Hindawi
Mathematical Problems in Engineering
Volume 2017, Article ID 3096917, 11 pages
https://doi.org/10.1155/2017/3096917

Research Article

K-Line Patterns' Predictive Power Analysis Using the Methods of Similarity Match and Clustering

Lv Tao,[1] Yongtao Hao,[1] Hao Yijie,[2] and Shen Chunfeng[3]

[1]*College of Electronics and Information Engineering, Tongji University, Shanghai 200092, China*
[2]*Rabun Gap-Nacoochee School, Rabun Gap, GA 30568, USA*
[3]*Shanghai Baosight Software Co., Ltd., Shanghai 200092, China*

Correspondence should be addressed to Lv Tao; superlvtao@163.com

Received 23 December 2016; Revised 31 March 2017; Accepted 5 April 2017; Published 22 May 2017

Academic Editor: Anna M. Gil-Lafuente

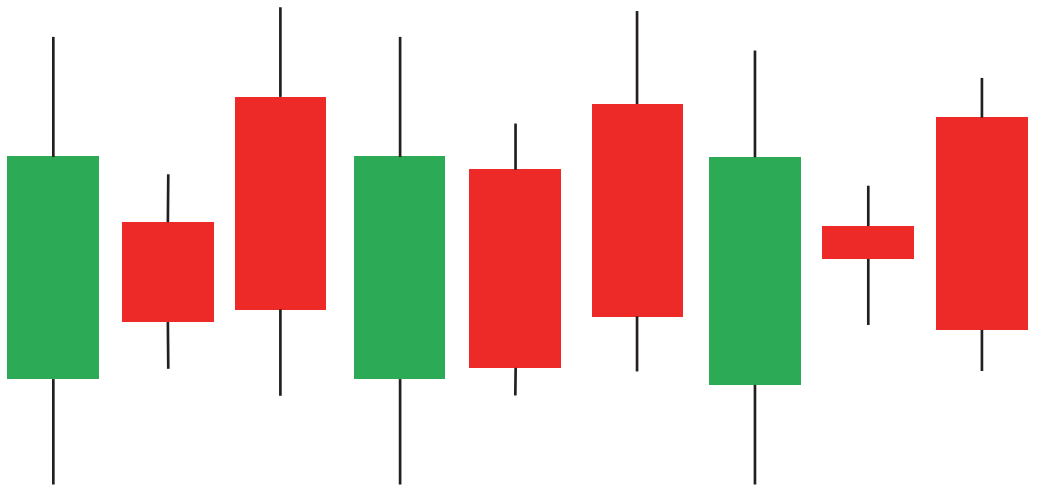
Stock price prediction based on K-line patterns is the essence of candlestick technical analysis. However, there are some disputes on whether the K-line patterns have predictive power in academia. To help resolve the debate, this paper uses the data mining methods of pattern recognition, pattern clustering, and pattern knowledge mining to research the predictive power of K-line patterns. The similarity match model and nearest neighbor-clustering algorithm are proposed for solving the problem of similarity match and clustering of K-line series, respectively. The experiment includes testing the predictive power of the Three Inside Up pattern and Three Inside Down pattern with the testing dataset of the K-line series data of Shanghai 180 index component stocks over the latest 10 years. Experimental results show that (1) the predictive power of a pattern varies a great deal for different shapes and (2) each of the existing K-line patterns requires further classification based on the shape feature for improving the prediction performance.

1. Introduction

A time series is a series of observations listed in time order. It is the most commonly encountered data type, touching almost every aspect of human life [1], for example, the meteorological time series, the time series of stock prices (stock time series for short) which are composed of stock price observations, and the time series of personal health that are consisted of the observation of blood pressure, temperature, white corpuscle, and so forth.

Researches show that the time series have two import features. (a) The historical information will affect the future trend [2]. That is, the historical values of observations will exert an influence on the future values in the time series. The influence can be described by time series' period, nonstationarity, varying volatility, and so on. (b) History repeats itself [3]. That is to say, some special time subseries will repeat in the entire time series. Because of the two features, all kinds of time series forecasting have become a present hot research, one of which is the prediction of stock time series, stock prediction for short. As a typical time series, not only have stock time series the features of time series, but also the trend of stock prices is directly related to the people's vital interests. Therefore, stock prediction has aroused the interest of a wide variety of researchers.

There are many technical analysis methods about stock prediction, the best known of which is candlestick technical analysis that is also called K-line technology analysis in Asia. In the stock market, in order to learn and study the fluctuation of stock prices in a more intuitive way, people invent a candlestick chart (also called K-line) to represent stock time series graphically. Taking a daily K-line, for example, a K-line represents the fluctuation of stock prices in one day, it not only shows the close price, open price, high price, and low price for the day but also reflects the difference and size between any two prices (all K-lines given in the paper refer to daily K-line, unless otherwise indicated). If the K-line of a stock lists in time order, then a series used to reflect the fluctuation of the stock price for some time can be formed, which can be called K-line series. As each K-line consists of four prices, the essence of K-line series is stock series with four observations.



(a) Shape A (b) Shape B (c) Shape C

Figure 1: Three kinds of shapes of TIU pattern.

In K-line series, if a K-line subseries contains some knowledge used to predict stock, then this subseries is called a K-line pattern series, a K-line pattern for short. For instance, when a subseries appears, the stock price will often rise or descend. Then, this subseries is a typical pattern series. Stock prediction based on K-line patterns is the essence of K-line technology analysis. How to mine the K-line patterns and how to make use of these patterns for predicting are main research contents of K-line technology analysis.

By the artificial methods of observing the K-line series of stock market (or Japanese rice market), people (the leading character is the founder of K-line, Munehisa Honma, who was a Japanese rice trader in the 18th century) have found many K-line patterns. The literatures [4, 5] introduce the existing patterns and their features in detail, such as Three Inside Up (TIU), Three Inside Down (TID), and Doji. Some papers [6–10] conclude from the experiment that the existing K-line patterns have a good forecasting capability for forecasting stock trends. Some other papers [11–15] have studied the stock prediction based on these patterns and have achieved some research results. However, there are also a number of papers [5, 16–18] challenging these patterns' predictive power. They argue that K-line technology analysis violates the efficient market hypothesis, so it is not feasible for stock investment based on K-line patterns. They also did some experiments, which show that the existing K-line patterns have no predictive power.

