ĐẠI HỌC QUỐC GIA TP HỒ CHÍ MINH

TRƯỜNG ĐẠI HỌC KHOA HỌC TỰ NHIÊN

KHOA CÔNG NGHỆ THÔNG TIN

BÁO CÁO MÔN HỌC MÁY HỌC

**Đề tài :**

**ỨNG DỤNG SUPPORT VECTOR MACHINE**

**VÀO VIỆC DỰ ĐOÁN XU HƯỚNG CỔ PHIẾU**

**Dựa trên bài báo :**

An SVM-based Approach for Stock Market Trend Prediction (Yuling Lin, Haixiang Guo and Jinglu Hu)

**GIẢNG VIÊN HƯỚNG DẪN :** TS. Trần Thái Sơn

**HỌC VIÊN THỰC HIỆN :**

Võ Quốc Bửu - 16C12010

Nguyễn Nguyên Khang - 16C12011

Đặng Văn Thành - 16C12012

**KHÓA :** 26

**NGÀNH : Khoa học máy tính**

TP.HCM, 04-2021

**Mục Lục**

[**Mục Lục** 1](#_Toc6945910)

[**Bảng tóm tắt** 1](#_Toc6945911)

[**Danh sách bảng** 2](#_Toc6945912)

[**Danh sách hình ảnh minh họa** 2](#_Toc6945913)

[**1.** **Mục tiêu bài báo** 3](#_Toc6945914)

[**1.1.** **Giới thiệu bài toán** 3](#_Toc6945915)

[**1.2.** **Mục tiêu bài toán** 3](#_Toc6945916)

[**1.3.** **Bài toán chính của bài báo** 3](#_Toc6945917)

[**1.4.** **Cách tiếp cận bài toán** 4](#_Toc6945918)

[**2.** **Support Vector Machines** 6](#_Toc6945919)

[**2.1.** **Tổng quan về SVM** 6](#_Toc6945920)

[**2.2.** **SVM bán tuyến tính** 7](#_Toc6945921)

[**3.** **Phương pháp đề xuất của bài báo** 8](#_Toc6945922)

[**3.1.** **Giới thiệu phương pháp** 8](#_Toc6945923)

[**3.1.1.** **Bộ lọc SVM tương quan lựa chọn đặc trưng** 9](#_Toc6945924)

[**3.1.2.** **SVM bán tuyến tính dự đoán xu hướng thị trường** 10](#_Toc6945925)

[**3.1.3.** **Hàm kernel bán tuyến tính thành phần** 13](#_Toc6945926)

[**3.2.** **Bảng tóm tắt công thức** 14](#_Toc6945927)

[**4.** **Thực nghiệm** 14](#_Toc6945928)

[**4.1.** **Dataset** 14](#_Toc6945929)

[**4.2.** **Mô tả dữ liệu đầu vào** 15](#_Toc6945930)

[**4.3.** **Kết quả thực nghiệm và phân tích** 16](#_Toc6945931)

[**4.3.1.** **Kết quả thử nghiệm lựa chọn và đánh giá đặc trưng** 16](#_Toc6945932)

[**4.3.2.** **Kết quả thực nghiệm của các mô hình dự đoán khác nhau** 18](#_Toc6945933)

[**4.3.3.** **Kết quả thực nghiệm của các hệ thống dự đoán khác nhau** 19](#_Toc6945934)

[**4.3.4.** **Kết quả thực nghiệm của các hệ thống có trọng số** 20](#_Toc6945935)

[**5.** **Demo** 21](#_Toc6945936)

[**6.** **Kết luận** 22](#_Toc6945937)

[References 24](#_Toc6945938)

**Bảng phân công nhiệm vụ**

|  |  |  |
| --- | --- | --- |
| **Họ và tên** | **Mã số** | **Nhiệm vụ** |
| Võ Quốc Bửu | 16C12010 | Giới thiệu bài báo và tổng quan SVM: mục 1, 2 |
| Nguyễn Nguyên Khang | 16C12011 | Phương pháp đề xuất của bài báo: mục 3 |
| Đặng Văn Thành | 16C12012 | Thực nghiệm, demo và kết luận: mục 4, 5, 6 |

**Bảng tóm tắt**

SME : Subject Matter Expert

AP : Affinity Propagation

OBV : On Balance Volume

PER : Price-earning ratio

ROI : Return On Investment

RSI : Relative Strength Index

MACD : Moving Average Convergence-Divergence

TAPI : Total amount per weighted stock price index

PSY : Psychological line

QP : Quadratic programming

SRM : Structural risk minimization

**Danh sách bảng**

[Bảng 1. Một số hệ thống tài chính chuyên gia dự đoán xu hướng cổ phiếu 5](file:///D:\Huai%20Ying\KHTN\Cao%20Hoc%202016-2018\K28\MayHoc_01\MayHoc_01.docx#_Toc6945860)

[Bảng 2. Bảng tóm tắt công thức 14](#_Toc6945861)

[Bảng 3. Dữ liệu nghiên cứu 15](file:///D:\Huai%20Ying\KHTN\Cao%20Hoc%202016-2018\K28\MayHoc_01\MayHoc_01.docx#_Toc6945862)

[Bảng 4. Kết quả thực nghiệm lựa chọn và đánh giá đặc trưng 18](file:///D:\Huai%20Ying\KHTN\Cao%20Hoc%202016-2018\K28\MayHoc_01\MayHoc_01.docx#_Toc6945863)

[Bảng 5. Kết quả thực nghiệm của các mô hình dự đoán khác nhau 19](file:///D:\Huai%20Ying\KHTN\Cao%20Hoc%202016-2018\K28\MayHoc_01\MayHoc_01.docx#_Toc6945864)

[Bảng 6. Kết quả thực nghiệm của các hệ thống dự đoán khác nhau 20](file:///D:\Huai%20Ying\KHTN\Cao%20Hoc%202016-2018\K28\MayHoc_01\MayHoc_01.docx#_Toc6945865)

[Bảng 7. Kết quả thực nghiệm của các hệ thống có trọng số khác nhau 21](file:///D:\Huai%20Ying\KHTN\Cao%20Hoc%202016-2018\K28\MayHoc_01\MayHoc_01.docx#_Toc6945866)

**Danh sách hình ảnh minh họa**

[Hình 1-1. Hệ thống dự đoán xu hướng thị trường điển hình 5](file:///D:\Huai%20Ying\KHTN\Cao%20Hoc%202016-2018\K28\MayHoc_01\MayHoc_01.docx#_Toc6931504)

