
XÂY DỰNG DANH MỤC TỐI ƯU VỚI KỸ THUẬT PHÂN CỤM CHUỖI THỜI GIAN - TRƯỜNG HỢP THỊ TRƯỜNG CHỨNG KHOÁN VIỆT NAM

Bùi Quốc Hoàn

Trường Đại học Kinh tế Quốc dân

Email: buiquochoan@neu.edu.vn

Mã bài: JED - 703

Ngày nhận bài: 02/06/2022

Ngày nhận bài sửa: 27/07/2022

Ngày duyệt đăng: 02/08/2022

Tóm tắt:

Việc xây dựng danh mục tối ưu trong thực tế không chỉ quan tâm đến mối quan hệ giữa lợi nhuận – rủi ro mà còn tính đến chi phí quản trị, được thể hiện bởi số lượng cổ phiếu trong danh mục. Bài viết này tiến hành xây dựng danh mục tối ưu cho thị trường chứng khoán Việt Nam theo cách tiếp cận mới, gồm hai giai đoạn. Đầu tiên, kỹ thuật phân cụm với chuỗi thời gian sẽ được thực hiện để phân cụm các cổ phiếu. Từ các cụm này, các cổ phiếu đại diện sẽ được chọn dựa trên chỉ số Sharpe. Ở giai đoạn tiếp theo, thuật toán tối ưu Markowitz được áp dụng cho nhóm các cổ phiếu đại diện để lựa chọn ra được danh mục tối ưu. Kết quả thực nghiệm cho thấy, danh mục tối ưu tìm được bằng phương pháp này tốt hơn đáng kể so với thị trường (đại diện bởi VN-Index) trong giai đoạn nghiên cứu cả về lợi suất trung bình và tỷ suất lợi suất trên rủi ro.

Từ khóa: Phân cụm K-means, Tối ưu hóa danh mục đầu tư, Phân cụm chuỗi thời gian, VN-Index.

Mã JEL: C02, C38, G11

Portfolio optimization based on time series clustering method – An empirical study of the Vietnamese stock market

Abstract:

In practice, portfolio optimization does not only deal with the risk-return relationship, but also involves management costs, as shown by the number of stocks in the portfolio. This paper proceeds to build the optimal investment portfolio for stocks listed on the Vietnamese stock market according to a new approach consisting of two phases. First, time series clustering technique will be implemented to cluster the stocks. Then, the stocks with the highest Sharpe ratios from each cluster are selected to create a stock market portfolio. In the next stage, the Markowitz optimization algorithm will be applied to the representative group of stocks to choose the optimal portfolio. The results of our analysis show that the optimal portfolio found by this method outperforms the market (represented by the VN-Index) over the study period in terms of both average return and return after risk adjustment.

Keywords: K-means clustering, Portfolio optimization, Time series clustering, VN-Index.

JEL codes: C02, C38, G11

1. Giới thiệu và tổng quan nghiên cứu

Lý thuyết danh mục đầu tư hiện đại của Markowitz được sử dụng rộng rãi trong việc xác định danh mục đầu tư. Phương pháp này nhằm tối đa hóa lợi suất kỳ vọng trong khi đảm bảo rủi ro không quá một mức độ nào đó, hoặc cực tiểu hóa rủi ro với một mức lợi nhuận nào đó (Markowitz, 1952). Markowitz đã chỉ ra rằng rủi ro có thể được giảm thiểu bằng việc đa dạng hóa danh mục. Tuy nhiên, việc có quá nhiều cổ phiếu trong danh mục là không cần thiết (Fisher & Lorie, 1970). Số lượng cổ phiếu trong danh mục càng nhiều, thì chi phí theo dõi danh mục càng lớn. Ngoài ra, các kết quả cũng đã chỉ ra rằng sai số dự báo của danh mục tăng theo số lượng cổ phiếu (DeMiguel & cộng sự, 2007). Điều này có thể là do việc ước lượng các tham số càng trở nên kém chính xác nếu số lượng cổ phiếu được đưa vào tính toán khi tối ưu hóa bằng mô hình Markowitz là lớn (Black & Litterman, 1992). Do đó gần đây, người ta quan tâm đến việc giới hạn lượng cổ phiếu được xem xét, và thường là dựa vào kỹ thuật phân cụm.

Trong các nghiên cứu về phân cụm cổ phiếu có thể phân ra thành hai cách tiếp cận là sử dụng dữ liệu tĩnh và sử dụng chuỗi thời gian. Phân cụm với dữ liệu tĩnh dựa vào các thuộc tính của dữ liệu được cố định, không biến động theo thời gian. Phân cụm với chuỗi thời gian dựa trên các đặc điểm động của dữ liệu. Chẳng hạn, khi xem xét dữ liệu là lợi suất của một cổ phiếu, nếu ta sử dụng giá trị trung bình của lợi suất để phân cụm, thì giá trị này là cố định và không biến động trong cả giai đoạn. Trong khi đó để sử dụng chuỗi thời gian, ta sẽ xem xét giá trị lợi suất của cổ phiếu biến động theo mỗi ngày. Gần đây, các nghiên cứu về phân cụm với chuỗi thời gian được sử dụng trong nhiều lĩnh vực. Các kết quả đáng chú ý được tổng hợp trong các bài viết của Liao (2005), Aghabozorgi & cộng sự (2015) và Ali & cộng sự (2020).

Sử dụng phân cụm với dữ liệu tĩnh, Nanda & cộng sự (2010) dùng phương pháp K-means, SOM (Self-Organizing Maps: bản đồ tự tổ chức) và FCM (Fuzzy C-means: C-means mờ) phân cụm các cổ phiếu trên thị trường chứng khoán Bombay (Ấn Độ). Các thuộc tính được đại diện cho các cổ phiếu là các giá trị tài chính được tính trong khoảng thời gian nghiên cứu. Ngoài ra, các tác giả còn sử dụng thêm các giá trị lợi suất trung bình theo các chu kỳ khác nhau, trong khoảng ngắn hạn (ngày, tuần, 30 ngày) và dài hạn (3 tháng, 6 tháng và 1 năm). Kết quả thực nghiệm cho thấy phân cụm bằng phương pháp K-means là tốt nhất trong 3 phương pháp được sử dụng. Cùng với đó, lợi suất của các danh mục đầu tư tìm được theo mô hình Markowitz (khi sử dụng kết quả phân cụm làm đầu vào) cao hơn lợi suất của chỉ số Sensex.

