Pattern Discovery of Fuzzy Time Series for Financial Prediction

Chiung-Hon Leon Lee, Alan Liu, *Member*, *IEEE*, and Wen-Sung Chen

Abstract—A fuzzy time series data representation method based on the Japanese candlestick theory is proposed and used in assisting financial prediction. The Japanese candlestick theory is an empirical model of investment decision. The theory assumes that the candlestick patterns reflect the psychology of the market, and the investors can make their investment decision based on the identified candlestick patterns. We model the imprecise and vague candlestick patterns with fuzzy linguistic variables and transfer the financial time series data to fuzzy candlestick patterns for pattern recognition. A fuzzy candlestick pattern can bridge the gap between the investors and the system designer because it is visual, computable, and modifiable. The investors are not only able to understand the prediction process, but also to improve the efficiency of prediction results. The proposed approach is applied to financial time series forecasting problem for demonstration. By the prototype system which has been established, the investment expertise can be stored in the knowledge base, and the fuzzy candlestick pattern can also be identified automatically from a large amount of the financial trading data.

Index Terms—Financial data processing, fuzzy sets, pattern recognition, time series.

1 Introduction

Financial predictions, such as predicting the variance of stock price or forecasting the fluctuation of future index, are one of the charming topics not only for research, but also for commercial applications. Oweing to its importance, a lot of concepts and techniques are established in the fundamental and technical analysis [1]. However, due to the numerical and continuous nature of stock price and future index, they require special preparation and transformation of data, which are, in many cases, critical for successfully applying the techniques of data mining, machine learning, or pattern recognition to financial prediction [2].

There are many existing tools trying to help the user predict the stock price or the future index. The approaches used in those tools include Artificial Neural Network (ANN) [3], [4], NeuroFuzzy [5], Genetic Algorithm (GA), Classification and Regression tree, Naive Bayes [6], Support Vector Machines (SVM) [7], and fuzzy time series [8], [9], [10], [11], [12], [13], [14], [15], [16], [17], etc. We argue that these tools cannot be used or trusted directly by the investors for several reasons, but we claim a good knowledge representation method, such as the one proposed in this paper, can assist the investors in their decision making process. First, training the prediction system is a nontrivial task, and the training results cannot be further used to other target even in the same domain. For example, the fuzzy time series approach proposed in [15] uses the data set of Taiwan Futures Exchange (TAIFEX) [18] to derive a forecasting model which cannot be used to forecast the Standard & Poor's index [19] or other stocks in the Taiwan stock market. The same problem occurs also in other methods [3], [4], [5], [6], [7].

Second, the prediction results are incomprehensible. The prediction decision process is a black box and unknown mechanism to the investors because not all the investors have the background knowledge about ANN, NeuroFuzzy, SVM, or fuzzy time series. However, in many situations, tuning the parameters of forecasting models dynamically is necessary for deriving better prediction results. Therefore, it is better for the investors to understand the forecasting models.

Third, there is a gap between prediction result and investment decision. The investor prefers to find the reversal patterns of the stock price rather than predict the stock price every day. In many cases [3], [4], [5], [6], [7], [8], the researchers focus their efforts on improving the forecasting results in every point to come, but the investors are usually only interested in when to buy or sell their stocks to make profits.

Motivated by the observations above, we propose a knowledge-based method which transfers the financial data to comprehensible rules and visual patterns for financial prediction. The pattern with domain-specific knowledge can be used to cope with the shortcoming of traditional approaches to financial prediction. Preparing domain-specific knowledge for data mining or machine learning models to improve forecasting results is not a new idea. In [15], Huarng proved that using domain-specific knowledge can improve the fuzzy time series forecasting results [8], [9], [10], [11]. There are several techniques which can be used to represent the domain-specific knowledge—fuzzy logic [20], rule-based system [21], or the case-based reasoning (CBR), etc. [22].

- *C.-H.L. Lee is with the Department of Electrical Engineering, National Chung Cheng University, Min-Hsiung, Chia-Yi, 621, Taiwan. E-mail: d8842019@ccu.edu.tw.*
- *A. Liu is with the Department of Electrical Engineering and the Center for Telecommunication Research, National Chung Cheng University, Min-Hsiung, Chia-Yi, 621, Taiwan. E-mail: aliu@ee.ccu.edu.tw.*
- *W.-S. Chen is with the Department of Electrical Engineering, National Cheng Kung University, Ta-Hsueh Road, Tainan 701, Taiwan. E-mail: wensung0302@giga.net.tw.*

Manuscript received 21 Nov. 2004; revised 21 June 2005; accepted 8 Sept. 2005; published online 17 Mar. 2006.
For information on obtaining reprints of this article, please send e-mail to: tkde@computer.org, and reference IEEECS Log Number TKDE-0471-1104.

