TỔNG LIÊN ĐOÀN LAO ĐỘNG VIỆT NAM
TRƯỜNG ĐẠI HỌC TÔN ĐỨC THẮNG
KHOA CÔNG NGHỆ THÔNG TIN

ĐỒ ÁN CUỐI KÌ MÔN

PHÂN TÍCH XÁC SUẤT VÀ GIẢI THUẬT NGẪU NHIÊN

DỰ ĐOÁN CỔ PHIẾU VÀ GIÁ CỔ PHIẾU DỰA TRÊN DỮ LIỆU XU HƯỚNG VÀ KỸ THUẬT HỌC MÁY

Người hướng dẫn: **TS NGUYỄN CHÍ THIỆN**

Người thực hiện: **HỒNG QUANG VINH – 186005004**

NGUYỄN ĐẠI THỊNH – 186005035

Lớp**: 18600531**

Khoá**: 2018-2020**

THÀNH PHỐ HỒ CHÍ MINH, NĂM 2019

TỔNG LIÊN ĐOÀN LAO ĐỘNG VIỆT NAM
TRƯỜNG ĐẠI HỌC TÔN ĐỨC THẮNG
KHOA CÔNG NGHỆ THÔNG TIN

ĐỒ ÁN CUỐI KÌ MÔN

PHÂN TÍCH XÁC SUẤT VÀ GIẢI THUẬT NGẪU NHIÊN

DỰ ĐOÁN CỔ PHIẾU VÀ GIÁ CỔ PHIẾU DỰA TRÊN DỮ LIỆU XU HƯỚNG VÀ KỸ THUẬT HỌC MÁY

Người hướng dẫn: **TS NGUYỄN CHÍ THIỆN**

Người thực hiện: **HỒNG QUANG VINH – 186005004**

NGUYỄN ĐẠI THỊNH – 186005035

Lớp**: 18600531**

Khoá**: 2018-2020**

THÀNH PHỐ HỒ CHÍ MINH, NĂM 2019

LỜI CẢM ƠN

Nhóm chúng em xin chân thành cảm ơn Thầy Nguyễn Chí Thiện đã giúp đỡ chúng em hoàn thành đồ án. Những hướng dẫn của Thầy giúp chúng em có một nền tảng lý thuyết đủ để có thể ứng dụng và nghiên cứu phát triển đề tài này. Xin chân thành cảm ơn Thầy.

ĐỒ ÁN ĐƯỢC HOÀN THÀNH TẠI TRƯỜNG ĐẠI HỌC TÔN ĐỨC THẮNG

Tôi xin cam đoan đây là sản phẩm đồ án của riêng chúng tôi và được sự hướng dẫn của TS Nguyễn Chí Thiện;. Các nội dung nghiên cứu, kết quả trong đề tài này là trung thực và chưa công bố dưới bất kỳ hình thức nào trước đây. Những số liệu trong các bảng biểu phục vụ cho việc phân tích, nhận xét, đánh giá được chính tác giả thu thập từ các nguồn khác nhau có ghi rõ trong phần tài liệu tham khảo.

Ngoài ra, trong đồ án còn sử dụng một số nhận xét, đánh giá cũng như số liệu của các tác giả khác, cơ quan tổ chức khác đều có trích dẫn và chú thích nguồn gốc.

Nếu phát hiện có bất kỳ sự gian lận nào tôi xin hoàn toàn chịu trách nhiệm về nội dung đồ án của mình. Trường đại học Tôn Đức Thắng không liên quan đến những vi phạm tác quyền, bản quyền do tôi gây ra trong quá trình thực hiện (nếu có).

TP. Hồ Chí Minh, ngày tháng năm

Tác giả

(ký tên và ghi rõ họ tên)

Hồng Quang Vinh

Nguyễn Đại Thịnh

PHẦN XÁC NHẬN VÀ ĐÁNH GIÁ CỦA GIẢNG VIÊN

Phần xác nhận của GV hướng dẫn

_____

_____

_____

_____

_____

_____

_____

Tp. Hồ Chí Minh, ngày tháng năm
(kí và ghi họ tên)

Phần đánh giá của GV chấm bài

_____

_____

_____

_____

_____

_____

_____

Tp. Hồ Chí Minh, ngày tháng năm
(kí và ghi họ tên)

TÓM TẮT

Bài viết này đưa ra những vấn đề của việc dự đoán sự biến động của cổ phiếu và chỉ số giá cổ phiếu cho thị trừng chứng khoán Ấn Độ. Bài viết so sánh 4 mô hình dự đoán, bao gồm Mạng nơ - ron nhân tạo (Artificial Neural Network - ANN), Support Vector Machin (SVM), Rừng ngẫu nhiên (Random Forest) và Giải thuật Bayes (Naive - Bayes) với 2 cách tiếp cận cho đầu vào của các mô hình này.

Cách tiếp cận đầu tiên cho dữ liệu đầu vào liên quan đến việc tính toán 10 chỉ số kỹ thuật sử dụng dữ liệu huấn luyện về chứng khoán (mở, cao, thấp và giá đóng). Trong khi cách tiếp cận thứ hai tập trung vào việc thể hiện các chỉ số này như là dữ liệu xác định xu hướng.

Độ chính xác của mỗi mô hình dự đoán chi 2 các tiếp cận này được tính toán. Việc tính toán dựa trên 10 năm dữ liệu từ 2003 tới 2012 của 2 cổ phiếu có tên là Reliance Industries và Infosys Ltd. và 2 chỉ số giá cổ phiếu là CNX Nifty và S&P Bombay Stock Exchange (BSE) Sensex.

Kết quả thử nghiệm cho thấy đối với cách tiếp cập đầu tiên, khi mà mười chỉ số kỹ thuật được thể hiện như các giá trị liên tục, mô hình Random Forest vượt trội hơn so với 3 mô hình còn lại về hiệu suất. Kết quả thử nghiệm cũng chỉ ra rằng hiệu suất của tất cả các mô hình dự đoán được cải thiện khi dữ liệu được thể hiện dưới dạng dữ liệu xác định xu hướng.

MỤC LỤC

DANH MỤC KÍ HIỆU VÀ CHỮ VIẾT TẮT

CÁC KÝ HIỆU
CÁC CHỮ VIẾT TẮT

DANH MỤC CÁC BẢNG BIỂU, HÌNH VẼ, ĐỒ THỊ

DANH MỤC HÌNH

CHƯƠNG 1 – GIỚI THIỆU

1.1 Thị trường chứng khoán

Việc dự đoán cổ phiếu và chỉ số giá của phiếu là khó khăn chúng không liên quan nhiều. Có 2 loại phân tích mà các nhà đầu tư thường sử dụng trước khi đầu tư vào một cổ phiếu.

Đầu tiên là phương pháp phân tích cơ bản: Trong đó các nhà đầu tư thường nhìn vào giá trị thực sự của từng cổ phiếu, tình hình của ngành công nghiệp và nền kinh tế, môi trường chính trị, v.v… Sau đó họ sẽ đưa ra quyết định có nên đầu tư hay không.

Phương pháp thứ hai dựa trên các nghiên cứu thống kê, được tạo ra bởi các hoạt động trên thị trường chứng khoán. Trong phương pháp này, các nhà đầu tư sẽ nhìn vào giá cả và mức độ giao dịch của cổ phiếu trong quá khứ. Họ không cố gắng để đo lường một mức an toàn thực tế, thay vào đó họ sử dụng các biểu đồ chứng khoán để xác định dữ liệu mẫu và xu hướng có thể xảy ra của một cổ phiếu trong tương lai.

Giả thuyết thị trường hiệu quả của Malkie và Fama (1970) nói rằng có thể dự đoán được giá cổ phiếu dựa trên dữ liệu giao dịch. Điều này khá hợp lý khi nhiều yếu tố không chắc chắn như chính trị, các hoạt động kinh tế, công nghiệp có xu hướng phản ánh thông qua giá cổ phiếu.

