



e-ISSN: 2319-8753 | p-ISSN: 2347-6710

IJIRSET

International Journal of Innovative Research in
SCIENCE | ENGINEERING | TECHNOLOGY



INTERNATIONAL JOURNAL OF INNOVATIVE RESEARCH

IN SCIENCE | ENGINEERING | TECHNOLOGY

Volume 10, Issue 12, December 2021

ISSN INTERNATIONAL
STANDARD
SERIAL
NUMBER
INDIA

Impact Factor: 7.569



9940 572 462



6381 907 438



ijirset@gmail.com



www.ijirset.com

Portfolio Selection Method based on Pattern Matching with Dual Information of Direction and Distance: A Comprehensive Analysis of Online Portfolio Strategies

Gayatri Annasagaram¹, Vanitha G R², Nagendra Naik K³, Rekha R⁴, Jyothi⁵

Assistant Professors, Department of Mathematics, City Engineering College, Bengaluru, Karnataka, India^{1,2,3,4,5}

ABSTRACT: The pattern matching method has long been recognized as a traditional and effective classification approach within the domain of online portfolio selection strategies. This article delves into the core elements of this method, focusing on two key components: the assessment of similarity and the selection of similarity sets. To address the growing complexity of financial markets, this study introduces an enhanced approach known as the Portfolio Selection Method based on Pattern Matching with Dual Information of Direction and Distance (PMDI). The aim is to optimize portfolio selection by integrating a more comprehensive analysis of historical stock price patterns. PMDI leverages a dual-information framework, combining both direction and distance metrics to create a more accurate and refined pattern-matching process. The method incorporates various well-known techniques, such as the Euclidean distance, Chebyshev distance, and the correlation coefficient, to capture the essential information embedded in historical stock price movements. By analyzing both the directional trends (the way prices move over time) and the relative distances between stock prices, PMDI identifies a similarity set that closely mirrors the desired investment strategy. This similarity set forms the foundation of the portfolio selection algorithm. The novelty of PMDI lies in its ability to not only assess patterns based on price direction but also to quantify how far stock prices deviate from one another. This dual approach enhances the system's capability to filter out noise and irrelevant data, thereby selecting only the most relevant patterns for further processing. As a result, the system can make more informed decisions when constructing an optimal investment portfolio, balancing between maximizing returns and minimizing risks. To validate the effectiveness of PMDI, extensive experiments were carried out using two real-world stock market datasets. The results of these tests demonstrated that PMDI outperformed other traditional algorithms, achieving a superior balance between returns and risk management. In comparison to single-metric models, which may focus only on direction or distance, PMDI's dual-information framework proves to be more robust and adaptable to the ever-changing dynamics of financial markets. In the increasingly volatile world of finance, having a strategy that can dynamically respond to market shifts is crucial. The PMDI method offers a significant improvement over existing portfolio selection techniques by utilizing a richer set of data inputs. Its ability to handle both directional trends and price deviations makes it particularly well-suited for real-world financial environments, where adaptability and precision are key to achieving sustainable long-term returns. As financial markets evolve, methods like PMDI will likely become essential tools for investors seeking to maximize performance while effectively managing risk.

KEYWORDS: Portfolio Selection, Pattern Matching, Dual Information, Financial Markets, Risk Management

I. INTRODUCTION

The allocation of limited capital in global financial markets has consistently garnered significant interest from both investors and researchers, particularly in the quest to optimize returns. The intricate nature and high interconnectivity of these markets complicate the ability to forecast market movements. Additionally, elements such as financial innovation, diverse trading strategies, and regulatory frameworks contribute to the prevailing uncertainty within the market. Consequently, devising effective capital allocation strategies within this multifaceted environment has emerged as a pressing global issue. Portfolio Selection (PS) stands as a fundamental endeavour within the financial sector, focusing on the distribution of capital across a defined array of assets to meet specific investment objectives. The roots of this challenge can be traced back to the mean-variance theory introduced by Markowitz in 1952, which employs mean-variance analysis to identify the optimal investment portfolio by balancing anticipated returns against associated risks. Nevertheless, as market dynamics evolve and investor preferences diversify, conventional portfolio selection techniques have increasingly revealed their shortcomings.

