**PHƯƠNG PHÁP DỰ ĐOÁN XU HƯỚNG CỔ PHIẾU DỰA TRÊN VIỆC KẾT HỢP K-MEANS VÀ SVM VỚI ƯỚC LƯỢNG XÁC SUẤT LỚP**

**Vạn Duy Thanh Long, Lê Minh Duy, Nguyễn Hoàng Tú Anh**

Trường Đại Học Khoa Học Tự Nhiên, ĐHQG - HCM

TÓM TẮT: Trong bài báo này, chúng tôi đề xuất một mô hình kết hợp giữa K-Means và SVM với ước lượng xác suất lớp, gọi là K-SVMeans, để giải quyết bài toán phân lớp cho dự đoán xu hướng các mã cổ phiếu ở thị trường Việt Nam. Việc dự đoán được thực hiện không chỉ ở chu kỳ 1 ngày tiếp theo mà cho cả chu kỳ 5 ngày. Để đánh giá độ chính xác của mô hình đề xuất, chúng tôi so sánh hiệu quả dự đoán với mô hình mạng nơ-ron lan truyền ngược, mô hình SVM truyền thống và mô hình SVM với ước lượng xác suất lớp. Kết quả cho thấy, mô hình của chúng tôi vượt trội so với những mô hình trên, đồng thời cho thấy khả năng dự đoán tương đối hiệu quả ở chu kỳ 5 ngày.

*Từ khóa*: *SVM, ước lượng xác suât, K-Means, mạng nơ-ron nhân tạo, dự đoán xu hướng cổ phiếu.*

# GIỚI THIỆU

Trong thị trường chứng khoán, việc dự đoán được xu hướng của thị trường là một đề tài hấp dẫn đối với cả giới nghiên cứu và nhà đầu tư tài chính. Để dự đoán được bước đi của thị trường ở một tỷ lệ chính xác cao là rất khó [21], vì dữ liệu chuỗi thời gian biến động phi tuyến ở mức cao. Đối với các nhà nghiên cứu, việc chọn được một mô hình phù hợp cho mục đích trên là một vấn đề nhiều thách thức. Thông thường, trong những cách tiếp cận bài toán dự đoán xu hướng, người ta chọn xem xét nó như một bài toán phân hai lớp: xu hướng tăng và xu hướng giảm.

Gần đây, các nghiên cứu đạt được hiệu quả cao thường đi theo hướng tính toán mềm và khai thác dữ liệu [8]. Rất nhiều nghiên cứu sử dụng mạng nơ-ron nhân tạo (ANN) và sau đó là các cải tiến của nó cho dự đoán xu hướng [12], [19], [20]. Từ những thành tựu ban đầu, người ta xây dựng các hệ lai giữa ANN và tính toán mờ [1], [3], ANN và thuật giải di truyền [13]. Tuy nhiên, hướng tiếp cận này tồn tại một số nhược điểm. Mô hình mạng nơ-ron nhân tạo đặc biệt là mạng lan truyền ngược (BPNN) cần dữ liệu huấn luyện lớn và khó tổng quát hóa được dữ liệu vì thường xảy ra quá khớp bởi những giá trị tối ưu cục bộ mà mô hình đạt được. Cùng lúc đó, máy học véc tơ hỗ trợ (SVM), được phát triển bởi Vapnik (1995), là một hướng mới được ưa chuộng nhờ khả năng tổng quát hóa cao, hiệu quả đối với dữ liệu nhiều chiều và biến động phi tuyến. Các kết quả của mô hình SVM cùng những cải tiến cho thấy hiệu quả dự đoán rất tốt. Mô hình SVM cải tiến bằng việc kết hợp với tính toán mờ và thuật giải di truyền cho kết quả tốt hơn BPNN với cùng cải tiến trong việc dự đoán xu hướng chứng khoán ở thị trường Đài Loan [6]. Sự kết hợp giữa SVM với phương pháp trích chọn đặc trưng cũng đem lại kết quả vượt trội so với BPNN truyền thống trong việc dự đoán xu hướng chỉ số NASDAQ [14]. Bên cạnh đó, bằng cách áp dụng lý thuyết hộp vào mô hình SVM hồi quy (SVR), hệ thống hỗ trợ ra quyết định mua bán các cổ phiếu trong danh mục S&P 500 cũng đem về lợi nhuận cao hơn hẳn so với phương pháp mua và giữ [23]. Ngoài ra, với mục đích nâng cao độ chính xác [9], [10] đã áp dụng mô hình dự đoán giá qua hai giai đoạn bằng cách áp dụng SOM để chia dữ liệu đầu vào thành các cụm; sau đó, xây dựng nên các mô hình SVR ứng với từng cụm riêng biệt. Cả hai mô hình đề xuất đều cho kết quả tốt hơn mô hình SVM cơ bản.