Vẫn theo các tiếp cận này, Nejad (2014) đã sử dụng các phương pháp K-means, FCM và SOM để phân cụm các cổ phiếu trên thị trường chứng khoán Tehran của Iran. Các thuộc tính được lựa chọn là các chỉ số tài chính: doanh thu tài sản, tỷ suất lợi nhuận ròng, lợi nhuận trên mỗi cổ phần, tốc độ tăng tổng lợi nhuận, tốc độ tăng lợi nhuận trên mỗi cổ phần, tỷ lệ giá trị thị trường trên giá trị sổ sách, tỷ lệ giá trên lợi nhuận, doanh thu của vốn cổ đông. Kết quả thực nghiệm cho thấy việc phân cụm làm giảm thời gian lựa chọn danh mục nhưng vẫn đảm bảo được tính đa dạng, rủi ro của danh mục được giảm thiểu. Trong nghiên cứu khác của Tavakoli & cộng sự (2020), các tác giả đưa ra một cách tiếp cận phân cụm chuỗi thời gian theo thuật toán autoencoder của học sâu (deep learning). Thuật toán K-means được sử dụng trong bước thứ nhất để xác định cụm mà các chuỗi thời gian sẽ được gán vào. Khi thực hiện phân cụm theo K-means, chuỗi thời gian được đại diện bởi thuộc tính lợi suất và độ rủi ro (tính trên toàn bộ khoảng thời gian). Các chuỗi thời gian với các nhãn là số cụm của chuỗi thuộc vào sau đó được tiếp tục sử dụng đưa vào mô hình phân cụm học sâu. Các tác giả nhận thấy việc sử dụng học sâu có thể phát hiện ra những thuộc tính ẩn mà K-means có thể không phát hiện ra. Tuy nhiên, kết quả ở giai đoạn sử dụng học sâu không quá khác biệt so với kết quả khi sử dụng phân cụm K-means.

Một cách tiếp cận khác, là phân cụm dựa trên tính chất chuỗi thời gian của cổ phiếu. Trong đó, mỗi đặc tính của cổ phiếu không chỉ được đại diện bởi một giá trị duy nhất cho cả giai đoạn nghiên cứu, mà được đại diện bởi nhiều thông tin dọc theo cả giai đoạn đó. Chẳng hạn, Iorio & cộng sự đã phân cụm dựa theo biến động của chuỗi giá đóng cửa. Trong nghiên cứu này, các tác giả đã sử dụng các hệ số ước tính theo phương pháp hàm ghép trơn (B-spline) để đại diện cho đường cong giá đóng cửa. Phân cụm K-means được thực hiện trên hệ số ước tính B-spline từ các chuỗi. Các cụm gồm các cổ phiếu có biến động giá tương tự nhau. Kết quả thực nghiệm sử dụng 86 cổ phiếu trên thị trường chứng khoán Italia. Bài viết gợi ý nhà đầu tư về việc xây dựng danh mục sau khi phân cụm, các nhà đầu tư có thể lựa chọn cổ phiếu từ các cụm theo tiêu chí của cá nhân.

Trong một nghiên cứu khác, Kim & Kim (2020) đã thực hiện so sánh 10 độ đo khoảng cách giữa các chuỗi thời gian khi thực hiện phân cụm theo hai phương pháp là phân cụm phân cấp và K-means. Các tác

giả sử dụng phân loại khoảng cách đo sự tương đồng của các chuỗi thời gian theo Montero & Vilar (2014). Sự tương đồng của hai chuỗi có thể xác định bởi sự gần nhau của các giá trị tại một vài thời điểm cụ thể. Khoảng cách cũng có thể xác định từ việc sử dụng các tham số nhận được từ mô hình khi áp dụng trên các chuỗi thời gian. Phức tạp hơn, khoảng cách có thể dựa trên việc đo lường mức độ thông tin được chia sẻ bởi cả hai chuỗi thời gian. Các tác giả thực hiện hai phương pháp phân cụm là phân cụm phân cấp và K-means với 10 khoảng cách cho chuỗi thời gian. Kết quả thực nghiệm cho thấy không có cách phân cụm nào có ưu thế rõ rệt.

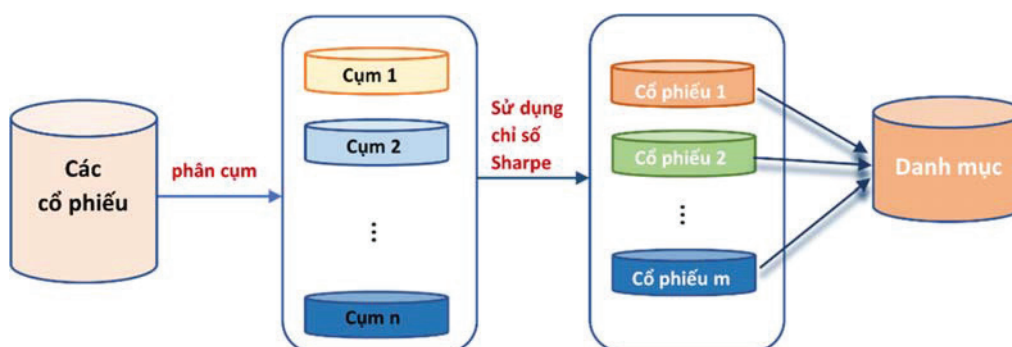
Ở Việt Nam, thị trường chứng khoán đã khá phát triển, với số lượng mã cổ phiếu trên sàn giao dịch thành phố Hồ Chí Minh (HOSE) tính đến ngày 29/04/2022 theo Ủy ban chứng khoán nhà nước công bố là 408 cổ phiếu. Các nghiên cứu về vấn đề lựa chọn danh mục tối ưu còn khá hạn chế, một số ít các nghiên cứu đã được thực hiện. Chẳng hạn, Võ Thị Thúy Anh (2012) đã ứng dụng lý thuyết trung bình - phương sai Markowitz để lựa chọn cổ phiếu thống trị. Cổ phiếu thống trị được tác giả định nghĩa là “những cổ phiếu mà không có cổ phiếu nào có tỷ suất sinh lợi kỳ vọng cao hơn hoặc bằng nó nhưng lại có rủi ro thấp hơn”. Trần Văn Trí (2015) nghiên cứu về vấn đề xây dựng và quản lý danh mục đầu tư trên thị trường chứng khoán Việt Nam. Tác giả ứng dụng mô hình định giá tài sản vốn và lý thuyết danh mục đầu tư Markowitz với số liệu từ Sở Giao dịch Chứng khoán Thành phố Hồ Chí Minh và Hà Nội giai đoạn 2008-2013. Tuy nhiên, hầu như chưa có nghiên cứu nào sử dụng kỹ thuật phân cụm để lọc các cổ phiếu trước khi thực hiện bài toán tối ưu danh mục trên thị trường chứng khoán Việt Nam.

