The 11th International Seminar on
Linear Algebra and its Applications
Feb 2022

Partitioning the stocks of a portfolio by k-medoids clustering approach

F. Soleymani *

Department of Mathematics, Institute for Advanced Studies in Basic Sciences (IASBS), Zanjan 45137–66731, Iran

Abstract

Machine learning is mainly used in practice because of the existence of large set of data. The target of this article is to study partitioning a large set of stocks inside a portfolio by the simple yet efficient k-medoids procedure. An algorithm is developed for this purpose. The unsupervised model is capable to receive financial returns and to illustrate the most and least risky clusters of stocks to manage the risk.

Keywords: Clustering; efficient portfolio; risk measure; financial returns; k-medoids
Mathematics Subject Classification [2010]: 91C20; 91G70

1 Introduction

Recently, there has been a proliferation of machine learning (ML) techniques and growing interest in their applications in finance, where they have been applied to sentiment analysis of news, trend analysis, portfolio optimization, and risk modeling for supporting investment management, see [4] for a general literature review and reading about several applications.

ML - as a sub-field of artificial intelligence (AI) - uses statistical methods to train machines from a given data set. After 'learning', the systems produce optimized models that express the data in the best way and restrict the potential biases, and further enabling better assessments and making appropriate decisions. Thus, such models are also broadly employed for predictions. In fact, clustering analysis and classification as data mining techniques can be employed for prediction of future outcome, [1, Chapter 22].

This paper explores the potential of ML to enhance the investment process. In this paper, an algorithm based on ML is discussed for grouping large portfolio of risky stocks specially when the numbers of stocks is high. Precisely, an application of ML in financial mathematics is discussed, see the book [1, Chapter 16] for some background.

We provide an ML-based clustering approach to narrow down the list of risky stocks in a portfolio. The advantages of this study comprising:

- Considering many stocks in a portfolio, then the taxonomy analysis gives us several different groups of stocks which have the most similarities and dissimilarity in the preceding trading day.

*Speaker. Email address: fazlollah.soleymani@gmail.com & soleymani@iasbs.ac.ir

- A risk measures is employed on each cluster of financial data to obtain the level entailed risk.
- The groups of data having the most and least risky performance are clarified and paid attention in upcoming trades.

The rest of this work is unfolded as follows. Section 2 is devoted to recall the definitions of a risk measure in finance that will be used later in work. Then, in Section 3, some discussions about the k-medoids are reminded and then a new algorithm for partitioning financial returns is brought forward as the novelty of this work. Next, the results are furnished in Section 4 along some notes. A conclusion is drawn in Section 5.

2 A risk measure

It is known that the volatility or the standard deviation (S.D.) of returns were adopted as a measure for risk in the seminal work by Markowitz in [6]. Having the pre-determined confidence level α, the risk measure of value-at-risk (VaR) can be defined as comes next:

$$\mathrm{VaR}_\alpha(X) := \inf\{z \in \mathbb{R} | F_X(z) \geq \alpha\}, \tag{1}$$

wherein X is a random variate, and $F_X(\cdot)$ is the cumulative distribution function (CDF). To discuss about (1), we recall that there is occasionally a chance of an adverse market movement that can lead a high loss. Note that VaR does not satisfy the sub-additivity property, and hence it is not coherent.

3 k-medoids procedure with VaR

At the presence of a large set of financial data, it will be observed that the regression-type methods which are classical statistical tools cannot anymore be employed to tackle problems. In fact, reasonable CPU times and prediction ability too make ML a promising approach than the traditional regression-like methods, [3, Chapter 1].

One of the fundamental clustering procedures is the k-medoids problem which is somewhat similar to k-means approach. These approaches break the large data sets into several groups (considered to be partitioned) and try to minimize the distance between nodes labeled to be in a cluster and a point designated as the center of the specified group, [5].

The algorithm of k-medoids works by choosing the true actual data nodes as centers (which we call sometimes as exemplars or medoids), unlike the standard k-means algorithm at which the group's center is not necessarily one of the input data nodes (it is the average between the groups' nodes.) This permits for greater interpretability of the cluster centers than in k-means. Furthermore, k-medoids could be employed with arbitrary dissimilarity measures, whereas k-means in general needs the Euclidean distance for efficient solutions.

