

## Quantitative Strategies Based on an Improved K-means Algorithm

Xinyu Wang<sup>1, a</sup> Lian Xue<sup>1, b\*</sup> Ruiyu Yu<sup>1, c</sup> Yike Wu<sup>1, d</sup> and Qunfang Yu<sup>1, e</sup>

<sup>1</sup> School of Computer and Computing Science, Zhejiang University City College  
Hangzhou, 310015, P.R. China

<sup>a</sup>31401056@stu.zucc.edu.cn, <sup>b\*</sup> corresponding author: xuel@zucc.edu.cn ,  
<sup>c</sup>31401064@stu.zucc.edu.cn, <sup>d</sup>31401057@stu.zucc.edu.cn, <sup>e</sup>31401063@stu.zucc.edu.cn

**Keywords:** An improved K-means algorithm; Rolling stock; Quantitative strategy

**Abstract.** In this paper, a kind of improved k-clustering stock-picking method is adopted to cluster the 300 indexes of the CSI 300 indexes, and investigate the average yield of all kinds, and select the class with the highest average rate of return to be held as our portfolio, thus achieving the excess return. The empirical evidence shows that, using the strategy of this paper to select the stocks and roll it, the excess accumulated yield of the portfolio is better than that of the CSI 300.

### Introduction

After several years of rapid development, quantitative investment funds have become the most important force in global asset management. With the rapid expansion of scale, the average performance of international quantitative investment funds has declined in the past two years.

Morgan Stanley said in a recent research report, compared with 10 to 15 years ago, is now more stringent regulation and fierce competition, many popular quantitative investment strategy is no longer valid, lead to quantify excess investment profit decline. For investors, perhaps the use of a single strategy and traditional strategy is unsustainable, and hybrid strategy and strategic innovation are the future direction of quantitative investing.

This paper establishes quantitative investment strategy through the improvement of cluster analysis.

### An Improved K-means Algorithm

K-means is based on clustering analysis, according to the characteristics of the individual or the sample will be a number of sample data classification, make the individuals within the same category as high as possible homogeneity while category between heterogeneity should have as high as possible.

The k-means algorithm usually use the optimize objective function, which can be defined as follows:

$$E = \sum_{i=1}^k \sum_{X \in C_i} \|X - V_i\|_p \quad (1)$$

Specific interpretation:

Step 1: after get k center point, Calculate the distance from the data objects in the data set N to the cluster centers of the K centers, Divide the data object into the cluster that is the smallest distance from it.

Step 2: Individually calculate the sum of distance between each and every other data object, select the smallest element as the new cluster center.

Step 3: Until the clustering center is no longer changed, the clustering ends.

However, the clustering results in k-means algorithm are vulnerable to the influence of the initial clustering center, so an improved algorithm of initial cluster center selection is proposed [2]. Before the simple k-means algorithm, the selection of the initial cluster center was increased:

Step 1: establishes K empty class clusters and initializes one of them into the original data set N.

Step 2: calculates the number of data objects contained in each class cluster, and selects the largest cluster of data objects, which is called M.

Step 3: calculate the distance between the two data objects in M and find the two largest data objects,  $x_i$  and  $x_j$ , and fill them in the empty cluster.

Step 4: take  $x_i$  and  $x_j$  as cluster centers, calculate the distance between the remaining data objects in M and  $x_j$ , and divide them into the smallest classes.

Step 5: checks whether the data set N is divided into K classes, which is the end of the initialization cluster center, or else to Step 2.

The algorithm constantly looking for the biggest clustering, and use the biggest distance of the two data object as the clustering center to split the clustering. so repeatedly, until they get the specified number of clustering centers. This improved algorithm can get better clustering results.

### The Establishment of Quantitative Stock Selection Strategy

This article takes the scroll holding pattern, that is, according to the previous period of time (sampling period) of stock yield could be divided into K classes, stock selection of K class average yield the largest class held for a period of time (holding period), sell stock holding period after the end, and according to a new sampling period calculation under the holdings of shares for a period of time, repeat cycle, specific steps as follows [3]:

Step 1: the sequence  $\{P_{n,t}\}$  is the price of the NTH stock at t time, and calculates the rate of return matrix of the stock sampling period. The formula is as follows:

$$X_{n,t} = \frac{P_{n,t+1} - P_{n,t}}{P_{n,t}} \quad (2)$$

Get the  $\{X_{n,t}\}$  for the yield sequence of the NTH stock, the sequence is the element we want to classify.

Step 2: Calculate the Manhattan distance[4] of each two stock sequences in the rate. The formula is

$$d(X, Y) = \sum_{i=1}^n |x_i - y_i| \quad (3)$$

And use the distance you get to use the improved clustering method to cluster.

Step 3: calculate the average yield per class after clustering.

Step 4: select the class with the highest average yield, buy the stock with all the funds of the account, and assign equal funds to each stock.

Step 5: when the holding period expires, then jump to step 1 to re-select the stock, until the time is over, and break out of the loop.

Step 6: calculate the accumulated excess return rate, and compare the cumulative yield of the portfolio selection with the cumulative yield of the CSI 200 index.

### The Empirical Results

This strategy takes different parameters in the time period from 2013-01-05 to 2016-01-25. The initial capital of the account is set to 1 million, parameter K is the number of central points of clustering, parameter S is the time length of the sampling period, parameter Hold is the holding number, and change the parameters to see the cumulative return rate of the selected stock strategy. and the selection of parameters and the final cumulative return results are shown in the Table 1. The trading platform used here is Mindgo Quantitative Trading Platform.