Nghiên cứu này khác với các nghiên cứu trước là, trong nghiên cứu sẽ sử dụng phương pháp phân cụm để chia các cổ phiếu vào các cụm, các cổ phiếu trong cụm sẽ có sự tương quan cao với nhau. Thêm vào đó, kỹ thuật phân cụm được sử dụng ở đây là kỹ thuật phân cụm theo chuỗi thời gian, chứ không phải là phân cụm thông thường, là một kỹ thuật còn được hiếm sử dụng trong các nghiên cứu về thị trường chứng khoán tại Việt Nam. Sau khi các cổ phiếu được phân vào các cụm, chỉ số Sharpe được sử dụng để lấy từ mỗi cụm các cổ phiếu có chỉ số cao nhất và thiết lập danh mục. Các cổ phiếu có chỉ số Sharpe cao được chọn ngụ ý với cùng mức rủi ro việc đầu tư vào các cổ phiếu này sẽ đem lại lợi suất tốt hơn các cổ phiếu có chỉ số thấp hơn. Cuối cùng, trọng số của danh mục tối ưu được lập bằng cách sử dụng mô hình Markowitz với mục tiêu tối thiểu hóa rủi ro. Kết quả sẽ được so sánh với danh mục cơ sở là VN-Index.

Các giai đoạn xây dựng danh mục tối ưu có thể được tóm tắt trong Hình 1.

Bài viết này sử dụng phương pháp phân cụm với chuỗi thời gian để đề xuất danh mục trước khi thực hiện

Hình 1. Các giai đoạn xây dựng danh mục tối ưu



Nguồn: Theo Nanda & cộng sự (2010).

tối ưu hóa. Tác giả hy vọng đóng góp được một số điểm mới về nghiên cứu thực nghiệm trong việc xây dựng danh mục tối ưu trên thị trường chứng khoán tại Việt Nam.

Bố cục của bài viết như sau: phần đầu tiên là Giới thiệu và tổng quan nghiên cứu, phần tiếp theo là Phương pháp hai giai đoạn trong lựa chọn doanh mục đầu tư, phần thứ ba là Mô tả dữ liệu, phần thứ tư là Kết quả thực nghiệm, và cuối cùng là phần Kết luận.

2. Sử dụng hai giai đoạn trong lựa chọn danh mục đầu tư

2.1. Giai đoạn 1: Sử dụng K-means với chuỗi thời gian để phân cụm cổ phiếu và lựa chọn các cổ phiếu đưa vào danh mục

Phân cụm là việc nhóm các đối tượng thành các cụm, sao cho sự tương đồng của các đối tượng trong một cụm cao hơn so với các đối tượng ngoài cụm đó. Nhiều thuật toán phân cụm đã được đưa ra, chúng khác nhau về cách xác định sự tương đồng và cách thiết lập các cụm. Sự tương đồng giữa các đối tượng thường được thể hiện bởi khoảng cách giữa chúng. Nếu khoảng cách giữa hai đối tượng nhỏ thì hai đối tượng được coi là tương đồng cao (giống nhau về đặc điểm đang xem xét), ngược lại nếu khoảng cách giữa hai đối tượng lớn thì hai đối tượng được coi là ít tương đồng.

2.1.1. Phương pháp K-means

Phương pháp K-means phát biểu bởi MacQueen (1967) và Lloyd (1982). K-means là một thuật toán phân cụm phân vùng đơn giản và được sử dụng khá phổ biến. Các đối tượng tương tự được nhóm trong cùng một cụm bằng cách sử dụng tinh chỉnh lặp đi lặp lại với mục tiêu là giảm thiểu hàm tổn thất được tính bởi tổng các bình phương sai số giữa mỗi tâm cụm và các phần tử trong cụm.

Phân cụm K-means phân phần tử vào cụm, mỗi phần tử được phân vào cụm có tâm (trung bình của các phần tử) gần nó nhất. Trong mỗi lần tinh chỉnh, mục tiêu là tối thiểu hóa tổng bình phương khoảng cách:

$$\sum_{i=1}^k \sum_{x \in S_i} d(x, \mu_i)^2$$

trong đó x là quan sát, $S = \{S_1, S_2, \dots, S_k\}$ là các cụm và μ_i là trung bình của các điểm trong S_i (còn được gọi là tâm cụm S_i)

Các bước thực hiện thuật toán K-means như sau:

1. Tìm k tâm cụm ban đầu (có thể chọn bằng cách lấy ngẫu nhiên k phần tử).
2. Tính toán khoảng cách của mỗi phần tử tới các tâm. Phần tử sẽ được phân vào cụm có tâm gần nó nhất.
3. Tính lại k tâm cụm bằng cách lấy trung bình tất cả các phần tử được gán cho mỗi cụm.
4. Lặp lại các bước 2 và 3 cho đến khi các tâm cụm không còn thay đổi.

Trong phương pháp K-means có hai yếu tố cần quan tâm, đó là hàm khoảng cách và số lượng cụm khi thực hiện phân cụm.