Based on the above analysis, it is obvious that there are some disputes on whether the K-line patterns have predictive power in academia. However, there are few papers analyzing the reason why there are two different positions regarding the patterns' predictive power. Paper [19] also pays attention to the debate, while it does not analyze the K-line patterns themselves but attempts to obtain an answer to the following question: are the trend reversals accompanied more often by some types of candlesticks than by others? Finally, paper [19] has found that there exist types of candlesticks that frequently tend to appear close to the trend-reversal regions and others that cannot be found in such regions. Although the paper's research shows that the K-line patterns exist, it does not give the answer that why there is a debate on the K-line patterns' predictive power.

Through reviewing the relevant literatures, this paper considers that the main reason is that the existing K-line patterns are lack of rigorous mathematical definition. For example, the shadow length and body size are not defined clearly in the definition of K-line patterns, which means that a K-line pattern has many different shapes. Because the predictive power of a pattern may vary a lot for different shapes. If we ignore the shape difference and research the predictive power of a pattern by taking all patterns with various shapes as a whole instead of classifying the pattern further based on its shape feature, then the study result of K-line patterns' predictive power may produce deviations. For instance, a TIU pattern has three shapes: shape A, shape B, and shape C, as shown in Figure 1, where shape A is the generic form of TIU pattern, and shape B and C are infrequent form of which. Suppose that shape A has predictive power, and shape B and C do not have predictive power. When studying the predictive power of TIU pattern, if we ignore the shape difference between the three patterns and research them as a whole, then we will come to the wrong conclusion that TIU pattern has no predictive power. However, if the three patterns are classified further based on shape features and researched separately, then we can get the correct conclusion that TIU pattern has predictive power only at shape A.

In addition, another reason is that, as the existing K-line patterns are mined by artificial means, there may be some spurious pattern in them.

In order to resolve the debate and verify the two inferences, this paper presents the research of K-line patterns'

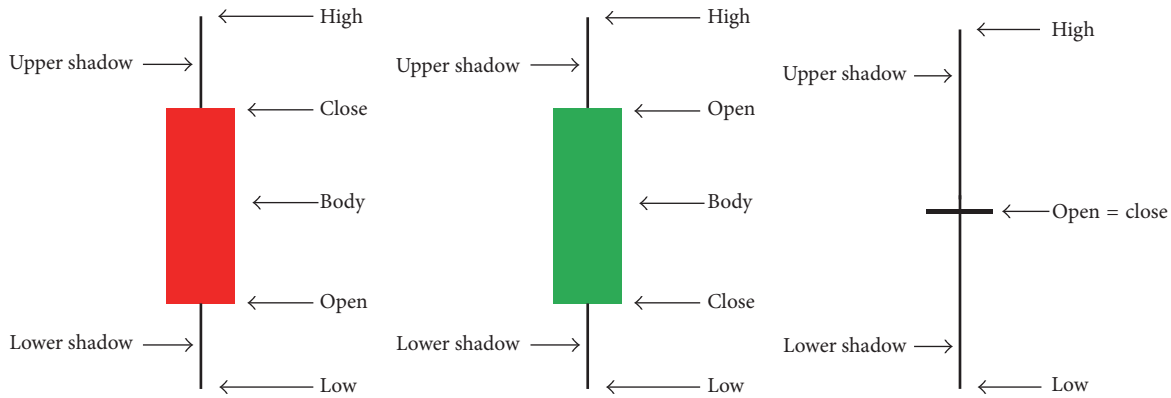


Figure 2: K-line chart.

predictive power using the data mining related method, such as pattern recognition, pattern clustering, pattern knowledge mining, and statistical analysis. The rest of this paper is organized as follows. In Section 2, we firstly shortly introduce K-line, K-line technology analysis, and K-line patterns. Then we define the similarity match model and nearest neighbor-clustering algorithm of K-line series. In Section 3, we define the mining method of patterns' predictive power. Section 4 presents the experimental result and discussion. Section 5 concludes the paper.

2. K-Line and K-Line Series Clustering

Firstly, we give the mathematic definition of K-line series. Let KS^i represent the i-th K-line series of any stock, and let D_t^i represent t-th K-line in KS^i; then

$$\mathrm{KS}^i = \begin{bmatrix} C_1^i & C_2^i & \cdots & C_{|\mathrm{KS}^i|} \\ O_1^i & O_2^i & \cdots & O_{|\mathrm{KS}^i|}^i \\ H_1^i & H_2^i & \cdots & H_{|\mathrm{KS}^i|}^i \\ L_1^i & L_2^i & \cdots & L_{|\mathrm{KS}^i|}^i \end{bmatrix}, \quad \left(\left|\mathrm{KS}^i\right| \in N^+\right), \tag{1}$$

$$D_t^i = \left[C_t^i, O_t^i, H_t^i, L_t^i\right]^T, \quad \left(1 \le t \le \left|\mathrm{KS}^i\right|\right),$$

where $|\mathrm{KS}^i|$ is the number of elements in KS^i, which is also called the length of KS^i. C_t^i, O_t^i, H_t^i, and L_t^i are the t-th day's close price, open price, high price, and low price in KS^i, respectively. In this paper, "| |" symbol indicates the number of elements in the set or series.

2.1. K-Line

2.1.1. K-Line Introduction. As defined in literature [4–6], the K-line is drawn by four basic elements: close price, open price, high price, and low price, where the part between the close price and open price is drawn into a rectangle called body of K-line and the part between the high price and body is drawn into a line called upper shadow of K-line. Moreover, the part between the lower price and body is drawn into a line called lower shadow of K-line. This kind of very personalized lines consisting of upper shadow, lower shadow, and body is called K-line.

In the K-line, if open price is lower than close price, K-line also called Yang line, the body is usually filled with white or green color, as shown in Figure 2(a). And if open price is higher than close price, K-line also called Yin line, the body is usually filled with black or red color, as shown in Figure 2(b). Moreover, if open price is equal to close price, K-line also called Doji line, the body then collapses into a single horizontal line, as shown in Figure 2(c). It is important to note that the body color of Yin line and Yang line is different in Chinese stock market and stock markets of European and American. In Chinese stock market, the body color of Yang line and Yin line is red and green, respectively. However, the body color of Yang line and Yin line is green and red, respectively, in the stock markets of European and American.

2.1.2. K-Line Technology Analysis. Firstly, we introduce and define some key concepts of K-line technology analysis. Let D_t represent the t-th day's K-line of any stock.

