Dự Đoán Xu Hướng Cổ Phiếu Ở Thị Trường Việt Nam Bằng Phương Pháp Hai Giai Đoạn Dựa Trên Viêc Kết Hợp K-Means Và SVM Với Ước Lượng Xác Suất Lớp

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

***Tóm tắt.*** Dự đoán xu hướng cổ phiếu là một đề tài hấp dẫn giới nghiên cứu trong nhiều năm qua. Tuy nhiên, việc xây dựng một phương pháp dự đoán với độ chính xác cao là một vấn đề thách thức vì dữ liệu chuỗi thời gian của cổ phiếu biến động phức tạp. Trong bài báo này, chúng tôi đề xuất một phương pháp kết hợp giữa K-Means và SVM với ước lượng xác suất lớp, gọi là K-SVMeans, để giải quyết bài toán dự đoán xu hướng cổ phiếu. Việc dự đoán được thực hiện ở chu kỳ 1 ngày và chu kỳ 5 ngày tiếp theo trên sáu mã cổ phiếu ở thị trường Việt Nam. Để đánh giá độ chính xác của phương pháp đề xuất, chúng tôi so sánh hiệu quả dự đoán với mô hình mạng nơ-ron lan truyền ngược, bộ phân lớp SVM truyền thống và bộ phân lớp SVM với ước lượng xác suất. Kết quả thử nghiệm cho thấy, phương pháp của chúng tôi vượt trội so với những mô hình trên ở cả hai chu kỳ dự đoán, đồng thời cho thấy khả năng dự đoán hiệu quả ở chu kỳ 5 ngày. Bên cạnh đó, chúng tôi cũng khảo sát kích thước của tập dữ liệu huấn luyện để phương pháp đề xuất cho kết quả dự đoán tốt nhất.

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

Giới Thiệu

Trên thị trường chứng khoán, việc dự đoán xu hướng của các cổ phiếu ở một tỷ lệ chính xác cao là rất khó ], để đưa ra một hướng tiếp cận phù hợp cho mục đích trên là một vấn đề nhiều thách thức vì thị trường thường bị tác động bởi các yếu tố kinh tế, các sự kiện chính trị cũng như các thị trường nước ngoài. Bên cạnh đó, dữ liệu chuỗi thời gian của cổ phiếu thường biến đổi phức tạp và phi tuyến. Đối với bài toán dự đoán xu hướng, các nghiên cứu thường tiếp cận nó như một bài toán phân hai lớp: xu hướng tăng và xu hướng giảm.

Những nghiên cứu gần đây đang phát triển theo hướng khai thác dữ liệu và tính toán mềm ], [2]. Các phương pháp áp dụng khai thác dữ liệu như giải thuật phân lớp K láng giềng gần nhất [22], hay K-Means kết hợp hồi quy tuyến tính [10] chưa cho kết quả cao. Những nghiên cứu sử dụng mạng nơ-ron nhân tạo (ANN) để dự đoán xu hướng ], ], ] và những cải tiến bằng các hệ lai giữa ANN và tính toán mờ ], ], ANN và thuật giải di truyền ] cho kết quả tốt hơn. Tuy nhiên, mô hình mạng nơ-ron nhân tạo đặc biệt là mạng lan truyền ngược (BPNN) cần dữ liệu huấn luyện lớn và khả năng tổng quát hóa thấp vì thường xảy ra quá khớp bởi những giá trị tối ưu cục bộ mà nó đạt được. Trong khi đó, máy học véc tơ hỗ trợ (SVM) đang được đánh giá tốt hơn nhờ khắc phục được những nhược điểm trên và tính hiệu quả đối với dữ liệu nhiều chiều và biến động phi tuyến của cổ phiếu. Phương pháp này cùng những cải tiến đã đem lại nhiều kết quả tốt trên các thị trường khác nhau. Sự kết hợp giữa SVM với tính toán mờ và thuật giải di truyền cho kết quả tốt hơn BPNN trong việc dự đoán xu hướng chứng khoán ở thị trường Đài Loan[6]. SVM lai 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 của Mỹ ]. Bằng cách áp dụng lý thuyết hộp (Box theory) 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 ]. Ngoài ra, các phương pháp hai giai đoạn áp dụng thuật giải 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 ], ] cũng cho kết quả tốt hơn mô hình SVR cơ bản trong việc dự đoán giá.

