

TRƯỜNG ĐẠI HỌC BÁCH KHOA

KHOA KHOA HỌC & KỸ THUẬT MÁY TÍNH



ĐỀ CƯƠNG LUẬN VĂN TỐT NGHIỆP

---

Phát triển công cụ dự đoán xu hướng  
giá ngắn hạn các đồng tiền mật mã  
bằng kỹ thuật học máy

---

*SVTH:*

Vũ Quang Nam

*GVHD:*

Nguyễn An Khương

Nguyễn Lê Thành

Ngày 23 tháng 12 năm 2018

# Mục lục

|  |          |
|--|----------|
| <b>Danh mục hình vẽ</b>  | <b>5</b> |
| <b>1 Giới thiệu</b>  | <b>1</b> |
| 1.1 Giới thiệu đề tài . . . . .                                  | 1        |
| 1.2 Mục tiêu và phạm vi đề tài . . . . .                         | 1        |
| 1.2.1 Mục tiêu . . . . .   | 1        |
| 1.2.2 Phạm vi đề tài . . . . .                                   | 1        |
| 1.3 Tiến độ thực hiện . . . . .                                  | 2        |
| <b>2 Tổng quan về lĩnh vực nghiên cứu</b>                        | <b>3</b> |
| 2.1 Những yếu tố tác động đến giá trị đồng tiền mã hóa . . . . . | 3        |
| 2.1.1 Cung và cầu của thị trường . . . . .                       | 3        |
| 2.1.2 Tin tức trên các phương tiện thông tin đại chúng . . . . . | 3        |
| 2.1.3 Quy định của chính phủ . . . . .                           | 3        |
| 2.1.4 Chính sách của các tổ chức . . . . .                       | 4        |
| 2.1.5 Các vấn đề kỹ thuật . . . . .                              | 4        |
| 2.2 Nhu cầu sử dụng tiền mã hoá của mỗi hệ sinh thái . . . . .   | 4        |
| <b>3 Dữ liệu</b>   | <b>5</b> |
| 3.1 Chuẩn bị dữ liệu . . . . .                                   | 5        |
| 3.2 Mô tả dữ liệu . . . . .                                      | 6        |
| <b>4 Cơ sở lý thuyết</b>   | <b>7</b> |
| 4.1 Cây hồi quy và phân loại . . . . .                           | 7        |
| 4.1.1 Cấu trúc cây nhị phân cơ bản . . . . .                     | 7        |
| 4.1.2 Các luật tách thường dùng . . . . .                        | 7        |
| 4.1.3 Tiêu chí tách . . . . .                                    | 8        |
| 4.1.4 Tỉa cây . . . . .  | 9        |
| 4.2 Rừng ngẫu nhiên . . . . .                                    | 9        |

|          |  |           |
|----------|--|-----------|
| 4.3      | Lớp tích chập trong mô hình mạng nơ-ron tích chập (Convolution neural network) . . . . . | 10        |
| <b>5</b> | <b>Thí nghiệm với mô hình</b>  | <b>11</b> |
| 5.1      | Phân tích hiện thực mô hình tham khảo . . . . .  | 11        |
| 5.1.1    | Xử lý dữ liệu bài toán . . . . .   | 11        |
| 5.1.2    | Mô hình rừng ngẫu nhiên . . . . .  | 11        |
| 5.1.3    | Mô hình mạng nơ-ron tích chập . . . . .  | 11        |
| 5.2      | Kết quả . . . . .  | 12        |

## Danh mục hình ảnh



# Chương 1

## Giới thiệu

### 1.1 Giới thiệu đề tài

(Logistic Regression), cây hồi quy và phân loại (CART), rừng ngẫu nhiên (Random Forest), mạng nơ-ron (Neural NetWork), máy vectơ hỗ trợ (SVM). Nhưng ở Việt Nam lại chưa có nhiều nghiên cứu về đề tài này. Vậy nên tôi quyết định chọn đề tài **Dự đoán xu hướng giá ngắn hạn các đồng tiền mật mã bằng kĩ thuật học máy**.