(1) Moving Average. It is the average of stock price for some time. The three-day moving average at time D_t is defined by

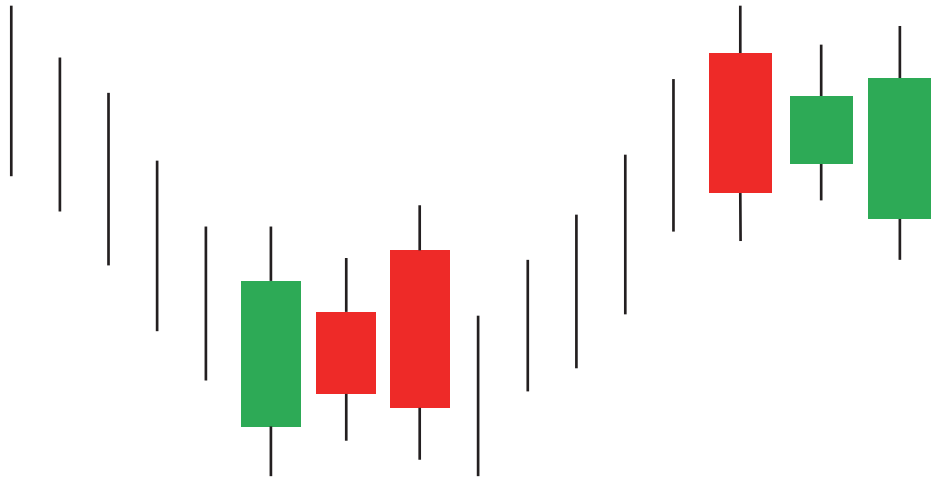
$$M_{\mathrm{avg}}\left(D_t\right) = \frac{1}{3}\left\{C_{t-2} + C_{t-1} + C_t\right\}, \tag{2}$$

where C_t denotes the close price of D_t.

(2) K-Line Trend. It is used to describe the K-line's trend, including uptrend and downtrend. D_t is said to be a downtrend if

$$M_{\mathrm{avg}}\left(D_{t-6}\right) > M_{\mathrm{avg}}\left(D_{t-5}\right) > \cdots > M_{\mathrm{avg}}\left(D_t\right) \tag{3}$$

with at most one violation of the inequalities. Uptrend is defined analogously.



(a) TIU (b) TID

Figure 3: THE standard K-line series chart of TIU and TID.

(3) Stock Price Trend. It is used to describe the general trend of stock prices for some time, including uptrend and downtrend. If the future trend of stock price is rising, it is called bullish market. In contrast, if the future trend of stock price is descending, it is called bearish market. Moreover, a more intense rising or descending trend indicates a more typical bullish or bearish market. The capability of a K-line patter for predicting the bullish market and bearish market is defined in formulas (17) and (18), respectively.

It is noted that the concepts of "moving average" and "K-line trend" are defined by the paper [6], while the concept of stock price trend is firstly defined by the paper.

2.1.3. K-Line Patterns. Many K-line patterns have been mined up to now, as shown in literatures [4–6]. Limited by space, only the patterns of TIU and TID will be introduced in the next content. Let KS = $[D_t, D_{t+1}, D_{t+2}]$ represent a three-day K-line series.

The conditions of KS becoming the TIU pattern are as follows: (1) D_t is a downtrend, and $C_t < O_t$. (2) $C_{t+1} > O_{t+1}$, $O_t \geq C_{t+1} > C_t$, and $O_t \geq O_{t+1} > C_t$, where at most one of the two equalities holds. That is, the second day $t + 1$ is Yang line and must be contained with the body of the first day. (3) $C_{t+2} > O_{t+2}$; $C_{t+2} > O_t$. That is, the third day $t + 2$ is Yang line and closes above the open of the first day. A standard TIU pattern is shown in Figure 3(a).

The predictive power of TIU pattern from the existing literature is that TIU is a trend-reversal pattern, which gives the bullish market signal. This means when the TIU pattern appears, the stock prices will be likely to be transferred from downtrend into uptrend or the stock market would be changed from bearish market to bullish market, and the stock prices would rise gradually.

The conditions of KS becoming the TID pattern are as follows: (1) D_t is an uptrend, and $C_t > O_t$. (2) $C_{t+1} < O_{t+1}$, $C_t \geq C_{t+1} > O_t$, and $C_t \geq O_{t+1} > O_t$, where at most one of the two equalities holds. That is, the second day $t + 1$ is Yin line and must be contained within the body of the first day. (3) $C_{t+2} < O_{t+2}$; $C_{t+2} < O_t$. That is, the third day $t + 2$ is Yin line and its close is lower than the first day's open. A standard TID pattern is shown in Figure 3(b).

The predictive power of TID pattern from the existing literature is that TID is a trend-reversal pattern, which gives the bearish market signal. That means, after the TID pattern appears, the stock prices will be likely to be transferred from uptrend into downtrend or the stock market would be changed from bullish market to bearish market, and the stock prices would fall gradually.

2.2. Similarity Match of K-Line Series. The similarity match of K-line series is an essential and basis task for K-line series clustering. In the literature, however, there are few papers focusing on the similarity match of K-line series. Only paper [20] studies the similarity match method and search algorithm of K-line series using image retrieval technology. In addition, paper [19, 21] proposes the similarity match model of K-line series based on the traditional Euclidean distance.

From the view of stock prediction, the K-line series' similarity refers to the trend similarity of K-line in the K-line series. However, the K-line trend is determined by the close price change, open price change, high price change, low price change, and the size relationship between close price and open price. Therefore, if we want to match the similarity between two K-line series, we should calculate the similarity of K-line price changes instead of the similarity of price values. As the changes of K-line price are not shown in the K-line chart, K-line prices distance rather than K-line price changes distance is used in the similarity match model of literature [19–21]. This means that these match models belong to similarity match methods based on K-line price values rather than K-line price changes. Therefore, they cannot accurately measure the similarity of stock prices trend in the K-line series.

For example, assuming that there are two K-line series KS^i and KS^j needed to match their similarity, where $KS^i = KS^j$ and $Sim^{i,j}$ indicate their similarity. Let $C_t^i = 10$, $C_{t+1}^i = 10.5$, $C_t^j = 20$, $C_{t+1}^j = 21$, and $RC_i^{t+1,t}$ indicate the close price change rate of KS^i at day $t+1$, which is calculated by $(C_{t+1}^i - C_t^i)/C_t^i$, $SRC_{i,j}^{t+1,t}$ denotes the similarity between $RC_i^{t+1,t}$ and $RC_i^{t+1,t}$, then $RC_i^{t+1,t} = RC_j^{t+1,t} = 5\%$, and $SRC_{i,j}^{t+1,t} = 1$. We cannot calculate the correct result of $SRC_{i,j}^{t+1,t} = 1$ by the similarity match model in literature [19–21]. Similarly, the same problems would occur for calculating the similarity of open price, high price, or low price.