Ở Việt Nam, thị trường chứng khoán còn đang trong giai đoạn phát triển, giá cổ phiểu trong rất nhiều giai đoạn biến động không có xu hướng. Do đó, khi giải quyết bài toán dự đoán xu hướng cổ phiếu, việc chỉ xem xét dự đoán với hai nhãn lớp tăng và giảm sẽ không phản ánh hết đặc trưng của thị trường này. Vì vậy, chúng tôi đề xuất hướng tiếp cận dựa trên việc giải quyết bài toán phân ba lớp (xu hướng tăng, xu hướng giảm và không có xu hướng) bằng ý tưởng của phương pháp hai giai đoạn đã từng thành công ở các thị trường khác cũng như áp dụng khai thác dữ liệu và tính toán mềm. Phương pháp này là sự kết hợp giữa giải thuật gom cụm K-Means và bộ phân lớp SVM với ước lượng xác suất lớp, gọi là K-SVMeans. Phương pháp ước lượng xác suất lớp đã được chứng minh là tốt hơn chiến lược bỏ phiếu trong mô hình SVM truyền thống khi giải quyết bài toán phân đa lớp nói chung ]. 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, số cụm k tối ưu sẽ được khảo sát trong phần thử nghiệm. Sau đó, mỗi cụm sẽ được huấn luyện để cho ra một bộ phân lớp SVM với việc ước lượng xác suất lớp. Quá trình dự đoán như sau: Sau khi xác định mẫu dữ liệu thử nghiệm thuộc vào cụm nào, ta dùng chính bộ phân lớp SVM của cụm đó để dự đoán. Để đánh giá phương pháp đề xuất, chúng tôi sẽ thực hiện dự đoán ở chu kỳ 1 ngày tiếp theo và 5 ngày tiếp theo.

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, phương pháp 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.

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) ]. Cho tập véctơ đầu vào và tập các giá trị nhãn lớp tuơng ứng cho bộ 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:

|  |  |
| --- | --- |
| sao cho:  . | (2) |

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 ] (Sequential Minimal Optimization). Phuơng pháp này đưa đến giải bài toán đối ngẫu quy hoạch toàn phương (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. Đối với dữ liệu chứng khoán biến đổi một cách phi tuyến ], hàm phi tuyến Gauss (RBF-Radial Basis Function) có thể chọn làm hàm nhân:

|  |  |
| --- | --- |
| . | (5) |

Ước lượng xác suất

Với việc áp dụng bộ phân lớp SVM, để phân m lớp, ta sử dụng cách tiếp cận một đối một, còn gọi là cặp lớp. Vì vậy, sẽ có m(m-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., Wu (2004) trong ] đề xuất phương pháp ước lượng xác suất cho việc phân m lớp và chứng minh cả trên mặt lý thuyết và thực nghiệm, phương pháp 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 xác suất thuộc về từng lớp của mẫu đó và từ đó xác định nhãn lớp ứng với xác suất lớn nhất.

Cho véc tơ đầu vào *x* và nhãn lớp *y*, mục tiêu của phương pháp là ước lượng xác suất mẫu *x* thuộc vào lớp *i* (ký hiệu là *pi*):

|  |  |
| --- | --- |
| . | (6) |

Theo cách tiếp cận một đối một, ta có 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*] 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 âm.

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:

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

với ràng buộc

Ta có thể viết lại (9) theo dạng:

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

với

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

|  |  |
| --- | --- |
|  | (11) |

Với *e* là véc tơ m×1 có các phần tử lấy giá trị một, *0* là véc tơ m×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

|  |  |
| --- | --- |
|  | (12) |

cho nên ***p*** tối ưu thỏa mãn

|  |  |
| --- | --- |
|  | (13) |

Thuật toán giải quyết (13) được mô tả như sau ]:

Bước 1. Khởi tạo *p* thỏa .

