Research on Comprehensive Analysis Method of Stock KDJ Index based on K-means Clustering

Baoyu Ding[1, a], Ling Li [2, b, *], Yunliang Zhu[1, c], Hui Liu[3], Junfeng Bao[1, d], Zezhu Yang[1]

[1]Engineering College of Field Engineering Army engineering university of PLA Nanjing, China

[2]National Key Laboratory on Environmental Electromagnetic Effects and Electro-Optic Army engineering university of PLA Nanjing, China

[3]Graduate College Army engineering university of PLA Nanjing, China

[a]baoyu199302@163.com, [b, *] leonleeust@hotmail.com, [c]15051801667@163.com, [d]2967116271@qq.com.

Abstract. This paper proposed a K-means measure to cluster stocks, and predicted the investment of strong profitability objects through comprehensive analysis of KDJ indicators. The paper analyzed the clustering hierarchy diagram, as well as the inter-cluster similarity structure diagram of different cluster numbers. It is found that the clusters can be effectively distinguished for each type of stock. The comprehensive prediction precision of KDJ are better than each single index. The feasibility and effectiveness of the suggested method are verified by the example of the constituents of the CSI 800 Index. The quantitative investment model established by the analytical method in this paper has better prediction effect.

Keywords: cluster analysis, K-means, KDJ index, stock analysis forecast.

1. Introduction

The earliest scholar of the effectiveness of securities technology analysis in academia is Alexander (1961), who in the study assumed that when the stock price rises by at least x% from a certain low level, the stock is bought and held; when the price falls from a high level, at least At x%, the stock is sold and continues to wait for the next buy signal. The study concluded that when tested using the Dow Jones Industrial Average and the Standard & Poor's Index, this strategy can yield significant excess returns compared to the buy-and-hold strategy. Based on Alexander's research, Fama and Blume (1966) further added the company's dividend effect and transaction cost factors, using the pre-recovery of the Dow Jones index constituent stock price data. The two studies have shown that this trading strategy does not lead to a better outcome than the buy-and-hold strategy [1].

With the advent and widespread use of computers, people have begun to consider using computers to test the effectiveness of technical analysis using more advanced statistical methods. Brock, Lakonishok, and Le Baron (1992) were among the first scholars to study technical analysis indicators. They examined the profitability of two technical trading rules, the moving average strategy and the support line resistance breakout strategy.

The neural network is a massively parallel complex nonlinear dynamic system that represents an extremely complex nonlinear model system. In recent years, many methods have been developed that use neural networks for stock analysis.

Stein used the 3-year data of the German stock market as a training set and one-year daily data as a test set to build and test an artificial neural network model. The results show that half of the 24 buy signals generated by the neural network are correct. Yao and Poh [1] used the BP neural network to predict the index of the Kuala Lumpur Stock Exchange. The results show that the prediction accuracy of the neural network model trained with daily data is higher than that of the neural network model trained with weekly data. The above research results have completed the analysis of stock analysis from different angles and using different data mining methods, and have good application value [2].

However, these results are mainly based on the study of the operating rules of one or several stocks, and lack of a comprehensive analysis mechanism for existing stock technical indicators. Therefore, the research in this paper starts with the basic stock technical analysis indicators, and carries out multi-dimensional comprehensive analysis through data mining technology to form the information needed for trading decision [3].

2. The Principle

2.1 Definition of Clustering

Clustering is a process of dividing a data set into thousands of subsets, and makes the data objects in the same set have higher similarity, while the data objects in different sets are not similar, similar or dissimilar metrics. It is determined based on the value of the data object description attribute, which is usually described by using the distance between each cluster. The basic guiding ideology of analysis is to maximize the similarity of objects in the class and the smallest similarity between objects.

2.2 K-means Clustering Algorithm

Clustering is one of the important tasks of data mining. Given a set of elements D, each of which has n observable attributes, the clustering algorithm divides D into k subsets (called a cluster), requiring similarities between elements within each subset. It may be high, and the dissimilarity of elements in different subsets is as high as possible.

