỉ người thụ hưởng chính là địa chỉ của người tạo ra block.

Ngoài ra, phần thưởng khích lệ khi tạo ra được một block còn bao gồm cả phí giao dịch từ các giao dịch đã được đóng trong block. Phí giao dịch thường rất nhỏ và không đáng kể ở thời điểm hiện tại.

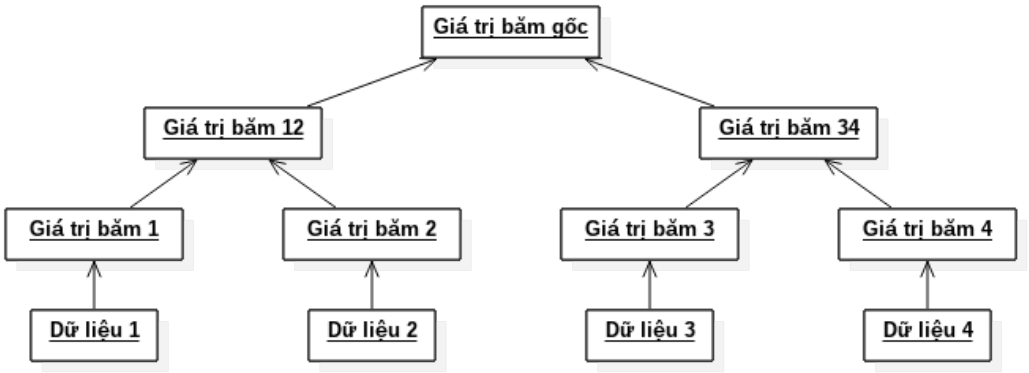
3.1.8 Tổ chức lưu trữ thông tin giao dịch

Đối với các node là các hệ thống máy tính lớn, khả năng lưu trữ và xử lý mạnh thì việc lưu một block với đầy đủ các thông tin không gặp nhiều vấn đề. Nhưng đối với các thiết bị di động hoặc các thiết bị khác với tài nguyên lưu trữ và xử lý tương đối hạn hẹp thì việc lưu một blockchain đầy đủ là khá khó khăn.

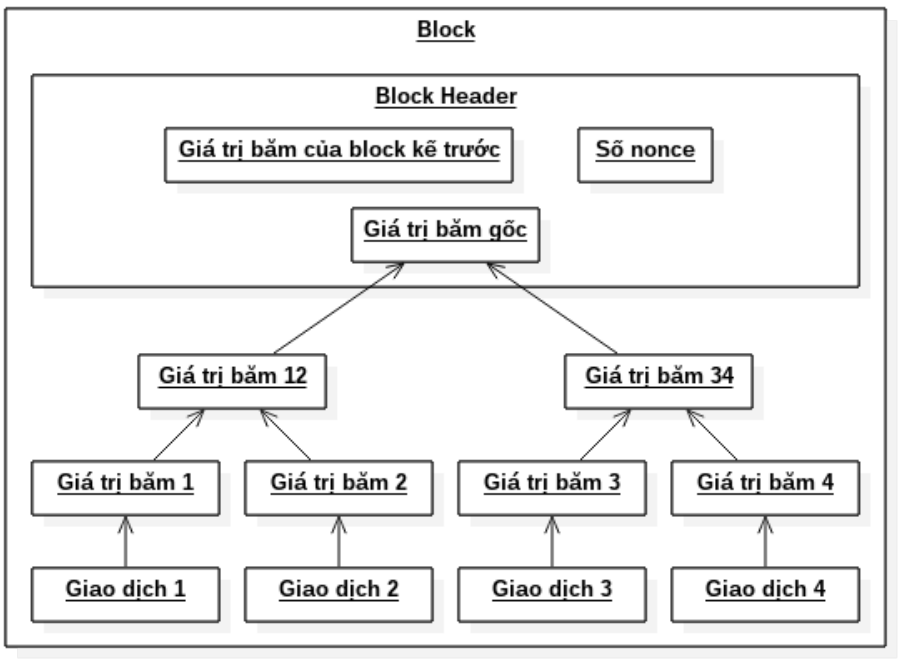
Cây Merkle là một cấu trúc tổ chức dữ liệu, trong đó giá trị của node cha sẽ là kết quả hàm băm tất cả các giá trị (nhãn hoặc dữ liệu) của nốt con. Các nốt không phải lá thì giá trị sẽ là nhãn - kết quả hàm băm, các nốt lá sẽ có giá trị là dữ liệu cần được tổ chức.

Bitcoin sử dụng cây Merkle để tổ chức các giao dịch, nốt cao nhất của cây được gọi là giá trị băm gốc (Root hash) và giá trị này sẽ được lưu vào block. Ở đây, ta thấy việc thay đổi bất kỳ một giá trị nào trong cây cũng sẽ dẫn đến việc thay đổi giá trị băm gốc, vì thế giá trị băm gốc khi được lưu vào block nó có chức năng dùng để kiểm tra lại các giao dịch trong block đó là

toàn vẹn hay không.



Hình 3.4: Cây Merkle



Hình 3.5: Cấu trúc tổ chức giao dịch trong một block

Một node với tài nguyên hạn chế khi muốn lưu blockchain không nhất thiết phải lưu đầy đủ thông tin của từng block trong blockchain, thay vào đó node có thể lược bỏ các thông tin về giao dịch được đóng trong block và chỉ lưu giá trị băm gốc của các giao dịch này. Điều này làm giảm chi phí về lưu trữ

nhưng vẫn đảm bảo được tính toàn vẹn, khi muốn xác minh bất kỳ giao dịch nào, node chỉ cần yêu cầu các giao dịch trong block và tính toán lại giá trị băm gốc, nếu giá trị băm nào giống với giá trị băm gốc được lưu nghĩa là các giao dịch hoàn toàn hợp lệ.

3.2 Sàn giao dịch Poloniex

3.2.1 Giới thiệu về sàn giao dịch

Sàn giao dịch là một thị trường có tính tổ chức cao, nơi các tài sản giao dịch có thể trao đổi bằng vật ngang giá (tiền tệ, hợp đồng, ...). Trong phạm vi đề tài này khi nhắc đến sàn giao dịch ta hiểu rằng đây là sàn giao dịch với tài sản giao dịch là BTC và vật ngang giá là USD.

Hiện nay, có rất nhiều sàn giao dịch, trong đó, Poloniex là một sàn giao dịch tương đối nổi tiếng trong cộng đồng Bitcoin bên cạnh các sàn khác như BTCMarkets, AltaPolis, Remitano... Bảng so sánh các sàn giao dịch:

Tên sàn	Thời gian hình thành	Quy mô	Tính đa dạng	Tính thanh khoản	API Dữ liệu
Poloniex	01/2014	Rất lớn	Cao	Cao	Có
BTCMarkets	07/2013	Lớn	Trung bình	Trung bình	Có
AltaPolis	10/2016	Nhỏ	Trung bình	Thấp	Không
Remitano	Chưa rõ	Nhỏ	Thấp	Thấp	Không

Bảng 3.1: Bảng so sánh sàn giao dịch

(Tính đa dạng biểu diễn số lượng nhiều hay ít các loại tài sản giao dịch khác nhau được sàn giao dịch hỗ trợ, tính thanh khoản tham khảo mục 3.3.2)

Vì Poloniex là một sàn lớn và cung cấp đầy đủ các API cho việc thu thập dữ liệu, ngoài ra, dữ liệu lịch sử giao dịch với số lượng lớn, nên Poloniex được chọn là sàn giao dịch đặc trưng dùng để phân tích trong bài dự thi này.

3.2.2 Giao dịch (trên sàn giao dịch)

Quá trình mua/bán BTC được thực hiện trên sàn giao dịch Poloniex được diễn ra bằng quá trình đặt cọc đầu tiên. Người dùng cần phải chuyển một lượng BTC hoặc USD vào một địa chỉ mà Poloniex cung cấp khi đăng ký tài khoản giao dịch. Khi đã chuyển một lượng BTC hoặc USD vào địa chỉ này,

người dùng có thể bắt đầu giao dịch.

Quá trình mua/bán được mô tả bằng một lệnh mua/bán của người dùng trên Poloniex, nếu lệnh này thành công, Poloniex là bên trung gian đại diện thực hiện việc chuyển USD từ tài khoản người mua sang tài khoản người bán và đồng thời, chuyển tài BTC từ tài khoản người bán sang tài khoản người mua. Quá trình chuyển BTC phức tạp hơn so với quá trình chuyển USD, lượng BTC cần chuyển được ghi vào một giao dịch (trên Blockchain), giao dịch này được truyền đi cho toàn mạng và chờ quá trình xác nhận từ các node - được ghi vào một block. Quá trình mua/bán thật sự hoàn tất khi giao dịch (trên Blockchain) được đóng trong block thuộc blockchain dài nhất.

Trên mỗi giao dịch, Poloniex tính phí dựa theo vai trò người dùng trong giao dịch đó, đối với người đặt lệnh phí là 0.15%, đối với người khớp lệnh (người đặt lệnh bán/mua để khớp với lệnh mua/bán đã có sẵn) phí là 0.25%. Phí này có thể giảm theo tổng số lượng BTC người dùng đã giao dịch trong lịch sử.