Bước 2. Lặp (t = 1, …, m, 1, …)

|  |  |
| --- | --- |
| . | (14) |

Chuẩn hóa ***p***

Cho đến khi (11) 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 dự đoán xu hướng 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 ]. 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 chuẩn hóa về [-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 ] 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.

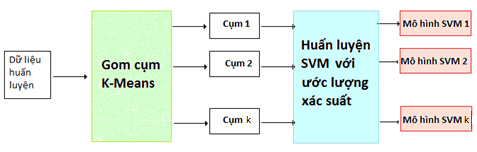
Nếu không thể thỏa mãn tất cả điều kiện trên để được đánh nhãn lớp có xu hướng tăng (giảm) thì ngày hiện tại được đánh nhãn không có xu hướng. Ở đây, chỉ số trung bình trượt n ngày được tính bằng trung bình cộng giá đóng cửa của n ngày trước, trong đó bao gồm cả ngày hiện tại.

Kiến trúc mô hình

Trong quá trình huấn luyện, phương pháp đề xuất được xây dựng qua hai giai đoạn. Ở giai đoạn đầu, thuật toán K-Means ] gom các mẫu dữ liệu huấn luyện có đặc điểm gần giống nhau về từng cụm. Giai đoạn hai là bước xây dựng bộ phân lớp SVM với ước lượng xác suất lớp: Ứng với mỗi cụm 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 chuẩn hóa về [-1, 1].

Bước 2. Gom các véc tơ đầu vào thành k cụm (k sẽ được khảo sát trong phần 4) bằng K-Means dựa trên khoảng cách Manhattan. Ở đây, không sử dụng nhãn lớp trong khi gom cụm.

Bước 3. Ứng với k cụm, ta huấn luyện k bộ phân lớp SVM với ước lượng xác suất, với các tham số chuẩn tắc C và tham số hàm nhân γ tối ưu được quyết định trong quá trình đánh giá chéo.

**Hình 1.** Phương pháp K-SVMeans với ước lượng xác suất lớp

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

Bước 1. Từ dữ liệu cần dự đoán, tiền xử lý để có được véc tơ đầu vào. Các phần tử trong véc tơ này cũng được chuẩn hóa về [-1, 1].

Bước 2. Tìm cụm có khoảng cách từ trung tâm cụm đến véc 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. Giá trị đầu ra là ba giá trị xác suất mà mẫu đó thuộc về các lớp: tăng, giảm và không có xu hướng. Kết quả phân lớp sẽ được quyết định bằng cách chọn ra xác suất lớn nhất trong ba giá trị trên.

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

Lựa chọn tham số

Vấn đề quan trọng nhất của bộ phân lớp SVM là việc tìm được các tham số phù hợp. Ở đây, hai tham số cần phải quyết định khi huấn luyện là C và γ. Chúng tôi sử dụng phương pháp thông dụng Grid search và thông qua đánh giá chéo (5-fold cross validation) để tìm các giá trị tối ưu cho hai tham số này. Giá trị của chúng được giới hạn trong khoảng : và .

Kết Quả Thử Nghiệm

Đánh giá mô hình

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

Với những yêu cầu đó, bên cạnh độ chính xác phân lớp Acc (Accuracy):

|  |  |
| --- | --- |
|  | (15) |

chúng tôi đề xuất một độ đo để đánh giá tỷ lệ lỗi dự đoán sai, còn gọi là tỷ lệ lỗi về dấu SERate (Sign Error Rate):

|  |  |
| --- | --- |
|  | (16) |

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 6 mã cổ phiếu của thị trường Việt Nam là BT6[[1]](#footnote-2), GIL[[2]](#footnote-3), GMD[[3]](#footnote-4), HAP[[4]](#footnote-5), REE[[5]](#footnote-6) và SAM[[6]](#footnote-7). Đây là các cổ phiếu của các công ty lớn thuộc những nhóm ngành nghề khác nhau. Dữ liệu được rút trích thành 3 bộ như mô tả trong bảng 1 để phục vụ các thử nghiệm khác nhau. Mỗi bộ gồm tập huấn luyện và tập thử nghiệm. Trong đó, số mẫu tập kiểm thử của 3 bộ là hoàn toàn giống nhau. Thống kê về nhãn lớp của tập thử nghiệm được mô tả trong bảng 2. Có thể thấy do thị trường chứng khoán Việt Nam còn trong giai đoạn phát triển nên nhãn không xu hướng chiếm một tỷ trọng tương đối lớn.