Therefore, this paper proposes a new similarity match model based on K-line price changes to measure the trend similarity between two K-line series. In this model, the similarity of K-line series is composed of two parts: one is the shape similarity of K-line, which is the similarity of the corresponding K-line's shape features in the two K-line series; the other is the position similarity of K-line, which is the similarity of the corresponding K-line's position features in the two K-line series. Therefore, this paper will define K-line series' shape similarity model and position similarity model, respectively. Then based on these two kinds of similarity models, the similarity model of the entire K-line series could be built.

2.2.1. The Shape Similarity of K-Line Series. According to the shape feature of K-line, this paper proposes using the shape distance to measure the shape similarity between two K-lines. Firstly, based on the shape structure of K-line, three components of K-line shape are extracted: the upper shadow shape, the lower shadow shape, and the body shape. Secondly, the similarity match methods of three shapes are defined, respectively. Finally, the shape similarity of K-line can be calculated by summing the three shapes' similarity. Assuming that D_t^i and D_t^j denote the t-th day's K-line of KS^i and KS^j, respectively, the shape similarity model of K-line series is defined as follows:

(1) Let $US^i[t]$ denote the upper shadow length of D_t^i, as defined in the following formula:

$$US^i[t] = \begin{cases} \dfrac{H_t^i - O_t^i}{C_{t-1}^i * 0.1}, & O_t^i \geq C_t^i, \\ \dfrac{H_t^i - C_t^i}{C_{t-1}^i * 0.1}, & O_t^i < C_t^i, \end{cases} \tag{4}$$

where $C_{t-1}^i * 0.1$ is used to normalize the upper shadow length. According to the related regulation of Chinese A-share market, the range of daily fluctuations of stock prices cannot exceed 10% of the previous day's close price. So $C_{t-1}^i * 0.1$ can be used to normalize the length of the K-line's upper shadow, lower shadow, and body.

Let $Sim_{US}^{i,j}(t)$ denote the upper shadow similarity between D_t^i and D_t^j, as defined by

$$Sim_{US}^{i,j}(t) = \begin{cases} \dfrac{\min\left(US^i[t], US^j[t]\right)}{\max\left(US^i[t], US^j[t]\right)}, & US^i[t] * US^j[t] > 0, \\ 0, & US^i[t] * US^j[t] = 0,\ US^i[t] \neq US^j[t], \\ 1, & US^i[t] = US^j[t] = 0. \end{cases} \tag{5}$$

(2) Let $LS^i[t]$ denote the lower shadow length of D_t^i, as defined in the following formula:

$$LS^i[t] = \begin{cases} \dfrac{C_t^i - L_t^i}{C_{t-1}^i * 0.1}, & O_t^i \geq C_t^i, \\ \dfrac{O_t^i - L_t^i}{C_{t-1}^i * 0.1}, & O_t^i < C_t^i. \end{cases} \tag{6}$$

Let $Sim_{LS}^{i,j}(t)$ denote the lower shadow similarity between D_t^i and D_t^j, as defined by

$$Sim_{LS}^{i,j}(t) = \begin{cases} \dfrac{\min\left(LS^i[t], LS^j[t]\right)}{\max\left(LS^i[t], LS^j[t]\right)}, & LS^i[t] * LS^j[t] > 0, \\ 0, & LS^i[t] * LS^j[t] = 0,\ LS^i[t] \neq LS^j[t], \\ 1, & LS^i[t] = LS^j[t] = 0. \end{cases} \tag{7}$$

(3) Let $B^i[t]$ denote the body length of D_t^i, as defined in the following formula:

$$B^i[t] = \frac{C_t^i - O_t^i}{C_{t-1}^i * 0.1}. \tag{8}$$

$$\mathrm{Sim}_{\mathrm{Body}}^{i,j}(t) = \begin{cases} \dfrac{\min\left(\left|B^i[t]\right|, \left|B^j[t]\right|\right)}{\max\left(\left|B^i[t]\right|, \left|B^j[t]\right|\right)}, & B^i[t] * B^j[t] > 0, \\ 0, & B^i[t] * B^j[t] < 0, \\ 0, & B^i[t] * B^j[t] = 0,\ \left|B^i[t]\right| \neq \left|B^j[t]\right|, \\ 1, & B^i[t] = B^j[t] = 0. \end{cases} \tag{9}$$

(4) Let $\mathrm{Sim}_S^{i,j}(t)$ denote the shape similarity between D_t^i and D_t^j, as defined by

$$\begin{aligned} w_{\mathrm{Body}} + w_{\mathrm{LS}} + w_{\mathrm{US}} &= 1, \\ w_{\mathrm{Body}} &\geq 0, \\ w_{\mathrm{US}} &\geq 0, \\ w_{\mathrm{LS}} &\geq 0, \\ \mathrm{Sim}_S^{i,j}(t) &= w_{\mathrm{Body}} * \mathrm{Sim}_{\mathrm{Body}}^{i,j}(t) + w_{\mathrm{US}} \\ &\quad * \mathrm{Sim}_{\mathrm{US}}^{i,j}(t) + w_{\mathrm{LS}} \\ &\quad * \mathrm{Sim}_{\mathrm{LS}}^{i,j}(t), \end{aligned} \tag{10}$$

where w_{Body}, w_{US}, and w_{LS} represent the weight of $\mathrm{Sim}_{\mathrm{Body}}^{i,j}(t)$, $\mathrm{Sim}_{\mathrm{US}}^{i,j}(t)$, and $\mathrm{Sim}_{\mathrm{LS}}^{i,j}(t)$, respectively.

(5) Let $\mathrm{SSim}^{i,j}$ denote the shape similarity between KS^i and KS^j, as defined by

$$\mathrm{SSim}^{i,j} = \sum_{t=1}^{n} \mathrm{Sim}_S^{i,j}(t) * w_S^t, \quad \left(n = \left|\mathrm{KS}^i\right|,\ \sum_{t=1}^{n} w_S^t = 1\right), \tag{11}$$

where w_S^t represent the weight of $\mathrm{Sim}_S^{i,j}(t)$. Thanks to the idea that each K-line can be given different weight, the K-line series having special shape features could be identified well.

2.2.2. The Position Similarity of K-Line Series. For computing the similarity between two K-line series, we not only consider the shape similarity of K-line series but also the position similarity. If we only consider the shape similarity, then it will cause the problem that two K-line series having same shape features but different position features will have the same similarity.