Chính vì vậy, nếu thông tin thu được từ giá cổ phiếu được xử lý trước một cách hiệu quả và áp dụng các thuật toán thích hợp thì xu hướng của cổ phiếu và chỉ số giá cổ phiếu có thể được dự đoán.

Từ nhiều năm, các kỹ thuật đã được phát triển để dự đoán xu hướng chứng khoán, ban đầu các phương pháp hồi quy cổ điển được sử dụng để dự đoán xu hướng chứng khoán. Vì dữ liệu chứng khoán có thể được phân loại thành dữ liệu chuỗi thời gian không cố định, các kỹ thuật học máy phi tuyến tính cũng đã được sử dụng. Trong đó, hai thuật toán học mý được sử dụng rộng rãi để dự đoán biến động của các cổ phiếu và chỉ số giá là Mạng nơ-ron nhân tạo (ANN) và Support Vector Machin (SVM).

1.2 Các phương pháp được đề xuất

Hassan, Nath và Kirley (2007) đề xuất và thực hiện một mô hình phản ứng tổng hợp bằng cách kết hợp các mô hình Hidden Markov (HMM), ANN và thuật toán di truyền (GA) để dự báo hoạt động của thị trường tài chính. Sử dụng ANN, giá cổ phiếu hàng ngày được chuyển đổi thành các bộ giá trị độc lập làm đầu vào cho HMM.

Wang và Leu (1996) đã phát triển một hệ thống dự đoán hữu ích trong việc dự báo xu hướng giá giữa kỳ tại thị trường chứng khoán Đài Loan. Hệ thống của họ dựa trên mộ mạng lưới thần kinh tái phát được huấn luyện bằng cách sử dụng các tính năng được trích xuất từ các phân tích ARIMA. Kết quả thực nghiệm cho thấy, các mạng được huấn luyện sử dụng dữ liệu hàng tuần trong vòng 4 năm có khả năng dự đoán xu hướng lên đến 6 tuần với độ chính xác có thể chấp nhận.

Abraham, Nath và Mahanti (2001) giới thiệu các kỹ thuật điện toán mềm lai hoá để dự báo thị trường chứng khoán tự động và phân tích xu hướng. Họ sử dụng chỉ số Nasdaq của thị trường chứng khoán Nasdaq với mạng nơ-ron cho một ngày trước dự báo và một hệ thống nơ-ron (neuro-fuzzty) để phân tích xu hướng. Các kết quả dự đoán có kết quả rất khả quan.

Chen, Leung và Daouk (2003) đã nghiên cứu về mạng lưới thần kinh song phương (PNN) để dự báo hướng của chỉ số sau khi được huấn luyện. Kết quả thực nghiệm cho thấy các chiến lược đầu tư dựa trên PNN thu lợi nhuận cao hơn các chiến lược đầu tư khác như các chiến lược sử dụng dự báo được dự đoán trên mô hình đi bộ ngẫu nhiên (random walk) và mô hình GMM tham số.

Vapnik (1999) đã phát triển thuật toán SVM rất nổi tiếng tìm kiếm một siêu phẳng ở không gian nhiều chiều để phân lớp. SVM là một loại thuật toán học các đặc trưng sử dụng các kernel function và scarcity solution.

Huang, Nakamori và Wang (2005) đã nghiên cứu tính khả thi của hướng di chuyển khi sử dụng mô hình SVM thông qua dự báo hướng di chuyển hàng tuần của chỉ số NIKKEI 225. Họ đã so sánh SVM với Linear Discriminant Analysis, Quadratic

Discriminant Analysis and Elman Backpropagation Nerral Networks. Kết quả thí nghiệm cho thấy SVM vượt trội so với các phương pháp phân loại khác.

SVM được Kim (2003) sử dụng để dự đoán hướng thay đổi giá cổ phiếu hàng ngày trong chỉ số giá cổ phiếu Hàn Quốc (KOSPI). Mười hai chỉ số kỹ thuật đã được chọn để tạo nên các thuộc tính ban đầu. Nghiên cứu đã so sánh SVM với Elman Backpropagation Neural Networks (BPN) và lý luận dựa trên trường hợp (CBR). Các kết quả đều cho thấy SVM vượt trội hơn hẳn.

Tsai, Lin, Yen và Chen (2011) đã nghiên cứu hiệu suất dự đoán sử dụng phương pháp phân loại tập hợp để phân tích lợi nhuận chứng khoán. Các phương pháp kết hợp giữa voting và bagging đã được cân nhắc.

Sun và Li (2012) đã đề xuất phương pháp dự đoán tình trạng khó khăn tài chính (FDP) dựa trên tập hợp SVM. Kết quả cho thấy nhóm các SVM vượt trội hơn hẳn so với từng SVM đơn lẻ.

Ou và Wang (2009) đã sử dụng tổng số mười kỹ thuật khai thác dữ liệu để dự đoán biến động giá chỉ số Hang Seng của thị trường chứng khoán Hong Kong. Các phương pháp tiếp cận bao gồm Linear discriminant analysis (LDA), Quadratic Discriminant Analysis (QDA), K-nearest neighbor classification, Naive Bayes dựa trên ước lượng kernel, mô hình Logit, phân loại dựa trên cây Tree, mạng nơ-ron, phân loại Bayes. Kết qủa thực nghiệm cho thấy SVM và LS-SVM tạo ra hiệu suất vượt trội so với các mô hình khác.

Các phần sau của bài viết tập trung so sánh hiệu suất dự đoán của các thuật toán ANN, SVM, Random forest vs Naive Bayes cho biến động của cổ phiếu và chỉ số giá cổ phiếu. Mười thông số kỹ thuật được sử dụng làm đầu vào cho các mô hình này.

Lớp chuyển đổi có nhiệm vụ chuyển đổi dữ liệu đầu vào có giá trị liên tục thành các đầu vào riêng biệt. Mỗi tham số vào ở dạng riêng cho thấy xu hướng tăng hoặc giảm có thể được xác định dựa trên các đặc tính vốn có của nó. Trọng tâm là để so sánh hiệu suất của các mô hình khi sử dụng dữ liệu giá trị thực và dữ liệu xu hướng.

Tất cả thử nghiệm được thực hiện dựa trên dữ liệu 10 năm của 2 cổ phiếu Reliance Industries và Infosys Ltd. và 2 chỉ số S&P BSE Sensex và CNX Nifty. Cả 2 cổ phiếu và chỉ số đều rất mạnh và được giao dịch mạnh vì vậy chúng phản ảnh toàn bộ nền kinh tế Ấn Độ.

CHƯƠNG 2 – DỮ LIỆU NGHIÊN CỨU

Nghiên cứu này sử dụng 10 năm dữ liệu của 2 cổ phiếu Reliance Industries và Infosys Ltd. cũng như 2 chỉ số giá cổ phiếu CNX Nifty và S&P BSE Sensex từ tháng 1 năm 2003 đến tháng 12 năm 2012. Tỷ phần trăm tăng giảm của mỗi năm được thể hiện trong bảng dưới đây.

Table 1
The number of increase and decrease cases percentage in each year in the entire data set of S&P BSE SENSEX.

Year	Increase	%	Decrease	%	Total
2003	146	58.63	103	41.37	249
2004	136	54.18	115	45.82	251
2005	147	59.04	102	40.96	249
2006	148	59.92	99	40.08	247
2007	139	55.82	110	44.18	249
2008	114	46.72	130	53.28	244
2009	127	52.70	114	47.30	241
2010	134	53.39	117	46.61	251
2011	116	47.15	130	52.85	246
2012	128	51.82	119	48.18	247
Total	1335	53.94	1139	46.06	2474

Hình 1: Tỷ lệ phần trăm tăng giảm của cổ phiếu S&P

Nghiên cứu này sử dụng 20% toàn bộ dữ liệu làm dữ liệu lựa chọn tham số. Dữ liệu này được sử dụng để xác định các tham số thiết kế của các mô hình dự đoán. Tập dữ liệu lựa chọn tham số được xây dựng bằng cách lấy tỷ lệ dữ liệu bằng nhau từ mỗi mười năm. Tỷ lệ phần trăm tăng giảm trong mỗi năm cũng được duy trì. Phương pháp lấy mẫu này cho phép tập dữ liệu cài đặt tham số trở thành đại diện tốt hơn cho toàn bộ tập dữ liệu.