Phương pháp thử nghiệm

Chúng tôi tiến hành thử nghiệm trên ba nội dung. Trong thử nghiệm đầu, chúng tôi khảo sát để tìm tham số tối ưu cho phương pháp K-SVMeans. Cụ thể ở đây, ta sẽ khảo sát số cụm k trong K-Means để tìm ra giá trị phù hợp. Tham số k sẽ được khảo sát trên đoạn [2; 5]. 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 bằng 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 kế tiếp, chúng tôi sẽ so sánh K-SVMeans với BPNN, SVM và SVM có ước lượng xác suất lớp (SVM-Prob) trên chu kỳ 1 ngày và 5 ngày tiếp theo. Hai thử nghiệm đầu sẽ được thực hiện trên bộ dữ liệu 1. Cuối cùng, chúng tôi sẽ khảo sát số mẫu tối ưu của tập huấn luyện cho phương pháp K-SVMeans dựa vào việc thử nghiệm trên 3 bộ dữ liệu có cùng 200 mẫu thử nghiệm.

**Bảng 1.** Mô tả các bộ dữ liệu cho quá trình kiểm thử

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | Tập huấn luyện | Tập thử nghiệm | Số mẫu huấn luyện | Tổng số mẫu |
| Bộ 1 | 04/01/2005-17/3/2008 | 18/3/2008-31/12/2008 | 797 | 997 |
| Bộ 2 | 05/01/2004-17/3/2008 | 18/3/2008-31/12/2008 | 1047 | 1247 |
| Bộ 3 | 02/01/2003-17/3/2008 | 18/3/2008-31/12/2008 | 1294 | 1494 |

**Bảng 2.** Các nhãn xu hướng của 200 điểm dữ liệu thử nghiệm ứng với từng mã cổ phiếu

|  |  |  |  |
| --- | --- | --- | --- |
| Mã | Giảm | Không xu hướng | Tăng |
| BT6 | 36 | 113 | 51 |
| GIL | 116 | 79 | 5 |
| GMD | 122 | 64 | 14 |
| HAP | 132 | 59 | 9 |
| REE | 125 | 60 | 15 |
| SAM | 134 | 58 | 7 |

BPNN là mô hình mạng nơ-ron nhân tạo ba lớp, quá trình học được thực hiện thông qua việc lan truyền ngược ]. Ở đây, hàm kích hoạt (Activation Function) được chọn là Tanh vì hàm này cho kết quả chính xác cao hơn so với các hàm kích hoạt khác đối với dữ liệu biến đổi phi tuyến ]. Trong bảng 3 là các tham số của BPNN được sử dụng trong quá trình thử nghiệm.

**Bảng 3.** Các tham số cài đặt cho 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

Bảng 4 thể hiện các kết quả dự đoán với số cụm k = 2, 3, 4, 5 tương ứng. Ta thấy K-SVMeans với k = 2 cho kết quả tốt nhất. Càng tăng k, độ chính xác càng giảm và độ lỗi càng tăng. Việc chia nhỏ các cụm gây thiếu thông tin cho việc huấn luyện bộ phân lớp hiệu quả.