Đối với thị trường chứng khoán non trẻ và đang trong giai đoạn phát triển như Việt Nam, có rất ít nghiên cứu theo hướng tính toán mềm được áp dụng thành công. Việc giải quyết bài toán dự đoán xu hướng cổ phiếu với hai nhãn lớp tăng và giảm cũng không phản ánh hết đặc trưng của xu hướng cổ phiếu Việt Nam, do giá cổ phiểu trong rất nhiều giai đoạn được nhận định là không có xu hướng. Vì vậy, trong nghiên cứu này, chúng tôi sẽ giải quyết bài toán phân ba lớp: xu hướng tăng, xu hướng giảm và không có xu hướng. Vận dụng ý tưởng của phương pháp hai giai đoạn đã từng thành công ở các nghiên cứu trước chúng tôi đề xuất mô hình dự đoán xu hướng dựa trên việc kết hợp phương pháp phân cụm K-Means và bộ phân lớp SVM với ước lượng xác suất lớp, gọi là K-SVMeans. Đầu tiên, các mẫu dữ liệu đầu vào sẽ được gom cụm bằng K-Means để quy những dữ liệu gần giống nhau vào cùng một cụm. Sau đó, mỗi cụm sẽ được huấn luyện để cho ra một mô hình SVM với việc ước lượng xác suất lớp. Khi có được mô hình đã huấn luyện, quá trình dự đoán được thực hiện như sau: Sau khi xác định được mẫu dữ liệu thử nghiệm thuộc vào cụm nào, ta dùng chính mô hình SVM của cụm đó để dự đoán. Giá trị đầu ra chính là ba giá trị xác suất mà mẫu đó thuộc về các lớp. Kết quả phân lớp sẽ được quyết định bằng cách chọn ra xác suất lớn nhất trong ba giá trị trên. Việc lựa chọn phương pháp ước lượng xác suất lớp thay cho chiến lược bỏ phiếu trong mô hình SVM truyền thống là vì nó được chứng minh là tốt hơn trên cả lý thuyết lẫn thực nghiệm [24]. Ngoài ra, trong nghiên cứu này, chúng tôi sẽ thực hiện dự đoán ở hai chu kỳ: 1 ngày tiếp theo và 5 ngày tiếp theo. Để đánh giá hiệu quả của mô hình đề xuất, chúng tôi so sánh kết quả với SVM truyền thống, SVM có ước lượng xác suất lớp, và mô hình BPNN, đồng thời khảo sát để chọn số cụm tối ưu cho bước phân cụm dùng K-Means.

Các phần tiếp theo của bài báo được tổ chức như sau. Phần 2 trình bày về mô hình SVM có ước lượng xác suất lớp. Trong phần 3, mô hình đề xuất sẽ được mô tả chi tiết. Phần 4 là thực nghiệm và đánh giá kết quả. Phần cuối cùng, chúng tôi sẽ đưa ra kết luận và hướng phát triển sắp tới.

# MÔ HÌNH SVM VỚI ƯỚC LƯỢNG XÁC SUẤT

## **Bô phân lớp SVM**

SVM là một phuơng pháp phân lớp dựa trên lý thuyết học thống kê, đuợc đề xuất bởi Vapnik (1995) [22]. Cho tập véctơ đầu vào và tập các giá trị nhãn lớp tuơng ứng cho mô hình phân lớp nhị phân. Hàm tuyến tính phân biệt hai lớp như sau:

(1)

trong đó, w là véctơ chuẩn của siêu phẳng phân cách, b là độ lệch, và (x) là hàm ánh xạ từ không gian đầu vào sang không gian đặc trưng, (M > D). Mục tiêu của SVM là tìm một siêu phẳng tối ưu sao cho khoảng cách lề giữa hai lớp đạt giá trị cực đại. Bên cạnh đó, để đảm bảo tính tổng quát hóa cao, một biến lỏng (slack variable) được đưa vào để nới lỏng điều kiện phân lớp. Bài toán đưa đến việc giải quyết tối ưu có ràng buộc:

(2)

sao cho:

trong đó, C > 0 là tham số chuẩn tắc (regularization parameter), là biến lỏng.