In recent years, the swift advancement of artificial intelligence and machine learning has introduced innovative approaches to address portfolio selection challenges. These technologies possess the capability to scrutinize historical data, uncovering patterns and trends that facilitate more accurate predictions of future market behaviour. By processing

extensive datasets, machine learning models can discern intricate relationships among variables, thereby enhancing predictive accuracy. This advancement not only aids in refining investment strategies but also equips investors and decision-makers with critical insights, enabling them to make more informed and judicious choices.

II. PATTERN-MATCHING BASED APPROACHES

This section introduces benchmark strategy (which will be used for comparative experiments in subsequent sections) and pattern-matching strategy.

2.1 Standards for comparison

- BAH. Buy-and-Hold strategy, in which assets are bought and held until the end of the period with the initial weight b , and the warehouse is not just during the trading period; Market.
- A uniform BAH strategy with initial weight Best-stock. It is also a BAH strategy, a hindsight strategy that invests all the money into assets that perform best in hindsight; CRP.
- Constant Rebalanced Portfolios (CRP) maintain the portfolio's initial weight through constant warehouse adjustments, as the value of each asset fluctuates over time;
- UCRP. Uniform Constant Rebalanced Portfolios (UCRP) rebalance to the same uniform weight in each trading period;
- BCRP. It is also a CRP strategy, a hindsight strategy with portfolio $\ast b$ to maximize the final wealth.

2.2 A Portfolio Selection Method Based on Pattern Matching with Dual Information of Direction and Distance

Investors often face challenges when attempting to invest in a vast array of stocks simultaneously, as it is impractical to manage thousands of options at once. Typically, they must select a limited number of stocks, often ranging from a few to several dozen, which are considered risk assets, and allocate their investments accordingly. Consequently, the fundamental issue in portfolio management revolves around the scientific and proportional distribution of investment capital across various assets, aiming to optimize returns while minimizing risks. In this chapter, we introduce an innovative strategy that aligns with this framework.

The method for selecting an investment portfolio is based on a matching approach. By identifying a sequence of time intervals characterized by a cycle denoted as w , we analyse stock price fluctuations that closely resemble those observed in the most recent w periods on each trading day. This analysis yields a set of stocks exhibiting similarity, which is then subjected to optimization algorithms to derive the desired investment portfolio for the forthcoming trading day. Additionally, this process allows for the extraction of valuable insights, including directional trends and distances relative to historical stock prices. This approach is referred to as the Portfolio Selection Method based on Pattern Matching with Dual Information of Direction and Distance (PMDI).

2.3 Basic Idea

Early non-parametric learning techniques, such as BK and BNN, primarily relied on Euclidean distance to assess the similarity between current and past market windows. A significant limitation of this approach is its failure to consider the directional information inherent in market window movements, which may lead to the inclusion of irrelevant or detrimental relative prices. To address this issue, CORN employs the Pearson product-moment correlation coefficient as a measurement tool. This coefficient essentially represents the cosine of the angle between two standardized data sets, serving to evaluate the directional relationship between the data points.

Despite its advantages, the CORN strategy is not without flaws. The historical similarity set it establishes encompasses all market vectors that maintain a correlation coefficient with the current market vector above a predetermined threshold, denoted as c . An inappropriate selection of this threshold, whether excessively high or low, can result in the erroneous inclusion of vectors that significantly diverge from the current market vector, thereby undermining the effectiveness of the CORN strategy. To enhance the previous algorithm, PMDI proposes a more sophisticated method for selecting similar sets to yield improved outcomes. Consequently, this paper investigates various distance measurement techniques, including Euclidean distance, Chebyshev distance, and the correlation coefficient, alongside the K-Nearest Neighbours (KNN) algorithm.