**Bảng 4.** Kết quả K-SVMeans theo số cụm theo độ đo Acc và SERate

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Mã | 2-SVMeans | | 3-SVMeans | | 4-SVMeans | | 5-SVMeans | |
| Acc (%) | SERate (%) | Acc (%) | SERate (%) | Acc(%) | SERate (%) | Acc (%) | SERate (%) |
| BT6 | 76.5 | 0 | 77.5 | 0 | 64 | 0 | 71.5 | 0 |
| GIL | 60.5 | 1 | 66.5 | 1.5 | 58.5 | 1 | 60.5 | 0 |
| GMD | 82 | 0 | 80.5 | 0 | 81 | 0 | 77.5 | 0 |
| HAP | 88.5 | 0 | 73 | 0 | 74 | 1 | 81 | 0 |
| REE | 85 | 0 | 73.5 | 0 | 74.5 | 0 | 59.5 | 0 |
| SAM | 75.5 | 0.5 | 70 | 1.5 | 65.5 | 8 | 58.5 | 8.5 |

Trong kết quả dự đoán 1 ngày ở bảng 5, ta nhận thấy phương pháp 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 cao. Bên cạnh đó, SVM-Prob tỏ ra nhỉnh hơn SVM truyền thống. Chính vì điều này nên khi cải tiến SVM-Prob bằng cách áp dụng thêm K-Means, K-SVMeans cho kết quả tốt hơn.

**Bảng 5.** So sánh kết quả dự đoán 1 ngày giữa các mô hình theo độ đo Acc và SERate

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Mã | BPNN | | SVM | | SVM-Prob | | K-SVMeans | |
| *Acc (%)* | *SERate (%)* | *Acc (%)* | *SERate (%)* | *Acc (%)* | *SERate (%)* | *Acc (%)* | *SERate (%)* |
| BT6 | 68.5 | 0 | 71.5 | 0 | 73.5 | 0 | 76.5 | 0 |
| GIL | 57 | 3.5 | 60 | 0.5 | 60 | 0.5 | 60.5 | 1 |
| GMD | 80 | 0 | 77.5 | 0 | 81 | 0 | 82 | 0 |
| HAP | 85 | 0 | 87 | 0 | 88 | 0 | 88.5 | 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 6, kết quả dự đoán với chu kỳ 5 ngày cũng cho nhận xét tương tự: 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. Mặc dù độ chính xác so với dự đoán chu kỳ 1 ngày có giảm và tỷ lệ lỗi về dấu cũng tăng lên, nhưng độ chính xác dự đoán trên 60% của phương pháp K-SVMeans là một kết quả khả quan đối với bài toán dự đoán xu hướng cho 5 ngày tiếp theo.

**Bảng 6.** So sánh kết quả dự đoán 5 ngày giữa các mô hình theo độ đo Acc và SERate.

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Mã | BPNN | | SVM | | SVM-Prob | | K-SVMeans | |
| *Acc (%)* | *SERate (%)* | *Acc (%)* | *SERate (%)* | *Acc (%)* | *SERate (%)* | *Acc (%)* | *SERate (%)* |
| BT6 | 56 | 0 | 57 | 0 | 59 | 0 | 66 | 0 |
| GIL | 39 | 21.5 | 60 | 0 | 59.5 | 1 | 60.5 | 0 |
| GMD | 34.5 | 0 | 69 | 0 | 67.5 | 0 | 68 | 0 |
| HAP | 76 | 0 | 61 | 0.5 | 65 | 0 | 67.5 | 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 6 mã cổ phiếu, độ chính xác dự đoán theo chiều tăng như sau: BPNN < SVM < SVM-Prob < K-SVMeans. Hình 2 thể hiện đồ thị so sánh độ chính xác trung bình của các mã cổ phiếu theo chu kỳ 1 ngày và chu kỳ 5 ngày.

**Hình 2.** Đồ thị so sánh độ đo Acc trung bình trong chu kỳ 1 ngày và 5 ngày.

Trong việc khảo sát số mẫu huấn luyện cho phương pháp đề xuất, kết quả được thể hiện ở bảng 7. Nhìn chung, khi tăng kích thước dữ liệu huấn luyện, độ chính xác tăng theo. Tuy nhiên, khi tăng đến một giới hạn nhất định điều này không còn chính xác. Ở đây, đối với cả hai chu kỳ, bộ 2 cho kết quả tốt hơn hẳn so với bộ 1 xét về bình quân hiệu suất. Độ chính xác tăng 4 đến 6.67 điểm phần trăm, và độ lỗi giảm đáng kể. Trong khi đó, dù thời gian lấy mẫu huấn luyện bộ 3 nhiều hơn bộ 2 một năm, độ chính xác hầu như không cải thiện, thậm chí độ lỗi ở cả hai chu kỳ đều cao so với kết quả của bộ 2.