3.2.3 Bitcoin và sự công nhận của pháp luật

Hiện tại, Bitcoin đang dần dần được thông qua ở nhiều quốc gia, nó được xem là hợp pháp và là tương lai của kinh tế. Bắt đầu từ các nước châu Âu, từ khoảng tháng 10 năm 2015, các nước châu Âu chính thức công nhận đồng tiền này và chính thức công nhận thông qua giao dịch hàng hóa, dịch vụ và chuyển khoản [4]. Nổi bật nhất, Hoa Kỳ và Vương quốc Anh là một trong những nước lớn thể hiện thái độ tích cực và công nhận Bitcoin.

Ngoài những sự ủng hộ, đối với một số quốc gia, việc công nhận Bitcoin trở thành một rủi ro, nhà nước lo sợ các hành vi rửa tiền, mua bán vũ khí, chất cấm... Đồng thời, còn cân nhắc ảnh hưởng tới nền kinh tế vì thế nhiều quốc gia đang đặt ra lệnh cấm đối với Bitcoin, đáng kể nhất có thể liệt kê là Nga, Trung Quốc.

Đối với Viêt Nam, vào tháng 2/2014, Ngân hàng Nhà nước đã cho biết trong một tuyên bố: "Việc sở hữu, mua bán, sử dụng tiền ảo như là một loại tài sản tiềm ẩn rất nhiều rủi ro cho người dân và không được pháp luật bảo vệ. Ngân hàng Nhà nước khuyến cáo các tổ chức, các nhân không nên đầu tư, nắm giữ, thực hiện các giao dịch liên quan đến bitcoin và các loại tiền ảo tương tự khác". Ngân hàng Nhà nước thể hiện rõ quan điểm không ủng hộ Bitcoin, nhưng trong tình trạng đặt ngoài vòng pháp luật, điều đó có nghĩa là không ủng hộ cũng như không cấm.

3.3 Một số khái niệm về tài chính

3.3.1 Phiên giao dịch và các giá trị cơ bản

Gọi T là một mốc thời gian bất kỳ, P là khoảng thời gian được chọn là một phiên giao dịch. Ta có thể nói một cách đơn giản là phiên giao dịch được mở tại thời điểm T và được kết thúc tại thời điểm $T + P$.

Cụ thể, giả sử chọn mốc mở phiên là 9:00am và phiên giao dịch có thời hạn là 30 phút, điều đó có nghĩa là kết thúc phiên giao dịch sẽ là 9:30am.

Các thông tin của một phiên giao dịch:

- Giá mở phiên: là giá bán của một giao dịch gần nhất sau thời điểm T. Ví dụ tại thời điểm 9:01am có một giao dịch bán 1 BTC là \$779 và trong khoảng thời gian 9:00am đến 9:01am không hề có bất kỳ giao dịch nào khác ngoại trừ giao dịch này, thì ta có thể nói giá mở phiên sẽ là \$779.
- Giá đóng phiên: là giá bán của một giao dịch gần nhất trước thời điểm $T + P$.
- Giá phiên cao nhất: là giá bán cao nhất của một giao dịch trong khoảng thời gian diễn ra phiên giao dịch, cụ thể là từ thời điểm T đến thời điểm $T + P$. Ví dụ, trong khoảng thời gian 9:00am (thời điểm mở phiên) đến thời gian 9:30am (thời điểm đóng phiên) có một giao dịch BTC với giá là \$801 và là giao dịch có giá trị cao nhất. Vậy ta có thể nó giá phiên cao nhất là \$801.
- Giá phiên thấp nhất: là giá bán thấp nhất của một giao dịch trong khoảng thời gian diễn ra phiên giao dịch, cụ thể là từ thời điểm T đến thời điểm $T + P$.
- Lượng giao dịch: tổng giá trị USD được dùng để mua/bán BTC trong một phiên giao dịch.
- Trung bình giao dịch: giá trị USD trung bình của tất cả các giao dịch diễn ra trong khoảng thời gian một phiên giao dịch.

3.3.2 Tính thanh khoản

Tính thanh khoản là đặc trưng chỉ mức độ mà một tài sản bất kì có thể được mua hoặc bán trên thị trường mà không làm ảnh hưởng đến giá thị trường của tài sản đó. Một tài sản có tính thanh khoản cao nếu có thể được bán nhanh chóng mà giá bán giảm không đáng kể. Trong kế toán, tài sản lưu

động có thể được chia làm 5 loại và được sắp xếp theo tính thanh khoản từ cao đến thấp như sau: tiền mặt, đầu tư ngắn hạn, khoản phải thu, ứng trước ngắn hạn và hàng tồn kho. Trong đó, tiền mặt có tính thanh khoản cao nhất vì luôn luôn dùng được trực tiếp để thanh toán, lưu thông, tích trữ.

Cách gọi thay thế khác cho tính thanh khoản đó là tính lỏng hoặc tính lưu động.

3.3.3 Quá mua và Quá bán

Quá mua (Overbought) dùng để định nghĩa trường hợp thị trường có mức cầu cao hơn mức cung, điều này làm đẩy giá của tài sản giao dịch lên cao vượt qua mức chính đáng. Ví dụ, trên tất cả các giao dịch có hơn 80% lệnh là đặt mua và dưới 20% là lệnh đặt bán, trường hợp này được xem là quá mua.

Ngược lại, quá bán (Oversold) dùng để định nghĩa trường hợp thị trường có mức cung cao hơn mức cầu, điều này làm kéo giá của tài sản giao dịch xuống thấp vượt mức chính đáng.

3.3.4 Chỉ số dao động ngẫu nhiên

Chỉ số dao động ngẫu nhiên (Stochastic Oscillator) là đại lượng dùng để đo xu hướng mua/bán của thị trường tại thời điểm phiên x thông qua n phiên trước đó. Giả sử:

L_n = giá phiên thấp nhất trong n phiên
H_n = giá phiên cao nhất trong n phiên
$P(x)$ = giá của ngày x

$$\%K = \frac{P(x) - L_n}{H_n - L_n}$$

Nếu $\%K$ nhỏ hơn 20 thì thị trường đang có xu hướng quá bán và nếu lớn hơn 80 thì thị trường đang có xu hướng quá mua.

3.3.5 Tỉ lệ thay đổi

Tỉ lệ thay đổi (Rate of Change) là đại lượng đo sự khác nhau của giá tại phiên thứ x so với n phiên trước đó. Giá sử $P(x)$ là giá của phiên thứ x thì:

$$ROC_n(x) = \frac{P(x) - P(x-n)}{P(x-n)}$$

Nếu $ROC > 0$ thì giá thị trường đang có xu hướng đi lên (tăng giá). Ngược lại, với $ROC < 0$ thì giá thị trường đang có xu hướng giảm xuống.

3.4 Học máy

3.4.1 Khái niệm cơ bản

3.4.1.1 Học máy

Học máy có hai cách định nghĩa chính và đang được chấp nhận phổ biến:

- Theo Arthur Samuel: *"Là một lĩnh vực nghiên cứu mà nó cung cấp cho máy tính khả năng học hỏi mà không cần lập trình một cách tường minh."*
- Theo Tom Mitchell: *"Một chương trình máy tính được chấp nhận là học hỏi được kinh nghiệm E bằng cách thực hiện một vài tác vụ T theo phép đo hiệu năng P, nếu và chỉ nếu việc thực thi các tác vụ trong T được đo bởi phép đo P đem lại kết quả là kinh nghiệm E được cải thiện."*

3.4.1.2 Học có giám sát

Học có giám sát (Supervised Learning) được định nghĩa, chúng ta được cho một tập dữ liệu đã biết với các đầu vào và đầu ra tương ứng nhau. Ý tưởng là chúng ta sẽ đi tìm mối quan hệ giữa đầu vào và đầu ra, đó chính là học có giám sát.

Vấn đề của học có giám sát được phân loại thành hai vấn đề chính là hồi quy (Regression) và phân lớp (Classification). Trong vấn đề hồi quy, chúng ta sẽ cố gắng dự đoán kết quả đầu ra tiếp theo một cách liên tục, nghĩa là chúng ta đi tìm ra một hàm đầu ra liên tục tổng quát với biến là các đặc trưng đầu vào. Còn với vấn đề phân lớp, chúng ta thay vì cố gắng dự đoán kết quả liên tục thì ta sẽ đi dự đoán chúng theo hướng rời rạc, hiểu theo một cách khác là chúng ta đi tìm một phép phân loại rời rạc cho các biến đầu ra với các biến đầu vào.

3.4.1.3 Học không giám sát

Học không giám sát (Unsupervised Learning) cho phép chúng ta tiếp cận các vấn đề mà ta chưa hề hoặc biết rất ít kết quả của chúng ta sẽ trông như thế nào. Chúng ta có thể xây dựng cấu trúc của dữ liệu mà không cần thiết phải biết mối quan hệ của các biến đó.

Chúng ta thực hiện việc này dựa trên ý tưởng gom cụm dữ liệu bằng cách xem xét mối quan hệ giữa các đặc trưng của dữ liệu. Các hướng tiếp cận dựa trên nhưng phương pháp như vậy thường được gọi là gom cụm (Clustering).