The K-means algorithm proposed by JB. MacQueen in 1967 is a classical clustering algorithm widely used in scientific research and industrial applications. It is simple and fast. In particular, it has been widely used for high scalability and noise immunity in large data set processing. The core idea of the K-means algorithm is to divide n data objects into n clusters, so that the sum of the squares of the data points in each cluster to the cluster center is the smallest [4]. The algorithm processing flow chart is as follows:

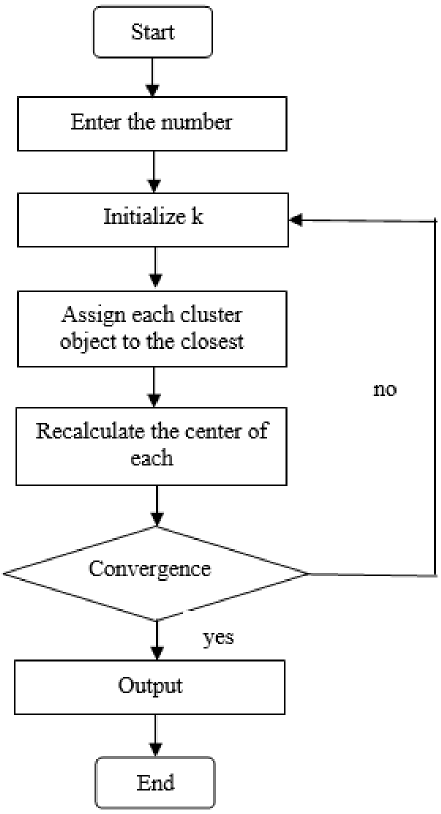

Start
Enter the number
Initialize k
Assign each cluster object to the closest
Recalculate the center of each
Convergence
no
yes
Output
End

Fig .1. The algorithm processing flow chart

First, k objects are arbitrarily selected from the n data objects as the initial cluster center: for the remaining other objects, they are assigned to be most similar to them according to their similarity (distance) with these cluster centers. Clustering (represented by cluster centers).

Then calculate the cluster center of each new cluster (the mean of all objects in the cluster) and repeat the process until the standard measure function begins to converge. Generally, the mean square error is used as a standard measure function, which is defined as follows:

$$E = \sum_{i=1}^{k} \sum_{p\in_{c_i}} \left|p - m_i\right|^2$$

Where *e* is the sum of the mean squared deviations of all objects in the database;

p is a point in the space representing the object;

m is the average value of cluster C (*p* and *m*, both are multidimensional).

The clustering criteria shown in the formula are intended to make the obtained *k* clusters have the following characteristics: each cluster itself is as compact as possible, and the clusters are separated as much as possible. There are 3 key issues:

1) Point-to-point, point and cluster, and cluster-to-cluster similarity calculation.

2) calculation of the center point of the cluster.

3) Iteration end condition setting.

2.3 KDJ Indicator Set

The KDJ index, also called stochastic index, is a fairly novel and practical technical analysis index. It was first used in the analysis of the futures market, and was widely used in the short-term trend analysis of the stock market. It is the most commonly used in the futures and stock markets. Technical analysis tools. The stochastic indicator KDJ is calculated based on the highest price, the lowest price and the closing price. The obtained K value, D value and J value are respectively formed at a point on the coordinate of the index, and an infinite number of such points are connected. Form a complete KDJ indicator that reflects the trend of price fluctuations. It is mainly a technical tool that uses the true volatility of price fluctuations to reflect the strength of price movements and overbought and oversold, and to issue trading signals before prices have risen or fallen. In the design process, it mainly studies the relationship between the highest price, the lowest price and the closing price. It also combines some of the advantages of the momentum concept, the strength indicator and the moving average [5]. Therefore, it can be judged quickly, quickly and intuitively. Quotes. Since the KDJ line is essentially a concept of random fluctuations, it is more accurate for grasping the short- and medium-term market.

The KDJ index is to obtain the immature stochastic index value by calculating the proportional relationship between the highest price, the lowest price and the closing price that have appeared in a specific period, and then smooth the corresponding data to obtain a series of index values, and draw the graph is used to judge the trend of securities.

The calculation of KDJ is more complicated. First, calculate the RSV value of the period (n, n, etc.), that is, the immature random index value, and then calculate the K value, D value, J value, and so on. Taking the calculation of the daily KDJ value as an example, the calculation formula is:

n day RSV=(Cn-Ln) ÷(Hn-Ln) ×100

In the formula, Cn is the closing price on the nth day; Ln is the lowest price in n days; Hn is the highest price in n days. The RSV value always fluctuates between 1 and 100.