**Bảng 7.** Kết quả dự đoán của K-SVMeans ở hai chu kỳ trên ba bộ dữ liệu.

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| Chu kỳ | Mã | Bộ 1 | | Bộ 2 | | Bộ 3 | |
| *Acc(%)* | *SERate(%)* | *Acc(%)* | *SERate(%)* | *Acc(%)* | *SERate(%)* |
| 1 ngày | BT6 | 76.5 | 0 | 76 | 0 | 76 | 0 |
| GIL | 60.5 | 1 | 80.5 | 0 | 78 | 0 |
| GMD | 82 | 0 | 83 | 0 | 83 | 0 |
| HAP | 88.5 | 0 | 80 | 0 | 86.5 | 0 |
| REE | 85 | 0 | 87.5 | 0 | 85 | 0 |
| SAM | 75.5 | 0.5 | 85 | 0 | 85 | 1 |
| Trung bình | 78 | 0.25 | 82 | 0 | 82.25 | 0.17 |
| 5 ngày | BT6 | 66 | 0 | 78 | 0 | 78.5 | 0 |
| GIL | 60.5 | 0 | 64.5 | 0.5 | 69 | 0.5 |
| GMD | 68 | 0 | 66.5 | 0 | 70 | 0 |
| HAP | 67.5 | 0 | 71.5 | 0 | 68.5 | 0.5 |
| REE | 76 | 0 | 79 | 0 | 75.5 | 0 |
| SAM | 61 | 13 | 79.5 | 1 | 78 | 1 |
| Trung bình | 66.5 | 2.17 | 73.17 | 0.25 | 73.25 | 0.33 |

Kết Luận

Trong nghiên cứu này, chúng tôi đề xuất phương pháp K-SVMeans để giải quyết bài toán dự đoán xu hướng cổ phiếu của thị trường chứng khoán Việt Nam trên chu kỳ 1 ngày và 5 ngày. Phương pháp đề xuất K-SVMeans sử dụng thuật toán K-Means để gom cụm các dữ liệu đầu vào. Sau đó, từ mỗi cụm, huấn luyện bộ phân lớp SVM với ước lượng xác suất để dự đoán ba giá trị nhãn lớp: xu hướng tăng, xu hướng giảm và không có xu hướng. Kết quả cho thấy phương pháp K-SVMeans có độ chính xác cao hơn so với BPNN truyền thống, bộ phân lớp SVM và SVM với ước lượng xác suất. Kết quả này khẳng định việc tiếp cận bài toán dự đoán xu hướng cổ phiếu trên nền tảng SVM là một giải pháp hiệu quả. Bên cạnh đó, việc chọn kích thước bộ huấn luyện cũng ảnh hưởng đến kết quả dự đoán. Đối với phương pháp đề xuất, dữ liệu học phù hợp nằm trong khoảng thời gian từ 4 đến 5 năm.

Để nâng cao tỷ lệ dự đoán cho chu kỳ 5 ngày, chúng tôi dự định sẽ tăng số chiều của không gian đầu vào bằng cách phối hợp thêm nhiều chỉ số kỹ thuật và xây dựng mô hình trích chọn đặc trưng trên không gian đó để chọn ra những đầu vào phù hợp nhất. Ngoài ra, số ngày tham chiếu trong quá khứ đối với công thức tính các chỉ số kỹ thuật cũng cần được khảo sát trong nghiên cứu sau.