[Hình 2-1. Lược đồ phân lớp SVM. 7](file:///D:\Huai%20Ying\KHTN\Cao%20Hoc%202016-2018\K28\MayHoc_01\MayHoc_01.docx#_Toc6931509)

[Hình 3-1. Nền tảng tổng quát của hệ thống dự đoán xu hướng chứng khoán dựa trên SVM 8](file:///D:\Huai%20Ying\KHTN\Cao%20Hoc%202016-2018\K28\MayHoc_01\MayHoc_01.docx#_Toc6931523)

[Hình 5-1. Tập dữ liệu HSX 22](file:///D:\Huai%20Ying\KHTN\Cao%20Hoc%202016-2018\K28\MayHoc_01\MayHoc_01.docx#_Toc6931532)

1. **Mục tiêu bài báo**
2. **Giới thiệu bài toán**

Từ xưa đến nay việc dự đoán về thị trường chứng khoán là một chủ đề hấp dẫn nhưng đầy thách thức đối với các nhà nghiên cứu trong nhiều lĩnh vực khác nhau bao gồm kinh tế,tài chính,lịch sử,toán học và máy tính.Vì thị trường chứng khoán là một hệ thống phức tạp,nhiều biến động và phi tuyến nên dự báo xu hướng thị trường chứng khoán được xem là một thách thức [1]. Dự báo thị trường được đặc trưng bới các yếu tố: cường độ dữ liệu,độ nhiễu,bất ổn và rất nhạy cảm với thông tin bên ngoài.Chẳng hạn thị trường châu Á đều chìm trong sắc đỏ ngay sau khi hội nghị thượng đỉnh Mỹ-Triều lần 2 đột ngột kết thúc.Điều này minh chứng giá cổ phiếu rất nhạy cảm với thông tin bên ngoài.

1. **Mục tiêu bài toán**

Mục tiêu của báo cáo này là đề xuất một mô hình dựa trên SVM để dự đoán xu hướng của thị trường chứng khoán.Đề xuất bao gồm hai phần : lựa chọn tính năng và mô hình dự đoán. Trong phần chọn lựa tính năng, phương pháp lọc SVM tương quan được áp dụng để xếp hạng và chọn lựa một tập con có các chỉ số tài chính tốt.Và các chỉ báo chứng khoán được đánh giá dựa trên bảng xếp hạng.*Trong phần mô hình dự đoán, một SVM bán tuyến tính được sử dụng để dự đoán xu hướng của thị trường chứng khoán dựa theo dữ liệu lịch sử bằng cách sử dụng tập hợp con của các chỉ số tài chính làm đầu vào có trọng số*.

1. **Bài toán chính của bài báo**

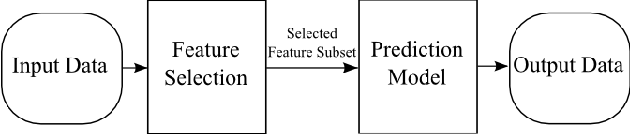
Mặc dù những hệ thống dự đoán xu hướng cổ phiếu hiện tại đã được chứng minh hiệu quả và có hiệu suất dự đoán xu hướng tốt trên thị trường chứng khoán. Tuy nhiên cũng có ba khía cạnh có thể được cải thiện. Đầu tiên, trong mô hình dự đoán SVM thường gặp rắc rối lớn trong vấn đề over-fitting khi các hàm kernel phi tuyến linh hoạt được sử dụng trong các tập dữ liệu có số chiều cao và độ nhiễu cao. Thứ hai, trong lựa chọn đặc trưng mặc dù các phương pháp chọn đặc trưng này có thể hiệu quả trong việc giảm số chiều và độ nhiễu của tập dữ liệu tài chính. Tuy nhiên lại có những hạn chế như không thể chọn chính xác một tập đặc trưng chứa các đặc trưng có tương quan cao với đầu ra mà không tương quan với nhau.Thứ ba, các hệ thống dự đoán xu hướng thị trường chứng khoán hiện có thường chỉ tập trung vào lựa chọn đặc trưng và mô hình dự đoán mà lại bỏ qua các đánh giá đặc trưng. Mặc dù đã có một số nghiên cứu quan tâm về đánh giá đặc trưng nhưng các phương pháp này bị ảnh hưởng bởi yếu tố chủ quan và tính đa cộng tuyến. Hiện tượng đa cộng tuyến là hiện tượng các biến độc lập trong mô hình phụ thuộc lẫn nhau và thể hiện được dưới dạng hàm số.

Để giải quyết những vấn đề tồn tại trên mà nhóm tác giả đã đề xuất một hệ thống dự đoán giá cổ phiếu dựa trên SVM ở Hình 3-1.Giải pháp này được diễn giải chi tiết ở mục 1.2 và mục 3.1

*Kết quả thực nghiệm trên bộ dữ liệu của thị trường chứng khoán Đài Loan chứng minh rằng mô hình dự đoán xu hướng chứng khóan dựa trên SVM có hiệu suất dự đoán chính xác hơn so với các phương pháp thông thường. Hơn nữa, kết quả thực nghiệm cũng cho thấy hệ thống dự đoán xu hướng thị trường chứng khoán dựa trên SVM được đề xuất có thể tìm ra một tập hợp con tốt và tìm ra các chỉ báo chứng khoán có thể cung cấp những thông tin hữu ích cho các nhà đầu tư*.

