A Novel high-dimensional clustering algorithm and its application in financial data forecasting

Yanshan Qian, Zhuohao Du, Chuanan Li*, Junda Qiu
School of Computer Engineering Jiangsu University of technology Changzhou213001, China
* Corresponding author's email: 1594779583@qq.com

ABSTRACT

The purpose of this paper is to study the nonlinear characteristics of stock market price rise and fall, and optimize the characteristics of high-dimensional data for the challenges brought by the sparsity and complexity of high-dimensional data to the traditional clustering algorithms. Explore the impact of the multidimensional characteristics and sparsity of stock data on clustering analysis, and propose a high-dimensional clustering algorithm based on Pareto-optimal Steiner points. Traditional clustering algorithms are deficient in dealing with nonlinear data and spatial clustering, while randomly selected clustering centers are easily interfered by market noise, which affects the accuracy of clustering results. In this paper, more accurate clustering results are achieved by selecting Steiner points with Pareto optimality as clustering centers. And the algorithm is applied to the analysis of the stock market, which can effectively help investors choose the appropriate investment portfolio and has important practical application value. Meanwhile, the distance matrix is constructed as the basis of clustering analysis through the preprocessing of stock data and distance metric. The innovation of this paper is that the Steiner points can better represent the center of the clusters thus effectively reducing the computational complexity and storage overhead and improving the efficiency of the algorithm. And the principle of Pareto optimality is applied to the high-dimensional clustering algorithm, which improves the accuracy of stock market analysis.

Keywords: high-dimensional data, Pareto-optimal Steiner points, high-dimensional clustering

1. INTRODUCTION

1.1 Background of the study

First, the volatility and uncertainty of the stock market pose challenges to investors' decision-making, and more accurate and efficient algorithms are needed for market analysis and prediction. Second, traditional clustering algorithms have limitations and challenges in dealing with high-dimensional data, and need to be optimized to fit the data characteristics of the stock market. Finally, the development of big data and artificial intelligence technologies provides opportunities to apply them to stock market analysis and prediction, which has become one of the hot issues in current research.

1.2 Research status

The stock market is a complex and nonlinear system and is affected by many factors. In addition, the random selection of clustering centers in traditional clustering algorithms brings great difficulties in stock stratification. Therefore, how to find a scientific and accurate clustering center and perform cluster analysis is an urgent problem in stock market analysis.

2. RELATED WORK

This paper reviews a large amount of related literature and analyzes the research results of scholars at home and abroad. Majd et al [1] proposed a hybrid PSO-K-mean clustering algorithm that uses PSO to find three-dimensional locations and finalize their two-dimensional locations. However, this is only a two-dimensional clustering algorithm that randomly initializes the clustering centers. Menneer et al [2] used k-mean clustering, which iteratively updates the affiliation of the clusters based on the nearest centers of mass, among other operations, to ultimately reduce the distance between the site and the clustering center. However, the selection of the clustering centers for this algorithm is still random. Qiao et al [3] first ranked the affinities of the variables, and then aggregated the observations one by one to classify the standardized variables by a fast clustering method. However, it is difficult to extend the clustering algorithm to high dimensions in the validation of functional and qualitative clustering. Liu et al [4] proposed a clustering algorithm for Cloud-Cluster that can characterize the object's vagueness and randomness at the same time in order to preserve uncertain information and describe the clusters as concepts. However, this clustering algorithm cannot solve the problem of optimal selection of clustering centers still has limitations. Du et al [5] proposed a non-Euclidean distance fuzzy clustering algorithm incorporating weights. The algorithm improves the classification performance and computational efficiency under noise interference and high dimensional datasets. This provides ideas for the innovation and improvement of the clustering algorithm in this project. Miguel et al [6] proposed a network co-evolutionary model that promotes intra-group cooperation and enhances isolation, and ultimately facilitates the selection of a Pareto-optimal equilibrium. This inspired the search for new clustering centers in this paper. Ni [7] utilized a density-based clustering algorithm that generates a "decision map" to identify the cluster centers, which is affected by the characteristics of the selected data and fails to capture the nonlinear relationships among the stock data. Gao [8] utilized the k-means algorithm to study the problems associated with equity funds. However, the algorithm could not be extended to high dimensions and the clustering centers were random in nature. Zhai [9] utilized a soft-DTW K-medoids clustering model to analyze stocks using soft-DTW distances to improve traditional clustering. However, the method could not find a better clustering center. Huang [10] et al. used a hierarchical clustering algorithm based on quadratic programming to estimate the stock portfolio composition and weights, which is prone to missing important features and has a high degree of randomness in the clustering centers.