Trong phát biểu ban đầu của phương pháp, K-means được sử dụng với dữ liệu tĩnh, mỗi phần tử được đại diện bởi các giá trị cố định (giống như các điểm m chiều). Tuy nhiên, phương pháp này có thể sử dụng được với dữ liệu chuỗi thời gian. Trong bài viết này, việc phân cụm thực hiện bởi phương pháp K-means với khoảng cách được lựa chọn để đo mức độ tương quan giữa các cổ phiếu.

2.1.2. Khoảng cách sử dụng

Golay & cộng sự (1998) đã đề xuất khoảng cách đo sự tương đồng giữa các chuỗi thời gian là khoảng cách dựa trên hệ số tương quan Pearson (còn được gọi là khoảng cách Pearson):

$$d_{cor}(x, y) = \sqrt{2(1 - \rho(x, y))} \quad (1)$$

trong đó ρ là hệ số tương quan Pearson, đo sự tương quan tuyến tính giữa x và y :

$$\rho(x, y) = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^n (x_i - \bar{x})^2} \sqrt{\sum_{i=1}^n (y_i - \bar{y})^2}} \quad (2)$$

với μ_x, μ_y là giá trị trung bình và σ_x, σ_y là độ lệch chuẩn của x và y .

Khi hai chuỗi có tương quan dương cao, có nghĩa là xu thế biến động của chúng là tương đồng lớn với nhau, thì khoảng cách giữa hai chuỗi là bé. Và ngược lại, khi hệ số tương quan là thấp, thì sự tương đồng là bé đi, sự khác biệt về xu thế biến động là lớn hơn – khoảng cách sẽ là lớn hơn. Khi hệ số tương quan âm, hai chuỗi có xu thế biến động ngược chiều nhau – khoảng cách giữa hai chuỗi sẽ càng lớn.

Trong nghiên cứu này, chúng tôi sử dụng các đặc trưng cho mỗi cổ phiếu là thông tin về cổ phiếu dưới dạng chuỗi thời gian, gồm lợi suất cổ phiếu và độ biến động của cổ phiếu đó. Đặc điểm quan sát của cổ phiếu sẽ được ghi nhận theo từng ngày giao dịch.

Nghiên cứu này sử dụng hai chuỗi thời gian thể hiện đặc điểm của cổ phiếu là chuỗi lợi suất và độ biến động của chuỗi lợi suất. Độ biến động được đo lường bởi phương sai có điều kiện thay đổi và được xác định

từ mô hình GARCH(1,1) (Engle & Rangle, 2008). Khi đó, khoảng cách giữa hai cổ phiếu được xác định như sau. Trước tiên, chúng tôi xác định khoảng cách giữa hai chuỗi lợi suất của hai cổ phiếu này, theo công thức (1), ký hiệu là d_1 , tương tự cho khoảng cách giữa hai chuỗi độ biến động, ký hiệu là d_2 . Khoảng cách giữa hai cổ phiếu sẽ là giá trị kết hợp bởi hai giá trị d_1 và d_2 theo công thức:

$$d = d_1 + d_2$$

Giá trị khoảng cách này nhằm đại diện cho sự tương đồng của hai cổ phiếu khi xem xét đồng thời theo đặc trưng: lợi suất và độ biến động theo thời gian. Tương tự như khoảng cách Pearson, khi khoảng cách giữa hai cổ phiếu lớn nghĩa là hai cổ phiếu có sự tương đồng thấp.

2.1.3 Lựa chọn số cụm

Một trong các cách lựa chọn số cụm k trong phương pháp K-means là sử dụng phương pháp khuỷu tay (Elbow method). Phương pháp này xác định độ phân tán trong cụm W_k cho các giá trị của k với:

$$W_k = \sum_{i=1}^k \frac{1}{2|S_i|} \sum_{x,y \in S_i} d(x,y)$$

trong đó $|S_i|$ là số lượng quan sát trong cụm i và $d(x,y)$ là khoảng cách giữa quan sát x và y . Số lượng cụm “tối ưu” là giá trị k mà khi ta tăng số cụm thêm nhưng hiệu quả giải thích của việc phân cụm không cải thiện nhiều. Biểu đồ giá trị của W_k tương ứng với giá trị của k thường được vẽ ra và giá trị được lựa chọn biểu hiện như một “khuỷu tay” trong biểu đồ.

2.1.4. Lựa chọn các cổ phiếu từ các cụm đưa vào danh mục

Sau khi các cổ phiếu được phân vào các cụm, chỉ số Sharpe của các cổ phiếu trong mỗi cụm được tính toán. Sau đó, từ 3 đến 5 cổ phiếu có chỉ số Sharpe cao nhất trong mỗi cụm được lấy ra để lập danh mục đầu tư. Tiếp theo trọng số của các cổ phiếu trong danh mục đầu tư sẽ được xác định qua việc giải bài toán tối ưu Markowitz.

Ở đây, khi thực hiện lựa chọn cổ phiếu theo lý thuyết danh mục đầu tư hiện đại của Markowitz, chúng ta phải đánh đổi giữa lợi suất và rủi ro. Nhà đầu tư có thể giảm rủi ro của toàn bộ danh mục đầu tư bằng cách kết hợp hai chứng khoán có hiệp phương sai âm, mà không làm giảm lợi tức kỳ vọng.

Chỉ số Sharpe được đặt tên theo kết quả công bố của Sharpe (1964). Chỉ số được tính theo công thức:

$$\text{Sharpe Ratio} = \frac{\text{Lợi suất trung bình} - \text{Lãi suất phi rủi ro}}{\text{Độ lệch chuẩn của lợi suất}}$$

Các cổ phiếu có chỉ số Sharpe cao phù hợp với mục tiêu tối thiểu rủi ro với lợi suất cố định. Cách lựa chọn chỉ số này tương tự như trong các nghiên cứu của Marvin (2015) và Iorio & cộng sự (2018).