Dữ liệu lựa chọn tham số này được chia thành tập training và holdout. Mỗi bộ bao gồm 10% toàn bộ dữ liệu. Hình 2 mô tả số lượng các trường hợp tăng và giảm cho tập dữ liệu lựa chọn tham số. Những thống kê này dành cho S & P BSE Sensex. Phân tích dữ liệu tương tự được thực hiện cho CNX Nifty, Reliance Industries và Infosys Ltd.

Table 2
The number of increase and decrease cases in each year in the parameter setting data set of S&P BSE SENSEX.

Year	Training			Holdout		
	Increase	Decrease	Total	Increase	Decrease	Total
2003	15	10	25	15	10	25
2004	14	11	25	14	11	25
2005	15	10	25	15	10	25
2006	15	10	25	15	10	25
2007	14	11	25	14	11	25
2008	11	13	24	11	13	24
2009	13	11	24	13	11	24
2010	13	12	25	13	12	25
2011	12	13	25	12	13	25
2012	13	12	25	13	12	25
Total	135	113	248	135	113	248

Hình 2: Dữ liệu 10 năm được trích 10% từ tổng

Các tham số tối ưu cho các mô hình dự đoán thu được bằng các thí nghiệm trên dữ liệu lựa chọn tham số. Sau đó, để so sánh ANN, SVM, Random forest và Naive-Bayes, bộ dữ liệu so sánh được đưa ra. Bộ dữ liệu này bao gồm toàn bộ mười năm dữ liệu. Nó cũng được chia trong đào tạo (50% toàn bộ dữ liệu) và giữ (50% toàn bộ dữ liệu) như bảng. Chi tiết về bộ dữ liệu này của S & P BSE SENSEX được hiển thị trong Hình 3.

Table 3
The number of increase and decrease cases in each year in the comparison data set of S&P BSE SENSEX.

Year	Training			Holdout		
	Increase	Decrease	Total	Increase	Decrease	Total
2003	73	52	125	72	52	124
2004	68	58	126	67	58	125
2005	74	51	125	73	51	124
2006	74	50	124	73	50	123
2007	70	55	125	69	55	124
2008	57	65	122	57	65	122
2009	64	57	121	63	57	120
2010	67	59	126	66	59	125
2011	58	65	123	58	65	123
2012	64	60	124	63	60	123
Total	669	572	1241	661	572	1233

Hình 3: Dữ liệu 10 năm được trích 50% từ tổng

Có một số chỉ số kỹ thuật thông qua đó người ta có thể dự đoán sự chuyển động của cổ phiếu trong tương lai. Ở đây trong nghiên cứu này, tổng cộng mười chỉ số kỹ thuật được sử dụng trong Kara et al. (2011) được sử dụng. Các chỉ số này được thể hiện trong Hình 4 cho thấy số liệu thống kê tóm tắt cho các chỉ số được chọn của hai chỉ số và hai cổ phiếu. Hai cách tiếp cận để biểu diễn dữ liệu đầu vào được sử dụng trong nghiên cứu này. Cách tiếp cận đầu tiên sử dụng biểu diễn giá trị liên tục, tức là chuỗi thời gian thực tế trong khi cách thứ hai sử dụng biểu diễn xác định xu hướng (có tính chất rời rạc) cho các đầu vào.

Indicator	Max	Min	Mean	Standard deviation
Nifty				
SMA	6217.37	935.38	3789.68	1047.47
EMA	6214.38	940.35	3789.69	1054.51
MOM	748.40	−1372.70	17.81	7.55
STCK%	99.14	1.84	60.51	76.33
STCD%	97.90	4.08	60.50	57.75
MACD	277.17	−357.33	13.52	−13.39
RSI	100.00	1.42	56.40	46.28
WILLR%	−0.86	−98.16	−39.49	−23.67
A/D Osc	98.24	1.91	53.31	86.74
CCI	333.33	−270.50	22.84	96.39
BSE-Sensex				
SMA	20647.77	2957.11	12602.94	3263.12
EMA	20662.52	2964.00	12603.00	3280.12
MOM	2362.24	−4139.84	59.22	17.58
STCK%	100.00	1.10	60.04	75.10
STCD%	97.79	5.17	60.02	56.60
MACD	921.17	−1146.29	45.05	−33.42
RSI	100.00	1.07	56.48	45.96
WILLR%	0.00	−98.90	−39.96	−24.90
A/D Osc	100.00	1.78	50.79	94.47
CCI	333.33	−247.49	23.09	97.32
Infosys				
SMA	3432.13	337.98	1783.73	543.61
EMA	3425.35	341.62	1783.74	550.85
MOM	340.10	−493.90	6.49	16.70
STCK%	100.00	0.67	55.40	92.65
STCD%	96.57	3.13	55.39	75.88
MACD	108.04	−145.99	5.05	−11.59
RSI	98.13	1.27	53.55	57.78
WILLR%	0.00	−99.33	−44.60	−7.35
A/D Osc	100.00	1.41	50.24	86.37
CCI	330.44	−314.91	16.65	140.39
Reliance				
SMA	3073.36	265.02	1102.55	278.16
EMA	3065.97	265.95	1102.55	281.48
MOM	483.90	−1122.20	2.03	2.40
STCK%	99.30	0.89	53.14	46.57
STCD%	98.02	2.88	53.13	41.97
MACD	162.91	−276.50	1.50	−4.54
RSI	100.00	4.31	53.22	42.58
WILLR%	−0.70	−99.11	−46.86	−53.43
A/D Osc	572.88	−350.48	46.36	44.00
CCI	333.33	−333.33	12.81	72.41

Hình 4: Mười chỉ số kỹ thuật được sử dụng

2.1 **Dữ liệu thể hiện dạng liên tục - chuỗi thời gian thực**

Mười chỉ số kỹ thuật được tính toán dựa trên công thức như được thảo luận trong Hình 5 được đưa ra làm đầu vào cho các mô hình dự đoán. Rõ ràng là mỗi chỉ số kỹ thuật được tính toán dựa trên công thức đã đề cập ở trên có giá trị liên tục. Các giá trị của tất cả các chỉ báo kỹ thuật được chuẩn hóa trong phạm vi giữa [-1, +1], để giá trị lớn hơn của một input params không áp đảo chỉ báo có giá trị nhỏ hơn. Hiệu suất của tất cả các mô hình đang nghiên cứu được đánh giá cho đại diện đầu vào này.