Second, calculate the K and D values:

The day K value = 2 / 3 × the previous day K value + 1/3 × the day RSV

Day D value = 2 / 3 × previous day D value + 1/3 × day K value

If there is no K value and D value on the previous day, 50 can be used instead.

j value = 3 × day K value - 2 × day D value

KDJ can characterize the fluctuations of short and medium-term markets by setting time periods of different durations. The duration can be set to the level of week, month, and so on [6].

Using MATLAB programming, KDJ clustering at k=5: There are five categories (overbought area, oversold area, squat area, long area, short area).

The stochastic indicator is three curves on the graph. Both K line, D line, J line. Using the relationship between these three curves, we can study the trend of stock prices. The stochastic indicator is mainly used to reflect the overbought and oversold phenomenon in the stock market, the trend of the relaxation phenomenon and the cross-breaking phenomenon of the K-line and the D-line. Thus, predicting the short-term trend to the top and bottoming process.

3. Stock Analysis

3.1 Overbought and Oversold Phenomenon

The value of KDJ ranges from 0 to 100 (the J line sometimes exceeds). According to the popular and commonly used judgment criteria, the 0-100 can be divided into overbought area, oversold area and squat area.

Overbought area: K value is above 80, D value is above 70, and J value is greater than 90 when it is overbought. Under normal circumstances, the stock price may fall. Investors should be cautious, and outsiders should not chase after them. The insiders should sell them in due course.

Oversold area: K value below 20, D value below 30 is the oversold area. Under normal circumstances, the stock price may rise and the possibility of rebound increases. Insiders should not easily throw stocks, and outsiders can look for opportunities to enter.

Hover area: The KD value is around 50. For example, in the long market, 50 is the back-support line; Short market50 is the rebound pressure line; if it is around 50, it means that the market is still finishing, should be based on wait and see, it is not appropriate to rush to decide to buy and sell.

Bulls are investors who are optimistic about the stock market and expect stock prices to be bullish, so buy stocks at low prices and sell them when stocks rise to a certain price to get the difference. Bulls are one of the speculative methods in futures exchanges.

Shorts are investors and stockists who believe that the current stock price is higher, but the stock market outlook is bad, the stock price is expected to fall, so the stock is sold, and the IU is sold at a high price. The trading method of buying and selling the difference first is called short position.

Speculators estimate that securities, commodities, etc. have a tendency to increase in price, buy in advance, and attempt to sell after price increases in order to obtain the benefit of the difference. This kind of speculation is based on the first purchase. Speculators have a lot of securities or commodities before they are sold, so they are called “long” and “short”.

It should be noted that since the J-line reaction is sensitive, the variation speed is faster and the amplitude is higher than the K-line. D line. Generally, only for reference.

3.2 The Phenomenon of Back Relaxation

When the stock price trend is higher than the peak, the J-line of the stochastic indicator is lower than the peak, or when the stock price trend is lower than the trough, the J-line is higher than the trough. This phenomenon is called divergence. When the stochastic indicator and the stock price trend are divergent, it is generally a signal of market turn, indicating that the medium-term or short-term trend has reached the top or has bottomed out. This is the time to buy and sell stocks [10].

3.3 Cross-breaking of the K-line and the D-line

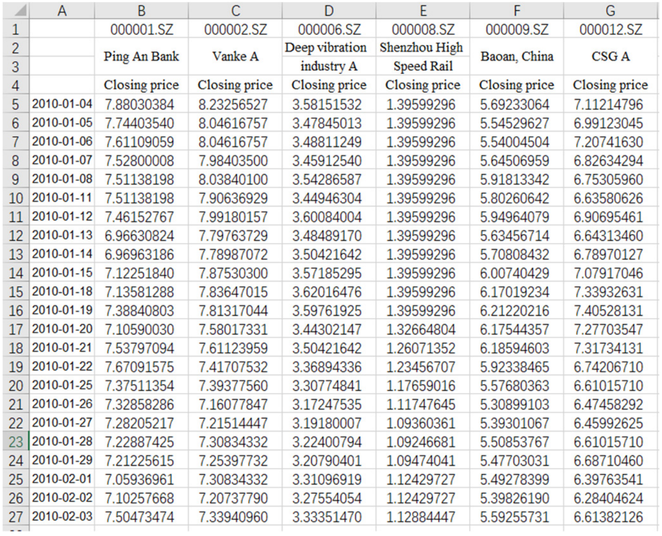
When the K value is greater than the D value, it indicates that the stock price is currently in an upward trend. Therefore, when the K line crosses the D line from the bottom to the top, it is the time to buy the stock. Conversely, when the K value is less than the D value, it indicates that the stock market is currently in a downward trend. Therefore, when the K line crosses the D line from top to bottom, it is the time to sell the stock.