3.4.2 Thông số đánh giá

Có ba tham số cơ bản dùng để xem xét và đánh giá giải thuật trong Học máy. Ký hiệu:

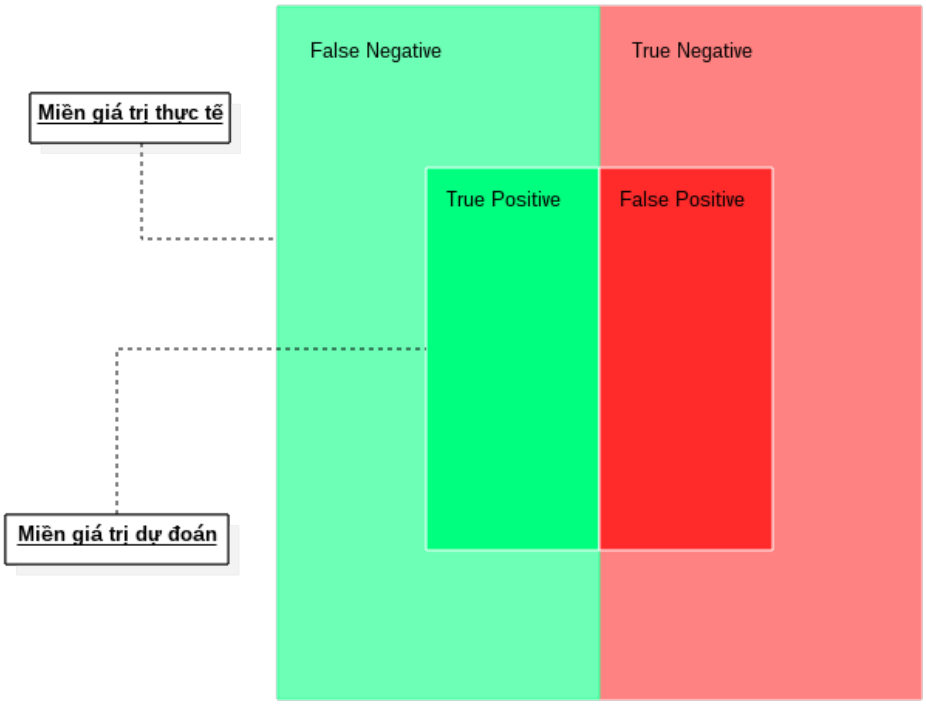
- True positive là TP
- False positive là FP
- True negative là TN
- False negative là FN

Ta có:

$$Accuracy = \frac{TP + TN}{TP + FP + TN + FN}$$

$$Precision = \frac{TP}{TP + FP}$$

$$Recall = \frac{TP}{TP + FN}$$



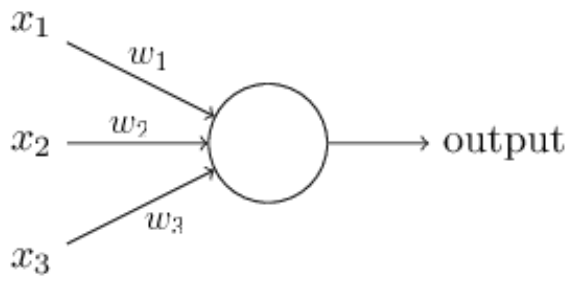
Hình 3.6: Thông số đánh giá

3.4.3 Mạng neural

Học sâu là một nhánh của Học máy, đại diện cho hướng tiếp cận gần với cái nhìn thực tế, học nhiều cấp và học từ bản chất dữ liệu. Học sâu thường giải quyết rất tốt với các loại dữ liệu mang tính "con người" như hình ảnh, âm thanh ... Để có được những tính chất này, Học sâu đã đưa ra các giải thuật mạng neural với cấu tạo nhiều lớp, mỗi lớp lại được cấu tạo từ nhiều phần tử nhỏ hơn. Mỗi phần tử bên trong đi giải quyết các phép toán rất đơn giản, nhưng khi được ghép nối lại thành một mạng neural hoàn chỉnh chúng có thể giải quyết các bài toán phức tạp hơn rất nhiều, điều này hoàn toàn tương tự với các hoạt động của bộ não người. [10]

3.4.3.1 Cấu trúc một perceptron

Một perceptron sẽ có các giá trị đầu vào $x_1, x_2, ...$, giá trị đầu ra sẽ là kết quả toán học của các giá trị đầu vào và là một giá trị nhị phân.



Hình 3.7: Perceptron

Một ví dụ, dựa vào hình trên, ta thấy perceptron này có 3 giá trị đầu vào là x_1, x_2, x_3 , giả sử đi kèm với mỗi giá trị đầu vào sẽ có một giá trị trọng số w_1, w_2, w_3. Giá trị đầu ra được định nghĩa là 0 và 1, nhận giá trị 0 khi $\sum_j w_j x_j$ nhỏ hơn giá trị ngưỡng và 1 khi lớn hơn giá trị ngưỡng.

Biểu diễn đại số:

$$output = \begin{cases} 1 & if \sum_j w_j x_j > threshold \\ 0 & if \sum_j w_j x_j \leq threshold \end{cases}$$

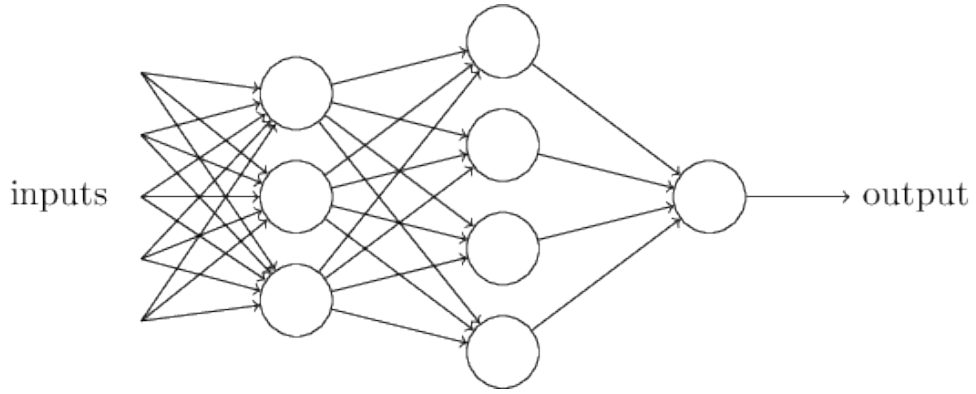
Các hàm số như trên được gọi là hàm hoạt động - activation function, có nhiều loại hàm hoạt động khác nhau như: $sigmoid, tang...$

3.4.3.2 MNN

Với một perceptron đơn lẻ, tự thân nó không thể giải quyết được một bài toán phức tạp như yêu cầu của mạng neural, vì thế, để giải quyết bài toán

phức tạp hơn, các perceptron được kết nối với nhau tạo thành một mạng neural.

MNN được cấu thành bằng cách sắp xếp các perceptron thành từng lớp. Các perceptron ở mỗi lớp sẽ kết nối với tất cả các perceptron ở các lớp liền kề, lớp những perceptron đầu tiên được gọi là lớp đầu vào (input layer), chúng có chức năng tiếp nhận các giá trị đầu vào để cung cấp cho các lớp tiếp theo. Các giá trị đầu ra ở lớp trước sẽ chính là các giá trị đầu vào cho các perceptron ở lớp tiếp theo. Các perceptron ở lớp cuối cùng được gọi là lớp đầu ra (output layer), trong trường hợp này đặc biệt chỉ có duy nhất một perceptron ở lớp đầu ra. Còn lại các lớp perceptron khác được gọi là lớp ẩn (hidden layer).

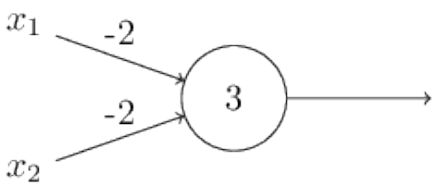


Hình 3.8: MNN

Giả sử đầu vào của perceptron là $x_1, x_2, ...$ tương ứng là đó là các trọng số $w_1, w_2, ...$ Thêm vào định nghĩa về *bias*, ở đây *bias* là một giá trị đại diện độ lệch của từng perceptron và được ký hiệu $b_1, b_2, ...$ Ta có biểu diễn của hàm hoạt động với *bias*:

$$output = \begin{cases} 1 & if \sum_j w_j x_j + b_i > 0 \\ 0 & if \sum_j w_j x_j + b_i \leq 0 \end{cases}$$

Ví dụ:



Hình 3.9: Ví dụ perceptron với giá trị *bias*

Ta có $w_1 = w_2 = -2$ và $b = 3$, khi đó nếu input $x_1 = 1, x_2 = 0$ suy

ra $w_1 * x_1 + w_2 * x_2 + b = (-2) * 1 + (-2) * 0 + 3 = 1$, ta có thể chọn $threshold = 0$ vì $1 > 0$ nên $output = 1$.

