POLITECNICO
MILANO 1863

SCUOLA DI INGEGNERIA INDUSTRIALE
E DELL'INFORMAZIONE

Executive Summary of the Whitepaper

Leveraging Cluster Analysis for Optimized Portfolio Construction in the SP 500

Laurea Magistrale in Computer Science and Engineering - Ingegneria Informatica

Author: Gloria Desideri

Academic year: 2023-2024

---

1. Introduction

In the dynamic landscape of financial markets, constructing a well-diversified portfolio that yields desirable returns while minimizing risk remains a paramount challenge for investors. Traditional approaches often rely on individual stock selection or passive index investing, both of which may fall short in providing an optimal balance between returns and risk management.
This whitepaper proposes an innovative approach to portfolio construction by harnessing the power of cluster analysis applied to the time series data of stocks within the SP 500 index. By grouping stocks based on their historical price movements and correlations, we aim to extract a more concentrated portfolio with similar expected results but at a lower and more accessible book price.
In this paper, we delve into the methodology behind cluster analysis and its application in portfolio optimization. We present the optimization problem formulation, elucidate the techniques for clustering time series data, and discuss the potential advantages and disadvantages of this approach.
Furthermore, we explore potential avenues for further improvement, including refining clustering algorithms, addressing limitations, and incorporating additional factors for a more comprehensive portfolio optimization strategy.

2. Related Work

In the work of Vasquez et al. the authors explore different clustering strategies based on closing prices and daily returns. They clusterize a bivariate time series using two metrics: the Dynamic Time Warp and the Fourier decomposition. They also employ the clustering of financial ratios from the fundamental analysis. The EROS is employed in the clustering process because it allows for the comparison of the set of series of each stock based on their interdependence and relationships over time. This measure is particularly useful when dealing with a large number of financial ratios features and when capturing the complex dynamics and dependencies present in financial data.
By using EROS, the clustering algorithm can effectively group stocks based on the relationships and patterns observed in their financial ratios time series, providing a comprehensive and insightful way to analyze and cluster stocks within the dataset. The Extended Frobenius Norm (EROS) measure is used in the paper for clustering stocks based on financial ratios. EROS was proposed by Yang and Shahabi in 2004 and is a distance measure that combines the results of Principal Component Analysis (PCA)

and the Frobenius Norm to compare time series data. After obtaining meaningful clusters they perform forecasting for the stocks in the clusters. They show how LSTM outperform ARIMA models at cost of computational complexity and training time. Due to the computational and time limitations of this whitepaper, I explored a univariate time series clustering using DTW and forecasting for the actions chosen as representatives for each cluster using ARIMA.

3. Data collection

To collect the necessary data I used the Yahoo Finnce API. I extracted the daily close prices and daily adjusted prices for all the stocks in SP500 from 2015 to 2022. To have a smoother function I used a 5 days Moving Average. I used a dataset containing sectors and industry of every stock to decorate the previous dataset. In this way I could Explore the differences in the time series of different sectors. To overcome the computational limitations I transformed the data in a monthly time series to perform the clustering.

4. Clustering

In this work we used Dynamic Time Warping to clusterize the time series this measure is based on curve similarity and is able to compute the similarity even if the two series vary in time and speed. The core idea behind DTW is to find an optimal alignment between two sequences, even if they are not perfectly synchronized, by minimizing the distance between them.
The process involves creating a "warping path" that aligns the two sequences in time. This path is found using a dynamic programming approach, which calculates the optimal alignment by minimizing the Euclidean distance between the aligned sequences. The warping path is a sequence of points that connects the start and end points of the two sequences, allowing for a flexible alignment that can accommodate differences in speed or timing between the sequences.