For example, supposing that the K-line series chart of KS^i and KS^j is shown in Figure 2, we can see that, according to the shape feature definition of K-line, all of the corresponding K-lines of KS^i and KS^j have the same shape features. These mean that KS^i and KS^j have identical shape features; that is, $\mathrm{SSim}^{i,j} = 1$. However, as is vividly shown in Figure 4, the relative positions of D_2^i and D_2^j are different though D_1^i and D_1^j have the same relative position in the K-line series. Therefor the stock price's overall trend of KS^i and KS^j are not identical, that is, $\mathrm{Sim}^{i,j} < 1$. If we only consider the shape similarity, we will draw the wrong conclusion that $\mathrm{SSim}^{i,j} = \mathrm{Sim}^{i,j} = 1$.

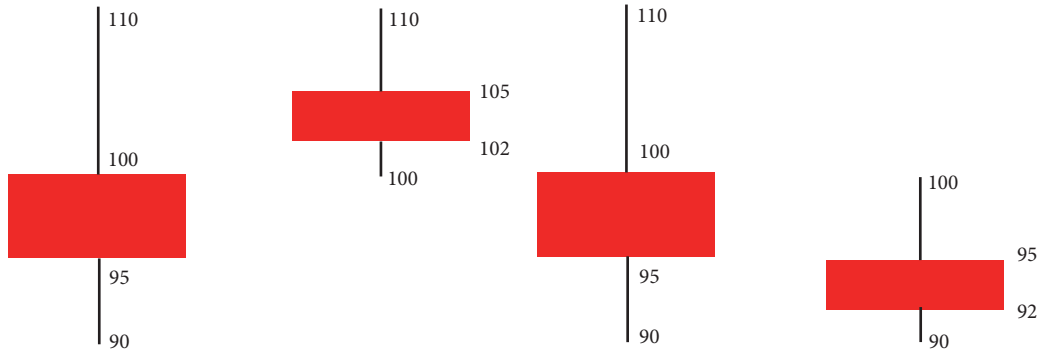
To solve this problem, the concept of K-line coordinate is introduced hoping to implement the position match of K-line by defining K-line's coordinate in the K-line series. In this paper, the sequence of K-line in the K-line series is called x coordinate of K-line; the increase range of close price is called y coordinate of K-line; in addition the first K-line's y coordinate is set to 1 in the K-line series. Therefore, the position similarity model of K-line series based on K-line coordinate is defined as follows.

(1) Let (x_t^i, y_t^i) denote the coordinate of D_t^i, which are defined in the following formula:

$$\begin{aligned} x_t^i &= t, \\ y_t^i &= \begin{cases} 1, & t = 1, \\ \dfrac{\left(C_t^i - C_{t-1}^i\right)}{C_{t-1}^i * 0.1}, & t > 1. \end{cases} \end{aligned} \tag{12}$$

Let $\mathrm{Sim}_P^{i,j}(t)$ denote the position similarity between D_t^i and D_t^j, as defined by

$$\mathrm{Sim}_P^{i,j}(t) = \begin{cases} \dfrac{\min\left(y_t^i, y_t^j\right)}{\max\left(y_t^i, y_t^j\right)}, & y_t^i * y_t^j > 0, \\ 0, & y_t^i * y_t^j = 0,\ y_t^i \neq y_t^j, \\ 1, & y_t^i = y_t^j = 0. \end{cases} \tag{13}$$



(a) KS^i (b) KS^j

Figure 4: K-line series chart.

(2) Let $PSim^{i,j}$ denote the position similarity between KS^i and KS^j, as defined by

$$PSim^{i,j} = \sum_{t=1}^{n} Sim_P^{i,j}(t) * w_P^t, \quad \left(n = \left|KS^i\right|, \sum_{t=1}^{n} w_P^t = 1\right), \tag{14}$$

where w_P^t represents the weight of $Sim_P^{i,j}(t)$. Thanks to the idea that each K-line can be given different weight, the K-line series having special coordinates could be identified well.

2.2.3. The Similarity of K-Line Series. Finally, based on the shape similarity and position similarity, the similarity of K-line series could be obtained. Therefore, the similarity match model between KS^i and KS^j is defined by

$$Sim^{i,j} = SSim^{i,j} * w_S + PSim^{i,j} * w_P, \tag{15}$$

where w_S and w_P represent the shape similarity weight and position similarity weight of K-line series, respectively.

2.3. Cluster of K-Line Series. The more accurate classification result of K-line patterns can be gotten by clustering them using the nearest neighbor-clustering algorithm based on the similarity match model of K-line series. The K-line series' nearest neighbor-clustering algorithm (KNNCA) is described as shown in Algorithm 1.

In addition, $|Q^m|$ represents the number of elements in Q^m. As each K-line series will be matched once with all of the K-line series stored in the cluster, the time complexity and space complexity of KNNCA are both $O(n^2)$.

3. Mining of Patterns' Predictive Power

We can mine and analyze the patterns' predictive power according to the following steps.

(1) Pattern Recognition. Based on the definition of K-line patterns, we identify all the K-line series belonging to a pattern (such as TIU or TID), and then they form a set (KSet).

(2) Pattern Clustering. We use the KNSSC algorithm to cluster KSet; then the set of clusters (CSet) can be gotten, in which different clusters represent the same pattern's different shapes.

(3) Knowledge Mining. We define some statistical indicators about stock prices, which we use to mine stock prediction knowledge from each cluster.

The pattern's predictive power is gotten primarily by analyzing the trend of the pattern's consequent K-line series. Paper [22] found that K-line technology is suited for short-term investment prediction and that the most efficient time period for prediction is 10 days. Therefore, we mainly analyze the close price trend of the pattern's consequent K-line series in 10 days. Let $KS = [D_t, D_{t+1}, D_{t+2}]$ denote a three-day K-line pattern; its consequent K-line series is denoted by CKS. The statistical indicators of CKS are defined as follows.

(a) Let C_k denote the k-th close price of CKS, let $P_U(C_k)$ denote that the probability of the trend of C_k is uptrend, and let $P_D(C_k)$ denote that the probability of the trend of C_k is downtrend. $P_U(C_k)$ and $P_D(C_k)$ are calculated by

$$P_U(C_k) = \frac{\left|Q_U^{m,k}\right|}{|Q^m|}, \quad P_D(C_k) = \frac{\left|Q_D^{m,k}\right|}{|Q^m|}, \tag{16}$$

where $|Q_U^{m,k}|$ represents the number of patterns meeting the condition of $C_k \geq C_{t+2}$ in Q^m, $|Q_D^{m,k}|$ represents the number of patterns meeting the condition of $C_k \leq C_{t+2}$ in Q^m, and C_{t+2} represents the close price of D_{t+2}. $P_U(C_k) > 0.5$ indicates that the future trend of C_t is rising. $P_D(C_k) > 0.5$ indicates that the future trend of C_k is descending.