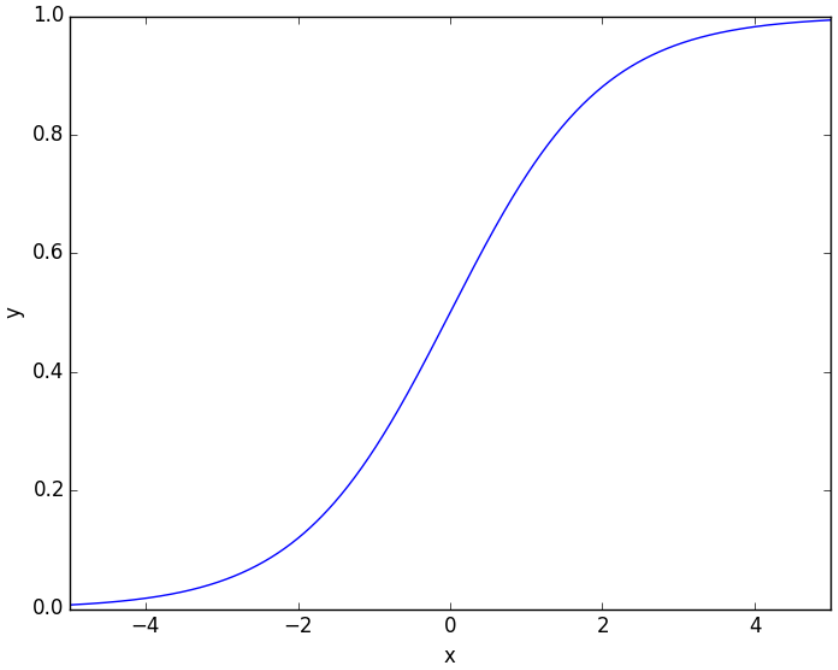
3.4.3.3 Hàm sigmoid

Với hàm hoạt động được định nghĩa như trên, giá trị của hàm hoạt động trên lý thuyết là không có giới hạn, nghĩa là $output \in \mathbb{R}$. Trong một trường hợp cụ thể, với việc sử dụng hàm hoạt động như trên có thể dẫn đến trường hợp đầu ra của một perceptron sẽ nhận giá trị rất lớn - giả sử là 1000, những một perceptron khác sẽ nhận giá trị rất bé - giả sử 0.001. Vì thế khi đến lớp tiếp theo thì gần như perceptron cho kết quả đầu ra là giá trị bé sẽ mất đi độ ảnh hưởng và làm mất cân đối cho toàn mạng.

Do đó để giới hạn giá trị đầu ra của hàm hoạt động chúng ta sẽ sử dụng hàm sigmoid. Hàm sigmoid được định nghĩa như sau:

$$\sigma(z) = \frac{1}{1 + e^{-z}}$$

Với biểu diễn đồ thị:



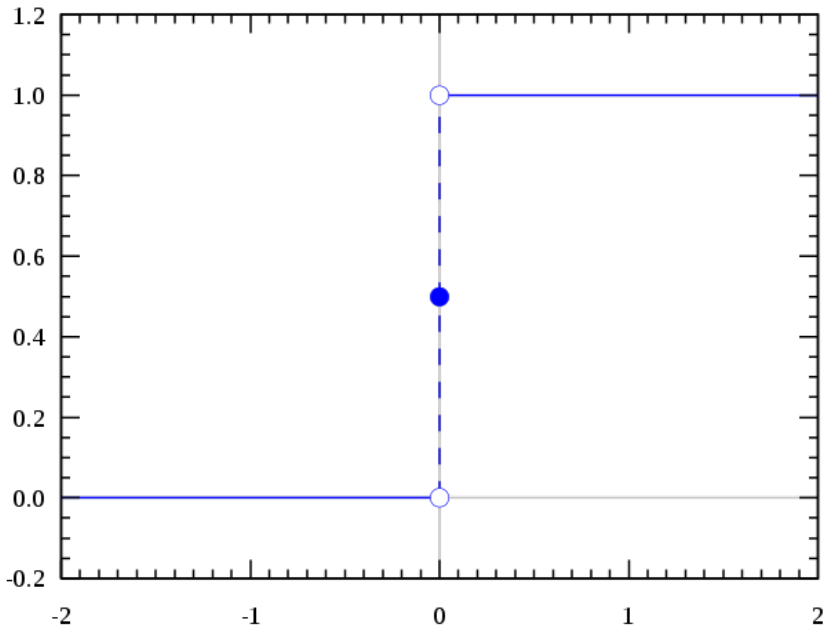
Hình 3.10: Đồ thị hàm sigmoid

Áp dụng hàm sigmoid vào hàm hoạt động ta có một hàm hoạt động dạng

sigmoid và khi đó hàm hoạt động của chúng ta của chúng ta sẽ có dạng:

$$\frac{1}{1 + exp(-\sum_j w_j x_j - b)}$$

Lúc này ta có một hàm hoạt động có giá trị được giới hạn trong khoảng từ 0 đến 1. Nhưng chú ý, giá trị đầu ra của hàm hoạt động vẫn là giá trị liên tục, để rời rạc hóa giá trị đầu ra của hàm hoạt động ta có thể sử dụng một phương pháp quen thuộc - sử dụng ngưỡng. Điển hình ta chọn ngưỡng $threshold = 0.5$, nếu lớn hơn ngưỡng thì giá trị đầu ra của hàm hoạt động sẽ nhận 1 và ngược lại sẽ nhận 0.



Hình 3.11: Đồ thị rời rác hóa hàm sigmoid

3.4.3.4 Giải thuật lan truyền ngược

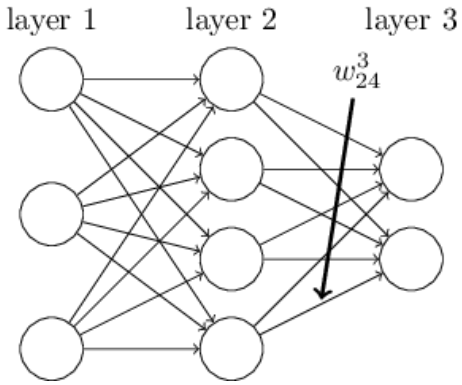
Sau khi đã có xây dựng thành công một mô hình MNN, công việc tiếp theo sẽ là cung cấp khả năng tự học hỏi từ đó để bản thân mạng có thể tự xây dựng mô hình quyết định và đưa ra các giá trị đầu ra tương ứng với từng trường hợp đầu vào cụ thể.

Cụ thể, khi nhìn lại một MNN với hàm hoạt động là một hàm hoạt động có dạng sigmoid và các tham số w, b, các tham số này chưa có giá trị. Việc cung cấp khả năng tự học hỏi cho mạng chính là cung cấp một giải thuật giúp mạng tìm được các tham số w, b với một tập kinh nghiệm - hay tập huấn

luyện - x, y cụ thể, trong đó x là giá trị đầu vào và y là giá trị đầu ra tương ứng với từng bộ x. Giải thuật lan truyền ngược là một giải thuật giúp giải quyết vấn đề trên.

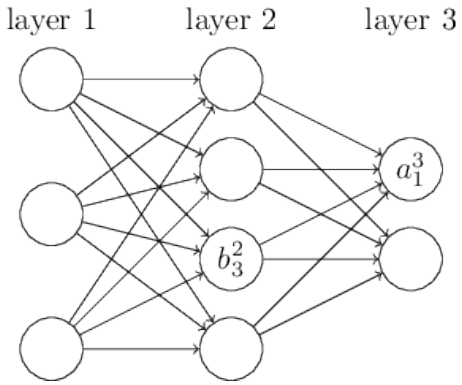
Biểu diễn trọng số, *bias* và hàm hoạt động:

Trọng số w^{ℓ}_{jk} là trọng số từ lớp perceptron thứ k của lớp $(\ell - 1)$ đến perceptron thứ j thuộc lớp thứ ℓ. Như hình trên, trọng số xuất phát từ perceptron thứ 4 thuộc lớp thứ 2 và kết thúc tại perceptron thứ 2 thuộc lớp thứ 3 được ký hiệu là w^3_{24}.



Hình 3.12: Ký hiệu trọng số

Tương tự như vậy với *bias* và hàm hoạt động của perceptron thứ j thuộc lớp thứ ℓ của mạng sẽ được kí hiệu thứ tự là b^{ℓ}_j, a^{ℓ}_j. Ví dụ, *bias* của perceptron thứ 3 thuộc lớp thứ 2 sẽ là b^2_3 và hàm hoạt động của perceptron thứ 1 thuộc lớp thứ 3 sẽ là a^3_1.



Hình 3.13: Ký hiệu *bias*

Biểu diễn ma trận trọng số và vector *bias*:

- w là ma trận của các giá trị trọng số.

$$w = \begin{bmatrix} w_{11} & \dots & w_{1k} \\ \vdots & \ddots & \vdots \\ w_{j1} & \dots & w_{jk} \end{bmatrix}$$

- b là vector của các giá trị *bias*.

$$b = \begin{bmatrix} b_1 \\ \vdots \\ b_k \end{bmatrix}$$

- σ là hàm sigmoid.
- $a_j = \sigma$ là hàm hoạt động có dạng sigmoid, a là vector của các hàm hoạt động.

$$a = \begin{bmatrix} a_1 \\ \vdots \\ a_k \end{bmatrix}$$

Lúc này ta có biểu diễn toán học đầy đủ của hàm hoạt động:

$$a_j^\ell = \sigma(\sum_k w_{jk}^\ell a_k^{\ell-1} + b_j^\ell) = \sigma(z_j^\ell)$$

$$z_j^\ell = \sum_k w_{jk}^\ell a_k^{\ell-1} + b_j^\ell$$

Tổng quát phát biểu với dạng:

$$a^\ell = \sigma(w^\ell a^{\ell-1} + b^\ell) = \sigma(z^\ell)$$

Để có thể tìm được giá trị của các biến w và b, giải thuật lan truyền ngược thực hiện bằng cách xuất phát w và b từ các giá trị ngẫu nhiên, thực hiện các phép lặp để đưa w và b về giá trị đúng, mỗi lần lặp sẽ có một hàm chi phí (cost function) đo đạc độ lệch của giá trị tính toán và giá trị thực tế để giúp giải thuật tìm được giá trị hội tụ.