### 1.2 Mục tiêu và phạm vi đề tài

#### 1.2.1 Mục tiêu

Mục tiêu của luận văn này là xây dựng một công cụ dự đoán xu hướng giá ngắn hạn các đồng tiền mật mã bằng kĩ thuật học máy. Dữ liệu đầu vào là các thông tin về lịch sử giá các đồng tiền ảo trong các phiên giao dịch.

#### 1.2.2 Phạm vi đề tài

- Tìm hiểu và nghiên cứu về lý thuyết học máy thống kê (statistical machine learning)
- Xây dựng mô hình dự đoán về xu hướng tăng giảm, dự đoán giá của các đồng trong thời gian ngắn hạn.

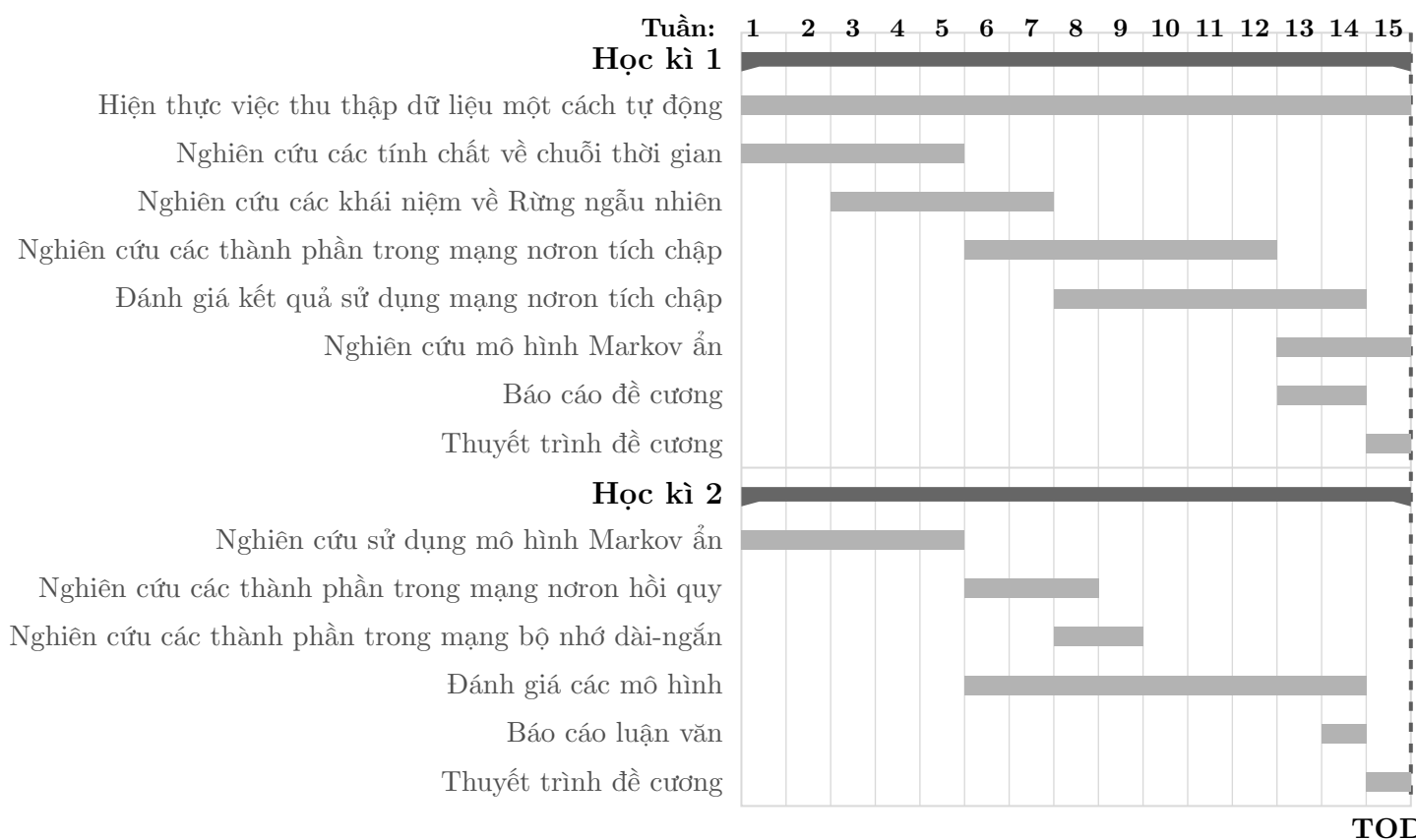
Các đối tượng nghiên cứu trong đề tài:

- Các tài liệu liên quan tới lý thuyết thống kê hiện đại
- Các mô hình trong học máy: hồi quy logistic, rừng ngẫu nhiên, mạng nơ-ron

- Sử dụng ngôn ngữ Python, R và một số thư viện để hiện thực mô hình.

### 1.3 Tiến độ thực hiện

Trong phần này, tác giả xin trình bày lịch trình công việc đã thực hiện đề tài trong học kỳ I và lịch trình dự kiến hiện thực đề tài trong quá trình làm luận văn chính thức ở học kỳ II dưới dạng biểu đồ Gantt sau đây.



TOD



## Chương 2

# Tổng quan về lĩnh vực nghiên cứu

## 2.1 Những yếu tố tác động đến giá trị đồng tiền mã hóa

### 2.1.1 Cung và cầu của thị trường

Trong nguyên tắc chính của kinh tế nếu người ta mua một đồng tiền, giá trị của đồng tiền sẽ tăng lên và nếu người ta bán đồng tiền, giá sẽ giảm.

### 2.1.2 Tin tức trên các phương tiện thông tin đại chúng

Các sự kiện chính trị và kinh tế trên toàn thế giới ảnh hưởng đến cách mà con người phản ứng với các dự đoán giá, tin tức cảnh báo về rủi ro tác động chính lên cung-cầu.

### 2.1.3 Quy định của chính phủ

Có 4 cấp độ quản lý tiền ảo hiện nay đang được các nước thực thi, cụ thể:

- Cấm trên diện rộng.
- Cấm trong lĩnh vực tài chính ngân hàng (trong đó có Trung Quốc).
- Cảnh báo rủi ro đối với người sử dụng, đầu tư,.
- Chấp nhận như một phương tiện thanh toán (trong đó có Hàn Quốc, Nhật Bản và Mỹ).

cập nhật ngày 14/4/2018.

### 2.1.4 Chính sách của các tổ chức

Facebook, Google và Twitter đã ngăn chặn khách hàng và người dùng sử dụng dịch vụ cryptocurrency.

### 2.1.5 Các vấn đề kỹ thuật

Vì đồng tiền mã hóa có thể bị hack thành công vào tài khoản hoặc tấn công máy chủ, có thể làm giảm tỷ giá hối đoái, dẫn đến giá giảm.

## 2.2 Nhu cầu sử dụng tiền mã hoá của mỗi hệ sinh thái

- Số thành viên tham gia vào hệ sinh thái (Số người đến khu vui chơi mua vé tham gia các trò chơi trong đó bằng tiền A).
- Số lượng dịch vụ trong hệ sinh thái (Khu vui chơi có càng nhiều trò chơi thì nhu cầu sử dụng tiền A càng tăng); Và các nền tảng như Ethereum luôn mở cho các đối tác tạo các dịch vụ gia tăng trên đó giống như khu vui chơi cho phép đối tác bên ngoài vào tổ chức trò chơi ở trong.
- Số người đầu cơ: Những người nhận thấy nhu cầu tiền mã hoá của một hệ sinh thái tăng dần sẽ mua để nắm giữ chờ tăng giá thì bán ra. (Giống như phe vé bóng đá ngày trước mua vé chờ sát trận nhu cầu tăng vọt thì bán ra. Khu vui chơi thì ít có nhóm này vì lượng vé không bị giới hạn).
- Số người bán bên ngoài chấp nhận tiền mã hoá: Một số người bán nhận thấy tính thanh khoản của tiền mã hoá và giá trị tăng dần của nó nên đã chấp nhận khách hàng thanh toán các hàng hoá dịch vụ của mình bằng loại tiền này (Nhà hàng bên cạnh khu vui chơi có thể chấp nhận khách hàng thanh toán bằng tiền A).

