Cluster 48 Industry portfolios through K-Means and Hierarchical Clustering

Xueer Zheng

Department of Computer Science, University College London

COMP0050 - Machine Learning with Applications in Finance

Dr. Fabio Caccioli

April 18, 2021

I. Introduction

Building a portfolio with stocks from different industries is a common way to reduce the investment risk. Through this measure, even parts of the portfolio might fail sometimes, the rest will succeed, and the overall performance of the portfolio will be balanced. Therefore, exploring the similarities among stocks through clustering analysis is crucial for investors, which can help them to build diversified portfolios and minimize risks. This study is based on the data composed of 48 different industry portfolios, aiming to find the similarities among these portfolios by two unsupervised learning methods (K-Means clustering and hierarchical clustering) Finally, K-Means and hierarchical clustering provided us with different ways to cluster the portfolios, both of their results were not interpretable

II. Data

The dataset named "48_Industry_Portfolios_daily.CSV" comes from Moodle and consists of 48,200 records of daily returns for 48 industry portfolios from July 1st, 1926 to Oct 31st, 2017. Specifically, an industry portfolio is composed of a New York Stock Exchange (NYSE) stock, an American Stock Exchange (AMEX) stock and a Nasdaq (NASDAQ) stock from the same industry. In the original dataset, the index of the rows is the date, and the index of the columns is the name of the industry. The original dataset is below:

	Agric	Food	Soda	Beer	Smoke	Toys	Fun	Books	Hshld	Clths	...	Boxes	Trans	Whlsl	Rtail	Meals	Banks	Insur	REst	Fin	Other
19260701	0.56	-0.07	NaN	-1.39	0.00	-1.44	0.62	-1.27	-0.90	0.12	...	-0.93	0.15	2.77	-0.02	0.27	-0.47	-0.56	-1.41	-0.72	-0.02
19260702	0.29	0.06	NaN	0.78	0.70	1.46	0.03	0.00	-0.34	-0.35	...	1.07	0.06	0.00	0.01	-0.10	0.35	0.80	0.72	1.66	0.83
19260706	-0.33	0.18	NaN	-1.74	0.50	-0.96	-0.06	4.27	-1.20	0.41	...	0.73	-0.19	0.77	-0.22	-0.67	-3.24	-3.24	-3.33	-3.38	-1.01
19260707	3.57	-0.15	NaN	-1.73	-0.12	-0.49	-0.06	-4.10	-0.22	0.16	...	2.22	0.18	-3.21	-0.57	-0.70	-0.68	-0.14	0.30	-0.25	0.44
19260708	0.30	1.12	NaN	-0.15	0.30	-0.49	0.24	0.00	-0.01	0.79	...	-0.39	0.46	-1.10	-0.38	0.33	-0.96	-0.37	-0.24	-0.80	-0.03

As the dataframe shows, the original dataset contained too many NaN values so that I decided to only focus on the last 10-year data points which did not have NaN values. Thus, a new dataframe with 2,520 records of daily returns from Oct 30th, 2007 to Oct 31st, 2017 was created, and its rows and columns were swapped to facilitate our analysis. The new dataset "trans_df" below has the date as the column index and the industries as the row indexes.

	20071030	20071031	20071101	20071102	20071105	20071106	20071107	20071108	20071109	20071112	...	20171025	20171026	20171027	20171030	20171031
Agric	-1.43	1.30	-4.09	-2.63	-0.86	1.19	-3.93	1.40	0.48	-2.28	...	0.60	0.36	0.69	-1.99	1.29
Food	-0.53	0.54	-2.64	-0.61	-1.20	-0.08	-1.81	0.70	-1.10	-0.04	...	0.02	-0.18	0.43	-1.74	1.42
Soda	-0.80	1.07	-2.18	0.58	-1.81	0.18	-2.29	-2.18	-1.98	-0.48	...	-1.05	0.72	-0.87	-1.43	1.18
Beer	-0.59	0.23	-2.84	0.19	-0.38	-3.51	-3.17	0.10	-0.72	-2.34	...	-0.64	0.18	-0.58	-0.54	2.55
Smoke	-0.02	0.19	-0.85	-0.35	0.41	0.98	-1.19	1.00	-1.43	-1.39	...	-1.30	1.25	-0.66	-1.58	-0.03
Toys	-0.36	0.42	-2.70	-0.19	-0.13	-0.08	-2.66	-0.95	-2.35	-1.14	...	-1.59	0.29	-1.56	-0.11	-2.52
Fun	-1.26	0.56	-1.69	-1.25	-1.39	0.84	-3.00	-0.03	-1.94	-0.44	...	-0.86	0.28	0.62	-1.06	0.98
Books	-1.58	1.27	-3.29	-0.82	-1.08	-0.69	-2.17	-0.30	-0.28	0.53	...	-0.06	0.67	-0.27	-1.15	0.59