4. The Experimental Process and Results Analysis and Verification

4.1 Sample Selection

This paper introduces the method of stock clustering by taking the constituents of CSI 800 as an example. The empirical study is conducted on the sample from the first trading day of 2010 to the last trading day of 2017 and the closing price of January 2018.

Stocks missing data are removed from the sample. And use this data to conduct a series of experiments to verify the validity of the proposed model. Export the raw data to Excel format by running the MATLAB program.



	A	B	C	D	E	F	G
1		000001.SZ	000002.SZ	000006.SZ	000008.SZ	000009.SZ	000012.SZ
2		Ping An Bank	Vanke A	Deep vibration industry A	Shenzhou High Speed Rail	Baoan, China	CSG A
3							
4		Closing price	Closing price	Closing price	Closing price	Closing price	Closing price
5	2010-01-04	7.88030384	8.23256527	3.58151532	1.39599296	5.69233064	7.11214796
6	2010-01-05	7.74403540	8.04616757	3.47845013	1.39599296	5.54529627	6.99123045
7	2010-01-06	7.61109059	8.04616757	3.48811249	1.39599296	5.54004504	7.20741630
8	2010-01-07	7.52800008	7.98403500	3.45912540	1.39599296	5.64506959	6.82634294
9	2010-01-08	7.51138198	8.03840100	3.54286587	1.39599296	5.91813342	6.75305960
10	2010-01-11	7.51138198	7.90636929	3.44946304	1.39599296	5.80260642	6.63580626
11	2010-01-12	7.46152767	7.99180157	3.60084004	1.39599296	5.94964079	6.90695461
12	2010-01-13	6.96630824	7.79763729	3.48489170	1.39599296	5.63456714	6.64313460
13	2010-01-14	6.96963186	7.78987072	3.50421642	1.39599296	5.70808432	6.78970127
14	2010-01-15	7.12251840	7.87530300	3.57185295	1.39599296	6.00740429	7.07917046
15	2010-01-18	7.13581288	7.83647015	3.62016476	1.39599296	6.17019234	7.33932631
16	2010-01-19	7.38840803	7.81317044	3.59761925	1.39599296	6.21220216	7.40528131
17	2010-01-20	7.10590030	7.58017331	3.44302147	1.32664804	6.17544357	7.27703547
18	2010-01-21	7.53797094	7.61123959	3.50421642	1.26071352	6.18594603	7.31734131
19	2010-01-22	7.67091575	7.41707532	3.36894336	1.23456707	5.92338465	6.74206710
20	2010-01-25	7.37511354	7.39377560	3.30774841	1.17659016	5.57680363	6.61015710
21	2010-01-26	7.32858286	7.16077847	3.17247535	1.11747645	5.30899103	6.47458292
22	2010-01-27	7.28205217	7.21514447	3.19180007	1.09360361	5.39301067	6.45992625
23	2010-01-28	7.22887425	7.30834332	3.22400794	1.09246681	5.50853767	6.61015710
24	2010-01-29	7.21225615	7.25397732	3.20790401	1.09474041	5.47703031	6.68710460
25	2010-02-01	7.05936961	7.30834332	3.31096919	1.12429727	5.49278399	6.39763541
26	2010-02-02	7.10257668	7.20737790	3.27554054	1.12429727	5.39826190	6.28404624
27	2010-02-03	7.50473474	7.33940960	3.33351470	1.12884447	5.59255731	6.61382126