## Chương 3

# Dữ liệu

### 3.1 Chuẩn bị dữ liệu

Có nhiều nguồn cung cấp dữ liệu cho bài toán dự đoán giá đồng tiền mã hóa, nghiên cứu này có sử dụng các lịch sử giao dịch các đồng với nhau (trading pair) được lấy từ API có sẵn từ 8 sàn giao dịch với cấu trúc bảng như sau:

| symbol    | market    | timeIndicator           | minTimestamp  | maxTimestamp  |
|-----------|-----------|-------------------------|---------------|---------------|
| SRN/BTC   | huobipro  | 2018-12-13 08:26:00 UTC | 1544689603614 | 1544689615119 |
| WTC/BTC   | huobipro  | 2018-12-13 08:26:00 UTC | 1544689570921 | 1544689590036 |
| EOS/PAX   | binance   | 2018-12-13 08:26:00 UTC | 1544689566328 | 1544689618905 |
| EKT/BTC   | huobipro  | 2018-12-13 08:26:00 UTC | 1544689562604 | 1544689611453 |
| NEO/USDT  | huobipro  | 2018-12-13 08:26:00 UTC | 1544689561044 | 1544689618173 |
| OMG/BTC   | bitfinex2 | 2018-12-13 08:26:00 UTC | 1544689588624 | 1544689588624 |
| XRP/BTC   | binance   | 2018-12-13 08:26:00 UTC | 1544689568501 | 1544689619519 |
| BTG/BTC   | binance   | 2018-12-13 08:26:00 UTC | 1544689577770 | 1544689577770 |
| HB10/USDT | huobipro  | 2018-12-13 08:26:00 UTC | 1544689569087 | 1544689617665 |
| ICX/BTC   | huobipro  | 2018-12-13 08:26:00 UTC | 1544689563889 | 1544689614203 |

| openPrice  | closePrice | highPrice  | lowPrice   | volume                |
|------------|------------|------------|------------|-----------------------|
| 1.331e-05  | 1.325e-05  | 1.331e-05  | 1.331e-05  | 0.0085977935000000009 |
| 0.00026948 | 0.00027049 | 0.00027049 | 0.00027049 | 5.3997e-06            |
| 1.9152     | 1.9152     | 1.9207     | 1.9207     | 204.88715399999998    |
| 1.24e-06   | 1.25e-06   | 1.25e-06   | 1.25e-06   | 0.0109388594          |
| 5.86       | 5.85       | 5.87       | 5.87       | 2191.216003           |
| 0.00036052 | 0.00036052 | 0.00036052 | 0.00036052 | 0.026111466704516802  |
| 8.898e-05  | 8.896e-05  | 8.9e-05    | 8.9e-05    | 1.1474145999999998    |
| 0.003396   | 0.003396   | 0.003396   | 0.003396   | 0.030564              |
| 0.2421     | 0.2419     | 0.2421     | 0.2421     | 56.687566             |
| 6.607e-05  | 6.625e-05  | 6.625e-05  | 6.625e-05  | 0.005767279581        |

## 3.2 Mô tả dữ liệu

Với phiên giao dịch dòng 1 SRN/BTC được ghi lại thành một dòng với thông tin như sau:

- symbol: Tên giao dịch giữa hai đồng với nhau cụ thể là đồng SRN so với đồng BTC
- Market: Tên sàn giao dịch cụ thể là sàn huopipro
- timeIndicator: Thời điểm mở phiên giao dịch 8 giờ 13/12/2018 UTC
- openPrice: Tỷ giá thời điểm mở phiên
- openPrice: Tỷ giá thời điểm đóng phiên
- highPrice: Tỷ giá cao nhất phiên giao dịch
- lowPrice: Tỷ giá thấp nhất phiên giao dịch
- volume: Khối lượng giao dịch (Ví dụ symbol là SRN/BTC volume có nghĩa là số đồng SRN)
- minTimestamp, maxTimestamp: do mỗi giao dịch cần một thời gian nhất định nên cần có thời gian bắt đầu và kết thúc giao dịch tính theo POSIX time.