III. Method

In this study, the research goal is to cluster the daily returns of these 48 industry portfolios. Since the ranges of these 48 industry portfolios are different, I standardized the dataset and made it have a mean of 0 and a variance of 1. Then, I used principal component analysis to reduce the dimensionality and finally applied K-means and hierarchical clustering to cluster the portfolios.

Standardization and Principal Component Analysis

The first thing I did was to standardize data through the build-in function "StandardScaler" in Python. The reason is that both K-means and hierarchical clustering are distance-based algorithms so that they are very sensitive to the range of features (Moore, 2001). That is, as both these two algorithms use distance metrics to measure the similarity between

clusters, if one feature has a very large range compared with other features, it will have a significant impact on the clustering and the result will be biased.

After standardizing the data, I applied principal component analysis (PCA) to reduce the dimensionality of the data (Syakur et al., 2018). In the Dataframe "trans_df", we have 2520 dimensions (2520 days), and PCA allows us to reduce it into 2 or 3 dimensions. Therefore, I drew a histogram of the percentage of variance explained by each PCA component and found the optimal number of components, and then performed a cluster analysis based on the low-dimensional data.

K-Means Clustering

The goal of K-Means clustering is to divide those 48 industry portfolios into k different non-overlapping groups by minimizing the Euclidean distance between the portfolios within a cluster. The Euclidean distance between two n-dimensional points a and b has a formula like this (Oyelade et al., 2010):

$$d(a,b) = \sqrt{\sum_{i=1}^{n}(a_i - b_i)^2}$$

To be specific, the algorithm of K-Means starts with randomly finding k points in the dataset as initial centroids, and then assigns all datapoints to their closest centroids based on the Euclidean distance between the datapoints and centroids. Then, the algorithm will continue to iterate until the assignment of data points is stable.

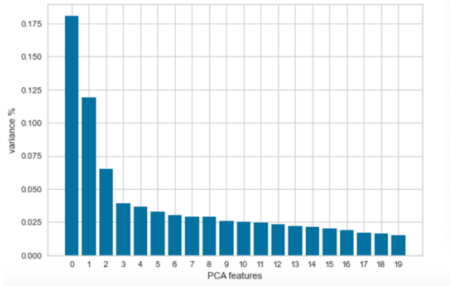
In K-Means clustering, the number of cluster k should be pre-defined (Pham et al.,2005) Therefore, I firstly generated an Elbow plot with k ranges from 2 to 20 in order to find the optimal value of k. The Elbow plot will show the sum of squared distances (SSD) between each data point and its assigned cluster center for each number of k, and the optimal value of k appears where the Elbow curve goes flat (Kodinariya & Makwana, 2013). Then, based on the K-Means clustering model with the optimal k, I built a Dataframe to show the results of K-Means clustering, and industry portfolios within the same cluster were assigned the same index.

Hierarchical Clustering

In this case, hierarchical clustering algorithms begin with 48 clusters which are these industry portfolios themselves, and then the dissimilarity metric between pairs of datapoints defined by Euclidean distance is extended to the dissimilarity metric between two clusters through the ward's linkage. In the beginning, the sum of squares is zero and it will increase when we start merging, and the ward's linkage aims to minimize the increase of the sum of squares (Murtagh & Legendre, 2014). In this study, I plotted a "dendrogram" to help us visualize the hierarchical relationship between these 48 industry portfolios (Jolliffe, 1989). After that, I also build a dataframe to show the results of K-Means clustering, and industry portfolios within the same cluster were assigned the same index.