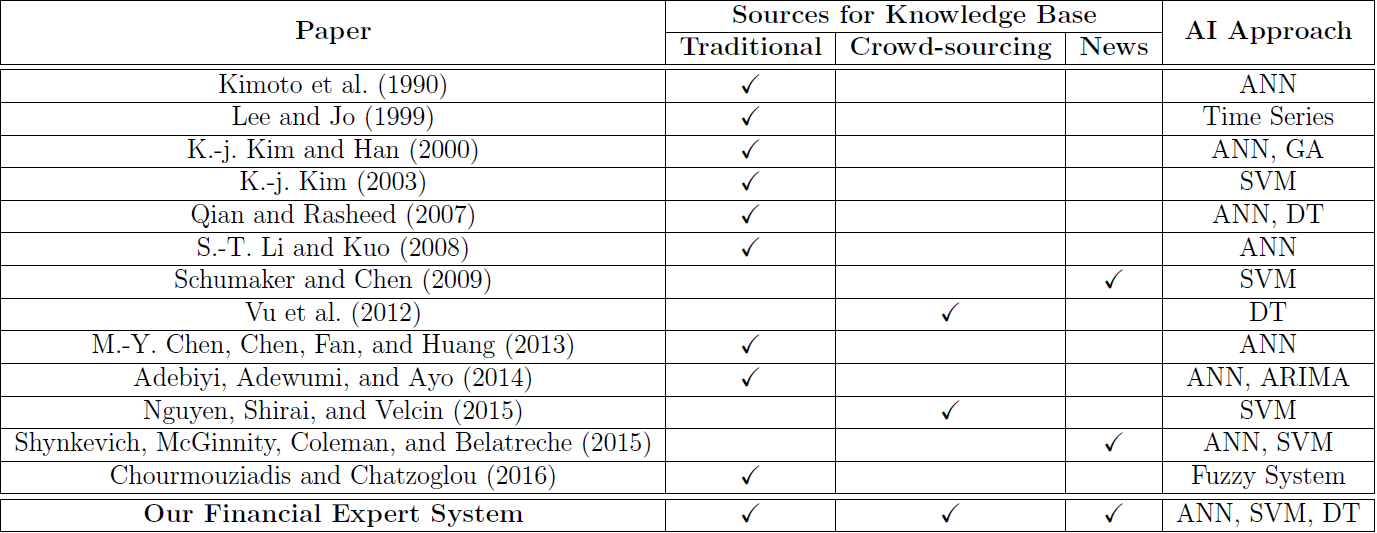
1. **Cách tiếp cận bài toán**

Một hệ thống dự đoán xu hướng thị trường chứng khoán điển hình thường bao gồm hai phần chính là lựa chọn đặc trưng và xây dựng mô hình dự đoán như trong Hình 1-1.



Hình 1-. Hệ thống dự đoán xu hướng thị trường điển hình

*Bảng 1 [1] giới thiệu một số hệ thống tài chính chuyên gia đã áp dụng các phương pháp máy học phổ biến như ANN,SVM,cây quyết định,thuật giải di truyền trong dự đoán xu hướng giá cổ phiếu trong thị trường chứng khoán*.



Bảng . Một số hệ thống tài chính chuyên gia dự đoán xu hướng cổ phiếu

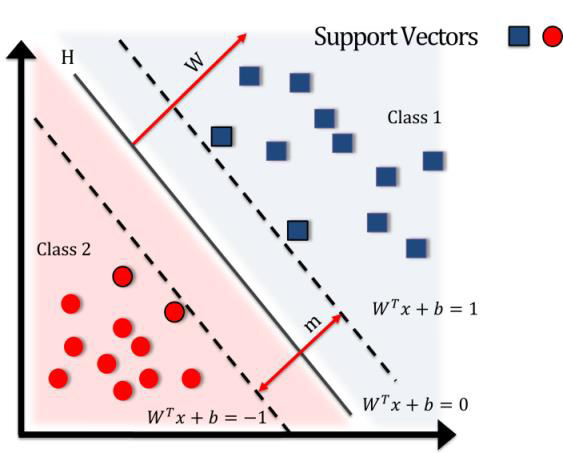
*Một cách tiếp cận dựa trên SVM được đề xuất để dự đoán xu hướng thị trường chứng khoán, kết hợp phương pháp lọc SVM tương quan và SVM bán tuyến tính. Phương pháp lọc SVM tương quan trước tiên được sử dụng để chọn và xếp hạng các đặc trưng đầu vào, tạo ra một tập hợp đặc trưng làm đầu vào có trọng số cho SVM bán tuyến tính để dự đoán xu hướng thị trường*.

1. **Support Vector Machines**
2. **Tổng quan về SVM**

Máy vectơ hỗ trợ (SVM) là một mô hình học có giám sát,một công cụ máy học mạnh mẽ được dùng để phân loại dữ liệu tuyến tính lẫn phi tuyến [2]. SVM có nguồn gốc từ thống kê xuất hiện từ những năm 1960. Bài báo đầu tiên về máy vectơ hỗ trợ đã được Vladimir Vapnik và các cộng sự Bernhard Boser và Isabelle Guyon trình bày vào năm 1992.

Một thuật toán SVM hoạt động như sau. Nó sử dụng ánh xạ phi tuyến để tăng số chiều của dữ liệu huấn luyện ban đầu. Trong kích thước mới này, nó tìm kiếm siêu phẳng phân tách tối ưu tuyến tính. Với ánh xạ phi tuyến riêng biệt đến kích thước đủ cao, dữ liệu từ hai lớp luôn có thể được phân tách bằng một siêu phẳng. SVM tìm thấy siêu phẳng này và lề bằng cách sử dụng các tập dữ liệu huấn luyện thiết yếu,các tập này được gọi là các vector hỗ trợ (Hình 2-1).

*Mặc dù thời gian đào tạo của các SVM nhanh nhất có thể cực kỳ chậm nhưng chúng có độ chính xác cao nhờ khả năng mô hình hóa các biên quyết định phi tuyến phức tạp. Các SVM ít bị over-fitting hơn nhiều so với các phương pháp khác. Các SVM được tìm thấy cũng cung cấp một mô tả vắn tắt về những mô hình đã học. Các SVM có thể được sử dụng để dự đoán số cũng như phân loại*. Chúng đã được áp dụng cho một số lĩnh vực bao gồm nhận dạng chữ số viết tay, nhận dạng đối tượng và nhận dạng người nói cũng như các bài kiểm chuẩn dự đoán chuỗi thời gian.



Hình 2-1. Lược đồ phân lớp SVM.

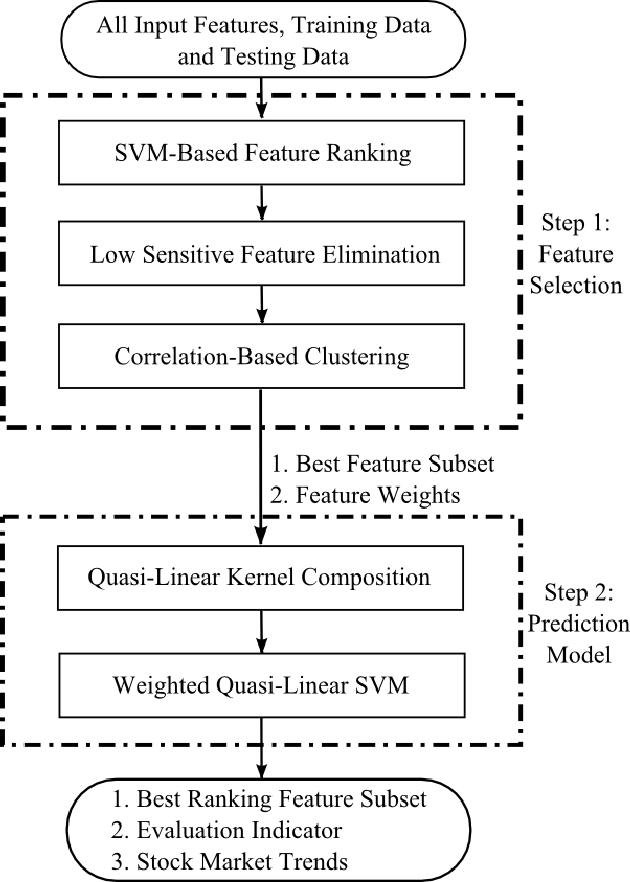
H là siêu phẳng phân lớp, W là vectơ bình thường của siêu phẳng, m là khoảng cách tối thiểu giữa các siêu phẳng dương và âm.