2.2. Giai đoạn 2: Giải bài toán Markowitz xác định tỷ trọng của các cổ phiếu trong danh mục

Lý thuyết danh mục đầu tư Markowitz xác định rủi ro danh mục đầu tư là phương sai của chuỗi lợi tức của danh mục đầu tư. Bài toán Markowitz để xác định danh mục tối ưu trong trường hợp này có thể biểu diễn dưới dạng toán học như sau:

$$\min(\sigma^2) = \min \sum_{i=1}^n w_i^2 \sigma_i^2 + \sum_{i=1}^n \sum_{j=1}^n w_i w_j \sigma_{ij}, i \neq j$$

với các ràng buộc:

1. $\sum_{i=1}^n w_i = 1$: tổng tỉ trọng (trọng số) trong danh mục bằng 1
2. $w_i \geq 0$ với $i = 1, \dots, n$: không được bán khống
3. $\sum_{i=1}^n w_i R_i \geq R_0$: lợi suất của danh mục lớn hơn hoặc bằng lợi suất cho trước

trong đó $\sigma_i^2 = E[(R_i - E(R_i))^2]$ là phương sai của cổ phiếu i và $\sigma_{ij} = E[(R_i - E(R_i))(R_j - E(R_j))]$

là hiệp phương sai giữa cổ phiếu i và cổ phiếu j .

Chúng tôi thực hiện tìm vectơ trọng số của danh mục đầu tư đảm bảo một mức lợi tức kỳ vọng nhất định (định trước) sao cho rủi ro là thấp nhất. Trong bài viết này, lợi suất kỳ vọng là lợi suất trung bình của toàn thị trường trong mỗi giai đoạn nghiên cứu. Trọng số của các cổ phiếu của danh mục tối ưu là lời giải của bài toán tối ưu Markowitz.

3. Mô tả dữ liệu sử dụng

Bài báo sử dụng số liệu giá đóng cửa của các cổ phiếu được giao dịch trên sàn chứng khoán thành phố Hồ Chí Minh, thị trường chứng khoán Việt Nam. Để đánh giá kết quả của phương pháp trong các khoảng thời gian khác nhau, chúng tôi thực hiện kết quả trên bốn khoảng thời gian: từ ngày 3/1/2018 đến ngày 28/12/2018 (năm 2018), từ ngày 2/1/2019 đến ngày 31/12/2019 (năm 2019), và từ ngày 3/1/2018 đến ngày 31/12/2029 (hai năm 2018 – 2019), từ ngày 2/1/2020 đến ngày 31/12/2021 (hai năm 2020 - 2021). Những cổ phiếu mới tham gia hoặc ngừng niêm yết không giao dịch đủ trong các khoảng thời gian này được loại bỏ.

Lợi suất của cổ phiếu i tại ngày t được xác định bởi công thức: $r_{i,t} = \ln \frac{p_{i,t}}{p_{i,t-1}}$

trong đó $p_{i,t}$ và $p_{i,t-1}$ lần lượt là giá của cổ phiếu i tại ngày t và ngày $t-1$

Chuỗi biến động cổ phiếu i trong ngày t được xác định theo mô hình GARCH(1,1):

Phương trình trung bình: $r_{i,t} = \mu + \sigma_{i,t} \varepsilon_{i,t}$ với $\forall i = 1, 2, \dots, N_t$

Phương trình phương sai: $\sigma_{i,t}^2 = \alpha_0 + \alpha(r_{i,t-1} - \mu)^2 + \beta\sigma_{i,t-1}^2$

với $\alpha_0 > 0, \alpha > 0, \beta > 0; \alpha + \beta < 1$

trong đó N_t là số lượng cổ phiếu trong ngày t ; $\varepsilon_{i,t}$ là các biến ngẫu nhiên độc lập, cùng phân phối chuẩn hóa.

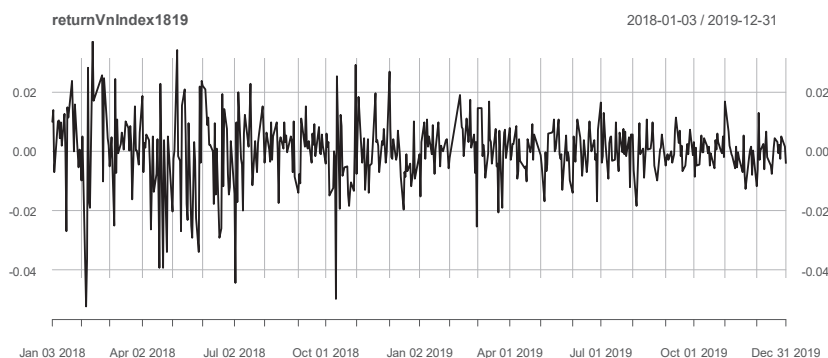
Giai đoạn đầu năm 2018 chỉ số Vn-Index có sự tăng và đạt mức cao nhất hơn 1200 điểm. Sau đó chỉ số thay đổi quanh ngưỡng 950 điểm. Trong năm 2019, sự dao động của VN-Index không quá chênh lệch, trong khoảng gần 900 tới hơn 1000 điểm.

Hình 2. Chỉ số VN-Index giai đoạn 2/1/2018 – 31/12/2019



Nguồn: Tác giả tính toán.

Hình 3. Lợi suất theo ngày của VN-Index giai đoạn 2018 – 2019



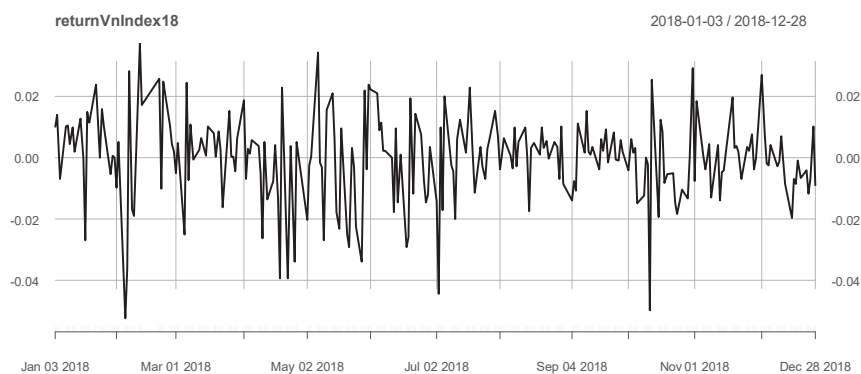
Nguồn: Tác giả tính toán.