Bài toán (2) có thể đựợc giải bằng phuơng pháp SMO [17] (Sequential Minimal Optimization). Phuơng pháp này đưa đến giải bài toán đối ngẫu quy hoạch toàn phương (QP-Quadratic Programming):

(3)

Thỏa mãn: và với αi là các nhân tử Lagrange.

Sau khi có được các giá trị từ bài toán (3), ta sẽ thu đươc các giá trị tối ưu w\* và b\* của siêu phẳng. Chỉ có các mẫu có mới tham gia vào các véc tơ hỗ trợ (support vector). Cuối cùng, hàm quyết định phân lớp có dạng:

(4)

Gọi là hàm nhân của không gian đầu vào. Theo đó, tích vô huớng trong không gian đặc trưng tuơng đương với hàm nhân K ở không gian đầu vào. Như vậy, thay vì tính trực tiếp giá trị tích vô huớng, ta thực hiện gián tiếp thông qua K. Với nhận định dữ liệu chứng khoán biến đổi một cách phi tuyến [5], ta chọn hàm nhân cho mô hình là hàm phi tuyến Gauss (RBF-Radial Basis Function):

(5)

## **Ước lượng xác suất**

Với việc áp dụng bộ phân lớp SVM, để phân k lớp, ta sử dụng cách tiếp cận một đối một (one-against-one, còn gọi pairwise). Vì vậy, sẽ có k(k-1)/2 bộ phân lớp được xây dựng để phân biệt mỗi cặp lớp. Trong SVM truyền thống, để dự đoán được nhãn lớp, ta áp dụng chiến lược bỏ phiếu. Nghĩa là, ta sẽ xây dựng một luật để phân biệt từng cặp lớp rồi chọn lớp thuộc về cặp chiến thắng nhờ hàm quyết định. Tuy nhiên, trong [24], Wu (2004) đã đề xuất mô hình ước lượng xác suất cho việc phân k lớp và chứng minh cả trên mặt lý thuyết và thực nghiệm, mô hình đề xuất tốt hơn chiến lược bỏ phiếu. Ý tưởng cơ bản của phương pháp này là thay vì quyết định trực tiếp nhãn lớp của một mẫu, nó thực hiện quá trình ước lượng để cho ra xác suất thuộc về từng lớp của mẫu đó để từ đó ta xác định được nhãn lớp ứng với xác suất lớn nhất.

Cho véc tơ đầu vào x và nhãn lớp y, mục tiêu là ước lượng xác suất

. (6)

Theo cách tiếp cận một đối một, đã tồn tại ước lượng xác suất cặp lớp (pairwise class probability) rij:

(7)

Từ lớp thứ i và thứ j của tập huấn luyện, Platt (2000) đã đưa ra công thức xấp xỉ rij [18] bằng hàm sigmoid:

, (8)

trong đó, f là hàm quyết định của x, A và B được xấp xỉ bằng cách cực tiểu hóa hàm log-likelihood (với N+ của giá trị yk dương, N- của yk âm):

(9)

trong đó, ,

và , n = 1, …, N.

Sau khi có tất cả giá trị rij, ta tính các giá trị pi trong (6) thông qua việc giải quyết bài toán tối ưu có ràng buộc:

(10)

với ràng buộc .

Ta có thể viết lại (10) theo dạng:

, (11)

với

Wu (2004) đã chứng minh ràng buộc không cần thiết. Từ việc bỏ đi ràng buộc trên, suy ra được điều kiện tối ưu tồn tại một vô hướng b (nhân tử Lagrange cho ràng buộc ) như sau:

(12)

với e là véc tơ k×1 có các phần tử lấy giá trị một, 0 là véc tơ k×1 với các phần tử có giá trị không. Thay vì giải hệ tuyến tính (12), ta đưa về phương pháp lặp đơn giản. Vì

(13)

cho nên **p** tối ưu thỏa mãn

(14)

Thuật toán giải quyết (14) được mô tả như sau [24]:

Bước 1. Khởi gán p thỏa .

Bước 2. Lặp (t = 1, …, k, 1, …)

(15)

Chuẩn hóa **p**

Cho đến khi (12) thỏa.

# PHƯƠNG PHÁP K-SVMEANS

Phương pháp K-SVMeans đề xuất dựa trên việc kết hợp thuật toán gom cụm K-Means và bộ phân lớp SVM với ước lượng xác suất lớp. Để đảm bảo hiệu suất của bộ phân lớp SVM, dữ liệu đầu vào được xây dựng ở dạng nhiều chiều và được tiền xử lý phù hợp.