Name of indicators	Formulas
Simple n(10here)-day Moving Average	$\frac{C_t+C_{t-1}+\ldots+C_{t-9}}{n}$
Weighted n(10here)-day Moving Average	$\frac{(10)C_t+(9)C_{t-1}+\ldots+C_{t-9}}{n+(n-1)+\ldots+1}$
Momentum	$C_t - C_{t-9}$
Stochastic $K\%$	$\frac{C_t - LL_{(n-1)}}{HH_{(n-1)} - LL_{(n-1)}} \times 100$
Stochastic $D\%$	$\frac{\sum_{i=0}^{n-1} K_{t-i}}{10}\%$
Relative Strength Index (RSI)	$100 - \frac{100}{1+\left(\sum_{i=0}^{n-1} UP_{t-i}/n\right)/\left(\sum_{i=0}^{n-1} DW_{t-i}/n\right)}$
Moving Average Convergence Divergence (MACD)	$MACD(n)_{t-1} + \frac{2}{n+1} \times (DIFF_t - MACD(n)_{t-1})$
Larry William's R%	$\frac{H_n - C_t}{H_n - L_n} \times 100$
A/D (Accumulation/Distribution) Oscillator	$\frac{H_t - C_{t-1}}{H_t - L_t}$
CCI (Commodity Channel Index)	$\frac{M_t - SM_t}{0.015D_t}$

Hình 5: Mười chỉ số kỹ thuật và công thức tính

2.2 Dữ liệu thể hiện dạng rời rạc - dữ liệu xác định xu hướng

Một lớp quyết định mới được sử dụng để chuyển đổi các tham số kỹ thuật có giá trị liên tục thành giá trị rời rạc, đại diện cho xu hướng cổ phiếu. Lớp này được gọi là ‘‘Lớp chuyển đổi dữ liệu xác định xu hướng”. Công việc của lớp mới này là chuyển đổi các giá trị liên tục này thành +1 hoặc -1 bằng cách xem xét thuộc tính này trong quá trình rời rạc. Bằng cách này, dữ liệu đầu vào của từng mô hình dự đoán được chuyển đổi thành +1 hoặc -1, trong đó +1 là chuyển động đi lên và -1 hiển thị chuyển động xuống. Chi tiết về từng chỉ số kỹ thuật được nêu ra dưới đây.

Đường trung bình động (MA) là công cụ phân tích kỹ thuật đơn giản giúp làm mịn dữ liệu giá bằng cách tạo ra giá trung bình được cập nhật liên tục. Trong bài báo này, 10 ngày trung bình động đơn giản (SMA) và trung bình động có trọng số (WMA) được sử dụng vì dự đoán tương lai ngắn hạn. Theo nguyên tắc chung, nếu giá cao hơn mức trung bình động thì xu hướng tăng. Nếu giá dưới mức trung bình động, xu hướng giảm. Vì vậy, bài viết đã rút ra ý kiến của cả hai chỉ số SMA và WMA cho mỗi ngày từ giá trị của SMA và WMA so với giá hiện tại. Nếu giá hiện tại cao hơn các giá trị trung bình di động thì xu hướng sẽ tăng lên và được biểu thị là +1, và nếu giá hiện tại nằm dưới các giá trị trung bình di chuyển thì xu hướng đó sẽ giảm xuống và được biểu thị dưới dạng -1.

STCK%, STCD% và Williams R% là các dao động ngẫu nhiên. Những dao động là chỉ số xu hướng rõ ràng cho bất kỳ cổ phiếu. Khi các dao động ngẫu nhiên đang tăng lên, giá cổ phiếu có thể sẽ tăng lên và tiếp tục. Vì vậy nếu giá trị của các bộ dao động ngẫu nhiên tại thời điểm ‘t’ lớn hơn giá trị tại thời điểm ‘t - 1’ thì ý kiến về xu hướng là lên và được biểu thị dưới dạng +1 và ngược lại.

MACD đi theo xu hướng của cổ phiếu, tức là nếu MACD tăng thì giá cổ phiếu cũng tăng và ngược lại. Vì vậy, nếu giá trị của MACD tại thời điểm ‘t’ lớn hơn giá trị tại thời điểm ‘t-1’, xu hướng về cổ phiếu sẽ là ‘tăng’ và được biểu thị là +1 và ngược lại.

RSI thường được sử dụng để xác định các điểm mua quá mức và bán quá mức. Nó nằm trong khoảng từ 0 đến 100. Nếu giá trị của RSI vượt quá 70, điều đó có nghĩa là cổ phiếu bị mua quá mức, do đó, nó có thể giảm trong tương lai gần (biểu thị -1) và nếu giá trị của RSI giảm xuống dưới 30 mức độ, điều đó có nghĩa là cổ phiếu bị bán quá mức, vì vậy, nó có thể sẽ tăng trong tương lai gần (biểu thị ý kiến +1). Đối với các giá trị nằm trong khoảng (30, 70), nếu RSI tại thời điểm ‘t’ lớn hơn RSI tại thời điểm ‘t-1’, thì xu hướng được biểu thị là +1 và ngược lại.

CCI đo lường sự khác biệt giữa thay đổi giá cổ phiếu và thay đổi giá trung bình. Chỉ số tích cực cao cho thấy giá cao hơn mức trung bình, đó là một sự thể hiện sức mạnh.

Chỉ số tiêu cực thấp cho thấy giá thấp hơn mức trung bình của họ, đây là một biểu hiện của sự yếu kém. CCI cũng được sử dụng để xác định mức mua quá mức và bán quá mức. Trong bài báo này, họ đã đặt 200 là mức mua quá mức và -200 là mức bán quá mức vì 200 là đại diện cho một thái cực thực sự. Điều này có nghĩa là nếu giá trị CCI vượt quá 200 cấp, xu hướng là -1 và nếu dưới mức 200 thì ý kiến cho xu hướng là +1. Đối với các giá trị trong khoảng (-200, 200), nếu CCI tại thời điểm 't' lớn hơn CCI tại thời điểm t-1, thì ý kiến về xu hướng là +1 và ngược lại.

Bộ tạo dao động A/D cũng tuân theo xu hướng chứng khoán có nghĩa là nếu giá trị của nó tại thời điểm 't' lớn hơn giá trị tại thời điểm 't - 1', thì xu hướng là +1 và ngược lại.

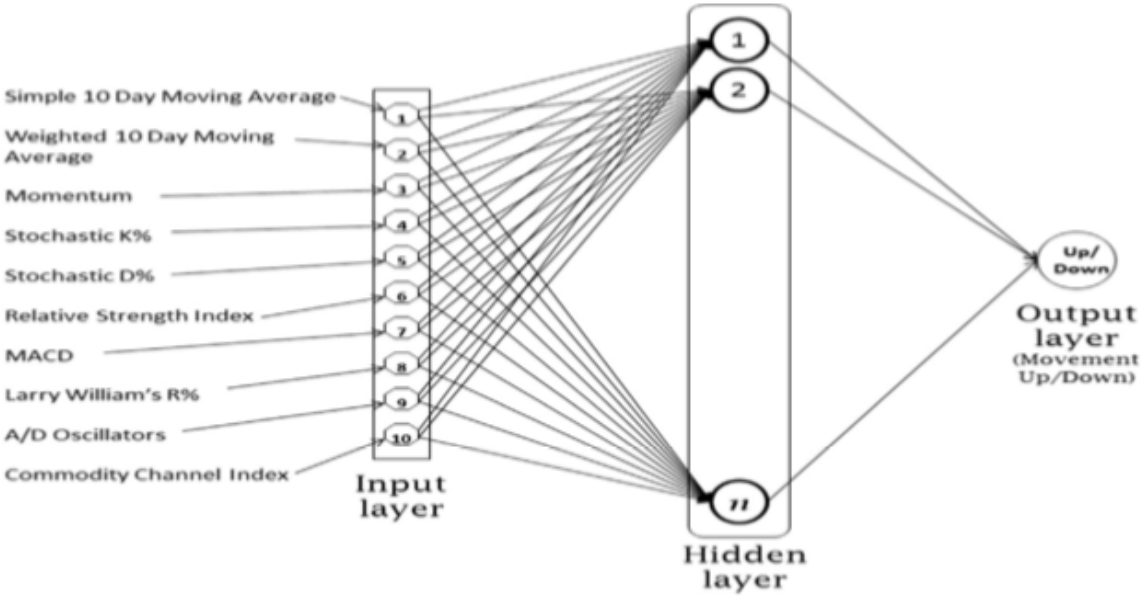
Momentum đo tốc độ tăng giảm của giá cổ phiếu. Giá trị dương của biểu thị xu hướng tăng và được biểu thị bằng +1 trong khi giá trị âm biểu thị xu hướng giảm và được thể hiện dưới dạng -1.