1. **SVM bán tuyến tính**

SVM bán tuyến tính là một SVM với hàm kernel bán tuyến tính đa hợp(công thức (7)), tính xấp xỉ một biên phân tách phi tuyến bằng các phân loại tuyến tính đa cục bộ với phương pháp nội suy [3](2).SVM bán tuyến tính được nhóm tác giả đề xuất để khắc phục hiện tượng over-fitting để cải thiện hiệu suất dự đoán. Over-fitting là một hiện tượng không mong muốn thường gặp trong máy học. Hiện tượng này xảy ra khi mô hình chỉ thể hiện sự chính xác trên tập huấn luyện nhưng lại kém chính xác trên tập kiểm tra. Hiện tượng over-fitting có thể được giải quyết bằng một hàm kernel thích hợp.Bằng phân đoạn tuyến tính,có thể linh hoạt điều chỉnh số lượng hàm phân đoạn tuyến tính [3](2)(4) để kiểm soát hiện tượng over-fitting

1. **Phương pháp đề xuất của bài báo**
2. **Giới thiệu phương pháp**

Hình 3-1 thể hiện một nền tảng tổng quát của hệ thống dự đoán xu hướng thị trường chứng khoán dựa trên SVM được đề xuất. Nó bao gồm hai phần: bộ lọc SVM tương quan để lựa chọn đặc trưng và một SVM bán tuyến tính để dự đoán xu hướng.



Hình 3-. Nền tảng tổng quát của hệ thống dự đoán xu hướng chứng khoán dựa trên SVM

1. **Bộ lọc SVM tương quan lựa chọn đặc trưng**

Nhiều chỉ số tài chính ảnh hưởng đến xu hướng thị trường chứng khoán nhưng không phải tất cả các chỉ số tài chính đều hữu ích cho dự đoán. Một số chỉ số tài chính có thể có ảnh hưởng hơn đến xu hướng thị trường chứng khoán trong khi những chỉ số khác lại ít liên quan hơn. Do đó, điều quan trọng là lựa chọn một tập con tốt của các chỉ số tài chính để thực hiện dự đoán tốt về xu hướng thị trường chứng khoán. Nói chung một tập con tốt không chỉ chứa sự đa dạng của các chỉ số tài chính được lựa chọn mà còn có tác động cao hơn đến xu hướng thị trường chứng khoán. Với mục đích này, nhóm tác giả áp dụng phương pháp lọc SVM tương quan được phát triển bởi họ để tìm ra một tập đặc trưng tốt chứa các đặc trưng có tương quan cao với đầu ra nhưng không tương thích với nhau.

Phương pháp lọc SVM tương quan bao gồm hai mô-đun chính: xếp hạng đặc trưng dựa trên SVM và phân cụm dựa trên tương quan. Xếp hạng đặc trưng dựa trên SVM không những thực hiện quy trình học có giám sát để hiện thực hóa xếp hạng đặc trưng dựa trên các độ nhạy Sensitivity trong SVM, giúp loại bỏ các chỉ số tài chính với số lượng ảnh hưởng thấp theo xu hướng thị trường chứng khoán mà còn đưa ra đánh giá cho chỉ báo tài chính được lựa chọn. Mặt khác,phân cụm dựa trên tương quan thực hiện thuật toán phân cụm AP để nhóm các chỉ số tài chính thành một số cụm dựa trên mối tương quan giữa các chỉ số tài chính. *Trong mỗi cụm, một chỉ số tài chính có tương quan cao với các chỉ số khác, nhưng tương quan thấp với các chỉ số tài chính trong các cụm khác. Vì vậy, các chỉ số tài chính từ cùng một cụm có đặc điểm tương tự. Từ mỗi cụm, một chỉ số tài chính làm đại biểu được lựa chọn dựa trên số lượng ảnh hưởng của nó trên đầu ra thu được trong mô đun xếp hạng đặc trưng dựa trên SVM*.

Theo cách này, phương pháp lọc SVM tương quan có thể lựa chọn một tập con tốt của các chỉ số tài chính và các chỉ số tài chính được lựa chọn được xếp hạng dựa trên ảnh hưởng của xu hướng thị trường chứng khoán. *Khi triển khai lọc SVM tương quan, hệ số tương quan tuyến tính được lựa chọn để đo lường mối tương quan của dữ liệu đầu vào và tham số ưu tiên p được xác định bởi phương pháp học Gunnar Rätsch để xác định số lượng cụm trong phương pháp phân cụm AP*.

Việc xếp hạng các chỉ số tài chính được lựa chọn cung cấp các chỉ số tài chính hữu ích cho các nhà đầu tư. Trong mô hình dự đoán, xếp hạng sẽ được sử dụng làm hệ số trọng số cho các đầu vào mô hình. Suy ra độ nhạy mô tả thứ hạng, trọng số có thể được tính bằng công thức :

|  |  |  |
| --- | --- | --- |
|  |  | (1) |

1. **SVM bán tuyến tính dự đoán xu hướng thị trường**

Vấn đề dự đoán xu hướng tăng và giảm trên thị trường chứng khoán đã được xem xét. Sau đó, các vấn đề phân loại nhị phân có thể được tập trung sự chú ý. Không gian đầu vào được ký hiệu là X trong đó kết hợp các trọng số đặc trưng,

Không gian đầu ra được ký hiệu là Y,. Thông thường, một tập dữ liệu huấn luyện minh họa N được biểu thị là:

Một biên tách phi tuyến có thể được xem như là một tập hợp của biên tuyến tính cục bộ M.

với phương pháp nội suy

|  |  |  |
| --- | --- | --- |
|  |  | (2) |

Trong đó x là trọng số đầu vào

Bằng cách tăng M, siêu phẳng phân tách phi tuyến có thể xấp xỉ bất kỳ độ mịn nào đủ do phương pháp nội suy với . Vai trò của hàm cơ sở cũng tương tự như hàm ngẫu nhiên.