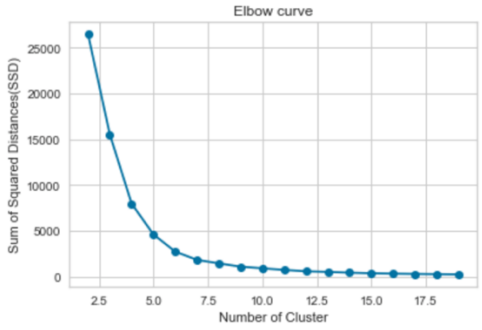
IV. Result

Principal Component Analysis (PCA)



In this PCA figure, the x-axis consists of the number of PCA components ranging from 0 to 19, and the y-axis consists of the percentage of variance explained by each PCA component. Moreover, as the first three components explained the majority of data's variance, we performed PCA with the first 3 components and applied K-Means and Hierarchical Clustering to fit these three components.

K-Means

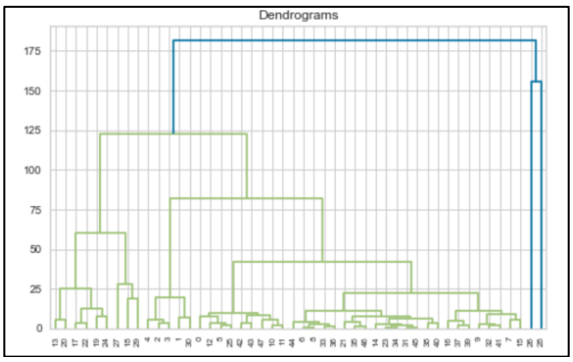


The Elbow plot that I generated is above has the number of clusters k ranging from 2 to 20 on the x-axis and has the within-cluster sum of squared distances (SSD) on the y-axis. Additionally, the plot shows that the value of SSD has decreased to a significantly small value after k=6. Thus, we regarded the k=6 as the optimal value of the number of clusters and built a dataframe to show its results.

	Industry	label																						
0	Agric	0	35	Chips	0	41	Rtail	0	21	ElcEq	0	11	MedEq	0	12	Drugs	0	3	Beer	4	24	Ships	5
25	Guns	0	36	LabEq	0	42	Meals	0	23	Aero	0	16	BldMt	0	27	Mines	1	2	Soda	4	17	Cnstr	5
31	Telcm	0	46	Fin	0	43	Banks	0	6	Fun	0	7	Books	0	29	Oil	1	1	Food	4	22	Autos	5
32	PerSv	0	37	Paper	0	44	Insur	0	8	Hshld	0	15	Txtls	0	28	Coal	2	4	Smoke	4	18	Steel	5
33	BusSv	0	39	Trans	0	45	RlEst	0	9	Clths	0	14	Rubbr	0	26	Gold	3	30	Util	4	19	FabPr	5
34	Comps	0	40	Whlsl	0	38	Boxes	0	10	Hlth	0	5	Toys	0	47	Other	4	20	Mach	5	13	Chems	5

As the dataframe shows above, "Agric", "Guns","Telcm", "PerSv", "BusSv", "Comps", "Chips", "LabEq", "Fin", "Paper", "Trans", "Whlsl", "Rtail", "Meals", "Banks", "Insur", "RlEst", "Boxes", "ElcEq", "Aero", "Fun", "Hshld", "Clths", "Hlth", "MedEq", "BldMt", "Rubbr", "Txtls", "Books", "Rubbr", "Txtls", "Toys", and "Drugs" are clustered together with index 0; "Mines" and "Oil" are clustered together with index 1; "Coal" itself is under index 2 and "Gold" itself is under index 3; "Other", "Beer", "Soda", "Food", "Smoke", and "Util" are under index 4; "Mach"," Ships","Cnstr","Autos","Steel","FabPr", and "Chems" are with index 5.