Dữ liệu xác định xu hướng được chuẩn bị bằng cách khai thác thực tế rằng mỗi chỉ số kỹ thuật đều thể hiện biến động về giá cổ phiếu. Khi các dữ liệu này được cung cấp làm đầu vào trái ngược với dữ liệu thực tế, thông tin xu hướng được nhập vào dựa trên từng chỉ số kỹ thuật riêng lẻ. Các mô hình dự đoán cần phải xác định mối tương quan giữa xu hướng đầu vào và xu hướng đầu ra.

CHƯƠNG 3 – CÁC MÔ HÌNH DỰ ĐOÁN

3.1 Mô hình Mạng nơ-ron nhân tạo (ANN)

Là một mạng lưới dày đặc các nơ-ron liên kết với nhau được kích hoạt dựa trên các đầu vào. Một mạng nơ-ron 3 lớp feed-forward được sử dụng trong nghiên cứu. Đầu vào cho mạng là 10 chỉ số kỹ thuật được biểu thị bằng 10 nơ-ron trong lớp đầu vào. Lớp đầu ra có một nơ-ron duy nhất ứng với log sigmoid như là hàm chuyển. Điều này dẫn đến đầu ra giá trị liên tục trong khoảng từ 0 - 1. Ngưỡng 0,5 được sử dụng để xác định dự đoán chuyển động lên hoặc xuống. Lớn hơn 0,5 dự đoán là chuyển động tăng, ngược lại là giảm. Mỗi tầng nơ-ron ẩn sử dụng tan sigmoid như là hàm chuyển.



Hình 6: Kiến trúc 3 lớp feedforward của mô hình ANN

Kiến trúc 3 lớp feed-forward ANN được minh hoạ ở hình trên. Độ dốc giảm với động lượng được sử dụng để điều chỉnh các trọng số, trong đó tại mỗi thời kỳ, trọng số được điều chỉnh sao cho có thể đạt mức tối thiểu toàn cục. Các thử nghiệm thiết lập tham số toàn diện để xác định tham số cho từng cổ phiếu và chỉ số giá cổ phiếu.

Các tham số cho mô hình ANN bao gồm: Số tầng nơ-ron ẩn (n), giá trị của tỉ lệ huấn luyện (lr), hằng số động lượng (mc), và số thời kỳ (ep). Để xác định chúng một

cách hiệu quả, 10 mức (n), 9 mức (mc), 10 mức (ep) được kiểm tra trong các thử nghiệm thiết lập tham số. Ban đầu, giá trị (lr) được cố định là 0,1. Các tham số này được thể hiện trong bảng dưới đây.

Table 6
ANN parameters and their levels tested in parameter setting.

Parameters	Level(s)
Number of hidden layer neurons (*n*)	10,20,. . .,100
Epochs (*ep*)	1000, 2000, . . ., 10,000
Momentum constant (*mc*)	0.1, 0.2, . . ., 0.9
Learning rate (*lr*)	0.1

Hình 7: Các thiết lập tham số cho mô hình ANN

Các thiết lập tham số mang lại tổng cộng 10x10x9 = 900 phương pháp ANN cho một cổ phiếu. Xét 2 cổ phiếu và 2 chỉ số giá cổ phiếu, tổng số 3600 phương pháp được thực hiện.

Ba cách kết hợp tham số hàng đầu được chọn là 3 mô hình ANN hàng đầu cho các thử nghiệm so sánh trên tập dữ liệu. Đối với mô hình này, tỉ lệ huấn luyện (lr) được thay đổi trong khoảng [0,1 - 0,9].

3.2 Mô hình Support Vector Machin (SVM)

Lần đầu tiên được giới thiệu bởi Vapnik (1999). Có 2 loại chính là Support Vector Classification (SVC) và Support Vector Regression (SVR). SVM là hệ thống học sử dụng không gian đặc trưng có số chiều lớn. Trong SVM, các điểm được phân loại bằng các gán chúng vào hai nửa không gian khác nhau trong không gian mẫu hoặc ánh xạ sang không gian có số chiều lớn hơn.

Mục tiêu của SVM là xác định siêu phẳng có biên tối đa. Ý tưởng là khoảng cách từ siêu phẳng này tới các điểm positive và negative là bằng nhau và lớn nhất. Sự đánh đổi giữa khoảng cách và lỗi phân loại được kiểm soát bởi tham số chính quy (c). Các

hàm nhân đa thức và cơ sở được sử dụng. Trong đó (d) là độ của hàm đa thức, (γ) là hằng số của hàm cơ sở.

Lựa chọn hàm kernel, mức độ (d) trong trường hợp kernel đa thức, gamma trong hàm kernel (γ) trong trường hợp kernel là hàm cơ sở và hằng số chính quy (c) là các tham số cho SVM.

Để xác định chúng một cách hiệu quả, 4 mức (d), 10 mức (γ) và 4 - 5 mức (c) được thử nghiệm trong các thử nghiệm thiết lập tham số. Các tham số được tóm tắt trong bảng dưới.

Table 7
SVM parameters and their levels tested in parameter setting.

Parameters	Levels (polynomial)	Levels (radial basis)
Degree of kernel function (d)	1, 2, 3, 4	–
Gamma in kernel function (γ)	–	0.5, 1.0, 1.5, . . . , 5.0, 10.0
Regularization parameter (c)	0.5, 1, 5, 10, 100	0.5, 1, 5, 10

Hình 8: Các thiết lập tham số cho mô hình SVM

Với một cổ phiếu, các thiết lập tham số này mang lại tổng cộng 20 và 40 phương pháp cho SVM sử dụng các hàm nhân cơ sở đa thức và hướng tâm tương ứng. Xem xét 2 cổ phiếu và 2 chỉ số giá cổ phiếu, tổng số 240 phương pháp cho SVM được thực hiện. Một kết hợp tham số cho mỗi SVM kernel đa thức và SVM kernel cơ sở xuyên tâm dẫn đến hiệu suất đào tạo và tổ chức trung bình tốt nhất được chọn làm hai mô hình SVM hàng đầu cho các thử nghiệm so sánh.

3.3 Mô hình Random Forest

Cây quyết định là một trong những kỹ thuật phổ biến nhất để phân loại. Độ chính xác tương đương với các phương pháp phân loại khác, và nó rất hiệu quả. Mô hình học được thông qua kỹ thuật này được biểu diễn dưới dạng cây và được gọi là cây quyết định.

Random forest thuộc về loại thuật toán học. Nó sử dụng cây quyết định làm cơ sở cho việc học. Ý tưởng của việc học là một bộ phân loại duy nhất không đủ để xác định lớp dữ liệu kiểm tra. Lý do là, dựa trên dữ liệu mẫu, bộ phân loại không thể phân biệt giữa nhiễu và mẫu. Vì vậy, nó thực hiện lấy mẫu với sự thay thế sao cho (n) cây được học dựa trên các mẫu tập dữ liệu này. Cũng trong các thí nghiệm, mỗi cây được học bằng 3 đặc trưng được chọn ngẫu nhiên. Điều này cũng tránh được vấn đề về quá khớp dữ liệu. Việc thực hiện thuật toán Random Forest được tóm tắt như dưới đây.

Algorithm 1. Our implementation of random forest

Input: training set D, number of trees in the ensemble k
Output: a composite model $M*$
1: **for** $i = 1$ to k **do**
2: Create bootstrap sample D_i by sampling D with replacement.
3: Select 3 features randomly.
4: Use D_i and randomly selected three features to derive tree M_i.
5: **end for**
6: return $M*$.