## **Đặc trưng đầu vào**

Đầu vào của mô hình là dữ liệu chuỗi thời gian quá khứ bao gồm giá đóng cửa và các chỉ số kỹ thuật trong chứng khoán. Để đảm bảo được số chiều vừa phải và tránh việc làm nhiễu dữ liệu đầu vào, chúng tôi chọn các chỉ số kĩ thuật thông dụng bao gồm: trung bình trượt giản đơn (SMA) 25 ngày và 65 ngày, Aroon up, Aroon down, dải biên độ biến động giá (Bollinger bands) với Bollinger upper, middle và lower, trung bình trượt hội tụ và phân kỳ (MACD), và MACD Histogram. Tất cả được mô tả chi tiết trong [7]. Giá đóng cửa cùng những chỉ số kỹ thuật tạo nên véc tơ đầu vào với số chiều là 10. Giá trị trong mỗi chiều được chuyển về đoạn [-1, 1].

## **Giá trị nhãn lớp**

Mô hình dự đoán xu hướng sẽ kết xuất ra ba giá trị nhãn lớp ứng với xu hướng tăng, giảm và không có xu hướng. Để quyết định xu hướng của ngày hiện tại, chúng tôi dùng một định nghĩa được mô tả chặt chẽ [16] như sau:

Một cổ phiếu được xác định có xu hướng tăng (giảm) trong ngày hiện tại khi tất cả những điều kiện sau được thỏa mãn:

* Giá đóng cửa phải cao hơn (thấp hơn) chỉ số trung bình trượt 25 ngày.
* Chỉ số trung bình trượt 25 ngày phải cao hơn (thấp hơn) chỉ số trung bình trượt 65 ngày.
* Đường trung bình trượt 25 ngày phải tăng (giảm) ít nhất trong 5 ngày.
* Đường trung bình trượt 65 ngày phải tăng (giảm) ít nhất 1 ngày.

Nếu không thể thỏa mãn tất cả điều kiện trên để được đánh nhãn lớp có xu hướng tăng (giảm) thì ngày hiện tại được đánh nhãn không có xu hướng. Ở đây, chỉ số trung bình trượt n ngày được tính bằng trung bình cộng giá đóng cửa của n ngày trước, trong đó bao gồm cả ngày hiện tại.

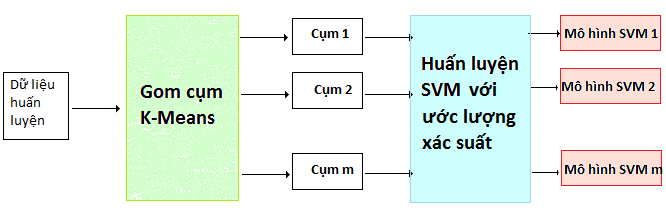
## **Kiến trúc mô hình**

Trong quá trình huấn luyện, mô hình đề xuất được xây dựng qua hai giai đoạn. Ở giai đoạn đầu, thuật toán K-Means [15] gom các mẫu dữ liệu huấn luyện có đặc điểm gần giống nhau về từng cụm. Giai đoạn hai là bước xây dựng mô hình SVM với ước lượng xác suất lớp: Ứng với mỗi cụm phân được sẽ có một bộ phân lớp SVM tương ứng được huấn luyện. Chi tiết các bước thực hiện được mô tả như sau:

Bước 1. Từ dữ liệu huấn luyện, tiền xử lý để có được véc tơ đầu vào kèm theo là một nhãn lớp như mô tả trong phần 3.1 và 3.2. Các giá trị đầu vào được quy đổi về đoạn [-1, 1].

Bước 2. Phân các véc tơ đầu vào thành m cụm (m sẽ được khảo sát trong phần 4) bằng K-Means dựa trên khoảng cách Manhattan. Ở đây, việc gom cụm không quan tâm đến nhãn lớp.

Bước 3. Ứng với m cụm, ta huấn luyện m bộ phân lớp SVM, với các tham số C và γ tối ưu được quyết định trong quá trình đánh giá chéo.



**Hình 1.** Kiến trúc K-SVMeans với ước lượng xác suất lớp

Sau bước 3, ta có được m mô hình phân lớp SVM ứng với m cụm. Quá trình dự đoán sẽ được thực hiện theo các bước sau:

Bước 1. Từ dữ liệu thử nghiệm, tiền xử lý để có được véc tơ đầu vào; nhãn lớp là giá trị cần dự đoán. Các phần tử trong véc tơ này cũng được quy về đoạn [-1, 1].

Bước 2. Tìm cụm có khoảng cách từ trung tâm cụm đến vec tơ đầu vào là gần nhất. Giả sử véc tơ đầu vào thuộc về cụm thứ i.