## Chương 4

# Cơ sở lý thuyết

### 4.1 Cây hồi quy và phân loại

Cây hồi quy và phân loại (CART) là một cây quyết định nhị phân được đề xuất bởi Breima [1].

#### 4.1.1 Cấu trúc cây nhị phân cơ bản

Ứng với một tập data ta cần tạo một cây nhị phân có đầu ra thành một chuỗi các lá, mục tiêu các lá có giá trị đầu ra tương đồng nhiều nhất. Khi bắt đầu từ nút cần chọn ra một thuộc tính và một giá trị sao cho giảm được "nhiều" nhiều nhất có thể. Ta có thể lựa chọn các độ đo khác nhau nhằm sinh ra cây nhị phân với ý tưởng này, với mỗi độ đo khác nhau tương ứng với một luật tách (splitting rule).

#### 4.1.2 Các luật tách thường dùng

Ta có thể chia thành 2 loại theo:

**Đối với Cây hồi quy**

- Least squares: phương pháp chọn tổng bình phương lỗi (SSE) nhỏ nhất giữa các quan sát với giá trị trung bình. Giá trị này tốt nhất khi đạt tới 0 nghĩa là tất cả các giá trị quan sát đều như nhau.

- Least absolute deviations: phương pháp chọn tổng trị tuyệt đối nhỏ nhất giữa các quan sát với giá trị trung bình, so với Least squares thì phương pháp này ít nhạy hơn đối với các dữ liệu ngoại lai (outlier).

### Đối với Cây phân loại

- Misclassification error: là tỉ lệ của các quan sát không cùng loại với loại chính.
- Gini index Entropy:
- Entropy index: hay cross-entropy

#### 4.1.3 Tiêu chí tách

CART sử dụng chỉ số Gini để làm tiêu chí tách với mô hình phân loại. Gọi  $RF(C_j, S)$  biểu diễn tần suất xuất hiện của lớp  $C_j$  trong các phần tử của tập  $S$ . Chỉ số Gini được xác định bằng công thức:

$$I_{gini}(S) = 1 - \sum_{j=1}^x RF(C_j, S)^2$$

Sau khi tập  $S$  được chia thành nhiều tập con  $S_1, S_2, \dots, S_t$ , bởi phép chia  $B$ , độ lợi thông tin  $G(S, B)$  được tính bằng công thức:

$$G(S, B) = I(S) - \sum_{i=1}^t \frac{|S_i|}{|S|} I(S_i)$$

Ta chọn phép chia  $B$  nào làm tối đa hóa độ lợi  $G(S, B)$ . Sau đó CART sẽ xây dựng các mô hình trên các tập  $S_i$ . Một cây phân loại sẽ dự đoán phân phối của một mẫu trên một lớp nhất định. Hiệu quả của mỗi cây phân loại sẽ được tính dựa trên sai số toàn phương trung bình. Với mỗi lớp  $j$ , gọi  $C_j(e)$  là chỉ báo có giá trị bằng 1 nếu mẫu  $e$  thuộc lớp  $j$  và bằng 0 nếu không. Sai số toàn phương trung bình  $MSE$  được tính bằng công thức:

$$MSE = E_e \left[ \sum_{j=1}^x (C_j(e) - P_j(e))^2 \right]$$

với kì vọng trên toàn bộ các mẫu,  $P_j(e)$  đại diện cho xác suất mẫu  $e$  thuộc lớp  $j$ . Đối với cây hồi quy, độ lệch  $R(S_i)$  là sai số toàn phương trung bình:

$$R(S) = \frac{1}{n} \sum_i (y_i - h(t_i))^2$$

với  $y_i$  là giá trị thực của biến mục tiêu trong mẫu  $t_i$  và  $h(t_i)$  là giá trị dự đoán của mô hình.