Hình 9: Thuật toán cho mô hình Random Forest

Số lượng cây trong toàn bộ (n) cây được coi là tham số của Random Forest. Để xác định nó một cách hiệu quả, nó được thay đổi từ 10 đến 200 với mức tăng 10 mỗi lần trong các thử nghiệm thiết lập tham số. Đối với một cổ phiếu, các cài đặt tham số này mang lại tổng cộng 20 phương pháp. Xem xét 2 cổ phiếu và 2 chỉ số giá cổ phiếu, tổng số 80 phương pháp điều trị được thực hiện. Ba cách kết hợp tham số là ba mô hình Random Forest hàng đầu cho các thử nghiệm so sánh.

3.4 Mô hình Naive Bayes

Phân loại Naive Bayes giả định lớp độc lập có điều kiện. Phân loại Bayes dự đoán xác suất dữ liệu thuộc về một lớp cụ thể. Để dự đoán xác suất, nó sử dụng khái niệm định lý Bayes. Định lý Bayes, hữu ích ở chỗ nó cung cấp cách tính xác suất, cụ thể qua thông thức bên dưới.

- $P(X|C) = \frac{P(X|C)P(C)}{P(X)}$

Trong đó P là xác suất giả thuyết C đúng với sự kiện X đã xảy ra. Trong trường hợp giả thuyết C là xác suất thuộc về lớp "tăng" = "giảm" và sự kiện X là dữ liệu thử nghiệm. P(X|C) là một xác suất có điều kiện xảy ra sự kiện X với giả thuyết C là đúng. Nó có thể được ước tính từ dữ liệu đào tạo.

Các bộ phân loại Bayes cũng đóng vai trò là lý thuyết cho các phân lớp khác không sử dụng rõ ràng định lý Bayes. Ví dụ, theo các giả định cụ thể, có thể chứng minh rằng nhiều mạng thần kinh và thuật toán khớp đường cong đưa ra giả thuyết posteriori tối đa, giống như phân lớp Naive-Bayes.

CHƯƠNG 4 – KẾT QUẢ THỰC NGHIỆM

Độ chính xác và độ đo F được sử dụng để đánh giá hiệu suất của các mô hình được đề xuất. Việc tính toán các biện pháp đánh giá này đòi hỏi phải ước tính Độ rõ ràng (Precision) và Thu hồi (Recall) được đánh giá từ True Positive (TP), False Positive (FP), True Negative (TN) và False Negative (FN). Các tham số này được xác định trong các phương trình dưới đây.

- $Precision_{positive} = \frac{TP}{TP+FP}$
- $Precision_{negative} = \frac{TN}{TN+FN}$
- $Recall_{positive} = \frac{TP}{TP+FN}$
- $Recall_{negative} = \frac{TN}{TN+FP}$

Trong đó, Precision là trung bình có trọng số của độ chính xác positive và negative trong khi Recall là trung bình có trọng số của thu hồi positive và negative. Độ chính xác và độ đo F được tính bằng các phương trình.

- $Accuracy = \frac{TP+TN}{TP+FP+TN+FN}$
- $F-measure = \frac{2 \times Precision \times Recall}{Precision+Recall}$

Giai đoạn thử nghiệm đầu tiên coi đầu vào là dữ liệu có giá trị liên tục. Các kết hợp tham số tốt nhất được xác định bằng các thử nghiệm trên tập dữ liệu cài đặt tham số cho từng mô hình dự đoán. Các kết hợp tham số này với Độ chính xác và Độ đo F đáp ứng trong các thử nghiệm cài đặt tham số được báo cáo trong bảng dưới.

Table 8
Best three parameter combinations of ann model and their performance on continuous-valued parameter setting data set.

epochs : *neurons* : *mc* & *lr* = 0.1			
Nifty			
	10,000:20:0.6	7000:10:0.7	7000:10:0.9
Accuracy	0.8434	0.8450	0.8558
F-measure	0.8614	0.8606	0.8686
BSE-Sensex			
	1000:80:0.1	2000:40:0.2	10,000:100:0.1
Accuracy	0.7968	0.7827	0.7723
F-measure	0.7743	0.7982	0.7862
Infosys			
	1000:70:0.7	8000:150:0.7	3000:10:0.3
Accuracy	0.7417	0.7023	0.6949
F-measure	0.7581	0.7098	0.7412
Reliance			
	8000:50:0.6	6000:40:0.4	9000:20:0.5
Accuracy	0.6356	0.6326	0.6898
F-measure	0.6505	0.6116	0.7067

Hình 10: Độ chính xác cho các kết hợp tham số tốt nhất của mô hình ANN

Table 9
Best two parameter combinations (one for each type of kernel) of SVM model and their performance on continuous-valued parameter setting data set.

	Kernel:Polynomial	Kernel:RBF
Nifty		
	c:100,degree:1	c:0.5,gamma:5
Accuracy	0.8427	0.8057
F-measure	0.8600	0.8275
BSE-Sensex		
	c:100,degree:1	c:1,gamma:5
Accuracy	0.8136	0.7823
F-measure	0.8321	0.8015
Infosys		
	c:0.5,degree:1	c:0.5,gamma:5
Accuracy	0.8139	0.7836
F-measure	0.8255	0.7983
Reliance		
	c:0.5,degree:1	c:1,gamma:5
Accuracy	0.7669	0.6881
F-measure	0.7761	0.7023

Hình 11: Độ chính xác cho các kết hợp tham số tốt nhất của mô hình SVM

Table 10
Best three parameter combinations of random forest model and their performance on continuous-valued parameter setting data set.

	ntrees		
Nifty			
	140	20	30
Accuracy	0.9148	0.9146	0.9099
F-measure	0.9186	0.9185	0.9162
BSE-Sensex			
	80	50	70
Accuracy	0.8819	0.8719	0.8786
F-measure	0.8838	0.8742	0.8802
Infosys			
	50	110	200
Accuracy	0.8138	0.8059	0.8132
F-measure	0.8202	0.8135	0.8190
Reliance			
	160	60	150
Accuracy	0.7368	0.7441	0.7450
F-measure	0.7389	0.7474	0.7478

Hình 12: Độ chính xác cho các kết hợp tham số tốt nhất của mô hình Random Forest

Mục đích của các thử nghiệm trên tập dữ liệu so sánh là để hoàn thành hiệu suất dự đoán của các mô hình này cho các kết hợp tương đương tốt nhất được báo cáo trong các thử nghiệm cài đặt tham số. Trong thí nghiệm so sánh này, mỗi mô hình dự đoán được học dựa trên các tham số tốt nhất được báo cáo bởi các thử nghiệm cài đặt tham số.

Table 11
Performance of prediction models on continuous-valued comparison data set.

Stock/Index	Prediction Models			
	ANN Kara et al. (2011)		SVM	
	Accuracy	F-measure	Accuracy	F-measure
S&P BSE SENSEX	0.7839	0.7849	0.7979	0.8168
NIFTY 50	0.8481	0.8635	0.8242	0.8438
Reliance Industries	0.6527	0.6786	0.7275	0.7392
Infosys Ltd.	0.7130	0.7364	0.7988	0.8119
Average	0.7494	0.7659	0.7871	0.8029
	Random forest		Naive-Bayes (Gaussian)	
	Accuracy	F-measure	Accuracy	F-measure
S&P BSE SENSEX	0.8775	0.8794	0.7354	0.7547
NIFTY 50	0.9131	0.9178	0.8097	0.8193
Reliance Industries	0.7420	0.7447	0.6565	0.6658
Infosys Ltd.	0.8110	0.8176	0.7307	0.7446
Average	0.8359	0.8399	0.7331	0.7461

Hình 13: Độ chính xác của từng mô hình

Bảng trên báo cáo độ chính xác trung bình và số đo F của từng mô hình trong quá trình thử nghiệm so sánh. Độ chính xác trung bình và số đo F được báo cáo trung bình trên các mô hình hoạt động hàng đầu. Có thể thấy rằng Naive-Bayes với quy trình Gaussian là ít chính xác nhất trong khi Random forest là chính xác nhất với độ chính xác trung bình gần 84%. Hình dưới mô tả quá trình dự đoán khi dữ liệu được định giá liên tục.