Hai tham số vector và được xác định bởi :

|  |  |  |
| --- | --- | --- |
|  | = | (3) |

phương trình (2) có thể được thể hiện thêm dưới dạng:

|  |  |  |
| --- | --- | --- |
|  |  | (4) |

Theo quy trình của một SVM chuẩn, cực tiểu hóa rủi ro có cấu trúc SRM công thức (4)đã được áp dụng sau đó vấn đề tối ưu hóa quy hoạch toàn phương(QP) được mô tả :

|  |  |  |
| --- | --- | --- |
|  | Phụ thuộc | (5) |

Vấn đề tối ưu hóa QP ở trên có thể được tính bằng thuật toán Lagrangian. Sau đó, giải thuật toán Lagrangian thì thu được công thức kép :

|  |  |  |
| --- | --- | --- |
|  | Phụ thuộc | (6) |

,được tính từ công thức(6)

Hàm kernel bán tuyến tính đa hợp được xác định :

|  |  |  |
| --- | --- | --- |
|  |  | (7) |
|  |  |  |

Cho i=1…N.Do đó, mô hình tuyến tính đa cục bộ với phương pháp nội suy được giảm xuống thành một SVM chuẩn được gọi là SVM bán tuyến tính dựa trên hàm kernel bán tuyến tính đa hợp được định nghĩa bởi công thức (7). Cuối cùng, mô hình có thể được xác định bằng phân lớp SVM phi tuyến với hàm kernel đa hợp :

|  |  |  |
| --- | --- | --- |
|  |  | (8) |

1. **Hàm kernel bán tuyến tính thành phần**

SVM bán tuyến tính được mô tả và định nghĩa ở 2.2, Hàm kernel đa hợp (7) được xây dựng bằng cách sử dụng hàm cơ sở phù hợp với thông tin phân phối của dữ liệu huấn luyện. Bởi vì các hàm cơ sở được sử dụng để ước tính nội suy các siêu phẳng phân đoạn tuyến tính và mỗi hàm cơ sở tương ứng với một siêu phẳng tuyến tính cục bộ. Do đó, các hàm cơ sở nên trừu tượng hóa thông tin phân phối của tập dữ liệu càng chính xác càng tốt. Trong bài báo này, các tác giả đã chọn một hàm Gaussian làm hàm cơ sở để biểu diễn thông tin phân phối của không gian đầu vào.

|  |  |  |
| --- | --- | --- |
|  |  | (9) |

:trung tâm của cụm dữ liệu

:chiều rộng của cụm dữ liệu

ʎ: tham số tỷ lệ

*Thông tin phân phối là dữ liệu trung tâm cục bộ và độ rộng cục bộ. Thông tin phân phối của không gian đầu vào và số lượng hàm cơ sở M được thu thập bằng cách sử dụng phương pháp phân vùng tập dữ liệu*. Phương pháp phân vùng dữ liệu xem xét phân phối mẫu và đặc trưng trong vùng lân cận. Nó có thể được sử dụng để phân vùng các mẫu vào các tập con cục bộ.Để có được các phân vùng dọc theo biên phân tách, phương pháp phát hiện biên phân tách được áp dụng trước tiên để tìm ra biên phân tách.Sau đó một thuật toán phân cụm k-means được sử dụng để phân vùng các mẫu được chọn thành các tập con để ước tính dữ liệu trung tâm cục bộ, độ rộng cục bộ và số lượng hàm cơ sở M của hàm kernel đa hợp. Số hàm cơ sở được xác định là 3 và tham số tỷ lệ ʎ được đặt thành 0.1 (ʎ =0.1).

1. **Bảng tóm tắt công thức**

|  |  |
| --- | --- |
| **Diễn giải** | **Công thức** |
| Trọng số | (1) |
|  | (2) |
| Hai tham số vector , | (3) |
| Cực tiểu hóa rủi ro có cấu trúc(SRM) | (4) |
| Tối ưu hóa quy hoạch toàn phương(QP) | (5) |
| Phương pháp nhân tử Lagrange | (6) |
| Hàm kernel bán tuyến tính đa hợp | (7) |
| được xác định bằng phân lớp SVM phi tuyến với hàm kernel đa hợp | (8) |
| Hàm cơ sở | (9) |
| Độ đo hiệu suất dự đoán Hit ratio | (10) |

Bảng . Bảng tóm tắt công thức

1. **Thực nghiệm**
2. **Dataset**

Tập dữ liệu thực nghiệm được dùng trong nghiên cứu này bao gồm 3 công ty và 1 chỉ số Index của thị trường chứng khoán Đài Loan,tập dữ liệu này được sử dụng để kiểm tra tính hiệu quả của mô hình được đề xuất. Đó là công ty Hồng Hải (HHC), công ty bán dẫn Đài Loan (TSC), công ty Evergreen (EC) và Taiwan 50 Index (TW50). Dữ liệu lịch sử là dữ liệu hàng ngày và được lấy từ Taiwan Economic Journal Database (TEJD). Các bộ dữ liệu bao gồm dữ liệu từ tháng 7 năm 2008 đến tháng 1 năm 2012 với tổng số 880 quan sát. Các tập dữ liệu được chia thành hai thời kỳ, thời ký đầu chiếm 95% tập dữ liệu còn thời kỳ sau chiếm 5% còn lại. Thời kỳ đầu được chỉ định cho mô hình học tức là tập huấn luyện còn thời kỳ sau được sử dụng cho mô hình xác thực tức là tập thử nghiệm. Dữ liệu nghiên cứu được liệt kê trong Bảng 3.