Bước 3. Xác định xu hướng bằng cách đưa vec tơ đầu vào đi qua bộ phân lớp SVM thứ i đã huấn luyện.

Trong nghiên cứu này, chúng tôi sử dụng thư viện LIBSVM [4] trong việc xây dựng mô hình đề xuất.

## Lựa chọn tham số

Vấn đề quan trọng nhất của mô hình SVM đề xuất là việc tìm được các tham số phù hợp. Ở đây, hai tham số cần phải quyết định khi huấn luyện mô hình là C và γ. Chúng tôi sử dụng phương pháp thông dụng Grid search và thông qua đánh giá chéo với dữ liệu huấn luyện được chia 5 phần để tìm các giá trị tối ưu cho hai tham số này. Giá trị của chúng sẽ được giới hạn trong đoạn nhất định, và .

# KẾT QUẢ THỬ NGHIỆM

## **Đánh giá mô hình**

Dựa trên yêu cầu thực tế của một bài toán dự đoán xu hướng cổ phiếu, mô hình đòi hỏi không chỉ ở độ chính xác mà độ sai về dự đoán phải ở một phạm vi cho phép. Việc dự đoán sai được xem là không chấp nhận được khi xu hướng thực tế là tăng nhưng kết quả dự đoán là giảm và ngược lại. Tỷ lệ này càng cao thì mô hình có độ tin cậy càng thấp.

Với những yêu cầu đó, bên cạnh độ chính xác phân lớp:

, (17)

chúng tôi đề xuất một độ đo để đánh giá tỷ lệ lỗi không chấp nhận được, còn gọi là tỷ lệ lỗi về dấu (SERate-Sign Error Rate):

(18)

trong đó, số mẫu phân lớp sai về dấu là các mẫu có nhãn là tăng (hoặc giảm) bị dự đoán giảm (hoặc tăng).

## **Dữ liệu**

Quá trình thử nghiệm được thực hiện trên 4 mã cổ phiếu của thị trường Việt Nam là BT6, KHA, REE và SAM. Dữ liệu được lấy từ ngày 4-1-2005 đến hết ngày 31-12-2008[[1]](#footnote-2), bao gồm 997 mẫu dữ liệu. Mỗi bộ dữ liệu của từng mã được chia thành hai tập, môt cho huấn luyện và một cho thử nghiệm. Trong đó, 797 mẫu huấn luyện được lấy từ ngày 4-1-2005 đến 17-3-2008 và 200 mẫu thử nghiệm từ ngày 18-3-2008 đến 31-12-2008. Thống kê về nhãn lớp của tập thử nghiệm được mô tả trong bảng 1.

**Bảng 1.** Thống kê các nhãn xu hướng của 200 điểm dữ liệu thử nghiệm ứng với từng mã cổ phiếu.

|  |  |  |  |
| --- | --- | --- | --- |
|  | **Giảm** | **Không xu hướng** | **Tăng** |
| **BT6** | 36 | 113 | 51 |
| **KHA** | 90 | 88 | 22 |
| **REE** | 125 | 60 | 15 |
| **SAM** | 134 | 58 | 7 |

## **Phương pháp thử nghiệm**

Chúng tôi tiến hành thử nghiệm trên hai nội dung. Trong thử nghiệm đầu, chúng tôi khảo sát để quyết định tham số tối ưu cho mô hình K-SVMeans đề xuất. Cụ thể ở đây, ta sẽ khảo sát số cụm m trong K-Means để tìm ra giá trị phù hợp. Tham số m sẽ được khảo sát trên 3 giá trị: 2, 3 và 4. Lý do chúng tôi chọn các giá trị này vì những nghiên cứu liên quan đều cho thấy số cụm cho kết quả tốt nhất thường nhỏ hơn 5. Hơn nữa, càng chia nhỏ dữ liệu, ta sẽ không có đủ thông tin cho bước huấn luyện bộ phân lớp SVM. Ở thử nghiệm sau, chúng tôi sẽ so sánh mô hình cài đặt với ba mô hình khác: BPNN, SVM và SVM có ước lượng xác suất lớp (SVM-Prob). Kết quả dự đoán được khảo sát trên hai chu kỳ là 1 ngày và 5 ngày tiếp theo.