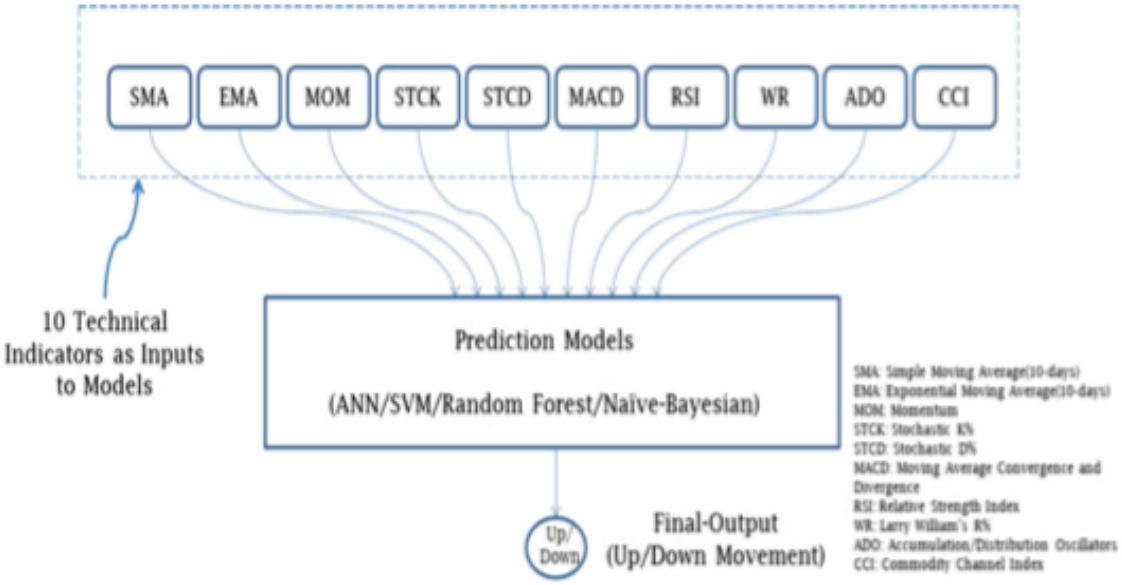


Fig. 2. Predicting with continuous-valued data.

Hình 14: Quá trình dự đoán với dữ liệu liên tục

Giai đoạn thử nghiệm thứ hai giống hệt với giai đoạn đầu tiên ngoại trừ đầu vào của các mô hình là dữ liệu xác định xu hướng. Ý tưởng được mô tả trong hình dưới đây.

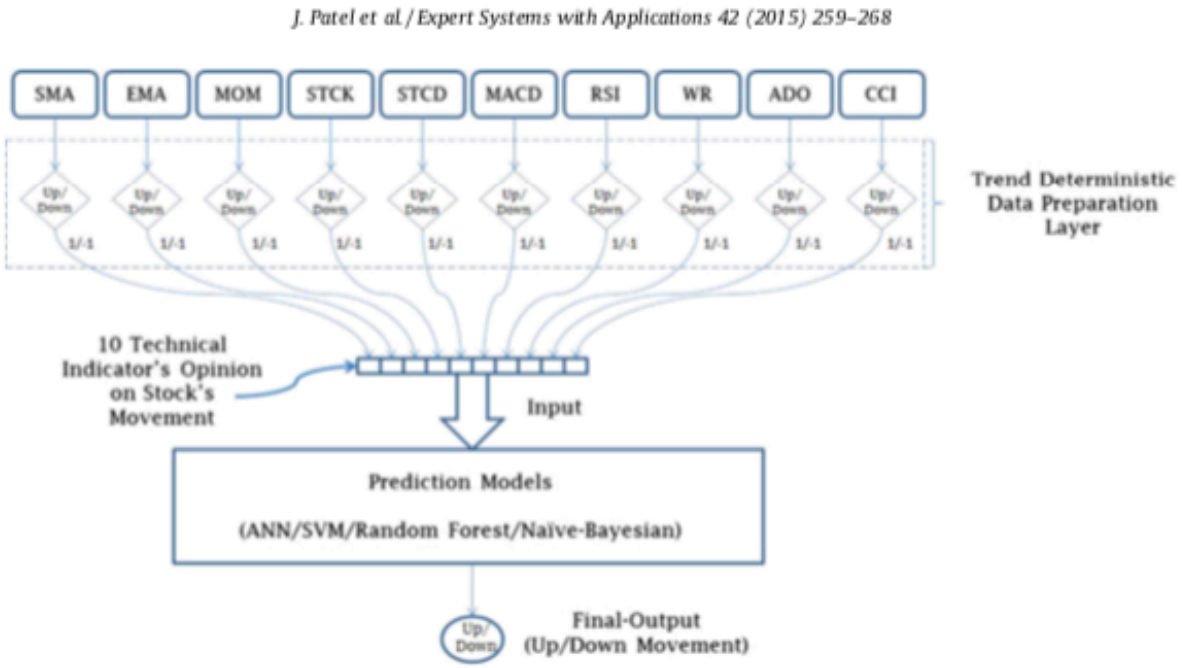


Fig. 3. Predicting with trend deterministic data.

Hình 15: Quá trình dự đoán với dữ liệu rời rạc

Bảng dưới cho thấy các kết quả của các kết hợp hình thành tốt nhất cho ANN, SVM và Random forest tương ứng trong các thí nghiệm cài đặt tham số.

Table 12
Best three parameter combinations of ann model and their performance on discrete-valued parameter setting data set.

Nifty			
	epochs : *neurons* : *mc* & *lr* = 0.2		
	4000:50:0.8	1000:100:0.6	3000:70:0.3
Accuracy	0.8703	0.8740	0.8729
F-measure	0.8740	0.8768	0.8801
BSE-Sensex			
	epochs : *neurons* : *mc* & *lr* = 0.1		
	6000:100:0.4	2000:30:0.3	4000:90:0.1
Accuracy	0.8563	0.8728	0.8717
F-measure	0.8632	0.8771	0.8759
Infosys			
	epochs : *neurons* : *mc* & *lr* = 0.1		
	6000:50:0.1	4000:70:0.2	9000:80:0.4
Accuracy	0.8531	0.8717	0.8468
F-measure	0.8600	0.8742	0.8503
Reliance			
	epochs : *neurons* : *mc* & *lr* = 0.2		
	1000:100:0.1	4000:90:0.9	8000:100:0.5
Accuracy	0.8573	0.8747	0.8808
F-measure	0.8620	0.8799	0.8826

Hình 16: Kết quả kết hợp tham số đối với mô hình ANN

Table 13
Best two parameter combinations (one for each type of kernel) of svm model and their performance on discrete-valued parameter setting data set.

	Kernel:Polynomial	Kernel:RBF
Nifty		
	c:1,degree:1	c:1,gamma:4
Accuracy	0.9010	0.8808
F-measure	0.9033	0.8838
BSE-Sensex		
	c:1,degree:1	c:5,gamma:1.5
Accuracy	0.8959	0.8780
F-measure	0.8980	0.8810
Infosys		
	c:0.5,degree:1	cc:1,gamma:3
Accuracy	0.8895	0.8865
F-measure	0.8916	0.8880
Reliance		
	c:1,degree:1	c:0.5,gamma:4
Accuracy	0.9221	0.8923
F-measure	0.9229	0.8932

Hình 17: Kết quả kết hợp tham số đối với mô hình SVM

Table 14
Best three parameter combinations of random forest model and their performance on discrete-valued parameter setting data set.

	ntrees		
Nifty			
	30	120	20
Accuracy	0.8913	0.8973	0.8969
F-measure	0.8934	0.8990	0.9005
BSE-Sensex			
	20	90	110
Accuracy	0.8886	0.8981	0.9011
F-measure	0.8914	0.9012	0.9028
Infosys			
	50	60	70
Accuracy	0.9035	0.8964	0.9004
F-measure	0.9051	0.8980	0.9019
Reliance			
	30	10	40
Accuracy	0.9079	0.9088	0.9070
F-measure	0.9085	0.9098	0.9078

Hình 18: Kết quả kết hợp tham số đối với mô hình Random Forest

Cần lưu ý rằng khi dữ liệu được biểu diễn dưới dạng dữ liệu xác định xu hướng, phân lớp Naive-Bayes được học bằng cách phù hợp với phân phối đa biến Bernoulli cho dữ liệu. Kết quả về tập dữ liệu so sánh cho tất cả các mô hình được đề xuất được báo cáo trong Bảng dưới đây.