Biểu diễn hàm chi phí với L lớp:

$$C = \frac{1}{2n} \sum_x \|y(x) - a^L(x)\|^2$$

Ta có thể thấy, dạng hàm số trên tương đồng với định nghĩa độ lệch chuẩn trong xác suất thống kê nhưng có một số biến đổi khác biệt. Thay vì giá trị

kỳ vọng và các giá trị xác suất, hàm chi phí sử dụng giá trị thực tế y của tập dữ liệu và giá trị $y = a$ là giá trị y tính toán được từ x với w và b. Vậy ta có thể hiểu được, hàm chi phí tính toán độ sai lệch của giá trị a so với y kỳ vọng thực tế. Do đó, hàm chi phí càng nhỏ thì biểu diễn giá trị của MNN sẽ càng gần với thực tế.

Để tìm được giá trị cực tiểu cho hàm chi phí ta sẽ thực hiện vòng lặp:

$$w_{jk}^{\ell} := w_{jk}^{\ell} - \eta \frac{\partial}{\partial w_{jk}^{\ell}} C(w, b)$$

$$b_{j}^{\ell} := b_{j}^{\ell} - \eta \frac{\partial}{\partial b_{j}^{\ell}} C(w, b)$$

Trong đó η là tỉ lệ học (learning rate), việc hội tụ về giá trị cực tiểu với tốc độ và độ chính xác phụ thuộc vào tỉ lệ này. Thực hiện được hai vòng lặp trên, ta sử dụng giải thuật lan truyền ngược để tính toán $\frac{\partial}{\partial w_{jk}^{\ell}} C(w, b)$ và $\frac{\partial}{\partial b_{j}^{\ell}} C(w, b)$.

Tham số lỗi δ_j^{ℓ} của neural thứ j thuộc lớp ℓ là biểu diễn của:

$$\delta_j^{\ell} \equiv \frac{\partial C}{\partial z_j^{\ell}}$$

Từ biểu diễn trên ta có:

$$\delta_j^L = \frac{\partial C}{\partial a_j^L} \sigma'(z_j^L)$$

Mặc khác $\partial C / \partial a_j^L = (a_j^L - y_j)$, suy ra:

$$\delta^L = \nabla_a C \odot \sigma'(z^L) = (a^L - y) \odot \sigma'(z^L)$$

Trong đó $\odot$ là tích Hadamard, ta gọi biểu thức trên là biểu thức (3.1). Ngoài ra, ta có biểu diễn khác của δ_j^L:

$$\begin{aligned}
\delta_j^L = \frac{\partial C}{\partial z_j^L} &= \sum_k \frac{\partial C}{\partial z_k^{\ell+1}} \frac{\partial z_k^{\ell+1}}{\partial z_j^{\ell}} \\
&= \sum_k \frac{\partial z_k^{\ell+1}}{\partial z_j^{\ell}} \delta_k^{\ell+1} \\
&= \sum_k \frac{\partial (\sum_j w_{kj}^{\ell+1} a_j^{\ell} + b_k^{\ell+1})}{\partial z_j^{\ell}} \delta_k^{\ell+1} \\
&= \sum_k \frac{\partial (\sum_j w_{kj}^{\ell+1} \sigma(z_j^{\ell}) + b_k^{\ell+1})}{\partial z_j^{\ell}} \delta_k^{\ell+1} \\
&= \sum_k w_{kj}^{\ell+1} \delta_k^{\ell+1} \sigma'(z_j^{\ell})
\end{aligned}$$

Gọi biểu thức trên là biểu thức (3.2). Với hướng tiếp cận tương tự ta có các biểu thức khác:

$$\begin{aligned}\frac{\partial C}{\partial b_j^\ell} &= \frac{\partial C}{\partial z_j^\ell}\frac{\partial z_j^\ell}{\partial b_j^\ell} \\ &= \delta_j^\ell \frac{\partial z_j^\ell}{\partial b_j^\ell} \\ &= \delta_j^\ell \frac{\partial(\sum_k w_{jk}^\ell a_k^{\ell-1} + b_j^\ell)}{\partial b_j^\ell} \\ &= \delta_j^\ell\end{aligned}$$

Ta có biểu thức trên là biểu thức (3.3).

$$\begin{aligned}\frac{\partial C}{\partial w_{jk}^\ell} &= \frac{\partial C}{\partial z_j^\ell}\frac{\partial z_j^\ell}{\partial w_{jk}^\ell} \\ &= \delta_j^\ell \frac{\partial(\sum_k w_{jk}^\ell a_k^{\ell-1} + b_j^\ell)}{\partial w_{jk}^\ell} \\ &= \delta_j^\ell a_k^{\ell-1}\end{aligned}$$

Và biểu thức (3.4). Từ (3.1), (3.2), (3.3) và (3.4) ta có tổng kết sau:

$$\delta^L = \nabla_a C \odot \sigma'(z^L) \tag{3.1}$$

$$\delta^\ell = ((w^{\ell+1})^T \delta^{\ell+1}) \odot \sigma'(z^\ell) \tag{3.2}$$

$$\frac{\partial C}{\partial b_j^\ell} = \delta_j^\ell \tag{3.3}$$

$$\frac{\partial C}{\partial w_{jk}^\ell} = a_k^{\ell-1}\delta_j^\ell \tag{3.4}$$

Sử dụng các biểu thức này, giải thuật lan truyền ngược dùng để tính toán $\frac{\partial}{\partial w_{jk}^\ell}C(w,b)$ và $\frac{\partial}{\partial b_j^\ell}C(w,b)$ được mô tả các bước như sau:

1. Cho giá trị đầu vào x, tính toán giá trị a^1 tương ứng với x.
2. Với mỗi $\ell = 2, 3, .., L$, ta lần lượt tính được $z^\ell = w^\ell a^{\ell-1} + b^\ell$ và $a^\ell = \sigma(z^\ell)$.
3. Tính giá trị của vector $\delta^L = \nabla_a C \odot \sigma'(z^L)$.
4. Từ giá trị của δ^L ta có thể tính được giá trị của các tham số lỗi còn lại, với mỗi $\ell = L-1, L-2, ..., 2$ ta thực hiện $\delta^\ell = ((w^{\ell+1})^T\delta^{\ell+1}) \odot \sigma'(z^\ell)$.
5. Kết thúc quá trình tính toán với $\frac{\partial C}{\partial w_{jk}^\ell} = a_k^{\ell-1}\delta_j^\ell$ và $\frac{\partial C}{\partial b_j^\ell} = \delta_j^\ell$.

Chương 4

Phân tích và thiết kế hệ thống

4.1 Thu thập dữ liệu

4.1.1 Nguồn dữ liệu

Poloniex là một sàn giao dịch tiền mã hóa trực tuyến và có trụ sở tại Mỹ. Được thành lập vào tháng 1 năm 2014, khi đi vào hoạt động, Poloniex định hướng cung cấp một môi trường thương mại an toàn, đồng thời còn cung cấp các dữ liệu về thị trường như biểu đồ, bảng xếp hạng và các công cụ phân tích dữ liệu để hỗ trợ khách hàng.

Ngoài ra, Poloniex còn cung cấp một số lượng dữ liệu liên quan đến các thống kê mua/bán của sàn giao dịch, các dữ liệu này được cung cấp thông qua API.

4.1.2 Mẫu dữ liệu

Nguồn dữ liệu được lấy thông qua API dữ liệu biểu đồ thị trường của sàn giao dịch Poloniex. Cụ thể API:

```
https://poloniex.com/public?command=returnChartData&currencyPair=
    BTC_XMR&start=1405699200&end=9999999999&period=14400
```

Tập dữ liệu về các phiên giao dịch Bitcoin được thu thập từ ngày 20/2/2015 đến ngày 29/10/2016 và có tổng cộng 29634 mẫu (mỗi mẫu là đại diện của một phiên giao dịch). Dữ liệu trả về là một mảng các phần tử JSON có dạng như sau:

```
[
    {"date":1424372400,"high":225,"low":225,"open":225,"close":225,"
        volume":0.999999,"quoteVolume":0.00444444,"weightedAverage
        ":225},
    {"date":1424374200,"high":225,"low":225,"open":225,"close":225,"
        volume":0,"quoteVolume":0,"weightedAverage":225},
```

```
    ...,
    {"date":1482575400,"high":900.79862142,"low":894.98864451,"open
        ":900.79862142,"close":895.56277339,"volume":2427.10998126,"
        quoteVolume":2.70065724,"weightedAverage":898.71085649}
]
```


Sau khi thu thập, dữ liệu được tiền xử lý để loại bỏ các thông tin không được sử dụng trong quá trình phân tích và xây dựng giải thuật. Giá trị có khóa là *close* là dữ liệu sẽ được sử dụng, các giá trị như: *date*, *high*, *low*, *open*, *volume*, *quoteVolume*, *weightedAverage* sẽ được lược bỏ.