Mô hình BPNN là mô hình mạng nơ-ron nhân tạo ba lớp, quá trình học được thực hiện thông qua việc lan truyền ngược [2]. Chi tiết về các tham số cài đặt của mô hình được mô tả trong bảng 2. Ở đây, ta lựa chọn hàm kích hoạt (Activation Function) là Tanh vì nó cho kết quả chính xác cao hơn so với các hàm kích hoạt khác đối với dữ liệu biến đổi phi tuyến [11].

**Bảng 2.** Các tham số cài đặt cho mô hình BPNN

|  |  |
| --- | --- |
| **Mô tả** | **Giá trị tham số và đặc tính** |
| Kiến trúc mạng BPNN | 3 lớp |
| Số nốt tầng nhập | 10 |
| Số nốt tầng ẩn | 4 |
| Số nốt tần xuất | 3 |
| Hàm kích hoạt | Tanh |
| Hệ số học | 0,3 |
| Hệ số Mometum | 0.1 |

## **Kết quả**

Các kết quả khảo sát số cụm m được thể hiện trong bảng 3.

**Bảng 3.** Kết quả K-SVMeans theo số cụm dựa trên độ chính xác phân lớp(A) và độ lỗi về dấu (SE).

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  | **2-SVMeans** | | **3-SVMeans** | | **4-SVMeans** | |
| **A(%)** | **SE(%)** | **A(%)** | **SE(%)** | **A(%)** | **SE(%)** |
| **BT6** | 76.5 | 0 | 77.5 | 0 | 64 | 0 |
| **KHA** | 70 | 0 | 56 | 0 | 51 | 0 |
| **REE** | 85 | 0 | 73.5 | 0 | 74.5 | 0 |
| **SAM** | 75.5 | 0.5 | 70 | 1.5 | 65.5 | 8 |

Ta nhận thấy, từ kết quả trên, mô hình K-SVMeans với m = 2 cho kết quả vượt trội. Càng tăng m, độ chính xác càng giảm và độ lỗi càng tăng. Ở phần thử nghiệm sau, ta sẽ dùng mô hình này với m = 2 để so sánh với SVM, SVM-Prob và BPNN.

Trong kết quả dự đoán 1 ngày ở bảng 4, ta nhận thấy mô hình đề xuất cho hiệu quả cao nhất so với BPNN, SVM và SVM-Prob xét về độ chính xác lẫn độ lỗi. Ở đây, BPNN vẫn cho thấy hạn chế của nó về khả năng tổng quát hóa dữ liệu khi hiệu suất dự đoán thấp nhất cùng với độ tin cậy kém khi tồn tại độ lỗi ở ba mã chứng khoán. Bên cạnh đó, SVM-Prob tỏ ra nhỉnh hơn SVM truyền thống. Chính vì điều này nên khi cải tiến mô hình SVM-Prob bằng cách áp dụng thêm K-Means, K-SVMeans cho kết quả tốt hơn.

**Bảng 4.** Kết quả so sánh dự đoán 1 ngày giữa các mô hình theo độ chính xác phân lớp (A) và độ lỗi về dấu (SE).

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | **BPNN** | | **SVM** | | **SVM-Prob** | | **K-SVMeans** | |
|  | ***A(%)*** | ***SE(%)*** | ***A(%)*** | ***SE(%)*** | ***A(%)*** | ***SE(%)*** | ***A(%)*** | ***SE(%)*** |
| **BT6** | 68.5 | 0 | 71.5 | 0 | 73.5 | 0 | 76.5 | 0 |
| **KHA** | 59.5 | 0.5 | 63.5 | 0 | 66.5 | 0 | 70 | 0 |
| **REE** | 73.5 | 1 | 85.0 | 0 | 85.5 | 0 | 85 | 0 |
| **SAM** | 74.5 | 3 | 65.5 | 7 | 71.5 | 2 | 75.5 | 0.5 |

Trong bảng 5, kịch bản cũng lặp lại đối với chu kỳ dự đoán 5 ngày: BPNN thể hiện tính không ổn định và hiệu suất thấp nhất, SVM-Prob vẫn cho kết quả tốt hơn SVM dù chênh lệch là không lớn, K-SVMeans vượt trội tất cả về độ chính xác lẫn độ lỗi. Mặt khác, phải thừa nhận rằng độ chính xác so với dự đoán 1 ngày giảm đi đáng kể và tỷ lệ lỗi về dấu cũng tăng lên khá cao. Tuy nhiên, độ chính xác dự đoán trên 60% của mô hình K-SVMeans là một kết quả khả quan đối với bài toán dự đoán xu hướng cho 5 ngày tiếp theo.