Hình 4. VN-Index giai đoạn 3/1/2018 – 28/12/2018



Nguồn: Tác giả tính toán.

Hình 5. Lợi suất theo ngày của VN-Index trong năm 2018



Nguồn: Tác giả tính toán.

Hình 6. VN-Index giai đoạn 2/1/2019 – 31/12/2019



Nguồn: Tính toán của tác giả.

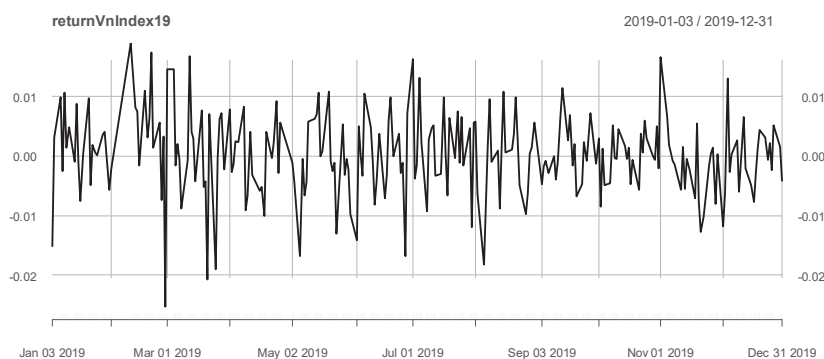
Giai đoạn này VN-Index giảm khá nhiều, mức lợi suất trung bình của VN-Index là giá trị âm (-0.000714) và mức rủi ro là 0.0110.

Trong năm 2018, VN-Index tăng trong khoảng đầu năm sau đó bước vào giai đoạn giảm. Chỉ số kết thúc năm 2018 với mức điểm ở ngưỡng 900 điểm (thấp nhất trong năm). Do đó, lợi suất trung bình của VN-Index năm 2018 là -0.00044 và mức rủi ro là 0.01408.

Trong năm 2019, VN-Index tăng trong giai đoạn đầu năm sau đó có giảm nhưng nhìn chung có sự tăng

trường. Lợi suất trung bình của VN-Index năm 2019 là 0.00030 và mức rủi ro là 0.00685.

Hình 7. Lợi suất theo ngày của VN-Index trong năm 2019



Nguồn: Tính toán của tác giả.

Lợi suất phi rủi ro được sử dụng để tính chỉ số Sharpe là trung bình lãi suất trái phiếu chính phủ 10 năm trong mỗi giai đoạn. Giá trị được tính toán theo số liệu công bố trên trang vn.investing.com.

Giai đoạn từ 2/1/2020 đến 31/12/2021 là một giai đoạn biến động mạnh của thị trường chứng khoán Việt Nam. VN-Index sụt giảm trong 4 tháng đầu năm 2020 từ ngưỡng 1000 điểm xuống mức 659 điểm. Sau đó là một giai đoạn tăng mạnh của VN-Index tới ngưỡng 1500 điểm vào cuối năm 2021.

4. Kết quả thực nghiệm

4.1. Số cụm lựa chọn

Để tìm ra số cụm tối ưu, chúng tôi thực hiện phân cụm K-means 20 lần. Mỗi lần các giá trị của số cụm được xem xét từ 2 đến 10 cụm và sử dụng phương pháp khuỷu tay để xác định số cụm lựa chọn. Số cụm tối ưu là giá trị có tần suất cao nhất trong số 20 lần thực hiện phân cụm.

Hình 8. VN-Index giai đoạn 2/1/2020 – 31/12/2021



Số cụm lựa chọn với dữ liệu trong các giai đoạn được thể hiện trong Bảng 1.

Trong năm 2018, các cổ phiếu tập trung trong Cụm 3, chiếm gần 50% tổng số cổ phiếu. Trong năm 2019, các cổ phiếu tập trung trong hai Cụm 3 và Cụm 4. Tương tự như vậy, trong các giai đoạn 2018 – 2019 và 2020 - 2021, cũng tồn tại một cụm tập trung nhiều cổ phiếu hơn hẳn các cụm còn lại. Điều này cho thấy rằng nhiều cổ phiếu trên thị trường có sự tương quan cao.

Sau khi tính toán chỉ số Sharpe của các cổ phiếu trong mỗi cụm, 3 – 5 cổ phiếu có chỉ số Sharpe cao nhất sẽ được lựa chọn để tạo danh mục. Tiếp theo, trọng số của các cổ phiếu trong danh mục sẽ được xác định để cho danh mục tối ưu.

4.2. Xác định danh mục tối ưu

Bài toán tối ưu theo mô hình Markowitz được sử dụng tìm véc tơ trọng số của danh mục sao cho mức rủi ro là thấp nhất với lợi suất không nhỏ hơn lợi suất trung bình của VN-Index.

Bảng 1. Số lượng cổ phiếu trong các cụm

Giai đoạn	Cụm 1	Cụm 2	Cụm 3	Cụm 4	Cụm 5	Cụm 6	Tổng
2018	62	73	150	19	28		332
2019	48	148	132	36			364
2018 - 2019	20	54	30	174	46		324
2020 – 2021	114	59	34	12	43	91	353

Nguồn: Tính toán của tác giả.

4.3. Nhận xét kết quả

Kết quả tính toán trên các giai đoạn được cho trong Bảng 3.