(b) Let $P_U^n \in [0, 1]$ denote the probability that the close price will rise in the next n days if the pattern appears. $P_D^n \in [0, 1]$ denotes the probability that the close price will

Input:
KSet = $\{KS^1, KS^2, \ldots, KS^n\}$ // the data set of K-line series
θ // Similarity threshold
Output:
CSet // the set of clusters
KNNCA Algorithm:
Assign initial value for parameters: $w_S, w_P, w_{\text{Body}}, w_{\text{US}}, w_{\text{LS}}, w_S^t, w_P^t$;
$m = 1$;
$Q^m = \{KS^1\}$; // Q^m represents the m-th cluster
CSet = $\{Q^m\}$;
FOR $i = 2$ **TO** n **DO**
{
SimMax = 0;
FOR EACH Q^{item} **IN** CSet
FOR EACH KS^j **IN** Q^{item}
Get $\text{Sim}^{i,j}$ based on formula (15);
If ($\text{Sim}^{i,j}$ > SimMax)
{
SimMax = $\text{Sim}^{i,j}$;
f = item; // f represents the ID of a cluster whose element is most similar to KS^i
}
End
End
IF (SimMax > θ) **THEN**
$Q^f = Q^f \cup KS^i$;
ELSE
{
$m = m + 1$;
$Q^m = \{KS^i\}$;
}
}
CSet = $\{Q^1, Q^2, \ldots, Q^m\}$;

Algorithm 1

fall in the next n days if the pattern appears. P_U^n and P_D^n are calculated by

$$P_U^n = \sum_{k=1}^{n} \frac{\text{Round}\left[P_U\left(C_k\right), 0.5\right]}{n},$$
$$\text{Round}\left[P_U\left(C_k\right), 0.5\right] = \begin{cases} 1, & P_U\left(C_k\right) > 0.5, \\ 0, & P_U\left(C_k\right) \le 0.5, \end{cases} \tag{17}$$

$$P_D^n = \sum_{k=1}^{n} \frac{\text{Round}\left[P_D\left(C_k\right), 0.5\right]}{n},$$
$$\text{Round}\left[P_D\left(C_k\right), 0.5\right] = \begin{cases} 1, & P_D\left(C_k\right) > 0.5, \\ 0, & P_D\left(C_k\right) \le 0.5, \end{cases} \tag{18}$$

where a higher value of P_U^n or P_D^n indicates a stronger capability for predicting bullish or bearish market.

(4) Analysis. Based on the statistical result, we analysis the pattern's predictive power.

4. Experiment and Result Analysis

4.1. Experiment Data and Method. As Yahoo provides the finance stock API used to download the transaction data of Chinese stock market, the stock transaction data of Chinese A-share market in any time can be acquired based on the API. To get a representative testing data, we select the K-line series data of Shanghai 180 index component stocks over the latest 10 years (from 2006-01-04 to 2016-08-24) as the test data. Limited by space, only the TIU and TID pattern's predictive power will be analyzed in the experiment. And the parameters of KNSSC algorithm are set as follows: $\theta = 0.75$, $w_S = 0.2$, $w_T = 0.8$, $w_{\text{Body}} = 0.6$, $w_{\text{US}} = 0.2$, $w_{\text{LS}} = 0.2$, and $w_S^t = w_P^t = 1/|KS^i|$ ($t = 1, 2, \ldots, |KS^i|$).

4.2. Experiment One. The aim of the first experiment is to analyze the TIU pattern's predictive power based on the method defined in Section 3. Firstly, based on the definition of TIU, 1516 TIU patterns are identified from the test data. Then we cluster these patterns using the KNSSC algorithm, and finally 554 clusters are obtained. We choose the top 20 clusters with the most elements to conduct statistical analysis, as shown in Table 1.

Table 1: The experiment result of TIU.

Q^m	$\lvert Q^m \rvert$	$P_U(C_1)$	$P_U(C_2)$	$P_U(C_3)$	$P_U(C_4)$	$P_U(C_5)$	$P_U(C_6)$	$P_U(C_7)$	$P_U(C_8)$	$P_U(C_9)$	$P_U(C_{10})$	P_U^{10}
Q^0	1516	0.493	0.486	0.518	0.495	0.475	0.488	0.481	0.479	0.484	0.479	**0.1**
Q^1	79	0.405	0.532	0.557	0.532	0.544	0.506	0.468	0.506	0.43	0.443	**0.6**
Q^2	42	0.405	0.429	0.5	0.452	0.405	0.452	0.429	0.476	0.524	0.476	**0.1**
Q^3	32	0.5	0.594	0.594	0.625	0.594	0.531	0.625	0.594	0.563	0.531	**0.9**
Q^4	25	0.6	0.6	0.52	0.48	0.48	0.4	0.48	0.48	0.6	0.64	**0.5**
Q^5	23	0.435	0.478	0.609	0.565	0.522	0.391	0.478	0.348	0.391	0.391	**0.3**
Q^6	22	0.455	0.636	0.545	0.5	0.545	0.545	0.591	0.591	0.636	0.636	**0.8**
Q^7	21	0.333	0.571	0.524	0.476	0.619	0.857	0.762	0.714	0.429	0.048	**0.6**
Q^8	21	0.524	0.476	0.524	0.429	0.476	0.476	0.429	0.476	0.429	0.429	**0.2**
Q^9	21	0.524	0.476	0.524	0.429	0.381	0.286	0.381	0.429	0.381	0.429	**0.2**
Q^{10}	19	0.105	0.368	0.421	0.421	0.368	0.421	0.474	0.474	0.579	0.421	**0.1**
Q^{11}	19	0.579	0.474	0.421	0.368	0.263	0.368	0.316	0.263	0.211	0.211	**0.1**
Q^{12}	17	0.588	0.294	0.471	0.353	0.412	0.471	0.471	0.412	0.471	0.412	**0.1**
Q^{13}	15	0.6	0.533	0.467	0.467	0.467	0.467	0.533	0.533	0.6	0.6	**0.6**
Q^{14}	15	0.533	0.6	0.667	0.6	0.533	0.6	0.6	0.667	0.6	0.6	**1**
Q^{15}	14	0.5	0.071	0.5	0.429	0.429	0.357	0.286	0.286	0.286	0.429	**0**
Q^{16}	14	0.643	0.571	0.714	0.714	0.714	0.5	0.643	0.571	0.571	0.714	**0.9**
Q^{17}	14	0.714	0.857	0.857	0.857	0.786	0.786	0.571	0.643	0.643	0.643	**1**
Q^{18}	14	0.357	0.571	0.429	0.5	0.357	0.571	0.429	0.429	0.5	0.5	**0.2**
Q^{19}	13	0.385	0.462	0.692	0.615	0.615	0.692	0.692	0.769	0.769	0.769	**0.8**
Q^{20}	12	0.5	0.667	0.667	0.667	0.75	0.583	0.583	0.667	0.583	0.583	**0.9**

In Table 1, Q^0 represents the cluster composed of 1516 TIU Patterns. Its P_U^{10} is only 0.5 which means that it may be a spurious pattern to predict bullish market. However, after further classifying the TIU patterns, we can see that (1) Q^3, Q^6, Q^{16}, and so forth have a strong capability for predicting bullish market, because their P_U^{10} both are above 0.8, (2) Q^1, Q^4, Q^7, and so forth have a moderate capability for predicting bullish market, as their P_U^{10} are only in 0.5~0.7, and (3) Q^2, Q^5, Q^8, and so forth have a weak capability for predicting bullish market, as their entireP_U^{10} are below 0.5. Particularly for Q^2, its P_U^{10} is only 0.1.