Hierarchical Clustering



The dendrogram above has the indexes of 48 industry portfolios on the x-axis and the Euclidean distance between the clusters on the y-axis, and it divided the portfolios into 3 clusters. The results are shown as a dataframe below:

	Industry	label																		
0	Agric	1	32	PerSv	1	38	Boxes	1	1	Food	1	7	Books	1	13	Chems	1	19	FabPr	1
25	Guns	1	33	BusSv	1	39	Trans	1	2	Soda	1	8	Hshld	1	14	Rubbr	1	20	Mach	1
27	Mines	1	34	Comps	1	40	Whlsl	1	3	Beer	1	9	Clths	1	15	Txtls	1	12	Drugs	1
29	Oil	1	35	Chips	1	41	Rtail	1	4	Smoke	1	10	Hlth	1	16	BldMt	1	47	Other	1
30	Util	1	36	LabEq	1	42	Meals	1	5	Toys	1	22	Autos	1	17	Cnstr	1	26	Gold	2
31	Telcm	1	37	Paper	1	43	Banks	1	6	Fun	1	11	MedEq	1	18	Steel	1	28	Coal	3

As shown in the dataframe, the hierarchical clustering groups those 48 industry portfolios into 3 clusters. “Gold” itself is under index 2 and “Coal” is under index 3, while all other industry portfolios are all under an index of 1.

V. Discussion

I roughly divide the 48 investment portfolios into four sectors: services, agriculture, resources and manufacturing,. The service sector contains 13 portfolios: “Telcm”, “PerSv”, “BusSv”, “Fin”, “Trans”, “Whlsl”, “Rtail”, “Meals”, “Banks”, “Insur “RlEst”, “Fun”, and “Hlth”. The agriculture sector has one portfolio “Agric”, and resource sector has 3 portfolios: “Gold”, “Mine”, and ”Coal”. All other portfolios (31) belong to the manufacturing sector.

The K-Means method grouped those 48 portfolios into 6 clusters. To be specific, the 31 industries with an index of 0 belong to the service(13 portfolios), manufacturing(17 portfolios), and agriculture(1 portfolio) sector; two industries with an index of 1 are non-renewable resources; "Coal" and "Gold" have their own clusters; 6 industries with an index of 4 are from the manufacturing sector related to people's daily life, while 7 industries with an index of 5 belong to the manufacturing sector. As clusters with index 0, index 4 and index 5 all contain the manufacturing sector, it’s not clear about the difference between clusters.

Comparing to K-Means, hierarchical clustering only clustered the 48 portfolios into 3 groups. Under index 1, there are 46 industry portfolios in total, 13 of them from the service sector, 30 of them from the manufacturing sector, 2 of them from non-renewable resources and one of them from the agriculture sector. As same as the result of k-means, "Coal" and "Gold" have their own clusters as well. Based on the result, under index 1, it’s still hard to find the common properties within the clusters.

Both of the clustering methods gave "Coal" and "Gold" their own clusters, which implied "Coal" and "Gold" are significantly different than others. Although K-Means and hierarchical clustering clustered the portfolios into the different number of clusters, neither of them provided us with meaningful interpretations of the clusters. One of the possible reasons is that the algorithms we used are too naive. Therefore, in the future study, the K-Means method should be improved by finding a better initial condition to start with. Also, other clustering methods, such as K-Medoids, might be able to create a more interpretable result.

Reference

Jolliffe, I. T., Allen, O. B., & Christie, B. R. (1989). Comparison of variety means using cluster analysis and dendrograms. *Experimental Agriculture*, *25*(02), 259-269.

Kodinariya, T. M., & Makwana, P. R. (2013). Review on determining number of Cluster in K-Means Clustering. *International Journal*, *1*(6), 90-95.

Moore, A. (2001). K-means and Hierarchical Clustering.

Murtagh, F., & Legendre, P. (2014). Ward's hierarchical agglomerative clustering method: which algorithms implement Ward's criterion?. *Journal of classification*, *31*(3), 274-295.

Oyelade, O. J., Oladipupo, O. O., & Obagbuwa, I. C. (2010). Application of k Means Clustering algorithm for prediction of Students Academic Performance. *arXiv preprint arXiv:1002.2425*.

Pham, D. T., Dimov, S. S., & Nguyen, C. D. (2005). Selection of K in K-means clustering. *Proceedings of the Institution of Mechanical Engineers, Part C: Journal of Mechanical Engineering Science*, *219*(1), 103-119.

Syakur, M. A., Khotimah, B. K., Rochman, E. M. S., & Satoto, B. D. (2018, April). Integration k-means clustering method and elbow method for identification of the best customer profile cluster. In *IOP Conference Series: Materials Science and Engineering* (Vol. 336, No. 1, p. 012017). IOP Publishing.