Table 1. The excess accumulation earnings of the stock strategy from 2013-01-04 to 2015-01-05

	S=30	S=60	S=90	S=120	S=150
Hold=30	-56.82%	11.53%	-25.75%	32.98%	93.02%
Hold=60	-38.25%	-12.24%	-37.21%	-18.73%	-40.67%
Hold=90	12.88%	-25.14%	-17.34%	12.59%	-23.14%
Hold=120	-14.27%	26.61%	-8.12%	33.26%	-39.75%

On the Table1, it can be found that several parameters have no obvious relevance to the stock-picking strategy. It's not as if the longer the sampling period is, the more stable the stock and the higher the yield. According to the data, we think the best parameter for the policy is  $k=10$ ,  $\text{hold}=30$ ,  $\text{sampling}=150$ , plot the portfolio of the optimal parameters and the performance of CSI 300 index during the inspection period.

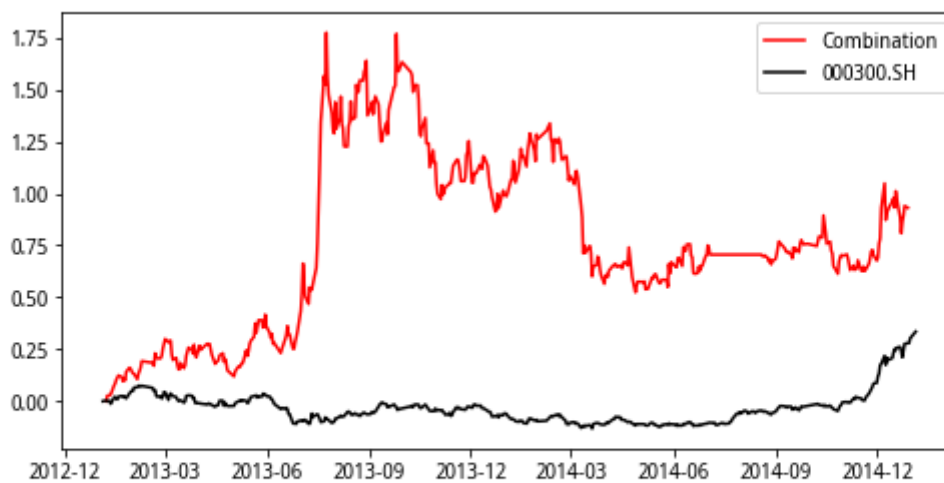


Figure. 1  $K=10$ ,  $\text{Hold}=30$ ,  $\text{sampling}=150$  the trend of strategic investment portfolio and CSI 300 index

As you can see in the figure above, the stock selection strategy is during the inspection period, the early yield is not obvious compared with the broader market, but the interim yield is far higher than the broader market. The late yield tends to be stable and higher than the broader market, overall, the yield is significantly higher than the market.

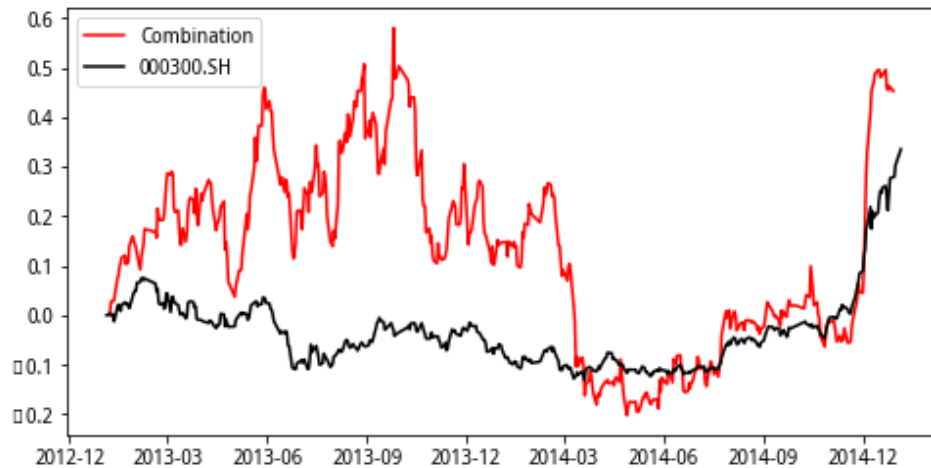


Figure. 2  $K=8, Hold=30, sampling=150$  The trend of strategic investment portfolio and csi 300 index

we also tried to change the clusters number ,and the result shown in Fig. 1, we can see that in the case of parameter, the early portfolio returns have been higher than the market, but the late suddenly yields fell sharply.

## References

- [1] X.Q.Chen:*Application Of Optimization Method In Clustering Algorithm*(University of electronic science and technology press,China 2014),p.12. (In Chinese)
- [2] G.P. Cheng, W.P. Wang and J.Huang: *Journal Of Chinese Computer Systems*, Vol. 33 No. 6 2012 (In Chinese)
- [3] Z.T.Cheng:*Cluster quantized stock selection strategy*(China 2010)
- [4] Information on <https://baike.baidu.com/item/%E6%9B%BC%E5%93%88%E9%A1%BF%E8%B7%9D%E7%A6%B/743092?fr=aladdin>