By comparing the predictive power of Q^3 and Q^2, as shown in Figure 5, we can see that the predictive result of Q^3 is bullish market while that of Q^2 is bearish market, which means that their predictive power is opposite. The result of experiment one shows that (1) the predictive power of TIU varies a great deal for different shapes and (2) to be a better pattern for predicting bullish market, the TIU pattern badly needs to be further classified, which are consistent with the expected analysis.

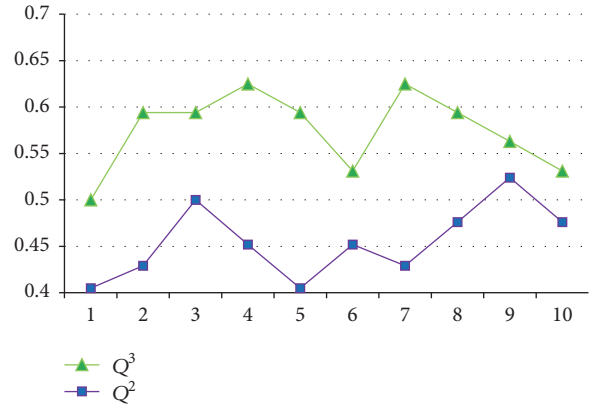


Figure 5: The comparison of predictive power between Q^3 and Q^2.

4.3. Experiment Two. The aim of the second experiment is to analyze the TID pattern's predictive power based on the method defined in Section 3. Firstly, based on the definition of TID, 1498 TID patterns are identified from the test data. Then we cluster these patterns using the KNSSC algorithm, and finally 572 clusters are obtained. We choose the top 20 clusters with the most elements to conduct statistical analysis, as shown in Table 2.

Similarly, the TID pattern may be also a spurious pattern to predict bearish market because P_D^{10} of Q^0 is 0, where Q^0 represents the cluster composed of 1498 TID patterns. Moreover, after further classifying the TID patterns, we can see that (1) except for Q^{11} and Q^{16}, almost all of the clusters have a weak capability for predicting bearish market, as their entire P_D^{10} are below 0.5 and (2) even Q^{11} and Q^{16} still not have a higher value of P_D^{10}, which are only 0.5. Therefore, we can

Table 2: The experiment result of TID.

Q^m	$\lvert Q^m \rvert$	$P_U(C_1)$	$P_U(C_2)$	$P_U(C_3)$	$P_U(C_4)$	$P_U(C_5)$	$P_U(C_6)$	$P_U(C_7)$	$P_U(C_8)$	$P_U(C_9)$	$P_U(C_{10})$	P_D^{10}
Q^0	1498	0.421	0.451	0.461	0.449	0.437	0.449	0.459	0.444	0.427	0.435	**0**
Q^1	78	0.449	0.5	0.436	0.436	0.372	0.359	0.449	0.397	0.359	0.346	**0.1**
Q^2	72	0.431	0.486	0.542	0.458	0.431	0.431	0.417	0.417	0.361	0.306	**0**
Q^3	44	0.364	0.364	0.364	0.386	0.273	0.386	0.364	0.386	0.386	0.341	**0.1**
Q^4	39	0.513	0.359	0.359	0.333	0.231	0.282	0.308	0.282	0.282	0.282	**0**
Q^5	23	0.174	0.261	0.304	0.391	0.348	0.391	0.391	0.435	0.435	0.391	**0.1**
Q^6	23	0.348	0.478	0.522	0.478	0.435	0.435	0.391	0.478	0.478	0.435	**0.4**
Q^7	21	0.429	0.429	0.476	0.429	0.619	0.524	0.524	0.476	0.476	0.571	**0**
Q^8	18	0.444	0.333	0.444	0.5	0.5	0.278	0.222	0.333	0.333	0.333	**0.2**
Q^9	18	0.333	0.667	0.611	0.389	0.444	0.444	0.389	0.389	0.389	0.389	**0.1**
Q^{10}	18	0.333	0.222	0.278	0.167	0.278	0.389	0.444	0.389	0.5	0.556	**0**
Q^{11}	18	0.444	0.444	0.278	0.333	0.389	0.222	0.278	0.333	0.444	0.389	**0.5**
Q^{12}	16	0.438	0.375	0.5	0.563	0.688	0.688	0.625	0.5	0.438	0.563	**0.2**
Q^{13}	15	0.267	0.267	0.467	0.4	0.267	0.6	0.533	0.467	0.4	0.4	**0**
Q^{14}	13	0.385	0.308	0.308	0.385	0.308	0.462	0.231	0.231	0.308	0.308	**0.4**
Q^{15}	12	0.25	0.333	0.583	0.667	0.667	0.583	0.5	0.5	0.417	0.5	**0.1**
Q^{16}	12	0.5	0.333	0.25	0.25	0.5	0.667	0.417	0.5	0.5	0.5	**0.5**
Q^{17}	12	0.25	0.833	0.75	0.667	0.583	0.5	0.5	0.417	0.417	0.583	**0.4**
Q^{18}	11	0.364	0.273	0.273	0.364	0.545	0.455	0.545	0.545	0.455	0.545	**0**
Q^{19}	11	0.364	0.455	0.273	0.364	0.364	0.364	0.455	0.455	0.455	0.364	**0**
Q^{20}	11	0.273	0.364	0.273	0.364	0.364	0.364	0.364	0.364	0.182	0.273	**0**

consider that the TID pattern is definitely a spurious pattern, which is also consistent with the expected analysis.