Table 15
Performance of prediction models on discrete-valued comparison data set.

Stock/Index	Prediction Models			
	ANN		SVM	
	Accuracy	F-measure	Accuracy	F-measure
S&P BSE SENSEX	0.8669	0.8721	0.8869	0.8895
NIFTY 50	0.8724	0.8770	0.8909	0.8935
Reliance Industries	0.8709	0.8748	0.9072	0.9080
Infosys Ltd.	0.8572	0.8615	0.8880	0.8898
Average	0.8669	0.8714	0.8933	0.8952
	Random forest		Naive-Bayes	
	Accuracy	F-measure	Accuracy	F-measure
S&P BSE SENSEX	0.8959	0.8985	0.8984	0.9026
NIFTY 50	0.8952	0.8977	0.8952	0.8990
Reliance	0.9079	0.9087	0.9222	0.9234
Infosys	0.9001	0.9017	0.8919	0.8950
Average	0.8998	0.9017	0.9019	0.9050

Hình 19: Kết quả so sánh cho tất cả mô hình

So sánh cuối cùng cho thấy tất cả các mô hình hoạt động tốt với dữ liệu đầu vào riêng biệt nhưng SVM, Random forest và Naive-Bayes hoạt động tốt hơn ANN. Độ chính xác của 3 mô hình trên là gần 90%.

CHƯƠNG 5 – KẾT LUẬN

5.1 Thảo luận

Dữ liệu thị trường chứng khoán là một ví dụ về dữ liệu không cố định. Tại thời điểm cụ thể có thể có xu hướng (trends), chu kỳ (cycles), "random walk" hoặc kết hợp cả ba. Người ta mong muốn rằng nếu một năm cụ thể là một phần của chu kỳ nói rằng một năm tăng thì mô hình của chúng ta nên theo mô hình này để dự đoán xu hướng. Tương tự có thể được xem xét cho một năm xu hướng.

Tuy nhiên, thông thường giá trị cổ phiếu của một năm cụ thể không bị cô lập và có những ngày "random walk". Giá trị cổ phiếu cũng bị ảnh hưởng bởi các yếu tố bên ngoài tạo ra xu hướng và tình trạng của nền kinh tế của đất nước. Chính trị cũng là yếu tố ảnh hưởng có thể dẫn đến chu kỳ.

Có thể thấy từ kết quả rằng tất cả các mô hình hoạt động tốt khi chúng được học từ các đầu vào có giá trị liên tục nhưng hiệu suất của từng mô hình được cải thiện hơn nữa khi chúng được học bằng cách sử dụng dữ liệu xác định xu hướng. Dữ liệu xác định xu hướng được chuẩn bị bằng cách phân biệt dữ liệu có giá trị liên tục. Ý tưởng này dựa trên thực tế là mỗi tham số có giá trị liên tục khi so sánh với giá trị ngày trước của nó cho thấy xu hướng tăng hoặc giảm trong tương lai. Các dữ liệu được rời rạc dựa trên các heuristic này. Khi dữ liệu này được cung cấp làm đầu vào cho mô hình, bài báo đã nhập xu hướng dựa trên từng tham số đầu vào.

Khi đưa ra các chỉ số kỹ thuật có giá trị liên tục làm đầu vào cho các mô hình, chúng sẽ bị tước đi, đây là xu hướng mà mỗi chỉ dẫn kỹ thuật đưa ra. Điều này khiến các mô hình dự đoán phân loại dựa trên giá trị của các chỉ số kỹ thuật này nhưng thông tin từ quá trình chuyển đổi giá trị của cổ phiếu bị mất và không được sử dụng bởi các mô hình dự đoán. Vì mục tiêu là dự đoán hướng di chuyển hoặc xu hướng, nên dữ liệu xác định xu hướng phù hợp hơn.

Ngoài ra, đối với bất kỳ cổ phiếu hoặc chỉ số nào cũng có kịch bản khi họ giao dịch ở một số giá trị là 200, sau đó do một số yếu tố bên ngoài, họ có thể bắt đầu giao

dịch ở mức giá cao hơn 400 và sau đó ổn định ở giá trị cao hơn đó. Nếu mô hình được cung cấp đầu vào có giá trị liên tục trực tiếp, thì có thể nó cố gắng thiết lập quan hệ giữa các giá trị trong 200 và trong 400 không yêu cầu xa hơn giá trị tuyệt đối của thay đổi. Do đó, dữ liệu xác định xu hướng về bản chất là rời rạc, là dấu hiệu thống kê về việc cổ phiếu được mua quá mức hay bán quá mức và độc lập với giá trị. Do đó, các tham số đầu vào này, khi được biểu diễn dưới dạng các xu hướng trong tương lai có thể đóng vai trò là thước đo tốt hơn về tình trạng cổ phiếu thay vì kịch bản khi chúng được gửi lại dưới dạng giá trị liên tục.

5.2 Kết luận

Nhiệm vụ tập trung trong bài viết này là dự đoán hướng di chuyển cho các cổ phiếu và chỉ số giá cổ phiếu. Hiệu suất dự đoán của bốn mô hình là ANN, SVM, Random forest và Naive-Bayes được so sánh dựa trên mười năm (2003 - 2012) dữ liệu của CNX Nifty, S & P BSE Sensex, Infosys Ltd. và Reliance Industries từ thị trường chứng khoán Ấn Độ.

Mười thông số kỹ thuật phản ánh tình trạng của chỉ số chứng khoán và giá cổ phiếu được sử dụng để tìm hiểu từng mô hình này. Lớp dữ liệu xác định xu hướng được sử dụng để chuyển đổi từng báo cáo kĩ thuật liên tục thành +1 hoặc -1 cho thấy khả năng tăng hoặc giảm tương lai có thể xảy ra trong tương lai.

Các thử nghiệm với dữ liệu liên tục cho thấy mô hình Bayes có hiệu suất thấp nhất (73,3%) và Random Forest cao nhất (83,56%). Hiệu suất của tất cả các mô hình này cải thiện đáng kể khi chúng được học thông qua dữ liệu xác định xu hướng. Độ chính xác của các mô hình ANN, SVM, Random Forest và Naive Bayes lần lượt là 86,69%, 89,33%, 89,98% và 90,19%.

Mười chỉ số kỹ thuật được sử dụng trong bài viết này để xây dựng nền tảng kiến thức, tuy nhiên, các biến động kinh tế vĩ mô khác như tỷ giá hối đoái, lạm phát, chính sách của chính phủ, lãi suất, v.v. cũng là một yếu tố có thể hữu ích trong việc quyết định xu hướng.

Thành công của phương pháp đề xuất dựa trên phương pháp đầu tư của con người, khuyến khích mô phỏng phương pháp ra quyết định của con người trong khi phát triển hệ thống chuyên gia và sử dụng thuật toán học máy cho các vấn đề trong các lĩnh vực khác nhau.

TÀI LIỆU THAM KHẢO

Tiếng Anh

1. Predicting stock and stock price index movement using Trend Deterministic Data Preparation and machine learning techniques - Jigar Patel, Sahil Shah, Priyank Thakkar ⇑, K Kotecha