Chuỗi lợi suất tích lũy của các danh mục được so sánh với chuỗi lợi suất tích lũy của VN-Index. Danh mục tìm được khi thực hiện phương pháp với dữ liệu trong năm 2019 cho lợi suất tích lũy cao hơn của VN-

Bảng 2. Tỷ trọng các cổ phiếu trong danh mục tối ưu

TT	2018		2019		2018 - 2019		2020 – 2021	
	Cổ phiếu	Tỷ trọng	Cổ phiếu	Tỷ trọng	Cổ phiếu	Tỷ trọng	Cổ phiếu	Tỷ trọng
1	ACL	0.03145	BCG	0.02265	ACL	0.00152	DCM	0.01028
2	ANV	0.02362	DAH	0.02614	ANV	0.03007	DHM	0.10829
3	APG	0.08528	DHG	0.13001	APG	0.06237	GAB	0.23533
4	CIG	0.05326	FIR	0.29532	CCL	0.05351	KDC	0.12361
5	CMX	0.02944	FIT	0.02869	CMX	0.02515	LBM	0.09277
6	FTM	0.28493	GTN	0.02923	DAH	0.09443	LCM	0.00162
7	NTL	0.08172	SAM	0.28633	HOT	0.02053	LHG	0.01221
8	OGC	0.03348	TCH	0.05804	IJC	0.06911	PDR	0.11711
9	SFG	0.12889	TNC	0.01972	LGC	0.04946	SHI	0.05115
10	SMA	0.04006	TRC	0.04932	NTL	0.05171	SVT	0.04194
11	SVI	0.05149	VCB	0.05456	PGD	0.07148	TGG	0.01600
12	TNI	0.13091			SVI	0.05483	TNT	0.00338
13	UDC	0.02547			TNI	0.12638	TVS	0.08609
14					VCB	0.04244	VPG	0.04161
15					VSH	0.24701	VSC	0.5862
Tổng		1		1		1		1

Nguồn: Tính toán của tác giả.

Index trong toàn bộ giai đoạn. Danh mục tìm được với dữ liệu trong năm 2018 có khoảng 4 tháng đầu thấp hơn nhưng sau đó vẫn thể hiện ưu thế trong giai đoạn tiếp theo. Tương tự trong năm 2018, khi sử dụng dữ liệu của giai đoạn 2018 – 2019, lợi suất tích lũy của danh mục vẫn cao hơn của VN-Index trong khoảng thời gian sau tháng 4/2018.

Trong giai đoạn 2020 – 2021, lợi suất tích lũy của danh mục đề xuất thể hiện ưu thế so với lợi suất tích

Bảng 3. Lợi suất trung bình và độ lệch chuẩn của lợi suất của các danh mục và của VN-Index

Giai đoạn	Lợi suất trung bình		Độ lệch chuẩn của lợi suất	
	VN-Index	Danh mục	VN-Index	Danh mục
2018	-0.00044	0.00136	0.01408	0.00806
2019	0.00030	0.00131	0.00685	0.00504
2018 - 2019	-0.00007	0.00078	0.011048	0.00818
2020 - 2021	0.00088	0.00332	0.01390	0.01040

Nguồn: Tính toán của tác giả.

lũy của VN-Index. Trong 4 tháng đầu năm 2020, VN-Index giảm điểm, nhưng trong khoảng thời gian này lợi suất tích lũy của danh mục tăng khá mạnh.

Có thể thấy rằng, trong những giai đoạn mà VN-Index tăng điểm thì các danh mục càng có lợi suất tích

Hình 9. Lợi suất tích lũy của VN-Index và của danh mục trong năm 2018



Hình 10. Lợi suất tích lũy của VN-Index và của danh mục trong năm 2019

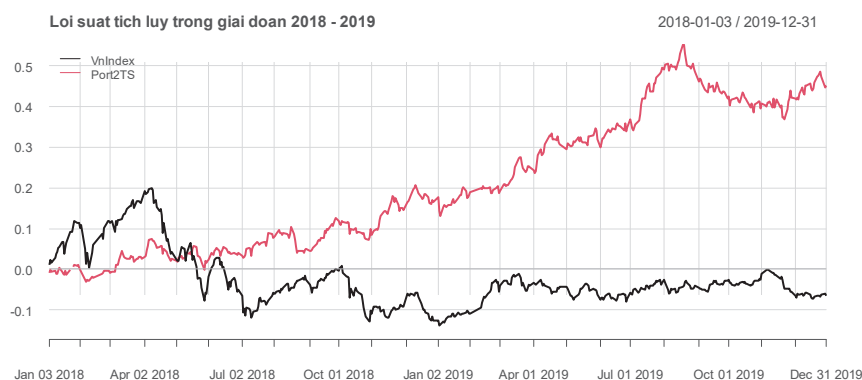


Nguồn: Tính toán của tác giả.

lũy tăng cao. Trong năm 2021, cùng với sự tăng điểm mạnh của VN-Index, danh mục tìm được cho lợi suất tích lũy tăng gần 7 lần so với tại thời điểm đầu năm 2020.

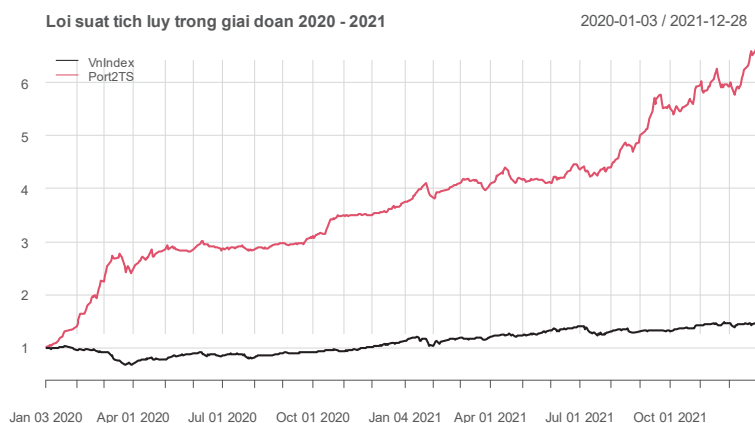
5. Kết luận

Hình 11. Lợi suất tích lũy của VN-Index và của danh mục trong giai đoạn 2018 - 2019



Nguồn: Tính toán của tác giả.

Hình 12. Lợi suất tích lũy của VN-Index và của danh mục trong giai đoạn 2020 - 2021



Nguồn: Tính toán của tác giả.

Mục đích của bài báo này xây dựng danh mục tối ưu kết hợp giữa phân cụm chuỗi thời gian và tối ưu hóa Markowitz. Phương pháp Kmeans thường được sử dụng với các véc tơ (bộ số) thể hiện những đặc điểm của đối tượng. Trong bài viết, tác giả sử dụng phân cụm K-means với chuỗi thời gian. Sự tương đồng giữa các cổ phiếu được đánh giá thông qua chuỗi lợi suất và độ biến động của chuỗi lợi suất của cổ phiếu. Các cổ phiếu có sự tương quan dương mạnh được nhóm thành các cụm. Sau đó, những cổ phiếu tốt nhất trong các cụm (theo chỉ số Sharpe) được chọn để lập danh mục đầu tư.