In an investment portfolio, the Euclidean distance metric is particularly responsive to variations in individual stock values, as it computes linear distances; thus, even minor fluctuations in any dimension can significantly influence the overall distance. When stock prices exhibit a high degree of correlation, employing Euclidean distance may be more

suitable. However, it is essential to recognize that stock price movements are often interconnected with broader market conditions, such as the effects of an epidemic on the financial landscape. Consequently, it is crucial to consider Euclidean distance when assessing similarity combinations.

Within an investment portfolio, the Pearson correlation coefficient serves as a metric for assessing the extent of linear dependence between two sets of relative price data. A coefficient nearing 1 signifies a robust positive linear correlation between the price data at two distinct time intervals, suggesting that an increase in one price vector is likely accompanied by an increase in the other. Conversely, a coefficient approaching -1 indicates a strong negative linear correlation, implying that an increase in one price vector corresponds with a decrease in the other. When the Pearson correlation coefficient is approximately 0, it suggests the absence of a significant linear relationship between the price vectors at the two time points. In the process of forming similarity combinations, it is essential to take into account the Pearson correlation coefficient. Our preference leans towards selecting relative price vectors that exhibit positive Pearson correlation coefficients, as these are likely to yield more meaningful relationships in the analysis.

The K-Nearest Neighbours (KNN) algorithm, while not a direct measure of distance, serves as an effective classification tool within the realm of machine learning, particularly for identifying similar sets in pattern recognition tasks. The fundamental operation of KNN involves predicting the category of a new sample by locating the K nearest sample points from the training dataset and determining the new sample's category based on the classifications of these neighbouring points. The parameter "K" plays a crucial role, as it dictates the number of neighbouring samples that will influence the classification decision.

Typically, the KNN algorithm employs distance metrics such as Euclidean distance or Manhattan distance to assess the similarity or distance between sample points. For illustration, consider a scenario where the small dashed circle represents $K = 3$ and the large dashed circle represents $K = 5$. In this context, the central red point signifies the sample to be predicted. When $K = 3$, the algorithm identifies two blue crosses and one green star in proximity, leading to the conclusion that the predicted point belongs to the "blue cross" category. Conversely, with $K = 9$, the surrounding points consist of four blue crosses and five green stars, resulting in the classification of the predicted point as a "green star." It is important to note that this algorithm may struggle with misclassification in domains with limited sample sizes, as evidenced by the classification results depicted in the diagram. KNN is generally more effective for automatic classification in domains characterized by larger sample sizes.

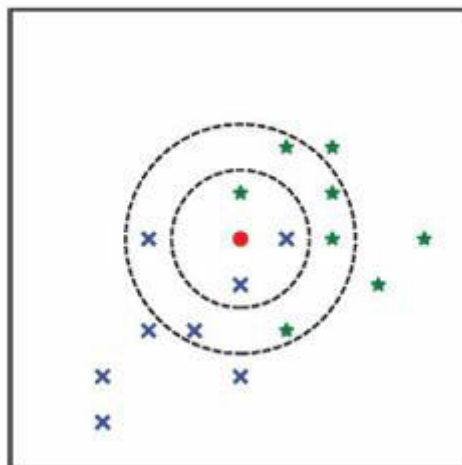


Figure 1. KNN Algorithm

The KNN algorithm can be employed to forecast the categories or trends of future price vectors by analysing historical data. In this context, the KNN distance represents the proximity between the current price vector and the K nearest neighbour points within the historical dataset. A small KNN distance indicates that the current price vector closely resembles a price vector from a specific historical period, suggesting that the market may be exhibiting similar behavioural patterns. This observation allows for the possibility of conducting trend analysis based on past market behaviours. 2. In light of this analysis, the article proposes a methodology that integrates various ranging techniques to establish a historical similarity set characterized by the highest degree of similarity across historical moments.