4.2 Xây dựng MNN

4.2.1 Xây dựng dữ liệu luyện tập

Một trong những yếu tố hết sức quan trọng trong Học máy đó chính đặc trưng - feature. Đặc trưng chính là các giá trị thuộc tính đại diện cho tập dữ liệu luyện tập, ví dụ chúng ta có tập dữ liệu về loài chim thì có thể nói đặc trưng chính là các thông số như: độ dài sải cánh, màu lông, vùng sinh sống... Một giải thuật có thể học được "kinh nghiệm" nhanh hay chậm, chính xác hay sai lệch phụ thuộc rất nhiều vào yếu tố đặc trưng. Vì vậy quá trình khai phá dữ liệu chính là đi tìm kiếm các đặc trưng có ý nghĩa cho giải thuật học máy và là công việc hết sức quan trọng.

Gọi S là đại diện cho một phiên giao dịch, các đặc trưng được xây dựng như sau:

- 10 feature RDP: $\{\, loop\{RDP_1(S_{i+j})\}_i \,\}_j$ Với $i \in [0:9]$, $j \in [0:29634]$
- 1 feature SO. Với $j \in [0:29625]$:

$$\{\%K_j = \frac{P(j+9) - L_{10}}{H_{10} - L_{10}}\}_j$$

- 1 feature ROC. Với $j \in [0:29625]$:

$$\{ROC_{10}(j) = \frac{P(j+9) - P(j)}{P(j)}\}_j$$

Ở đây, chúng ta chọn mỗi vector đặc trưng được hình thành bởi 10 phiên giao dịch. Các giá trị SO và ROC đều được tính trong thời gian là 10 phiên giao dịch. Sau khi đã có tập luyện tập (tập hợp các vector đặc trưng), ta cần nhãn - label để phân lớp tập luyện tập. Nhãn được định nghĩa như sau, nếu giá BTC ở phiên thứ 11 lớn hơn phiên thứ 10 thì nhãn sẽ là 1, ngược lại sẽ là 0 (Phiên 11 chính là phiên thứ 1 của nhóm 10 phiên liền sau nhóm 10 phiên

hiện đang xét).

$$label_i = \begin{cases} 1 & if\ P_i(10) > P_{i+1}(1) \\ \\ 0 & if\ P_i(10) \leq P_{i+1}(1) \end{cases}$$

Kết quả dữ liệu luyện tập:

```
0 0.06666666666666667 0.016666666666666666 0 0 0 0 0 0 0
   0.08444444444444445 1 0
0.06666666666666667 0.016666666666666666 0 0 0 0 0 0 0 0
   0.016666666666666666 1 0
...
0.002400968290280231 3.9565217470358025e-9 0.003910087158215986
   -0.00045140752064857115 -0.0013329209110243738
   0.0008262923440509297 0.018429551716218056 0.003236901952232515
   -0.004644611078085971 -0.0016267310238298762
   0.018318840535169866 0.7405268424216611 0
```

Mỗi hàng đại diện cho một vector đặc trưng và nhãn, chi tiết ở một vector đặc trưng 10 giá trị đầu sẽ là 10 đặc trưng RDP, 1 giá trị tiếp theo là đặc trưng ROC, 1 giá trị tiếp theo là đặc trưng SO và 1 giá trị cuối cùng là nhãn.

4.2.2 Học giải thuật

Bên cạnh chạy giải thuật MNN, chúng ta sẽ chạy các giải thuật khác nhằm so sánh và đánh giá giải thuật chính. Các giải thuật được chọn để so sánh với giải thuật chính: SVM, KNN, LR.

Các giải thuật được chạy bằng thư viện scikit-learn phiên bản 0.18.0 với các tham số được điều chỉnh như sau:

Giải thuật	Lớp giải thuật	Tham số điều chỉnh
KNN	neighbors.KNeighborsClassifier	n_neighbors = 101, weights = "uniform"
LR	linear_model.LogisticRegression	C = 1
SVM	svm.SVC	kernel = "linear", C = 1, probability = True
MNN	neural_network.MLPClassifier	solver = "adam", hidden_layer_sizes = (100), activation = "logistic", learning_rate_init = 0.001

Bảng 4.1: Bảng giải thuật

Tập dữ liệu đưa vào được thay đổi thang độ, tất cả các giá trị được nhân với 1000 nhằm giúp cho các giá trị nằm trong giới hạn kiểu biến Number của Python, việc làm này giúp cho quá trình tính toán của máy tính được nhanh hơn và giảm thiểu thời gian chạy giải thuật.

4.2.3 Đánh giá giải thuật

Để đánh giá mức độ ý nghĩa của từng giải thuật, chúng ta sẽ sử dụng 3 tham số đánh giá là Accuracy, Recall, Precision. Để có được kết quả đánh giá trên lý thuyết, tập dữ liệu sẽ được chia ra thành hai phần:

- Tập huấn luyện: chiếm 7/10 tổng số dữ liệu, dùng để chạy trong quá trình học của giải thuật.
- Tập đánh giá: chiếm 3/10 tổng số dữ liệu, dùng để chạy trong quá trình đánh giá giải thuật.

Mô hình đánh giá giải thuật được mô tả bằng cách, dùng tập huấn luyện để tạo ra mô hình học máy của các giải thuật, sau đó sử dụng tập đánh giá để làm đầu vào cho từng mô hình, kết quả dự đoán sẽ được so sánh với kết quả thực tế của từng vector đặc trưng đầu vào. Kết quả so sánh được ghi nhận lại.

Kết quả chạy giải thuật được xuất ra dưới dạng ma trận nhầm lẫn - confusion matrix, sau đó, sử dụng kết quả của ma trận nhầm lẫn để tính toán giá trị của các tham số đánh giá. Cụ thể, kết quả của 4 giải thuật được trình bày như sau:

	KNN	LR	SVM	MNN
Accuracy	62.93%	66.24%	66.40%	69.86%
Precision	44.69%	18.18%	0%	60.50%
Recall	43.62%	0.15%	0%	29.55%

Bảng 4.2: Bảng đánh giá

Trước tiên theo bảng đánh giá, ta có các giải thuật LR và SVM cho kết quả Accuracy là gần khoảng 66%, nhưng khi nhìn vào chi tiết các giá trị Precision và Recall ta nhận thấy kết quả điều cho ra rất thấp. Kết quả này cho thấy giải thuật LR và SVM đều có số lần True Positive là xấp xỉ bằng 0, đồng nghĩa với việc các giải thuật này hầu như chỉ dự đoán kết quả là nhãn *Down* cho tất cả trường hợp. Điều này hoàn toàn không có ý nghĩa trong dự đoán đầu tư.

Xét đến KNN và MNN, đối với KNN ta có thể thấy giải thuật có xu hướng cân bằng các giá trị Accuracy, Precision và Recall. Nhưng đối với MNN, giải thuật có xu hướng tối ưu hóa bộ thiêu chuẩn Accuracy và Precision. Vậy câu

hỏi đặt ra ở đây là kết quả nào có giá trị đầu tư hơn?

Chú ý đến Recall, dựa theo định nghĩa thì Recall có thể hiểu nếu trong thực tế có 10 phiên là Up thì KNN sẽ dự đoán đúng khoảng 4 lần và MNN sẽ dự đoán đúng khoảng 3 lần. Điều này có nghĩa là KNN sẽ chiếm ưu thế so MNN khi sử dụng tiêu chuẩn là Recall.

Xét đến Precision, ta có thể hiểu Precision như sau, với 10 lần dự đoán sẽ có phiên Up thì KNN sẽ đúng khoảng 4 lần và MNN sẽ dự đoán đúng 6 lần. Giả sử, mức độ tin tưởng của chúng ta vào hệ thống là 100%, cứ mỗi lần hệ thống dự đoán có phiên Up thì ta sẽ quyết định đầu tư. Điều đó đồng nghĩa, nếu theo KNN sẽ có 6 lần ta chịu lỗ vì hệ thống dự đoán sai và với MNN thì ta sẽ có 4 lần ta chịu lỗ.

Quay lại với Recall, giá trị này không đo đạt được việc chúng ta sẽ lợi nhuận hoặc thua lỗ ra sao mà thực ra là giá trị đo đạt khả năng tận dụng cơ hội của hệ thống.

Tới lúc này, ta có thể kết luận, bộ tiêu chuẩn chiếm ưu thế cao hơn sẽ là Accuracy và Precision. Điều đó cũng có nghĩa là giải thuật MNN cho kết quả ý nghĩa hơn so với KNN và sẽ được lựa chọn để giải quyết bài toán dự đoán xu hướng giá trị BTC.