Bảng 3. Dữ liệu nghiên cứu

1. **Mô tả dữ liệu đầu vào**

Trong nghiên cứu này, các chỉ báo kỹ thuật được sử dụng làm đặc trưng đầu vào vì nhóm tác giả muốn cố khai thác xu hướng giá cổ phiếu. *Những chỉ báo kỹ thuật có thể được thu thập thông qua các công thức toán học,những công thức toán học này cung cấp dấu vết về xu hướng của thị trường chứng khoán và thông tin ảnh hưởng đến giá vào thị trường trong một khoảng thời gian hữu hạn chứ không phải ngay lập tức.53 chỉ báo kỹ thuật được lựa chọn là các đặc trưng đầu vào được xác định bởi đánh giá của các chuyên gia nghiệp vụ(SME) và các nghiên cứu trước đó. Một số chỉ báo kỹ thuật quan trọng là chỉ báo cường độ dao động (OBV), hệ số P/E (PER), tỷ suất hoàn vốn (ROI), trung bình động 6 ngày (MA6), trung bình động 10 ngày (MA10), chỉ báo sức mạnh nội tại 6 ngày (RSI6), chỉ số sức mạnh nội tại 10 ngày (RSI10), chỉ báo dao động stochastic 9 ngày (K9) và (D9), phân kỳ và hội tụ của đường trung bình động (MACD), Chênh lệch (DIF), dự báo tự động (AR6), dự báo tự động 10 ngày (AR10), chỉ số BR 10 ngày (BR10), chỉ số BR 20 ngày (BR20), mức độ biến động 10 ngày (VR10), mức độ biến động 20 ngày (VR20), độ lệch 6 ngày (BIAS6), độ lệch 10 ngày ( BIAS10), cường độ dao động Williams% R 6 ngày (WMS6), cường độ dao động William R% 10 ngày (WMS10), (TAPI), dòng tâm lý 6 ngày (PSY6), dòng tâm lý 10 ngày (PSY10) [4] [5]*.

1. **Kết quả thực nghiệm và phân tích**

Trong các thực nghiệm, đầu tiên nhóm tác giả thể hiện kết quả của việc lựa chọn tập đặc trưng và đánh giá đặc trưng bằng phương pháp lọc SVM tương quan. Thứ hai, nhóm tác giả điều tra khả năng dự đoán của xu hướng giá thị trường chứng khoán bằng cách sử dụng SVM với năm loại hàm kernel. Bao gồm kernel tuyến tính (), kernel RBF (), kernel đa thức (), kernel sigmoid () và kernel bán tuyến tính (). Kernel tuyến tính đã được đề xuất trong nghiên cứu trước đó của họ. Thứ ba, tập hợp đặc trưng tốt nhất được chọn lựa được sử dụng cho năm mô hình dự đoán trên để đánh giá phương pháp lọc SVM tương quan (C-SF). Cuối cùng, nhóm tác giả so sánh hiệu suất của hệ thống dự đoán được đề xuất(C-SF+)với bốn hệ thống dự báo khác.Nhóm tác giả cũng giới thiệu SVM có trọng số để xây dựng 4 hệ thống dự báo được dùng để so sánh : C-SF+, C-SF+

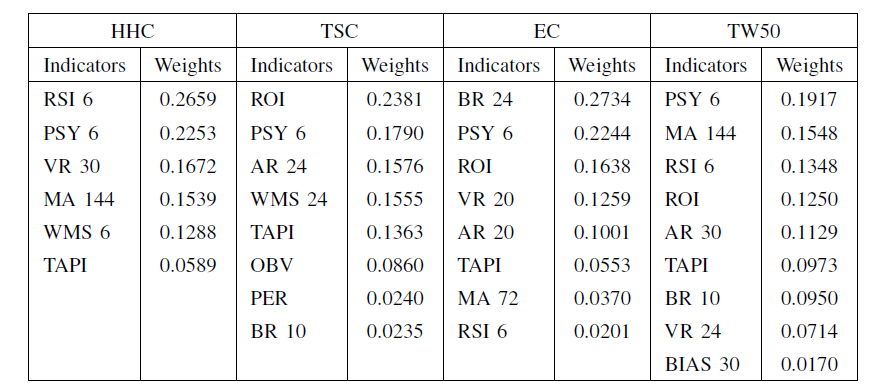
1. **Kết quả thử nghiệm lựa chọn và đánh giá đặc trưng**

Kết quả lựa chọn đặc trưng và đánh giá đặc trưng được tạo ra bằng phương thức lọc SVM tương quan. Trong phương thức lọc SVM tương quan, tham số ưu tiên p xác định có bao nhiêu đặc trưng được chọn lựa. Trong nghiên cứu này, phương pháp học của Gunnar Rätsch đã được áp dụng để xác định tham số ưu tiên p. Đối với HHC p = -2, TSC p = -3, EC p = -1,5 và TW50 p = -1.1. *Nhóm tác giả có tham số ưu tiên p để xây dựng từng tập hợp đặc trưng. Các tập hợp đặc trưng được chọn được hiển thị trong Bảng 4. Trong Bảng 4, sáu đặc trưng được chọn làm chỉ báo chính trong số 53 đặc trưng trong dữ liệu HHC. Điều này làm giảm số chiều của dữ liệu cho dữ liệu HHC. Đối với dữ liệu TSC và dữ liệu EC, phương thức lọc SVM tương quan chọn 8/53 đặc trưng. Lượng giảm dữ liệu đạt . Đối với dữ liệu TW50, lựa chọn 9/53 đặc trưng. chiều dữ liệu bị giảm. Điều này chỉ ra rằng chúng ta có thể giảm các yêu cầu thu thập, tính toán và lưu trữ dữ liệu thông qua phương pháp lọc SVM dựa trên tương quan*.

Nhìn vào Bảng 4 để thấy các đặc trưng chính quyết định xu hướng của giá cổ phiếu bằng các trọng số được chỉ định trên mỗi tập dữ liệu thị trường chứng khoán. Chúng tôi một phần thuộc tính phát hiện này cho mối tương quan mạnh mẽ giữa các tính năng liên quan đến chuyển động giá. Chúng tôi sắp xếp tìm kiếm của chúng tôi như sau :

* Hai chỉ báo PSY6 và TAPI được lựa chọn bởi tất cả các phân vùng dữ liệu và bốn tập dữ liệu thị trường chứng khoán được thử nghiệm. *Điều này có ý nghĩa đáng kể vì hai chỉ báo kỹ thuật này có liên quan đáng kể đến biến động giá cổ phiếu trong tương lai từ phân tích này. Do đó,những người ra quyết định đầu tư sẽ tăng thêm trọng lượng cho hai đặc trưng này để dự đoán trong tương lai. Hơn nữa, chỉ báo RSI6, MA144 và chỉ báo BR10 xuất hiện nhiều hơn một lần, phát hiện này cũng phải được chú ý*.
* Chín loại chỉ báo kỹ thuật PSY, TAPI, BR, VR, MA, AR, RSI , WMS, ROI rất quan trọng trong dự đoán xu hướng thị trường chứng khoán.

Những phát hiện này rất có ý nghĩa đối với những người ra quyết định và nhà đầu tư bởi vì phát hiện này có thể cho biết rằng họ nên phụ thuộc chủ yếu vào chỉ số kỹ thuật nào khi đầu tư vào thị trường chứng khoán.