3. KEY METHODS

3.1 The optiomal rally point in high dimensional coordinate system

There are $n(n \geq 3)$ weighted points in a bounded closed box in a three-dimensional plane,whose corresponding positive weights are $\xi_i \in [0,1](1 \leq i \leq n)$,and $\sum_{i=1}^{n} \xi_i = 1$ .If a point P^* exists,whose Euclidean distance to the other given points meet the following condition:

$$D = \min \sum_{i=1}^{n} \left| P^* P_i \right|$$

$$= \min(\sqrt{(x^* - a_1)^2 + (y^* - b_1)^2 + (z^* - c_1)^2} + ... + \sqrt{(x^* - a_n)^2 + (y^* - b_n)^2 + (z^* - c_n)^2}) \qquad (1)$$

then P^* can be defined as the optimal rally point.(see Fig 1)

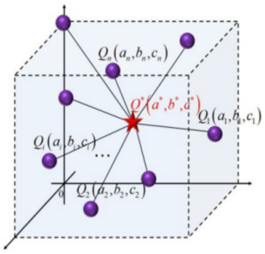

Figure. 1 Optimal set point in three-dimensional space.

3.2 optimal marshalling algorithm

Step 1: To set the initial growth point $x^0 \in x$ and set the step size λ (in this paper the step size is set to 1/1000), where $x = (x_1, x_2, ..., x_m)$ is the vector set of box E. Set $X_{\min} = x^0$ and $F_{\min} = f(x^0)$, where $f(x^0)$ is the backlight function of x^0 .

Step 2: Set x^0 as the centre of thinking and draw line segments parallel to the x- and y-axes, extending further on this $a_1 \leq x_1{}^0 \leq b_1$, $a_2 \leq x_2{}^0 \leq b_2$,..., $a_m \leq x_m{}^0 \leq b_m$ and make it a new branch. Find $S_{i_1 j_1}^0 (1 \leq i_1 \leq m, 1 \leq j_1 \leq k_1)$ from a branch of λ , where $S_{i_1 j_1}^0$ is the j_1 growth point in a branch of i_1 .

Step 3: Compare $f(S_{i_1 j_1}^0)$ with $F_{\min}$. If $f(S_{i_1 j_1}^0) \leq F_{\min}$ is satisfied, $X_{\min} = S_{i_1 j_1}^0$ holds; otherwise, $X_{\min}$ and $F_{\min}$ remain unchanged.

Step 4: If $f(x^0) \leq f(S_{i_1 j_1}^0)$ is satisfied, then the growth hormone concentration can be obtained as $C_{S_{i_1 j_1}^0} = 0$. Otherwise, bring in eq.(1) to calculate $C_{S_{i_1 j_1}^0}$:

$$C_{S^0_{i_1 j_1}} = \frac{f(x^0) - f(S^0_{i_1 j_1})}{\sum_{i_1=1}^{m}\sum_{j_1=1}^{t_1-1}[f(x^0) - f(S^0_{i_1 j_1})]} \tag{2}$$

Step 5: Calculate and record the growth hormone concentration at all growth points to create a state map of morpholino concentrations ranging from 0 to 1. A number δ_0 was randomly selected in that interval, and if

$$\sum_{i_1=1}^{r_1}\sum_{j_1=1}^{t_1-1} C_{S^0_{i_1 j_1}} < \delta_0 < \sum_{i_1=1}^{r_1}\sum_{j_1=1}^{t_1} C_{S^0_{i_1 j_1}} \tag{3}$$

Step 6: Repeat steps 2 through 5, stopping when the result is unchanged. At this point, $X^* = X_{\min}$ is the globally optimal solution; set it to $\dot{\mu} = x^*$,and end the calculation.

4. PRACTICAL EXAMPLE

4.1 Experimental data sources

Stock information for the years 2017-2022 is collected from http://quote.eastmoney.com and data processing is performed. By reviewing the related literature, this paper applies principal component analysis to dimensionality reduction of multi-attribute features of stocks and the dataset is shown in Table 1.

Table 1. Selected stock data from 2017 to 2022 after processing.

Year	Stock code	Equity multiplier	Total asset turnover	Net profit margin	Year	Stock code	Equity multiplier	Total asset turnover	Net profit margin
2017	688001.SH	1.4236	1.6512	15.3062	2018	688001.SH	1.3643	0.9166	24.2056
	688002.SH	1.3865	0.3859	41.3247		688002.SH	1.1842	0.4541	32.588
	688003.SH	1.5062	0.9823	16.1593		688003.SH	1.5147	0.9778	18.5869
2019	688001.SH	1.126	0.7442	14.0292	2020	688001.SH	1.1507	0.5802	15.8041
	688002.SH	1.0861	0.3666	29.5182		688002.SH	1.2098	0.5129	37.44
	688003.SH	1.1613	0.4269	15.373		688003.SH	1.3714	0.4796	11.1379
2021	688001.SH	1.4583	0.4594	15.5416	2022	688001.SH	1.4599	0.4338	14.269
	688002.SH	1.2927	0.4221	25.7803		688002.SH	1.4577	0.4718	10.8294
	688003.SH	1.6829	0.537	10.6008		688003.SH	1.7389	0.5758	9.5713