Việc lựa chọn một số cổ phiếu trong các cụm sẽ đưa ra một danh mục có số lượng nhỏ cổ phiếu. Điều đó giúp các nhà đầu tư dễ dàng kiểm soát và giảm chi phí trong việc quản trị. Nhà đầu tư cũng thuận lợi hơn trong việc thu thập thông tin bổ sung về các cổ phiếu trong danh mục. Ngoài ra, bài toán Markowitz phải sử dụng các ước lượng mẫu từ danh mục. Số lượng cổ phiếu nhỏ sẽ làm giảm bớt sai số của các ước lượng đầu vào này.

Kết quả thực nghiệm được thực hiện với dữ liệu lợi suất của các cổ phiếu trong giai đoạn hai giai đoạn: từ ngày 3/1/2018 đến ngày 28/12/2019 và từ ngày 2/1/2020 đến ngày 31/12/2021. Hai giai đoạn này đại diện cho hai thời gian “bình thường” và “chịu tác động khách quan”. Hai năm 2020 và 2021, các nhà đầu tư bị ảnh hưởng tâm lý bởi dịch Covid nên VN-Index thay đổi khá “cực đoan”: giảm sâu và sau đó tăng mạnh. Tuy nhiên, danh mục được đề xuất trong các giai đoạn thể hiện lợi thế so với tiêu chuẩn chung thị trường (đại diện bởi VN-Index). Khi so sánh với VN-Index, trong các giai đoạn danh mục tối ưu đều cho lợi suất

trung bình cao hơn và độ lệch chuẩn của chuỗi lợi suất thấp hơn. Trong những giai đoạn VN-Index có lợi suất trung bình dương, danh mục có lợi suất tích lũy cao hơn rõ rệt so với lợi suất tích lũy của VN-Index.

Dựa vào kết quả thực nghiệm có thể thấy kỹ thuật xây dựng danh mục tối ưu dựa trên phân cụm chuỗi thời gian có hiệu quả trong giai đoạn nghiên cứu. Chúng tôi hy vọng kết quả này có thể là một phương pháp để các nhà đầu tư tham khảo khi tham gia thị trường chứng khoán.

Tài liệu tham khảo

- Aghabozorgi, S., Shirkhorshidi, A. S. & Wah, T. Y. (2015), 'Time series clustering - A decade review', *Information System*, 53, 16 - 38.
- Black, F. & Litterman, R. (1992), 'Global portfolio optimization', *Financial Analysts Journal*, 48(5), 28 - 43.
- Demiguel, V., Garlappi, L. & Uppal R. (2007), 'Optimal versus Naive Diversification: How Inefficient Is the 1/N Portfolio Strategy?', *The Review of Financial Studies*, 22, 3411 - 3447.
- Engle, R.F. & Rangel, J.G. (2008), 'The spline-GARCH model for low-frequency volatility and its global macroeconomic causes', *The Review of Financial Studies*, 21(3), 1187-1222.
- Fisher, L. & Lorie, J. (1970), 'Some studies of variability of returns on investments in common stock', *Journal of Business*, 43(2), 99 - 134.
- Lloyd, S.P. (1982), 'Least squares quantization in PCM', *IEEE Transactions on Information Theory*, 28 (2), 129-137.
- Golay, X., Kollias, S., Stoll, G., Meier, D., Valavanis, A. & Boesiger, P. (1998), 'A new correlation-based fuzzy logic clustering algorithm for FMRI', *Magnetic Resonance in Medicine*, 40(2), 249-260.
- Iorio, C., Frasso, G., Dambrosio, A. & Siciliano, R. (2018), 'A p-spline based clustering approach for portfolio selection', *Expert Systems with Applications*, 95, 88 - 103.
- Javed, A., Lee, B.S. & Rizzo, D.M. (2020), 'A benchmark study on time series clustering', *Machine Learning with Applications*, DOI: 10.1016/j.mlwa.2020.100001.
- Kim, J. & Kim, J. (2020), 'Comparison of time series clustering methods and application to power consumption pattern clustering', *Communications for Statistical Applications and Methods*, 27(6), 589 - 602.
- Marvin, K. (2015), 'Creating diversified portfolios using cluster analysis', Independent Work Report Fall, Princeton University.
- Markowitz, H. (1952), 'Portfolio Selection', *The Journal of Finance*, 7(1), 77 - 91.
- MacQueen, J.B. (1967), '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
- Montero, P., & Vilar, J. A. (2014), 'TSclust: An R Package for Time Series Clustering', *Journal of Statistical Software*, 62(1), 1 - 43.
- Nanda, S., Mahanty, B. & Tiwari, M. (2010), 'Clustering Indian stock market data for portfolio management', *Expert Systems with Applications*, 37, 8793 - 8798.
- Nejad, M.K. (2014), 'Clustering stock exchange data by using evolutionary algorithms for portfolio management', *European Research Studies Journal*, 17(4), 55 - 66.
- Liao, T.W. (2005), 'Clustering of time series data - A survey', *Pattern Recognition*, 38(11), 1857 - 1874.
- Tavakoli, N., Siami-Namini, S., Adl Khanghah, M., Soltani, F.M. & Namin, A.S. (2020), 'An autoencoder-based deep learning approach for clustering time series data', *Springer Nature Applied Sciences*, 2, DOI: <https://doi.org/10.1007/s42452-020-2584-8>.
- Trần Văn Trí (2015), 'Xây dựng và quản lý danh mục đầu tư chứng khoán ở Việt Nam', Luận án Tiến sĩ, Học viện Ngân hàng.
- Sharpe, W.F. (1994), 'The sharpe ratio', *Journal of Portfolio Management*, 21(1), 49 - 58.
- Võ Thị Thủy Anh (2012), 'Đề xuất lựa chọn cổ phiếu thống trị theo lý thuyết Markowitz', *Công nghệ Ngân hàng*, số 75 tháng 6/2012, 53 - 55.