4.3 Xây dựng hệ thống

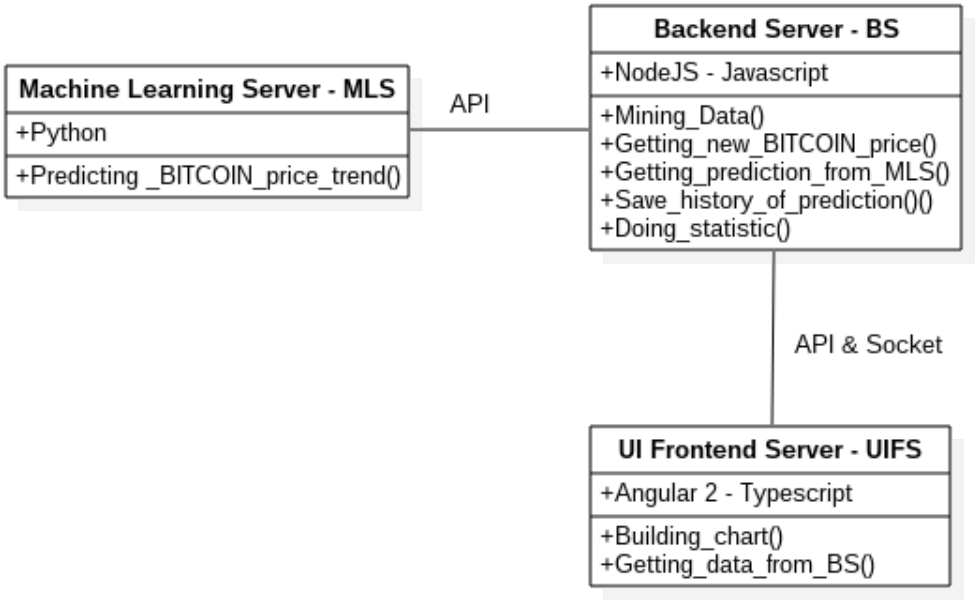
Hệ thống được xây dựng với mục tiêu là một ứng dụng nền web, cung cấp công cụ hỗ trợ cho việc dự đoán xu hướng giá trị BTC.

4.3.1 Tổng quan hệ thống

Hệ thống được xem xét và được thiết kế với 3 khối máy chủ , mỗi khối máy chủ đại diện cho một khối chức năng riêng biệt. Cấu trúc hệ thống như vậy nhằm dự trù và đảm bảo cho các hoạt động về quản lý hệ thống như phân phối tải, bảo trì và mở rộng sau này được thực hiện dễ dàng và tiết kiệm thời gian. Chi tiết các hệ thống máy chủ:

- Hệ thống máy chủ học máy - Machine Learning server
- Hệ thống máy chủ backend - Backend server
- Hệ thống máy chủ UI frontend - UI Frontend server

Các khối hệ thống giao tiếp với nhau bằng API và Socket - đối với các chức năng chạy thời gian thực.



Hình 4.1: Cấu trúc quan hệ và chức năng hệ thống

4.3.2 Hệ thống máy chủ học máy

Đây là hệ thống cốt lõi của của sản phẩm, nó đảm nhiệm khối chức năng chính liên quan đến các tác vụ học máy, cụ thể, dựa vào các tham số được truyền vào để thực hiện quá trình chạy giải thuật dự đoán từ đó đưa ra giá trị nhãn dự đoán tương ứng cho bộ tham số đó.

Lưu ý, để đưa ra một kết quả dự đoán, hệ thống yêu cầu các tham số đầu vào phải được xây dựng theo mô tả tại mục 4.1.1 với tổng số đặc trưng là 12.

Hệ thống máy chủ học máy được chia thành hai phần:

1. Prediction: bao gồm các chức năng đọc mô hình MNN đã xây dựng, chạy mô hình với tham số truyền vào và lấy các kết quả đầu ra.Kết quả đầu ra có giá trị nhãn $Up - Down$ và xác suất dự đoán.

2. Django: bao gồm các chức năng để trở thành một máy chủ giao tiếp thông qua API, các chức năng có thể ví dụ như tiếp nhận các yêu cầu thông qua API, phản hồi các yêu cầu dưới dạng các kết quả JSON...

Vì tính chất hỗ trợ tốt cho Học máy nên Python được lựa chọn là ngôn ngữ để phát triển hệ thống này.

4.3.3 Hệ thống máy chủ backend

Vì bản thân hệ thống máy chủ học máy không có các khối chức năng liên quan đến việc lấy dữ liệu giá BTC cũng như khai phá dữ liệu, nên hệ thống máy chủ backend được xây dựng để thực hiện các chức năng này. Đồng thời, máy chủ backend còn là cầu nối giữa trải nghiệm người dùng (hệ thống máy chủ UI frontend) và hệ thống máy chủ học máy.

Để thực hiện được công việc trên, hệ thống bao gồm được xây dựng các chức năng:

1. Cập nhật giá BTC: thông qua các public API được sàn giao dịch Poloniex cung cấp, các hàm lấy giá được chạy liên tục để cập nhật giá BTC mới nhất nhằm phục vụ cho quá trình dự đoán (trung bình 20 giây).

2. Khai phá dữ liệu: dữ liệu được các hàm cập nhật giá BTC lấy được vẫn còn ở dạng thô, chưa qua xử lý. Khai phá dữ liệu là biến đổi các dữ liệu này về các bộ tham số có ý nghĩa với Học máy, các giá trị này mới đích thực dùng để làm đầu vào dự đoán xu hướng giá trị BTC.

3. Giao tiếp với hệ thống máy chủ học máy: truyền tham số đi và nhận kết quả trả về từ hệ thống máy chủ học máy thông qua API.

4. Lưu trữ và thống kê dữ liệu: thực hiện việc lưu trữ dữ liệu, từ đó tạo nên một hệ thống các dữ liệu phục vụ cho việc phân tích, thống kê để cung cấp cho người dùng đầu cuối. Đó là các thông tin hết sức quý giá phục vụ cho các nhà đầu tư.

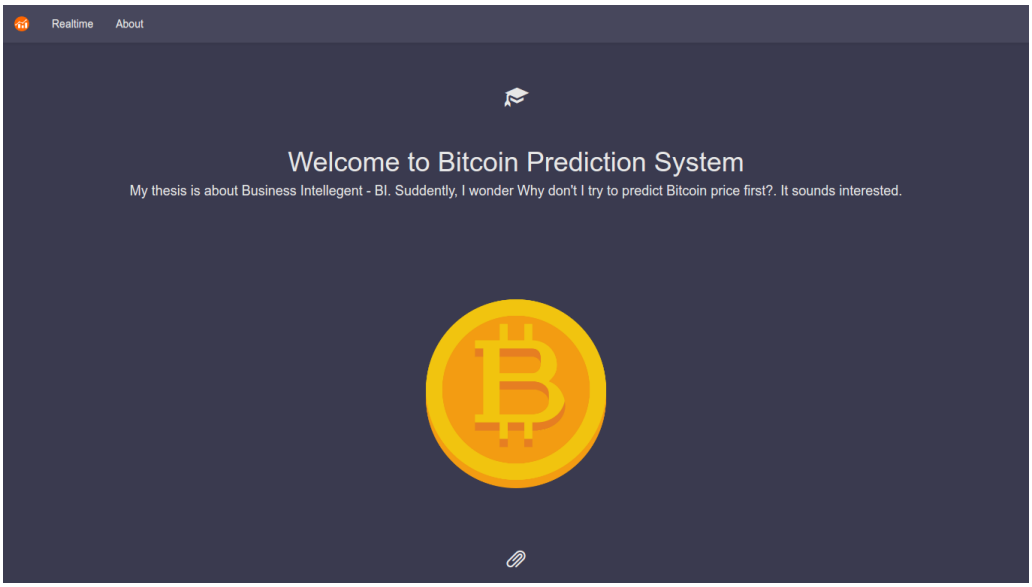
5. Giao tiếp với hệ thống máy chủ UI frontend: đưa ra những API chức năng nhằm phục vụ cho máy chủ UI frontend. Ví dụ như: yêu cầu dữ liệu dự đoán, yêu cầu thống kê đúng/sai, yêu cầu dữ liệu giá cho biểu đồ...

Với khả năng xử lý nhanh, được hỗ trợ tốt nên NodeJS được dùng để phát triển hệ thống. Đồng thời, cơ sở dữ liệu của hệ thống là MongoDB vì tính linh hoạt trong cấu trúc dữ liệu và khả năng mở rộng cao.