The algorithm primarily utilizes the Correlation Coefficient and Chebyshev Distance as its distance metrics. Initially, it filters historical stock prices that align closely with the directional information from the current trading period, as well as the distance metrics between individual stocks. Subsequently, the KNN algorithm is applied using Euclidean distance as a measure to refine the similar sets further, thereby extracting relevant distance and directional information from the historical stock prices.

III. PORTFOLIO OPTIMIZATION

The initial two sections introduced a methodology for identifying analogous sets within the context of pattern matching. Subsequently, by determining the similar set C for each trading day, an optimal investment portfolio is derived by maximizing the cumulative return of all relative price vectors contained within that similar set. The comprehensive procedure for portfolio optimization is illustrated in Algorithm 2. By employing this algorithm to generate b_t on a daily basis, it will serve as the investment portfolio for the subsequent trading day. Through this recursive approach, the ultimate investment return can be computed.

Algorithm 1: Algorithm of PMDI

Data: X_1^{t-1} : Historical market Windows
Data: w : Window size
Data: m : Chebyshev Distance threshold
Data: ρ : Correlation Coefficient threshold
Data: l : Number of neighbors
Result: C_2 : Similarity set
Initialize: $C_1 = \emptyset, C_2 = \emptyset$
for $i = w + 1$ **to** $t - 1$ **do**
 if $\|X_{t-w}^{t-1} - X_{i-w}^{i-1}\|_\infty \leq m$ **or** $\text{corrcoeff}(X_{t-w}^{t-1}, X_{i-w}^{i-1}) \geq \rho$ **then**
 $C_1 = C_1 \cup \{i\}$
 end
end
for $i \in C_1$ **do**
 if X_{i-w}^{i-1} **is among the** l **NNs of** X_{t-w}^{t-1} **then**
 $C_2 = C_2 \cup \{i\}$
 end
end

Algorithm 2: Portfolio optimization process

Data: X_1^{t-1} : Historical price series
Data: w : Window size
Data: t : Current trading day
Data: C_2 : Similarity set
Result: b_t : Portfolio
if $t \leq w + 1$ **then**
 $b_t = (\frac{1}{M}, \frac{1}{M}, \dots, \frac{1}{M})$
end
if $C_2 = \emptyset$ **then**
 $b_t = (\frac{1}{M}, \frac{1}{M}, \dots, \frac{1}{M})$
else
 $b_t = \arg \max_{b \in \Delta_M} \prod_{i \in C_2} b \cdot x_i$
end

The Sequential Quadratic Programming (SQP) algorithm is recognized for its super linear convergence rate and minimal requirement for function and gradient evaluations. This algorithm adeptly converts intricate nonlinear constrained optimization challenges into more manageable quadratic programming (QP) problems, making it one of the most efficient techniques for addressing nonlinear optimization issues that involve equality and bound constraints. Consequently, this paper will employ the SQP algorithm to enhance the investment portfolio.

The SQP method can be categorized into two distinct approaches based on how it resolves sub problems in quadratic programming: the line search SQP method and the trust region SQP method. Both approaches play a crucial role in ensuring global convergence and in determining the displacement and new iteration points during each iteration. The primary distinction between these two methods is that the line search method initially establishes the search direction before selecting the step size, whereas the trust region method directly determines both the direction and step size, subsequently generating new iteration points.

IV. EXPERIMENTS

4.1 Preparation

In order to test the empirical effect of the proposed strategy, we conducted extensive experiments on 2 real-world stock market data sets, which are summarised in Table 1.