**Bảng 5.** Kết quả so sánh dự đoán 5 ngày giữa các mô hình theo độ chính xác phân lớp (A) và độ lỗi về dấu (SE).

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | **BPNN** | | **SVM** | | **SVM-Prob** | | **K-SVMeans** | |
|  | ***A (%)*** | ***SE (%)*** | ***A (%)*** | ***SE (%)*** | ***A (%)*** | ***SE (%)*** | ***A (%)*** | ***SE (%)*** |
| **BT6** | 56 | 0 | 57 | 0 | 59 | 0 | 66 | 0 |
| **KHA** | 51.5 | 4.5 | 64 | 0 | 67 | 0 | 67 | 0 |
| **REE** | 71 | 4.5 | 70.5 | 0 | 74.5 | 0 | 76 | 0 |
| **SAM** | 48.5 | 22.5 | 51.5 | 21 | 56 | 18 | 61 | 13 |

Nhìn chung, nếu xét về bình quân hiệu suất trên cả 4 mã chứng khoán, hình 2 thể hiện độ chính xác dự đoán hướng theo chiều tăng BPNN < SVM < SVM-Prob < K-SVMeans.

**Hình 2.** Đồ thị thể hiện độ chính xác dự đoán trung bình trong chu kỳ 1 ngày và 5 ngày của các mô hình trên 4 mã cổ phiếu Việt Nam.

# KẾT LUẬN

Trong nghiên cứu này, chúng tôi đã đề xuất một phương pháp kết hợp giữa K-Means và SVM với ước lượng xác suất lớp để giải quyết bài toán dự đoán xu hướng của các cổ phiếu của thị trường chứng khoán Việt Nam. Kết quả cho thấy, mô hình đề xuất vượt trội hơn hẳn so với BPNN truyền thống, mô hình SVM và SVM với ước lượng xác suất. Kết quả này khẳng định việc tiếp cận bài toán dự đoán cổ phiếu trên nền tảng SVM cho kết quả cao nhất, đồng thời cũng cho thấy khả năng giải quyết bài toán dự đoán trên thị trường đang phát triển ở Việt Nam là khả quan.

Trong nghiên cứu tới, chúng tôi sẽ tập trung vào việc nâng cao tỷ lệ dự đoán cho chu kỳ 5 ngày bằng việc xây dựng mô hình trích chọn đặc trưng trong không gian đầu vào đồng thời khảo sát cách chọn số ngày tham chiếu trong quá khứ đối với công thức tính các chỉ số kỹ thuật.

**A HYBRID METHOD FOR STOCK TREND FORECASTING BY INTEGRATING K-MEANS AND SUPPORT VECTOR MACHINE WITH PROBABILITY ESTIMATES**

***ABSTRACT:*** *In this paper, we propose a hybrid method, named K-SVMeans, which combines K-Means and support vector machine with probability estimates for stock trend forecasting in Viet Nam. The hybrid method performs one-day-ahead forecasts and five-day-ahead forecasts. To evaluate the forecasting performance of our prosed method, we compare it with the traditional back -propagation neural network, the standalone SVM and the SVM with probability estimates. We show that K-SVMeans outperforms the others. In addition, our experimental results show that the proposed method provides efficient accuracies in five-day-ahead forecasts.*

***Keyword***: *SVM, probability estimates, K-Means, artificial neural network, stock trend forecasting.*

TÀI LIỆU THAM KHẢO

[] George S. Atsalakis, Kimon P. Valavanis, *Forecasting Stock Market Short-Term Trends Using A Neuro-Fuzzy Based Methodology*, Expert Systems with Applications, Vol. 36(7), 10696-10707, (September 2009).

[] Christopher M. Bishop, *Pattern Recognition and Machine Learning*, 2nd edition, Springer, (2007).

[] Melek Acar Boyacioglu, Derya Avci, *An Adaptive Network-Based Fuzzy Inference System (ANFIS) For The Prediction Of Stock Market Return: The Case Of The Istanbul Stock Exchange*, Expert Systems with Applications, Vol. 37(12), 7908-7912, (December 2010).

[] Chih-Chung Chang and Chih-Jen Lin*, LIBSVM: A Library for Support Vector Machines*, (2001). Software available at http://www.csie.ntu.edu.tw/~cjlin/libsvm.

[] Shiyi Chen, Kyho Jeong, Wolfgang K. Hardle, *Recurrent Support Vector Regression for a Nonlinear ARMA Model with Applications to Forecasting Financial Returns*, SFB 649 Economy Risk, Berlin, (2008).