Bảng 4. Kết quả thực nghiệm lựa chọn và đánh giá đặc trưng

1. **Kết quả thực nghiệm của các mô hình dự đoán khác nhau**

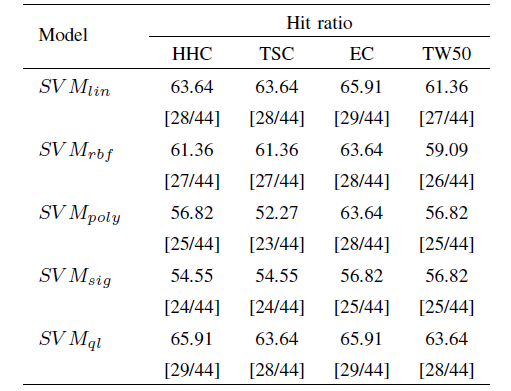
*Nhóm tác giả nghiên cứu khả năng dự đoán về xu hướng của thị trường chứng khoán bằng cách sử dụng SVM với năm loại hàm kernel theo các tập dữ liệu thị trường chứng khoán gốc. Hiệu suất tương đối của các mô hình được đo bằng Hit-Ratio.Hit-Ratio là tỷ lệ phần trăm chính xác của dự đoán xu hướng*. Hit-Ratio được sử dụng để đo hiệu suất dự đoán. Hit-Ratio được xác định bằng công thức (10) .Các kết quả thực nghiệm được thể hiện trong Bảng 5.

|  |  |  |
| --- | --- | --- |
|  |  | (10) |

N là số lượng ví dụ thử nghiệm

Các kết quả thực nghiệm cho thấy độ chính xác dự báo của hàm kernel tuyến tính vượt trội hơn so với các hàm kernel phi tuyến khác (hàm kernel RBF, hàm kernel đa thức, hàm kernel sigmoid).

Một lý do mà SVM với hàm kernel phi tuyến có thể xảy ra vấn đề over-fitting do dữ liệu tài chính được đặc trưng bởi tính biến động cao, bất quy tắc và nhiễu. Hiệu suất của hàm kernel bán tuyến tính cho thấy độ chính xác dự báo cao hơn các hàm kernel thông thường khác. Nó được thực hiện bằng hàm phân đoạn tuyến tính, có thể điều chỉnh linh hoạt số lượng hàm phân đoạn tuyến tính để kiểm soát hiện tượng over-fitting.



Bảng . Kết quả thực nghiệm của các mô hình dự đoán khác nhau

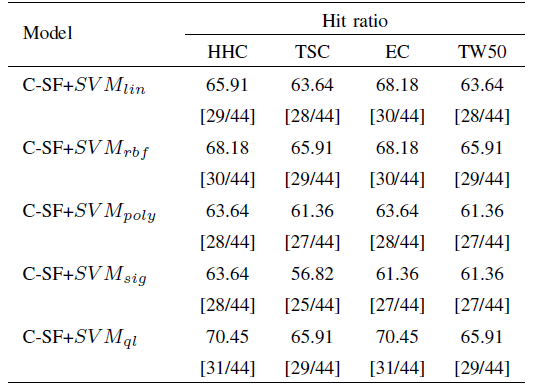
1. **Kết quả thực nghiệm của các hệ thống dự đoán khác nhau**

Tập hợp tính năng tốt nhất được giới thiệu để xây dựng các mô hình dự đoán, được tạo bằng phương pháp lọc SVM tương quan (C-SF). Các kết quả thực nghiệm được thể hiện trong Bảng 6. So sánh với kết quả của Bảng 5, hiệu suất của tất cả các hàm kernel được cải thiện đáng kể, phương pháp lọc SVM tương quan có tác động đáng kể đến tập dữ liệu thị trường chứng khoán, giúp giảm số chiều và độ nhiễu dữ liệu. Trong các thực nghiệm, nhóm tác giả cũng thấy rằng hiệu suất của các hàm kernel RBF có thể vượt quá hàm kernel tuyến tính, điều này thể hiện các tập dữ liệu gốc ban đầu là số chiều dữ liệu cao và độ nhiễu cao. Ngoài ra, hàm kernel thành phần vẫn tạo ra hiệu suất dự đoán tốt.

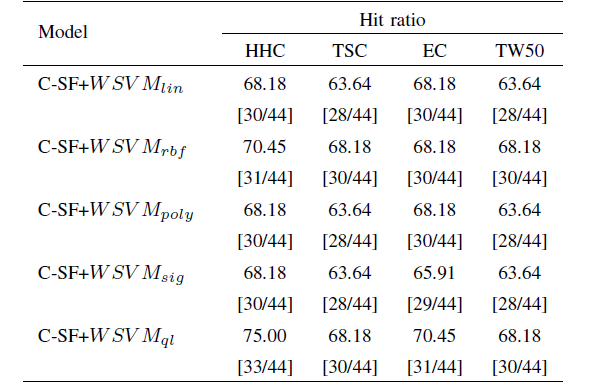
1. **Kết quả thực nghiệm của các hệ thống có trọng số**

*Nhóm tác giả giới thiệu tập con đặc trưng tốt nhất và các trọng số tương ứng cho phương pháp SVM bán tuyến tính để xây dựng hệ thống dự đoán xu hướng thị trường chứng khoán dựa trên SVM được đề xuất.*

*Sau đó, họ so sánh hiệu suất của hệ thống dự đoán được đề xuất với bốn hệ thống dự báo tương đương. Các kết quả được thể hiện trong Bảng 7. Kết quả thực nghiệm của các hệ thống có trọng số. Kết quả thực nghiệm cho thấy độ chính xác dự báo của hệ thống được đề xuất đạt được hiệu suất tốt nhất. Điều này xác minh rằng phương pháp đề xuất thực sự sở hữu khả năng không thiên vị và mạnh mẽ. Nó tránh các hiện tượng over-fitting của SVM thông thường trong tập dữ liệu tài chính*. *Hơn nữa, bằng cách so sánh kết quả với Bảng 6, họ biết rằng các trọng số đặc trưng thực sự có thể có liên quan nhiều hơn đến thông tin phân loại trên tập dữ liệu thị trường tài chính, đặc biệt là tại các hàm kernel phi tuyến*.