For the convenience of obtaining and easy to reproduce, this article uses daily closing price data from the stock market. One dataset is widely used by other scholars in the field. NYSE (O) was first used by Cover (1991) [9] and has been reused by many later scholars. It consists of the closing prices of 36 stocks on the New York Stock Exchange over 5651 trading days. In addition to the classic data sets NYSE (O), we collect an additional data set. The data set is the list of 19 high-market stocks in the China Securities 500 Index. The data spans 2405 trading days, or ten years, from December 25th 2012 to November 18th 2022. We call this data set “CSI500”. This is a very new dataset that also takes into account the pandemic period of the past three years. The purpose of collecting this dataset is to determine whether our algorithm can adapt to the economic development of the new era and handle the impact of the epidemic on the economy. Also since the data sets are stored as relative prices, we do not need to consider exchange rates.

Table 1. Summary of the 2 real data sets

Dataset	Region	Time Frame	Trading days	Assets
CSI500	CN	Dec. 25 th 2012 - Nov. 18 th 2022	2405	19
NYSE(O)	US	Jul. 3 rd 1962 - Dec. 31 st 1984	5651	36

The annualized standard deviation is an important indicator for measuring the volatility of investment portfolios, which can reflect the risk level of investment portfolios and help us better manage investment risks.

V. SR (SHARPE RATIO)

The Sharpe ratio [12] is one of the three classic indicators that can comprehensively consider both returns and risks. As a risk adjusted rate of return, it can eliminate the adverse effects of risk factors on performance evaluation. The formula for calculating the Sharpe ratio is: $APY \text{ Sharpe Ratio} = \frac{R - R_f}{\sigma}$ Wherein, R_f , R is the annualized risk-free interest rate, which generally adopts the interest rate of treasury bond in the same period. The purpose of this formula is to calculate how much excess return will be generated per unit of total risk undertaken by the investment portfolio.

5.1 Cumulative Wealth

The first experiment evaluates the CW obtained by various strategies without considering transaction costs. Table 2 summarizes the CW obtained by various algorithms on four datasets, with bold numbers indicating the top ranked achievement on each dataset. On the CSI500 dataset, the PMDI algorithm achieved the best results, which proves the superiority of the PMDI algorithm. Although the performance of the PMDI algorithm on the NYSE (O) dataset is not as good as other pattern matching algorithms, it is worth mentioning that the PMDI algorithm does not use expert algorithms, which greatly accelerates the running speed of the algorithm compared to other pattern matching algorithms. It is not easy to achieve a revenue of nearly 100 million dollars in this situation.

Table 2. The CW obtained through different strategies

Strategies	CSI500	NYSE (O)
Market	4.72	14.50
Best-stock	19.51	54.14
BCRP	26.79	250.60
B^K	11.06	1.08E+09
B^{NN}	22.50	3.19E+11
CORN	30.62	1.45E+13
PMDI	61.20	3.60E+07

5.2 APY

The performance of APY is almost consistent with that of CW, as shown in the Table 3, so it will not be further elaborated here.

Table 3. APYs of different Strategies

Strategies	CSI500	NYSE (O)
Market	18%	13%
Best-stock	37%	19%
BCRP	41%	28%
B^K	29%	153%
B^{NN}	39%	226%
CORN	43%	286%
PMDI	54%	117%

Table 4. Risks of different Strategies

Strategies	CSI500	NYSE (O)
Market	31%	15%
Best-stock	60%	24%
BCRP	47%	31%
B^K	36%	36%
B^{NN}	38%	40%
CORN	46%	49%
PMDI	48%	42%

5.3 Risk

In this section, we analyse the volatility of strategy returns, specifically focusing on the annualized standard deviation. The data presented in Table 4 illustrates a relationship between the volatility of returns for each strategy and both cumulative and annualized returns, thereby reinforcing the investment principle that suggests a correlation between high returns and high risks. By evaluating this metric, we gain a more comprehensive insight into the risk profiles associated with each strategy, which in turn supports more informed investment decisions. It is widely recognized that elevated risk can lead to substantial returns. Notably, the PMDI algorithm yielded the most favourable returns on the CSI500 dataset, although its risk indicator was not the highest among the strategies. Consequently, it is anticipated that the risk-adjusted returns discussed in the subsequent section will be the most favourable.