Noting that due to the fact that k-medoids does the minimization on a sum of pairwise dissimilarities in lieu of a sum of squared Euclidean distances, this method is better for noisy environment and outliers than k-means.

3.1 An algorithm

To increase the reliability of the clustering algorithm, only one price for each stock (ticker) is not considered and in fact all the adjusted open-high-low-closed (OHLC) prices are used. The values are different to raw prices. This leads to the point that list of data containing 4D arrays.

In addition, since the construction of an efficient portfolio is mainly based on financial returns, here they are used and computed as follows:

$$\text{Returns}(\text{OHLC}_i) = \frac{\log\ p_{\text{OHLC}}(2021-12-20)}{\log\ p_{\text{OHLC}}(2021-12-21)}, \tag{2}$$

for $i = 1, 2, \ldots, n$, where n is the total number of stocks and log stands for the natural logarithm. In this work, to check the generalizability of the clustering approach, we assume to have a portfolio of S&P500, which includes the 500 large companies (with 505 stocks) listed on stock exchanges in the United States. To show how the first 10 stocks of such a portfolio of companies acts along time, we compare their price trends in Figure 1.

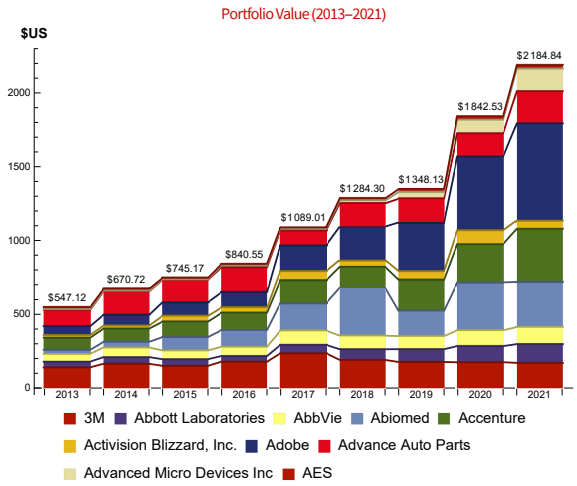

Figure 1: Price trends for the first 10 stocks of S&P500 over time.

The proposed algorithm of this work in order to tackle large portfolio by putting the stocks having the highest and lowest risk in similar groups are given in what follows.

Algorithm 3.1. Clustering large portfolio of stocks based on the k-medoids method and VaR.

1. Consider a large portfolio having stocks. And extract the financial returns corresponding to the stocks based upon two successive trading day. Here we used (2).

2. The missing data must be identified and the corresponding ticker from the list of stocks must be excluded. Here, only the data for one stock (position 303 out 505) is missing based on the trading days. Noting that the missing data can be tackled in different ways, see [3, Chapter 1].

3. Find the clusters on the large set of financial returns by k-medoids method. The number of partitions k must be given along with the distance function. Here, we use the squared Euclidean distance whenever required.

4. Find the size of each cluster. Besides, since most statistical features are required for decision making on each cluster, now flatten the data and compute their, mean, standard deviation, skewness, and kurtosis. This answers that does each cluster follow the normal distribution? Or does it have skewness with fat tails?

5. Compute the VaR measure for each cluster using (1). The strategy is to employ such values of VaR in forecasting the risk and trading in the next working day. The highest the

VaR value, the riskiest the group of stocks is. A negative value for VaR shows that the portfolio leads to profit with a high probability.

6. Make the decision via the obtained riskiest stocks. If the number of riskiest stocks is still high, one may repeat the procedure with higher values of k.

4 Computational results

In this section, the results of employing Algorithm 3.1 are furnished. The computational simulations are given for one-day 99% VaR in Table 1 and Figure 2 when $k = 6$. Noting that we set the seed of the random number generator to 1234 whenever needed. This means that due to randomness in employing unsupervised algorithms of clustering such as k-means and k-medoids, the results might change a bit each time, though the procedure of getting the most risky clusters would be similar. In implementations, we have used the command `FindCluster[]` in Mathematica 12 to call k-medoids algorithm properly.