[] Deng-Yiv Chiu, Ping-Jie Chen, *Dynamically Exploring Internal Mechanism Of Stock Market By Fuzzy-Based Support Vector Machines With High Dimension Inputspace And Genetic Algorithm*, Expert Systems with Applications,Vol.36, 1240–1248, 2009.

[] Robert W. Colby, *The Encyclopedia of Technical Market Indicators* 2nd Edition, McGraw-Hill, (2003).

[] Ehsan Hajizadeh, Hamed Davari Ardakani, Jamal Shahrabi, *Application Of Data Mining Techniques In Stock Markets: A Survey*. Journal of Economics and International Finance Vol. 2(7), 109-118, (July 2010).

[] Sheng-Hsun Hsu, JJ Po-An Hsieh, Ting-Chih Chih, Kuei-Chu Hsu, *A Two-stage Architecture For Stock Price Forecasting By Integrating Self-Organizing Map And Support Vector Regression*, Expert Systems with Applications, Vol. 36, 2009, 7947–7951.

[] Cheng-Lung Huang, Cheng-Yi Tsai, *A Hybrid SOFM-SVR With A Filter-Based Feature Selection For Stock Market Forecasting*, Expert Systems with Applications, Vol. 36, Issue 2, Part 1, 1529-1539, (March 2009).

[] Bekir Karlik and A. Vehbi Olgac, *Performance Analysis of Various Activation Functions in Generalized MLP Architectures of Neural Networks*, International Journal of Artificial Intelligence And Expert Systems (IJAE), Vol. 1: Issue 4, (2011).

[] W. Kreesuradej, D. Wunsch, and M. Lane, *Time-delay Neural Network For Small Time Series Data Sets*, in World Cong. Neural Networks, San Diego, CA, (June 1994).

[] Yung-Keun Kwon, Byung-Ro Moon, *A Hybrid Neurogenetic Approach for Stock Forecasting*, IEEE Transactions on Neural Networks, Vol. 18, No. 3, (May 2007).

[] Ming-Chi Lee, *Using Support Vector Machine With A Hybrid Feature Selection Method To The Stock Trend Prediction*. Expert Systems with Applications, Vol. 36, Issue 8, 10896–10904, (October 2009).

[] J. B. MacQueen, *Some Methods for classification and Analysis of Multivariate Observations*, Proceedings of 5-th Berkeley Symposium on Mathematical Statistics and Probability, Berkeley, University of California Press, 1:281-297, (1967).

[] Binoy B. Nair, V.P Mohandas, N.R. Sakthivel, *A Genetic Algorithm Optimized Decision Tree-SVM based Stock Market Trend Prediction System*, International Journal on Computer Science and Engineering Vol. 02, No. 09, 2981-2988, (2010).

[] J. C. Platt. *Fast Training Of Support Vector Machines Using Sequential Minimal Optimization*, MIT Press, Cambridge, MA, USA, (1999).

[] J. C. Platt. *Probabilistic Outputs For Support Vector Machines And Comparison To Regularized Likelihood Methods*. In A. Smola, P. Bartlett, B. Scholkopf, and D. Schuurmans, editors, Advances in Large Margin Classiers, Cambridge, MA, MIT Press, (2000).

[] E. Saad, D. Prokhorov, D. Wunsch, *Advanced Neural-Network Training Methods For Low False Alarm Stock Trend Prediction*, in Proc. IEEE Int. Conf. Neural Networks, Washington, D.C., (June 1996).

[] H. Tan, D. Prokhorov, and D. Wunsch, *Probabilistic And Time-Delay Neural-Network Techniques For Conservative Short-Term Stock Trend Prediction*, in Proc. World Congr. Neural Networks, Washington D.C., (July 1995).

[] Tsaih R., Hsu Y., Lai C. C, *Forecasting S&P 500 stock index futures with a hybrid AI system.* Decision Support Systems, Vol. 23(2), 161–174, (1998).

[] V.N.Vapnik, M. Jordan, S.L. Lauritzen, J.F. Lawless, *Nature of Statistical Learning Theory*. Berlin: Springer, (1999).

[] Qinghua Wen, Zehong Yang, Yixu Song, Peifa Jia, *Automatic Stock Decision Support System Based On Box Theory And SVM Algorithm*, Expert Systems with Applications, Vol. 37, Issue 2, 1015-1022, (March 2010).

[] T.-F. Wu, C.-J. Lin, and R. C. Weng, *Probability estimates for multi-class classication by pairwise coupling*. Journal of Machine Learning Research, 5:975-1005, (2004).

1. Dữ liệu chứng khoán, URL: http://www.cophieu68.com/export.php. [↑](#footnote-ref-2)