Bảng . Kết quả thực nghiệm của các hệ thống dự đoán khác nhau



Bảng . Kết quả thực nghiệm của các hệ thống có trọng số khác nhau

1. **Demo**

Demo này minh họa dùng SVM để dự đoán xu hướng giá cổ phiếu bằng cách đo độ chính xác của 4 hàm kernel RBF,linear,poly,sigmoid với các khoản thời gian giao dịch(7,14,21,30,60,90,120,180…..ngày) khác nhau qua các thang đo Precision,ReCall,F1-score,support.

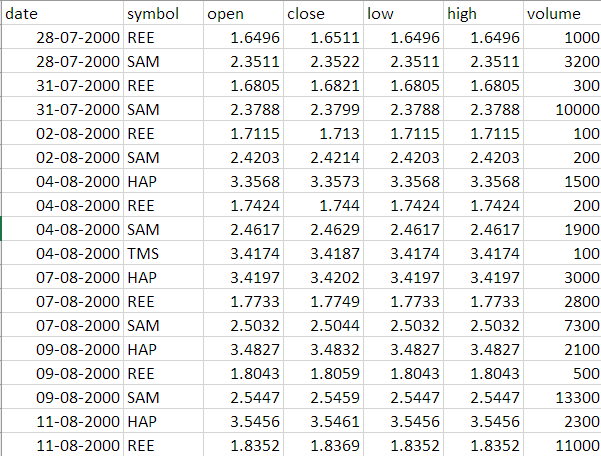
Nguồn:<https://www.kaggle.com/rosand/fork-of-predict-stock-prices-with-svm/>(Source.pdf)

Dataset(HSX.csv) : tập dữ liệu lưu trữ lịch sử giá 413 mã cổ phiếu trên sàn HOSE(HSX) trong giai đoạn 28-07-2000 đến 29-03-2019(Hình 5-1)

Input : mã cổ phiếu,các chỉ báo kỹ thuật được tính tự động từ tập dữ liệu trên, số ngày giao dịch(7,14,21,30,60,90,120,150,180….ngày).

Output(Result.pdf) : dự báo xu hướng Drop/Down của cổ phiếu bằng cách đánh giá độ chính xác của 4 hàm kernel bằng các độ đo precision,recall,f1-score,support

Hình 5-. Tập dữ liệu HSX



1. **Kết luận**

Nghiên cứu này đề xuất một hệ thống dựa trên SVM dự đoán xu hướng thị trường chứng khoán. Hệ thống này chọn lựa một tập hợp đặc trưng tốt, đánh giá các chỉ báo chứng khoán và kiểm soát hiện tượng over-fitting đối với dự đoán xu hướng thị trường chứng khoán. Hệ thống đề xuất có ba lợi thế khác biệt :

* *Đầu tiên là tập đặc trưng tốt được chọn, trong đó có các đặc trưng có tương quan cao với đầu ra, nhưng không tương quan với nhau*.
* *Hai là các đặc trưng được lựa chọn được đánh giá và xếp hạng. Việc lựa chọn đặc trưng và đánh giá đặc trưng được thực hiện đồng thời bằng phương pháp lọc SVM dựa trên tương quan.Nó làm giảm số chiều và độ nhiễu của dữ liệu tài chính cũng như cung cấp thêm thông tin đầu tư hữu ích cho các chuyên gia và nhà đầu tư trên thị trường chứng khoán để họ quyết định đầu tư*.
* *Cuối cùng hệ thống đề xuất sở hữu khả năng mạnh mẽ do phương pháp đề xuất được thực hiện bằng các hàm phân đoạn tuyến tính và các trọng số đặc trưng được kết hợp để xây dựng siêu phẳng tách tối ưu*.

Phương pháp đề xuất đã được thử nghiệm trên bộ dữ liệu của thị trường chứng khoán Đài Loan. Kết quả mô phỏng cho thấy hệ thống của nhóm tác giả tạo ra hiệu suất mạnh mẽ, vượt trội so với các hệ thống dự đoán khác. Nó tránh được các hiện tượng over-fitting của các hệ thống dự đoán xu hướng thị trường chứng khoán thông thường.

*Nghiên cứu của nhóm tác giả có các giới hạn cần được nghiên cứu thêm. Thứ nhất, có nhiều yếu tố ảnh hưởng đến giá cổ phiếu. Vì vậy, trong tương lai chúng ta có thể thử kiểm tra nhiều yếu tố ảnh hưởng hơn với hệ thống dự đoán này. Thứ hai, trong hệ thống dự đoán được đề xuất, việc thiết lập các tham số có tác động quan trọng đến hiệu suất của hệ thống kết quả. Chúng ta cũng cần nghiên cứu để phát triển một phương pháp có cấu trúc để chọn một giá trị tối ưu của các tham số trong đề xuất hệ thống dự đoán cho hiệu suất dự đoán tốt nhất*.

# References

|  |  |
| --- | --- |
| [1] | B. Weng, "Application of machine learning techniques for stock market prediction," Auburn, Alabama, May 6, 2017. |
| [2] | Jiawei Han,Micheline Kamber,Jian Pei, Data Mining: Concepts and Techniques, 3rd ed., MA 02451, USA: Morgan Kaufmann, 2012. |
| [3] | Benhui Chen, Feiran Sun,Jinglu Hu, "Local Linear Multi-SVM Method for Gene Function Classification," *Proceedings of the World Congress on Nature and Biologically Inspired Computing,* 2010. |
| [4] | S. B. Achelis, Phân tích Kỹ thuật từ A đến Z(Vietstock Media Company Ltd dịch), Nhà xuất bản Thông tấn, 2014. |
| [5] | S.-M. Jhao, Taiwan Stock Forecasting with the Genetic Programming, Kaohsiung 80424, Taiwan: Department of Computer Science and Engineering National Sun Yat-sen University, August 2011. |
| [6] | Rafael Rosillo,Javier Giner,David De la Fuente, "Stock Market Simulation using Support Vector Machines," *Journal of Forecasting,* vol. 33, no. 6, pp. 488-500, September 2014. |
| [7] | K.-j. Kim, "Financial time series forecasting using support vector machines," *Neurocomputing,* vol. 55, no. 1-2, pp. 307-319, September 2003. |
| [8] | J. L. Crowley, "Kernel Functions and Support Vector Machines," 17 February 2016. |
| [9] | S. Madge, "Predicting Stock Price Direction using Support Vector Machines," *Independent Work Report Spring 2015,* 2015. |
| [10] | Klaus-Robert Müller, Alexander J. Smola, Gunnar Rätsch, Bernhard Schölkopf, Jens Kohlmorgen, Vladimir Vapnik, "Predicting time series with support vector machines," in *ICANN '97 Proceedings of the 7th International Conference on Artificial Neural Networks*, 1997. |