4.4. Experiment Conclusion. Through the above experiment, we can draw the following conclusion. (1) The predictive power of a pattern varies a great deal for different shapes. Take TIU; for example, some shapes' TIU patterns have a strong capability for predicting bullish market, while some others have the opposite predictive power. Therefore, to analyze the predictive power of a pattern, we should make a concrete analysis of concrete shapes. (2) There are definitely some spurious patterns in the existing K-line patterns. Therefore, in order to improve the stock prediction performance based on K-line patterns, we need to further classify the existing patterns based on the shape feature, identify all the spurious patterns, and choose the patterns having stronger predictive power to predict the stock price.

5. Conclusion

Stock prediction is a popular research field in the time series prediction. As a primary technology analysis method of stock prediction, there is different option on the stock price prediction based on K-line patterns in the academic world, though it is widely used in reality. To help resolve the debate, this paper uses the data mining method, like pattern recognition, similarity match, cluster and statistical analysis, and so forth, to study the predictive power of K-line patterns. Experimental results show that one reason for the debate is that the definition of K-line patterns is more open and lack of mathematical rigor. The other is that there are some spurious patterns in the existing K-line patterns. In addition, the method presented in the paper can be used not only to test the predictive power of patterns but also for K-line patterns mining and stock prediction. Therefore, the future works as follows. (1) It will be a necessary and significant task to identify the entire spurious pattern using the proposed method. (2) On the basis of the proposed method, we can research an automatic pattern mining method to discover more useful patterns for stock prediction.

Conflicts of Interest

The authors declare that there are no conflicts of interest regarding the publication of this paper.

Acknowledgments

The Key Basic Research Foundation of Shanghai Science and Technology Committee, China (Grant no. 14JC1402203), and the Science and Technology Support Program of China (Grant no. 2015BAF10B01) financially supported this work.

References

[1] J. Lin, S. Williamson, K. D. Borne et al., "Advances in machine learning & data mining for astronomy," *Pattern Recognition in Time Series*, pp. 617–645, 2010.

[2] L. Lihui, T. Xiang, Y. Haidong et al., "Financial time series forecasting based on SVR," *Computer Engineering and Applications*, vol. 41, no. 30, pp. 221–224, 2005.

[3] R. D. Edwards, J. Magee, and W. H. C. Bassetti, *Technical Analysis of Stock Trends*, CRC Press, Boca Raton, Fla, USA, 10th edition, 2012.

[4] S. Nison, *Japanese Candlestick Charting Techniques: A Contemporary Guide to the Ancient Investment*, Technique of the Far East, Institute of Finance, New York, NY, USA, 1991.

[5] B. R. Marshall, M. R. Young, and L. C. Rose, "Candlestick technical trading strategies: can they create value for investors?" *Journal of Banking and Finance*, vol. 30, no. 8, pp. 2303–2323, 2006.

[6] G. Caginalp and H. Laurent, "The predictive power of price patterns," *Applied Mathematical Finance*, vol. 5, no. 3-4, pp. 181–205, 1998.

[7] K. H. Lee and G. S. Jo, "Expert system for predicting stock market timing using a candlestick chart," *Expert Systems with Applications*, vol. 16, no. 4, pp. 357–364, 1999.

[8] M. M. Goswami, C. K. Bhensdadia, and A. P. Ganatra, "Candlestick analysis based short term prediction of stock price fluctuation usingSOM-CBR," in *Proceedings of the 2009 IEEE International Advance Computing Conference (IACC '09)*, pp. 1448–1452, March 2009.

[9] T. H. Lu and J. Chen, *Candlestick charting in European stock markets*, no. 2, pp. 20–25, 2013.

[10] T. H. Lu and Y. M. Shiu, "Tests for two day candlestick patterns in the emerging equity market of Taiwan," *Emerging Markets Finance & Trade*, vol. 48, no. 1, pp. 41–57, 2014.

[11] H. Li, W. W. Y. Ng, J. W. T. Lee, B. Sun, and D. S. Yeung, "Quantitative study on candlestick pattern for shenzhen stock market," in *Proceedings of the 2008 IEEE International Conference on Systems, Man and Cybernetics, (SMC '08)*, pp. 54–59, October 2008.

[12] W. Xiao, W. W. Y. Ng, M. Firth et al., "L-GEM based MCS aided candlestick pattern investment strategy in the shenzhen stock market," in *Proceedings of the 2009 International Conference on Machine Learning and Cybernetics*, pp. 243–248, July 2009.

[13] T. Kamo and C. Dagli, "Hybrid approach to the Japanese candlestick method for financial forecasting," *Expert Systems with Applications*, vol. 36, no. 3, pp. 5023–5030, 2009.

[14] M. Jasemi, A. M. Kimiagari, and A. Memariani, "A modern neural network model to do stock market timing on the basis of the ancient investment technique of Japanese Candlestick," *Expert Systems with Applications*, vol. 38, no. 4, pp. 3884–3890, 2011.

[15] S. Barak, J. H. Dahooie, and T. Tichý, "Wrapper ANFIS-ICA method to do stock market timing and feature selection on the basis of Japanese Candlestick," *Expert Systems with Applications*, vol. 42, no. 23, pp. 9221–9235, 2015.

[16] J. H. Fock, C. Klein, and B. Zwergel, "Performance of candlestick analysis on intraday futures data," *The Journal of Derivatives*, vol. 13, no. 1, pp. 28–40, 2005.

[17] B. R. Marshall, M. R. Young, and R. Cahan, "Are candlestick technical trading strategies profitable in the Japanese equity market?" *Review of Quantitative Finance and Accounting*, vol. 31, no. 2, pp. 191–207, 2008.

[18] M. J. Horton, "Stars, crows, and doji: the use of candlesticks in stock selection," *Quarterly Review of Economics and Finance*, vol. 49, no. 2, pp. 283–294, 2009.

[19] L. J. Chmielewski, M. Janowicz, and A. Orłowski, "Prediction of trend reversals in stock market by classification of Japanese candlesticks," in *Proceedings of the 9th International Conference on Computer Recognition Systems (CORES '15)*, vol. 403, pp. 641–647.

[20] C.-F. Tsai and Z.-Y. Quan, "Stock prediction by searching for similarities in candlestick charts," *ACM Transactions on Management Information Systems*, vol. 5, no. 2, Article ID 2591672, 2014.

[21] L. Chmielewski, M. Janowicz, J. Kaleta, and A. Orłowski, "Pattern recognition in the Japanese candlesticks," *Advances in Intelligent Systems and Computing*, vol. 342, pp. 227–234, 2015.

[22] G. L. Morris and R. Litchfield, *Candlestick Charting Explained: Timeless Techniques for Trading Stocks and Futures*, McGraw-Hill, New Yourk, NY, USA, 2nd edition, 2006.