Fig.2. Part of the stock closing price Excel diagram

People can't directly observe and judge the high-dimensional data of hundreds of stocks. Therefore, it is necessary to analyze the data through clustering algorithm, and discover hidden information in the data, so as to select a class of stocks with large appreciation space. investment. The condensed algorithm is a bottom-up strategy. First, each data point is treated as a single cluster, and then they are merged to form a larger cluster until all data points are in the same cluster, or a termination condition is reached. This paper intends to use a cohesive algorithm to program through the built-in functions of MATLAB statistics and machine learning toolbox. [7]

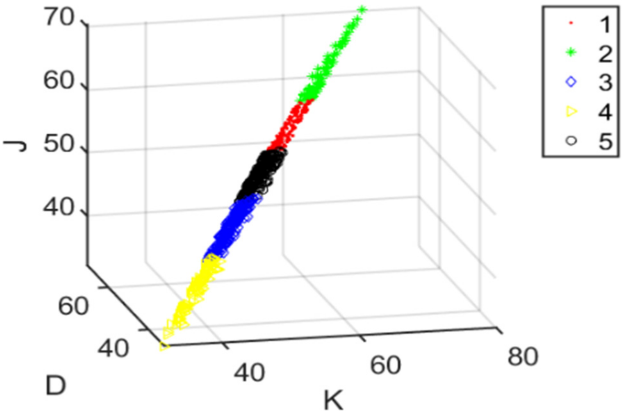


Fig.3. k is the cluster analysis result of the constituents of the CSI 800

4.2 Analysis of the Table

Table 1. Some Stock KDJ Indicator Values Indicate

Securities code	*Securities name*	*KDJ_K*	*KDJ_D*	*KDJ_J*
000001.SZ	Ping An Bank	54.66	49.85	53.85
000002.SZ	Vanke A	60.30	49.97	58.16
000006.SZ	Deep vibration industry A	53.22	50.03	54.15
000008.SZ	Shenzhou High Speed Rail	62.64	50.04	62.28
000009.SZ	Baoan, China	47.87	50.18	49.44
000012.SZ	CSG A	41.91	50.11	42.76
000021.SZ	Deep technology	45.47	49.99	46.57
000025.SZ	Special force A	56.41	49.89	57.01
000027.SZ	Shenzhen Energy	44.41	49.88	45.94
000028.SZ	National medicine	61.16	49.96	60.58
000031.SZ	COFCO Real Estate	45.92	49.84	47.44
000039.SZ	CIMC	45.06	50.18	46.07
000049.SZ	Desai battery	55.55	50.07	55.61
000060.SZ	Zhongjin Lingnan	38.76	50.05	40.28
……	……	……	……	……

4.3 Analysis of the Results

As shown in Figure 4, 800 stocks are clustered into 5 clusters. Typical characteristics of a cluster can be obtained by analyzing the characteristics of the data in each cluster. For example, for a stock with a stock code of 000008.sz, the stock has a k value of 62.64, a d value of 50.04, and a j value of 62.28, which is shown as the first cluster on the chart. According to the stock technical analysis experience, the stock is in the bullish area. You can also analyze the stock through the trend of the k-line d-line and the j-line. For example, select the stock code as 000001.SZ for the KDJ value analysis.

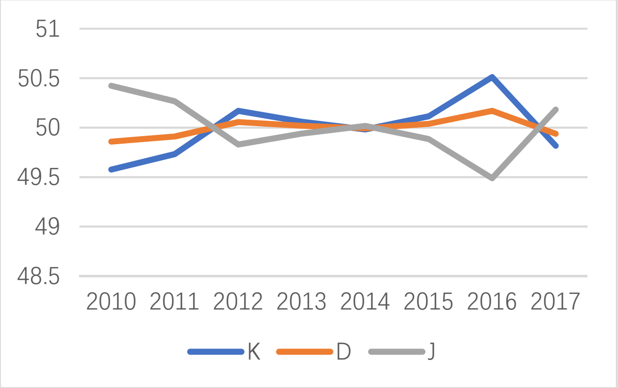


Fig.4. Pingtung Bank KDJ value trend chart

As shown in the above figure: When the blue k value is below the low level of 50, a phenomenon in which the bottom is higher than the bottom is formed, and the value of k is crossed from the bottom to the top twice, and the stock price will be larger.