Based on Figure 2-right, the second and the fifth clusters contain the riskiest stocks. The statistical properties for each cluster in this case based on our proposed algorithm are furnished in Table 1. Results reveal that $73 + 21$ stocks are in risky groups and none of the cluster follow normal distribution.

Table 1: The information of clusters using $k = 6$ and Algorithm 3.1.

	Cluster 1	Cluster 2	Cluster 3	Cluster 4	Cluster 5	Cluster 6
Size	97	73	228	84	21	1
Norm	0.174	0.554	0.402	0.471	0.601	0.216
Mean	-0.001	0.031	0.011	0.021	0.063	-0.096
S.D.	0.011	0.011	0.009	0.016	0.021	0.057
Skewness	-1.201	0.412	-0.841	-0.130	0.621	0.136
Kurtosis	5.275	3.329	4.967	4.009	4.831	1.335

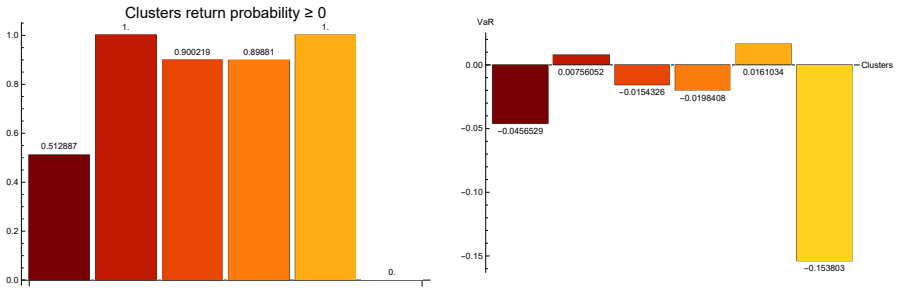

Figure 2: Probability of cluster returns being positive in left and the VaR measure values to assess the risk of each cluster in right.

4.1 Higher k

When the number of riskiest stocks is still high for the trader to make appropriate decisions, we may use a higher k to cluster the financial returns as much as possible. Following this, we may choose $k = 50$ as an instance. Selecting this, first yields to more number of clusters. However, for our real data from S&P500, only 16 clusters will be given. Results are given in Table 2. This reveals that Algorithm 3.1 does not necessarily put the items into useless groups and k is just an upper bound for the number of clusters.

Table 2: The information of clusters using $k = 50$ and Algorithm 3.1.

Cluster	1	2	3	4	5	6	7	8
Size	34	40	41	46	32	78	66	30
Norm	0.049	0.099	0.421	0.150	0.225	0.332	0.170	0.177
Mean	0.003	-0.	0.030	0.008	0.019	0.016	0.01	0.016
S.D.	0.005	0.01	0.015	0.009	0.006	0.010	0.006	0.006
Skewness	-0.086	-0.422	0.057	-0.197	-0.033	-0.575	-0.446	0.239
Kurtosis	2.892	3.797	2.788	2.482	2.883	2.580	3.904	3.532
Cluster	9	10	11	12	13	14	15	16
Size	33	13	32	24	11	16	7	1
Norm	0.538	0.209	0.299	0.153	0.142	0.193	0.445	0.216
Mean	0.046	0.018	0.026	0.002	-0.019	0.019	0.083	-0.096
S.D.	0.015	0.024	0.009	0.017	0.015	0.016	0.018	0.057
Skewness	0.285	-0.160	-0.301	-0.307	-0.241	-0.045	1.639	0.137
Kurtosis	2.838	2.965	2.725	2.058	2.432	1.960	6.053	1.336

Results in this case show that the most risky clusters are 5,8,9,11 and 15. Figure 3 is provided to show the associated VaR values. The highest VaR belongs to cluster 15 consisting of 7 stocks only. Hence, the riskiest stocks based on this higher VaR value can be then obtained as follows:

```
"NYSE:CCL", "NASDAQ:CTXS", "NYSE:DVN",
"NASDAQ:EXPE", "NYSE:LYV", "NASDAQ:MU", "NYSE:NCLH"
```