4.2 4.2 Experimental results and analysis

In this paper, we have collected data on three features of more than three hundred stocks in the 2017 Cochran edition, using the optimal set point model, combined with the simulated plant growth algorithm, as follows:

Initialize the given point as the initial seed point. Set the number of iterations n to 1000 to ensure its full convergence. Then initialize the optimal set point best_node and set the initial value as the first seed point. Initialize the objective function best_obj_function of the optimal set point and set its initial value to the value of the objective function of the initial seed point. Set the growth step length step_length and determine the final value through experimentation and tuning.

In each iteration, calculate the objective function value of each seed point and select the seed point with the smallest objective function value as the current optimal set point. By constantly updating the optimal set point, this paper determines the clustering center of the stock data as (12.7682,0.7900,1.8560), as shown in Figure 2.

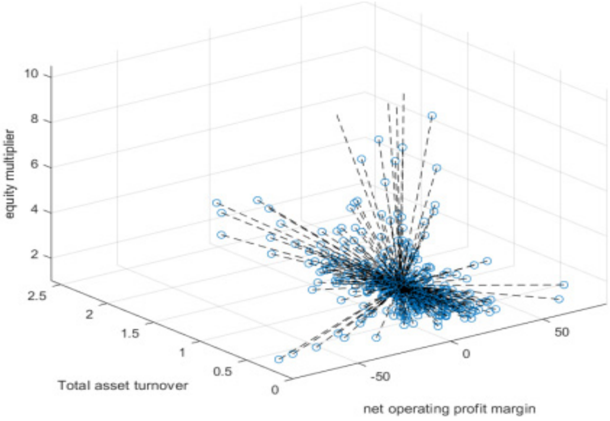

Figure. 2 Optimal assembly point selection map.

By replacing the randomly generated clustering centre in the k-means algorithm with the optimal clustering point calculated from the stock data in 2017, the data of more than three hundred stocks in the Kechuan version from 2017 to 2021 are clustered separately, and the distance between each stock and the optimal clustering point is calculated for each year, which is combined with the results of the calculations of the past few years, and the thresholds are classified, so as to obtain a more reasonable criterion to classify the level of the stocks. As show in figure 3.

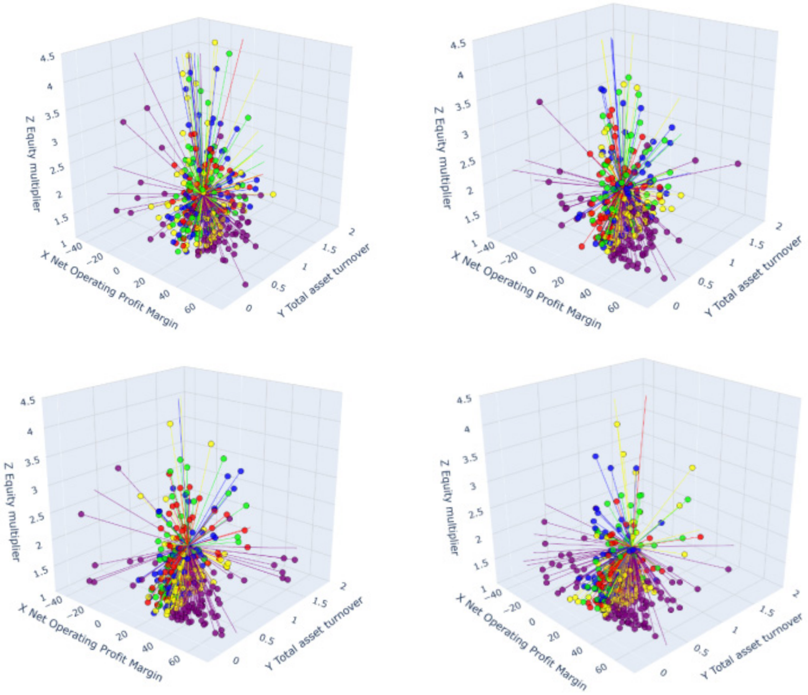

Figure 3. 2018-2020 stock clustering map.

Five thresholds of 3.213, 5.482, 8.689, 15.256, 100 were set by calculating the distance from the stock data points to the clustering center in 2017. In order to improve the accuracy, this paper adjusts the parameters with the clustering from 2018 to 2021, and finally sets the thresholds to 2.731, 6.115, 9.837, 16.412, 100 to better judge the degree of superiority or inferiority of the stocks. In this paper, this threshold is applied to the year 2022, so that it can better reflect the current status of the stock also in order to make the clustering results more accurate. This is shown in the figure 4 below.