When the blue K value is at a high level of 50 or higher, a phenomenon in which the top is lower than the top is formed, and the k value is crossed from the top to the bottom twice in orange, the stock price will have a large drop. Blue K line After crossing the orange d line from bottom to top, it failed to turn to the bottom. k line Cross the d line again, and the space between the two lines is called the upward reversal wind tunnel. As shown above, the stock price will rise when there is an upward reversal of the wind tunnel. Conversely, it is called downward reversal of the wind tunnel. The stock price will fall when there is a downward reversal of the wind tunnel.

Hover area: The KD value is around 50.For example, in the long market, 50 is the back support line; Short market50 is the rebound pressure line; if it is around 50, it means that the market is still finishing, should be based on wait and see, it is not appropriate to rush to decide to buy and sell [8].

4.4 Experimental Verification

To further validate the validity of the model, we used the average closing price of the stock in January 2018 to verify the previous stock price. Table 3 shows the results of partial data verification. The results of the rise and fall are identified by -1, 0, and 1, which represent the fall, level, and rise. The prediction accuracy is calculated as the percentage of the stock in a cluster that matches the actual rise and fall.

Table 2. Cluster Analysis Results

Value	***Accuracy%***
K	35.35
D	24.55
J	49.13
KDJ	75.66

The comparison results of the KDJ comprehensive analysis method and the K, D, J individual technical indicators analysis experiments are shown in Table 2 [9].

It can be seen from Table 4 that the accuracy of the KDJ cluster analysis method proposed in this paper is higher than that of the three indicators alone.

5. Conclusion

The KDJ-k-means model is a solution to the use of data mining technology based on the KDJ indicator set. It provides an idea and method for the in-depth analysis of stock data. The main contributions are summarized as follows:

1) Although the proposed model has only been analyzed and tested for the KDJ indicator set, it is also very useful for the comprehensive analysis of other stock technical analysis indicators.

2) Compared with neural networks and other methods, there are few researches on stock technology analysis methods using cluster mining technology, but this study shows that cluster mining application is not only feasible in stock analysis, but also has obvious features that are intuitive and easy to use.

The next research work is mainly to use the numerical analysis software MTLAB programming to carry out related research and experiments on various stock technical indicators and multiple data mining algorithms, and to expand the KDJ-k-means model and its scope of use.

Acknowledgments

This work is supported by Foundation of National Key Laboratory on Environmental Electromagnetic Effects and Electro-Optic Engineering. This paper is supported by the China Postdoctoral Fund Project (No. 2015M572730).

References

[1]. MENDELSSOHN L, STEIN J. Fundamental analysis meets the neural networks [J]. Futures, 1991, 20:22-24.

[2]. WingKeung Wong, Meher Manzur, Boon Kiat Chew. How rewarding is technical analysis? Evidence from Singapore stock market[J]. Applied Financial Economics,2002,13(7):543-551.

[3]. Vasiliou D, Eriotis N, Papathanasiou S. How rewarding is technical analysis? Evidence from Athens Stock Exchange[J]. Operational Research,2006, 6(2):85-102.

[4]. THUR AISINGHAM B M, CEUTIMG. Understanding data mining and applying it to command, control, communications and intelligence environments [C] Proceedings of the 24th Annual IEEE International Computer Software and Applications Conference.2013.

[5]. Chi-Jie Lu. Integrate independent component analysis-based denosing scheme with neural network for stock price prediction.2010.37(3).

[6]. Pedregosa F, Varoquaux, Gramfort A, et al. Scikit-learn: Machine Learning in Python[J]. Journal of Machine Learning Research, 2012,12(10):2825-2830.

[7]. Dea C, Heckler M, Grunwald G, et al. Java FX 8: Introduction by Example[J]. Springer Berlin, 2014.

[8]. Chang C C, Lin C J. LIBSVM: A library for support vector machines[J]. Acm Transactions on Intelligent Systems & Technology, 2011, 2(3):389-396.

[9]. Bollen J, Mao H, Zeng X. Twitter mood predicts the stock market[J]. Journal of Computational Science, 2010, 2(1):1–8.

[10]. Yamamoto R. Intraday technical analysis of individual stocks on the Tokyo Stock Exchange [J]. Journal of Banking&Finance,2012,36(11).