Table 5. Sharpe ratios of different Strategies

Strategies	CSI500	NYSE (O)
Market	43%	58%
Best-stock	55%	64%
BCRP	79%	78%
B^K	68%	409%
B^{NN}	92%	551%
CORN	86%	578%
PMDI	103%	269%

5.4 Sharpe Ratios

From the table, it can be seen that the Sharpe ratio of PMDI has always maintained a high level of over 100%.

VI. CONCLUSIONS

The approach known as Matching with Dual Information of Direction and Distance (PMDI) is designed to enhance pattern matching techniques within online portfolio selection algorithms. This method presents theoretical benefits, as it innovatively introduces a technique for deriving similarity sets through an analysis of various historical pattern matching algorithms. By integrating multiple similarity measurement techniques, the KNN algorithm is employed to identify the most effective method. Following a thorough examination of historical stock price data, a similarity set that incorporates both distance and directional information is established, upon which optimization algorithms are subsequently applied. In addition to its theoretical contributions, extensive empirical testing was conducted using two distinct datasets, which included real-world stock data. The PMDI method demonstrated superior performance compared to all other algorithms in one of the datasets. While the outcomes in the second dataset did not meet expectations, the runtime was significantly reduced by avoiding the use of expert algorithms. The differing timeframes and financial contexts of the two datasets underscore the robustness and adaptability of the PMDI algorithm across various market conditions. Consequently, this method is well-suited for practical applications in financial environments, such as high-frequency trading. Future research will continue to investigate the potential of pattern matching-based methodologies in addressing online portfolio selection challenges.

REFERENCES

1. Markowitz, H. (1952). Portfolio Selection. *Journal of Finance*, 7(1), 77-91.
2. Li, B., & Hoi, S. C. H. (2014). Online portfolio selection: A survey. *ACM Computing Surveys (CSUR)*, 46(3), 1-36.
3. Gyorfi, L., Ottucsak, G., & Vajda, I. (2006). Machine learning for financial market prediction. *International Journal of Theoretical and Applied Finance*, 9(03), 403-424.
4. Cover, T. M. (1991). Universal portfolios. *Mathematical Finance*, 1(1), 1-29.
5. Li, B., & Hoi, S. C. H. (2012). Pattern matching for online portfolio selection. *SIAM International Conference on Data Mining*.



6. Zhou, Y., & Li, B. (2016). Robust pattern matching for online portfolio selection. *Computational Economics*, 47(2), 227-254.
7. Tsang, E. P., Li, J., & Butler, J. M. (2013). Eddie for financial portfolio management. *ACM Transactions on Economics and Computation*, 1(2), 1-27.
8. Cesa-Bianchi, N., & Lugosi, G. (2006). *Prediction, learning, and games*. Cambridge University Press.
9. DeMiguel, V., Garlappi, L., & Uppal, R. (2009). Optimal versus naive diversification: How inefficient is the 1/N portfolio strategy? *Review of Financial Studies*, 22(5), 1915-1953.
10. Gao, X., & Zhang, S. (2017). Stock portfolio selection using genetic algorithms and pattern recognition. *Expert Systems with Applications*, 72, 128-136.
11. Maringer, D. (2006). *Portfolio management with heuristic optimization*. Springer.
12. Bodie, Z., Kane, A., & Marcus, A. J. (2014). *Investments* (10th ed.). McGraw-Hill Education.



**Impact Factor:
7.569**



INTERNATIONAL JOURNAL OF INNOVATIVE RESEARCH

IN SCIENCE | ENGINEERING | TECHNOLOGY

 **9940 572 462**  **6381 907 438**  **ijirset@gmail.com**



www.ijirset.com

Scan to save the contact details