Forecasting Stock Trend in Vietnam Using A Two-Stage Architecture by Integrating K-Means and Support Vector Machine with Probability Estimates

**Abstract.** Stock trend forecasting is considered an interesting financial subject that has attracted researchers for many years. However, developing a forecasting model with sufficient accuracy is a challenging problem because the time series data is dynamic and quite complex. In this paper, we propose a two-stage architecture, named K-SVMeans, which combines K-Means and support vector machine with probability estimates for stock trend forecasting. The proposed method performs one-day-ahead forecasts and five-day-ahead forecasts on six stock indexes of VietNam stock exchange market. 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 and provides efficient accuracies in five-day-ahead forecasts. In addition, we also examine the optimal number of observations in the training set for the best performance of our proposed method.

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

Tài Liệu Tham Khảo

1. Atsalakis G. S., Valavanis K. P., Forecasting Stock Market Short-Term Trends Using A Neuro-Fuzzy Based Methodology, Expert Systems with Applications, Vol. 36(7), 10696-10707 (2009).

. Atsalakis G. S., Valavanis K. P., Surveying stock market forecasting techniques – Part II: Soft computing methods. Expert Systems with Applications 36, 5932–5941 (2009).

3. Bishop C. M., Pattern Recognition and Machine Learning, 2nd edition, Springer (2007).

4. Boyacioglu M. A., Avci D., 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 (2010).

5. Chang C. C., Lin C. J., LIBSVM: A Library for Support Vector Machines, (2001). Software available at http://www.csie.ntu.edu.tw/~cjlin/libsvm.

6. Chen S., Jeong K., Hardle W. K., Recurrent Support Vector Regression for a Nonlinear ARMA Model with Applications to Forecasting Financial Returns, SFB 649 Economy Risk, Berlin (2008).

. Chiu D. Y., Chen P. J., 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).

8. Colby R. W., The Encyclopedia of Technical Market Indicators 2nd Edition, McGraw-Hill (2003).

9. Hajizadeh E., Ardakani H. D., Shahrabi J., Application Of Data Mining Techniques In Stock Markets: A Survey. Journal of Economics and International Finance Vol. 2(7), 109-118 (2010).

. He H. X., Chen J., Jin H. D., Chen S. H., Stock Trend Analysis and Trading Strategy. Joint Conference on Information Sciences 9 (2006).

11. Hsu S. H., Hsieh P. A., Chih T. C., Hsu K. C., A Two-stage Architecture For Stock Price Forecasting By Integrating Self-Organizing Map And Support Vector Regression, Expert Systems with Applications, Vol. 36, 7947–7951 (2009).

12. Huang C. L., Tsai C. Y., 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 (2009).

13. Karlik B., Olgac A. V., 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).

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

15. Kwon Y. K., Moon B. R., A Hybrid Neurogenetic Approach for Stock Forecasting, IEEE Transactions on Neural Networks, Vol. 18, No. 3 (2007).

16. Lee M. C., 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 (2009).

17. MacQueen J. B., 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).

18. Nair B. B., Mohandas V.P, Sakthivel N.R., 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).

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

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

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

. Teixeira L. A., Oliveira A. L. I., A method for automatic stock trading combining technical analysis and nearest neighbor classiﬁcation, Expert Systems with Applications 37, 6885–6890 (2010)

23. Tan H., D. Prokhorov, 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 (1995).

24. 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).

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

26. Wen Q., Yang Z., Song Y., Jia P., Automatic Stock Decision Support System Based On Box Theory And SVM Algorithm, Expert Systems with Applications, Vol. 37, Issue 2, 1015-1022 (2010).

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

1. Công ty cổ phần Bê-tông 620 Châu Thới. [↑](#footnote-ref-2)
2. Công ty cổ phần sản xuất kinh doanh xuất nhập khẩu Bình Thạnh [↑](#footnote-ref-3)
3. Công ty cổ phần Đại lý Liên hiệp Vận chuyển [↑](#footnote-ref-4)
4. Công ty cổ phần HAPACO [↑](#footnote-ref-5)
5. Công ty cổ phần cơ điện lạnh. [↑](#footnote-ref-6)
6. Công ty cổ phần cáp và vật liệu viễn thông. [↑](#footnote-ref-7)