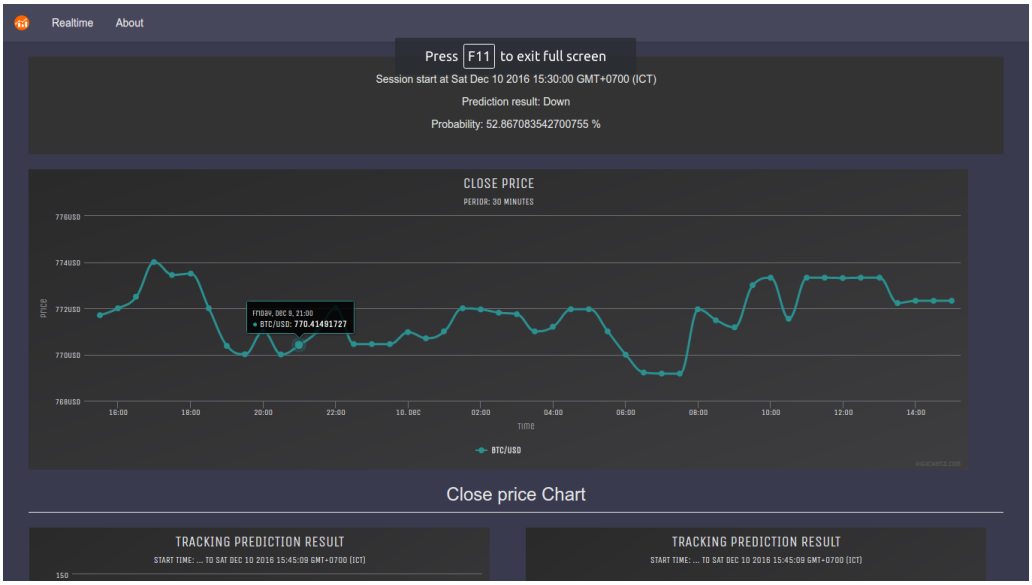
4.3.4 Hệ thống máy chủ UI frontend

Hệ thống máy chủ UI frontend là một giao diện người dùng, nó cho phép người dùng có thể tiếp cận với các chức năng của toàn bộ hệ thống một cách dễ dàng. Hệ thống bao gồm nhiều biểu đồ, cũng như tham số cung cấp các thông tin có ý nghĩa đầu tư - dự đoán xu hướng giá trị Bitcoin - đồng thời với đó, là các thông tin về độ tin cậy của hệ thống, các thống kê về lịch sử dự đoán...

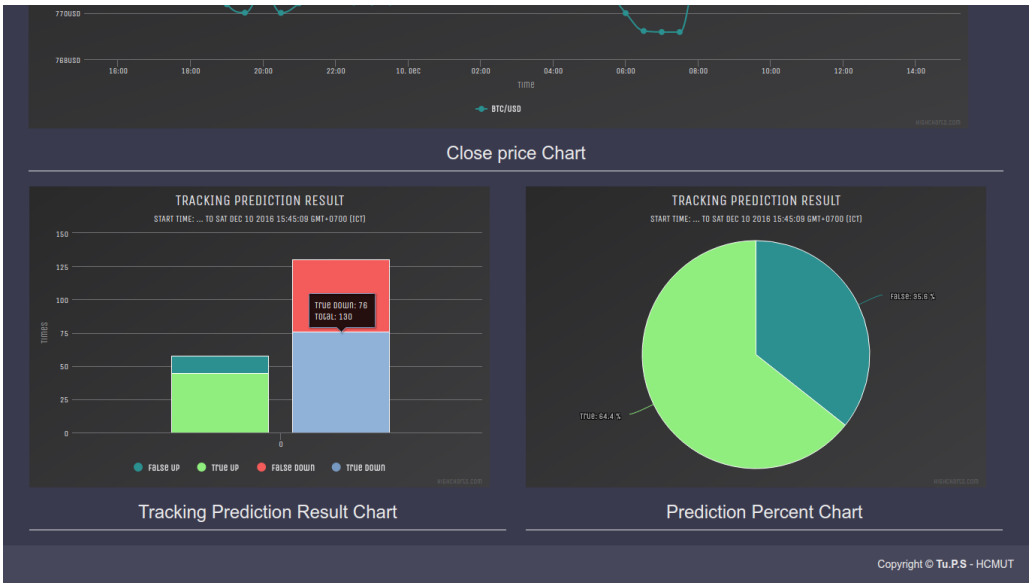
Một số hình ảnh về hệ thống thực tế.



Hình 4.2: Giao diện 1 máy chủ UI frontend



Hình 4.3: Giao diện 2 máy chủ UI frontend



Hình 4.4: Giao diện 3 máy chủ UI frontend

Hệ thống máy chủ UI frontend được xây dựng theo xu hướng one-page, cũng chính vì vậy mà Angular 2 là lựa chọn phù hợp, với khả năng phát triển nhanh, hỗ trợ tốt từ các bên thứ 3.

Chương 5

Kết luận và hướng phát triển

5.1 Kết luận

Kết thúc đề tài, sản phẩm cuối cùng được hoàn thiện là một công cụ nền Web hỗ trợ, cung cấp các thông tin có giá trị tham khảo để đầu tư Bitcoin. Dựa trên các con số lý thuyết, khả năng dự đoán chính xác là rất khả quan và đặc biệt, giải thuật được tối ưu cho phù hợp với góc nhìn của một người đầu tư.

Với không nhiều sai lệch khi so sánh bên cạnh các con số lý thuyết, khi hệ thống được cho chạy thực tế trong vòng 4 ngày liên tiếp (Cụ thể từ 22:30:00 13/11/2016 đến 20:30:00 17/11/2016) đã cho ra kết quả:

Accuracy	Precision	Recall
64.4%	77.6%	45.5%

Bảng 5.1: Bảng đánh giá hệ thống thực tế

Các tham số đánh giá chạy thực tế như vậy, có thể thấy với một lần đầu tư ta có tới hơn 70% là có lợi nhuận. Tuy vậy, bất kỳ một hệ thống cũng vẫn sẽ có những điểm thiếu sót.

Vì giới hạn của thời gian thực hiện đề tài, phạm vi của đề tài cũng được thu hẹp để phù hợp nên vì thế đã bỏ qua một số yếu tố thị trường ảnh hưởng khá lớn đối với hướng giải quyết. Trong lúc này, bản thân có thể nhận ra hai vấn đề:

- Phí giao dịch: ở tất cả các sàn giao dịch, đều có một khoảng phí trung gian từ 0.1% đến 0.3% và phí này được trừ trực tiếp vào các giao dịch. Hướng tiếp cận của đề tài bỏ qua hoàn toàn yếu tố này và có thể hiểu là phí bằng 0%

- Biên độ lợi nhuận và thua lỗ: chúng ta cũng đã bỏ qua yếu tố này, mặc dù dựa theo đánh giá thì số lần đầu tư lợi nhuận sẽ nhiều hơn thua lỗ. Nhưng, chúng ta không thể kết luận việc đầu tư sẽ chắc chắn đem về lợi nhuận. Hãy nói đến một trường hợp xấu, biên độ lợi nhuận chỉ có \$1 cho mỗi lần nhưng biên độ thua lô lại là \$100, tại đây chúng ta có thể thấy là việc đầu tư không hề có lợi.

Việc nhìn nhận được các vấn đề trên không hẳn là điều tồi tệ, mà ngược lại giúp chúng ta có thể hiểu rõ bài toán và đưa ra những hướng phát triển tiếp theo.

5.2 Hướng phát triển

Với các vấn đề còn tồn tại được nêu ra bên trên (Mục 5.1), giai đoạn tiếp theo của đề tài là đi giải quyết vẫn đề tài như hiện giờ nhưng thêm vào đó là yếu tố phí giao dịch. Tuy là một yếu tố nhỏ nhưng nó dẫn đến việc thay đổi hoàn toàn bộ dữ liệu ban đầu, điều này đồng nghĩa toàn bộ hệ thống hiện giờ sẽ không tương thích. Vì thế, cần thực hiện lại quá trình xây dựng giải thuật từ đầu.

Mặc khác, việc chỉ học duy nhất từ tập dữ liệu về giá BTC là không đủ để đưa ra một dự đoán chính xác cao. Ngày nay, mạng xã hội đang phát triển như vũ bão, đây là một kênh thông tin cực kỳ quý giá, chính vì vậy mà hệ thống ở giai đoạn phát triển tiếp theo sự tận dụng tài nguyên này.

Phát triển hệ thống xử lý ngôn ngữ tự nhiên, xây dựng hệ thống lắng nghe các thông tin tài chính, chính trị có ảnh hưởng tới giá trị BTC, phân tích, đánh giá và cho cân bằng với hệ thống học từ dữ liệu giá BTC để cho ra một dự đoán tổng quát và chính xác hơn.

Đồng thời, hệ thống có thể mở rộng ra cho nhiều loại tiền mã hóa khác như: Ethereum, Zcash, Monero...

Tài liệu tham khảo

[1] Hough, Jack (3/6/2011). "The Currency That's Up 200,000%". SmartMoney (Dow Jones & Company). Truy cập 24/12/2016.

[2] Wallace, Benjamin (23/11/2011). "The Rise and Fall of Bitcoin". Wired. Truy cập 24/12/2016.

[3] Marc Kenigsberg (26/10/2013). "BTC vs mBTC vs uBTC". Bitcoinchaser. Truy cập 24/12/2016.

[4] Đồng Bitcoin (04/2016). "Liên minh châu âu đã công nhận Bitcoin". DongBitcoin. Truy cập 25/12/2016.

[5] Satoshi Nakamoto (2008). Bitcoin: A Peer-to-Peer Electronic Cash System.

[6] I. Madan, S. Saluja và A. Zhao (2014). Automated Bitcoin Trading via Machine Learning Algorithms.

[7] Sean McNally (22/08/2016). *Predicting the price of Bitcoin using Machine Learning*. MSc Reseach Project, National College of Ireland.

[8] Megan Potoski (2013). Predicting Gold Prices.

[9] Yuqing Dai & Yuning Zhang (2013). Machine Learning in Stock Price Trend Forecasting.

[10] Michael Nielsen (1/2016). *Neural Networks and Deep Learning*. [Online]. Nguồn: http://neuralnetworksanddeeplearning.com/