#### 4.1.4 Tỉa cây

Khi xây dựng cây bằng cách "vét cạn", tối ưu tất cả các mẫu trong tập huấn luyện, dẫn đến các node lá trong cây mang ít các quan sát. Điều này làm kết quả xấu khi thử ở tập kiểm tra mặc dù tập huấn luyện có kết quả tốt. Nếu một cây được xây dựng quá nhỏ tức độ sâu quá ngắn thì chưa trích xuất được hết thông tin.

Ta có thể tùy chỉnh kích thước của cây theo các cách sau đây:

- Không nhất thiết node lá hoàn toàn đồng nhất, ta nên dừng việc tách nhánh khi độ đồng nhất trên mức chấp nhận được.
- Một cách khác là "vét cạn" cây đến khi đạt đến node lá nhỏ nhất (thường chỉ có một quan sát). Xác định độ sâu thích hợp dựa trên tập kiểm tra độc lập với tập huấn luyện hoặc dùng cross-validation, sau đó tỉa các nhánh đưa cây về độ sâu đã chọn.

**Tập huấn luyện - kiểm tra độc lập** Khi tập mẫu đủ lớn, ta chia tập thành 2 phần riêng, độc lập với nhau.

- Tập huấn luyện: dùng để sinh cây có độ dài lớn đủ để có thể tỉa cây.
- Tập kiểm thử: từ cây đã sinh ở trên ngẫu nhiên tỉa các nhánh để tạo ra nhiều cây con, thử các quan sát ở tập kiểm thử trên những cây con này từ đó xác định được số lỗi nhỏ nhất theo bài toán regression hoặc classification.

**Cross-Validation** Nếu dữ liệu chưa đủ cho việc tách riêng biệt thành hai tập với tỉ lệ như trên, nói cách khác chúng ta cần giữ lại tập train càng nhiều càng tốt nhưng vẫn cần sự độc lập giữa hai tập này.

## 4.2 Rừng ngẫu nhiên

Tuy trực quan, các cây quyết định đơn lẻ thường có phương sai cao khi tập kiểm tra khác với tập huấn luyện dẫn đến kết quả dự đoán kém trên dữ liệu thực tế. Mục đích của rừng ngẫu nhiên nhằm giảm overfitting. Các bước thực hiện của rừng ngẫu nhiên gồm[1]:

- Bootstrap: Chọn lựa các dữ liệu ngẫu nhiên .
- Chọn lựa các thuộc tính: chọn các thuộc tính ngẫu nhiên đối với mỗi dữ liệu vừa tạo.

- Xây dựng cây.
- Tính toán tỉ lệ lỗi của mỗi cây dựa trên dữ liệu kiểm thử không nằm trong dữ liệu vừa tạo.
- Tổng hợp các cây.

### 4.3 Lớp tích chập trong mô hình mạng nơ-ron tích chập (Convolution neural network)

Mô hình mạng nơ-ron truyền thống với nhiều lớp truyền ngược có thể học được dữ liệu phức tạp, nhiều chiều. Tuy nhiên khi sử dụng nhiều lớp kết nối đầy đủ (fully connected) với đầu vào dạng véc-tơ với kích thước lớn gây ra hiện tượng overfitting do quá nhiều tham số (high-variance) làm bùng nổ mô hình liên quan trực tiếp tới vấn đề hiệu năng, bộ nhớ, khả năng tính toán của phần cứng. Để tránh hiện tượng này, ta có thể xây dựng bộ trích xuất đặc trưng trước khi đưa vào các lớp kết nối đầy đủ để giảm số lượng trọng số này. Một cách tổng quát hơn ta có thể xây dựng những lớp gồm các bộ lọc để giảm được số chiều của đầu vào trước đó, trong quá trình huấn luyện mô hình, lớp này có khả năng tự điều chỉnh các trọng số của các bộ lọc tại bước truyền ngược. Đây chính là lý do chính của lớp tích chập, kiến trúc của phép tích chập gồm các tham số:

- Kích thước của bộ lọc (cửa sổ)
- Hệ số lẻ (padding)
- Hệ số bước trượt (stride)



## Chương 5

# Thí nghiệm với mô hình

### 5.1 Phân tích hiện thực mô hình tham khảo

#### 5.1.1 Xử lý dữ liệu bài toán

Yêu cầu của bài toán là dự đoán giá của các đồng với nhau. Bài toán có thể chia ra làm 2 loại như sau:

- Phân loại giá tăng hoặc giảm trong một khoảng thời gian kế tiếp.
- Dự đoán giá của các đồng xác định trong phạm vi liên tục tại thời gian kế tiếp.