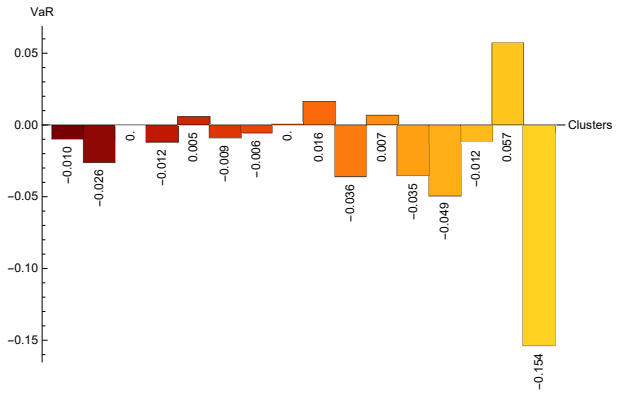

Figure 3: The VaR risk measure when $k = 50$ based on Algorithm 3.1.

4.2 Comparisons

Here a query may arise that applying a clustering technique for partitioning portfolio without comparison to the existing techniques does not bring novelty to the work. To respond this, it is pointed out that authors in the work [2] employed k-medoids or similar algorithms for partitioning while in this work we construct a k-medoids-VaR approach which works based on a risk measures. In addition, here k is the maximum number of clusters and a comparison along the existing techniques is done in Table 3 for the case $k = 50$. The time is reported on seconds. The results show that due to the type of our financial data and some other restrictions, some of the clustering techniques cannot even be used. k-means also relies on the choice of the initial seeds too much and thus the only competitor for k-medoids-VaR could be the "Optimize" method. This method also yields in many groups having positive VaR values which lead to difficulty in choosing the most risky stocks.

Table 3: Comparison of the existing clustering techniques for the experiment in Subsection 4.1.

Algorithm	Description	CPU time	Number of clusters
Agglomerate	Find clustering hierarchically	0.40	50
Optimize	Find clustering by local optimization	0.03	50
DBSCAN	Density-based spatial clustering of applications with noise	Cannot be used	-
Gaussian Mixture	Variational Gaussian mixture algorithm	Cannot be used	-
Jarvis Patrick	Jarvis-Patrick clustering algorithm	Cannot be used	-
KMeans	k-means clustering algorithm	0.12	16
KMedoids	Partitioning around medoids	0.15	16
Mean Shift	Mean-shift clustering algorithm	Cannot be used	-
Neighborhood Contraction	Displace examples toward high-density region	Cannot be used	-
Spanning Tree	Minimum spanning tree-based clustering algorithm	Cannot be used	-
Spectral	Spectral clustering algorithm	0.94	1

5 Conclusion

We have proposed an algorithm based on unsupervised k-medoids clustering approach in the category of ML models for unlabeled classifications of financial returns. The algorithm works on multi-dimensional data of any sizes and only gets an upper bound for k to work. The proposed procedure is able to pin down the riskiest stocks existing in the most risky clusters using a risk measure. Note that, the procedure does not say anything about asset allocation. If a trader wants to do such a thing, some nonlinear optimizations based on the well-known Markowitz portfolio construction can be done to find the allocations of assets.

References

[1] M.L. De Prado, *Advances in Financial Machine Learning*, Wiley, New Jersey, 2018.

[2] F.G. Duarte, L.N. De Castro, A framework to perform asset allocation based on partitional clustering, *IEEE Access*, 8 (2020), 110775-110788.

[3] G. Gan, C. Ma, J. Wu, *Data Clustering: Theory, Algorithms, and Applications*, SIAM, Philadelphia, 2007.

[4] B.M. Henrique, V.A. Sobreiro, H. Kimura, Literature review: Machine learning techniques applied to financial market prediction, *Expert Sys. Appl.*, 124 (2019), 226-251.

[5] L. Kaufman, P. Rousseeuw, *Finding Groups in Data-An Introduction to Cluster Analysis*, Wiley Series in Probability and Mathematical Statistics, New York: John Wiley & Sons, Inc., 1990.

[6] H. Markowitz, Portfolio selection, *J. Finance*, 7 (1952), 77-91.