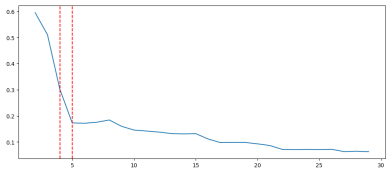

Figure 2: Elbow method

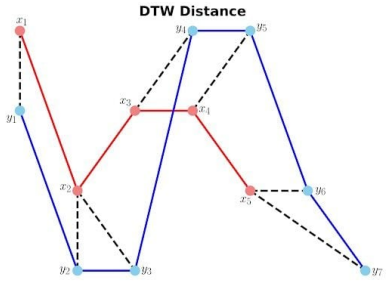

Figure 1: Dynamic time warping

Before applying the DTW algorithm, I standardized the time series data by scaling each series to have a mean of zero and a standard deviation of one. This preprocessing step ensures that all time series are on a comparable scale, mitigating the influence of magnitude variations on the clustering process. To determine the optimal number of clusters for our dataset, I conducted cross-validation of the k-means clustering algorithm. We varied the number of clusters (k) from 2 to 30 and evaluated the clustering performance using two commonly employed metrics: the silhouette score and the elbow method.
The silhouette score quantifies the cohesion and separation of clusters, ranging from -1 to 1, where a higher score indicates better-defined clusters. The elbow method, on the other hand, identifies the point where the rate of decrease in within-cluster variance slows down, suggesting the optimal number of clusters. After thorough cross-validation, I observed that the most promising result was obtained with five clusters, yielding a silhouette score of 0.1729. This indicates a moderate level of cluster separation and cohesion, suggesting a meaningful partitioning of the SP 500 stocks based on their adjusted price time series. To choose the representatives for each cluster I calculated the distance between every stock in the cluster and chose the k that were closer to the centroid where k is a proportionality constraint based on the cluster size. The total new representatives are 49 stocks.

5. Forecasting

For the forecasting I used an ARIMA model which is made of three parts:

- AutoRegressive (AR) Component: This component models the relationship between an observation and a lagged set of observations.
- Integrated (I) Component: This component accounts for differencing, which involves taking the difference between consecutive observations to make the time series stationary.
- Moving Average (MA) Component: This component models the relationship between an observation and a residual error from a moving average model applied to lagged observations.

The parameters of an ARIMA model are denoted by p, d and q representing the number of autoregressive terms, the degree of differencing, and the number of moving average terms, respectively. To ensure a better precision I used the whole series of data for each stock instead of the monthly series. To tune the order I used the autoArima library and with the fine tuned order I predicted the expected return of each stock with lots variance.

6. Optimization problem

To formulate the problem I started with a simple portfolio optimization problem in which we try to maximize the expected return of the portfolio while minimizing the variability of return. I added a few more constraints to the problem

- Proportionality constraint: $|x_i - k \cdot w_{\text{index}_i}| \leq \epsilon$. The new weights must me proportional to the ones of the index with a certain tolerance.
- Return similarity: $|E(R_{\text{portfolio}}) - E(R_{\text{index}})| \leq \delta$. The returns of old and new portfolio should be similar
- Price to book constraint: $\sum_{i=1}^{n} x_i \times \text{P/B}_i \leq \text{P/B}_{\text{index}}$. The price to book of the new portfolio should be less than the one of the index

Notice I am not saying anything on comparing the risks since it is implied that a more concentrated portfolio will have a higher risk. Also the maximal allowable risk in the original problem should be set higer than the variability of the initial index.

7. Conclusions

In this work, I explained a simplification approach to portfolio construction by leveraging cluster analysis of the time series data of stocks within the SP 500 index and forecasting future returns using the AutoRegressive Integrated Moving Average (ARIMA) model. This methodology aimed to extract a more concentrated portfolio with similar expected results but at a lower and more accessible book price.

There are several avenues for further exploration and improvement. Firstly, incorporating fundamental data alongside price data could enhance the clustering process, allowing for a better assessment of stock similarities based on both financial metrics and market behavior.

Secondly, exploring advanced forecasting techniques such as Long Short-Term Memory (LSTM) networks could offer additional capabilities for capturing nonlinear relationships and long-term dependencies in stock price movements. LSTM models excel in capturing complex patterns in time series data, potentially improving the accuracy of return forecasts.

Furthermore, considering bivariate time series, such as incorporating the relationship between stock prices and macroeconomic indicators, could offer a more comprehensive understanding of market dynamics and lead to more robust portfolio optimization strategies.