#### 5.1.2 Mô hình rừng ngẫu nhiên

Đề tài có sử dụng thư viện hỗ trợ scikit-learn để xây dựng rừng ngẫu nhiên. Rừng ngẫu nhiên có đầu vào là dữ liệu ở dạng bảng với mỗi thuộc tính(mỗi cột) có thể ở dạng loại (category) hoặc liên tục (numerical). Khi áp dụng dữ liệu thô, với mỗi quan sát là một dòng thuộc bảng và nhãn là giá lên hoặc xuống, đối với giá trong 1 phút sau ta có thể phân ra 2 loại tăng hoặc giảm. Tuy nhiên với mỗi quan sát trước có thể phụ thuộc vào nhiều quan sát trước đó, việc sử dụng các cây CART chỉ với một giao dịch là không phù hợp,ta có thể thêm các thuộc tính như giá trung bình trong 1 giờ trước đó, trong một ngày trước đó, ...

#### 5.1.3 Mô hình mạng nơ-ron tích chập

Đề tài có sử dụng thư viện hỗ trợ keras trên nền tensorflow một thư viện mã nguồn mở có hỗ trợ khả năng tính toán của các bộ xử lý đồ họa. Trong bước tiền xử lý dữ liệu để đưa vào trong mạng có sử dụng mỗi cửa sổ trượt làm một ảnh một chiều với số kênh là

số thuộc tính của mỗi giao dịch, độ dài của mỗi ảnh được định nghĩa trước là số giao dịch liên tục. Nhãn của những ảnh này là xu hướng tăng hoặc giảm của giao dịch cuối cùng với mỗi ảnh.

Khi dữ liệu được đưa vào mạng nơ-ron cần được chuẩn hóa để tránh hiện tượng các nơ-ron không cập nhật được khi hàm kích hoạt có họ RELU hoặc khó kích hoạt khi các hàm kích hoạt phi tuyến khác như hàm sigmoid hoặc hàm tanh. Các bước chuẩn hóa dữ liệu được mô tả như sau:

- Các thuộc tính về thời gian đổi về dạng số nguyên theo chuẩn UNIX. Các số này khá lớn nên được chuẩn hóa dạng logarit, sau đó chuẩn hóa theo standard score.
- Các thuộc tính còn lại như lượng giao dịch, giá mở, giá đóng theo dạng standard score

Khi áp dụng standard score, với mỗi thuộc tính có giá trị trung bình về 0 và phương sai về 1. Điều này ảnh hưởng tốt cho mạng nơ-ron tích chập khi hội tụ nhanh và khó bị ‘kẹt lại’ ở những điểm thung lũng hơn.

## 5.2 Kết quả

Các số liệu kết quả trong phần này được lấy từ các lần thực nghiệm huấn luyện trên google cloud, một nền tảng dịch vụ có hỗ trợ phần cứng và thư viện miễn phí. So sánh các mô hình rừng ngẫu nhiên và mạng nơ-ron tích chập với dữ liệu đầu vào từ 2012-01-01 đến 2018-11-11 với kết quả của rừng ngẫu nhiên là 53% và mạng nơ-ron tích chập là 71% trong bài toán phân loại.

# Bibliography

- [1] L. Breiman, J. Friedman, C. J. Stone, R. A. Olshen (1984). *Classification and regression trees*. CRC press.
- [2] Andrew Y. Ng, Michael I. Jordan *On Discriminative vs. Generative Classifiers: A comparison of logistic regression and naive Bayes*.
- [3] Yann LeCun, Yoshua Bengio *Convolutional Networks for Images, Speech, and Time-Series*.