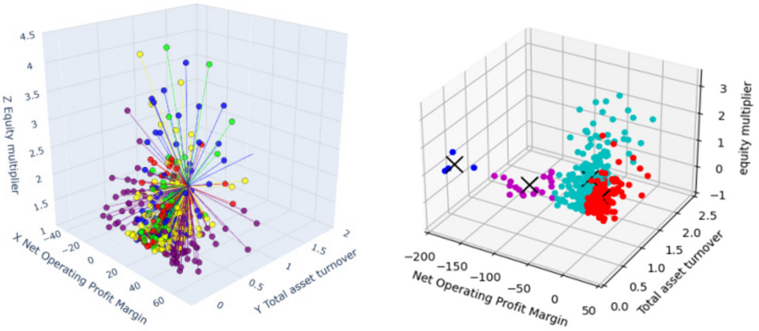

Figure 4. Clustering comparison diagram: new clustering algorithm (left), KNN clustering algorithm (right).

At the same time, this paper has done a comparison experiment to compare the clustering algorithm in this paper with the KNN clustering algorithm, as can be seen in Figure 5, the KNN clustering algorithm clusters by randomly selecting the clustering centers and forming clusters, and according to years of data comparison of the clustering effect is poorer (this paper is put in the year 2022 only), so compared with the traditional method, the speed of convergence is faster, and the degree of accuracy is higher. See in table 2.

Table 2. 2022 stock clustering partial results embodiment table.

Name	X	Y	Z	Distance	Threshold Category
Yupont Power	12.733	0.4883	1.6384	0.373646477	1
Golden Crown Electric	13.0056	0.5351	1.6401	0.40981164	1
Lyle Technologies	10.1528	0.4517	1.1282	2.735773545	2
Ricotta (brand)	15.5787	0.7191	1.5777	2.825135032	2
Nuotai Biologicals	18.8503	0.2802	1.3104	6.127765972	3
Liyuanheng	6.8872	0.5598	3.625	6.14560933	3
Canqin Technologies, Inc.	22.6307	0.1528	1.0983	9.912065344	4
Vigor Orthopedics	29.1731	0.3605	1.2015	16.42356796	5

5. CONCLUSIONS

In this paper, a new high-dimensional clustering algorithm is proposed, which puts forward the Steiner point with Pareto optimality as the new clustering center, solves the problem of randomly selecting the clustering center in the traditional clustering algorithm, improves the accuracy, and converges faster compared with the traditional method. At the same time, the interference of noise on the traditional clustering algorithm is reduced, and the algorithm is used for stock data analysis, which can provide a more scientific decision-making program for the majority of stockholders.

REFERENCES

[1] Shakhatreh, M., Shakhatreh, H., Ababneh, A., Efficient 3D Positioning of UAVs and User Association Based on Hybrid PSO-K-Means Clustering Algorithm in Future Wireless Networks[J]. Mobile Information Systems DOI:http://doi.org/10.1155/2023/6567897 (2023)

[2] Menneer, T., Mueller, M., Townley, S., A cluster analysis approach to sampling domestic properties for sensor deployment[J]. Building and Environment 231(1) ,110032 (2023)

[3] Qiao, L. F., Zhang, Y. C., Qi, A. G., Evaluation and classification of residential greenbelt quality based on factor analysis & clustering analysis: An example of Xinxiang City, China[J]. Journal of Forestry Research 19(4), 311-314 (2008)

[4] Liu, Y., Liu, Z. T., Li, S., Guo, Y. K., Liu, Q., Wang, G. Y., Cloud-Cluster: An uncertainty clustering algorithm based on cloud model[J]. Knowledge-Based Systems,2023,263,110261.

[5] Du, X. Z. A Robust and High-Dimensional Clustering Algorithm Based on Feature Weight and Entropy[J]. Entropy 25(3), 510-510 (2023)

[6] González, C. M. A., Sánchez, A., San, M. M., Network coevolution drives segregation and enhances Pareto optimal equilibrium selection in coordination games.[J]. Scientific reports 13(1), 2866-2866 (2023)

[7] Ni, Y. Y., Research on density peak clustering algorithm and its application in financial inclusion [D]. Northwest Normal University, 2022.DOI:10.27410/d.cnki.gxbfu.2022.001674

[8] G, J. N., Research on style refinement of equity funds[D]. Jiangxi University of Finance and Economics,2023.DOI:10.27175/d.cnki.gjxcu.2022.001271.

[9] Zhai, X.T., Cluster analysis based on soft-DTW distance and its application to A-share market[D]. Shandong University, 2023.DOI:10.27272/d.cnki.gshdu.2022.004329.

[10] Huang, L.S.,Zhang, Z.L.,Zhao, X.J., et al. A study on inferring stock portfolio positions based on hierarchical clustering dimensionality reduction model[J]. Financial Theory and Practice,2021(03), 7-13.