One of the advantages of using domain-specific knowledge for data mining is that the system decision and suggestion are more comprehensible and interpretable. The system might not need any training data and training process, and the system efficiency depends on the quality of the acquired knowledge. By checking the rule or case, the user can understand why the system makes such decision or gives certain suggestions. The system efficiency and correctness can be enhanced by updating the rule or case bases. However, to construct an efficient knowledge base or case base is nontrivial. The difficulty of these approaches is mainly the problem in acquisition, representation, and validation of domain expertise.

How to represent the time series is one of the most important challenges in time series researches. A suitable choice of representation will greatly increase the easiness and efficiency of time series processes. In this paper, we use symbolic fuzzy linguistic variables to represent the numerical financial time series data for knowledge discovery because the numerical computation load can be greatly reduced [23], [24], [25] and linguistic variables could be comprehensible to the investors.

An old, viable, and effective technical analysis theory for stock and commodity market timing and analysis called the Japanese candlestick theory [26] is employed as background knowledge to assist the representation of financial time series data. The financial prediction system based on our approach will become more acceptable to the investors because the investors can participate in the forecasting process or share their investment knowledge with others.

The contribution of our approach is to propose a richer data representation method for the pattern discovery in financial time series application. The paper is organized as follows: In Section 2, the Japanese candlestick theory and related work of time series database are introduced. The fuzzy-based Japanese candlestick pattern representation method is proposed in Section 3. Section 4 describes the proposed method and introduces how to apply our approach for investment decision making. A system prototype to implement the proposed method is depicted in Section 5. Finally, Section 6 provides the conclusion of this paper.

2 Background Knowledge and Related Work

In this section, the Japanese candlestick theory is introduced, and the research of symbolic time series [23], [24] and fuzzy time series [8], [9], [10], [11], [12], [13], [14], [15], [16], [17] is also discussed for comparing with our approach.

2.1 Japanese Candlestick Theory

Rich information exists in the financial time series database, but most of the traditional approaches only scratch the surface of the wealth of knowledge buried in the data. For example, many financial time series prediction approaches only use daily closing price as raw data to construct the forecasting model [3], [4], [5], [6], [7], [8].

Fig. 1 shows three general ways to represent the stock trading prices during a trading time period. The original

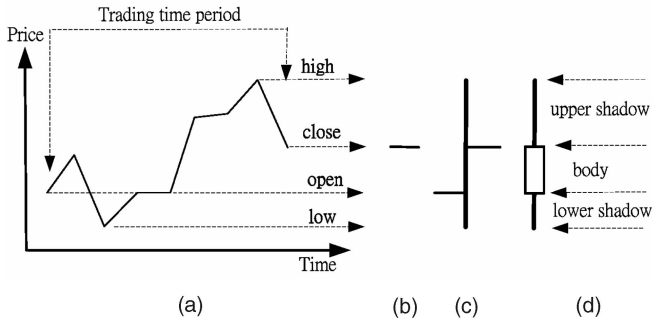

Fig. 1. Three ways to represent the stock price fluctuation.

stock price fluctuation is represented in Fig. 1a. The first trading price during the trading time period is called open price, the last trading price is called close price, the highest price is called high price, and the lowest price is called low price.

Fig. 1b indicates a single closing price. Fig. 1c represents the bar line with richer information than Fig. 1b. The data required to produce a standard bar chart consists of the open, high, low, and close prices for the time period under study. A bar chart consists of vertical lines representing the high to low prices in the trading time period. The open price is represented by the small tick mark extending from the bar out to the left and the close price is represented by another tick on the right side.

Fig. 1d illustrates the candlestick line which is similar to the bar line but uses a box to make up the difference between the open and close price. The box is called the body of a candlestick line. The height of the body is the range between a trading day's open price and the day's close price. When the body is black, it means that the closing price was lower than the opening price. When the closing price is higher than the opening, the body is white. The candlestick line may have small thin lines above and below the body. These lines are called shadows and represent the highest and lowest prices reached during the trading time period. The height of the upper shadow is the range between the high price and the higher price among the day's open and close prices. The height of the lower shadow is the range between the low price and the lower price among the day's open and close prices. Compared with the bar chart, the candlestick chart provides more visual information for the investor to identify specific patterns from the financial time series.

The Candlestick theory assumes that the candlestick patterns reflect the psychology of market and an experienced investor can make the investment decision by the observation of the candlestick chart. In other words, this theory provides a prediction mechanism to the investors who prefer to pattern recognition rather than real number calculation.

Fig. 2 shows an example of the daily candlestick chart for the stock market. Daily open, close, high, and low prices are recorded in the candlestick lines from d1 to d10. On day d3, the price closes at a highest price and still keeps the uptrend from d1 to d2. On day d4, the opening price is higher than previous closing price, but the price closes at the lowest price and leaves a long upper shadow. This situation might

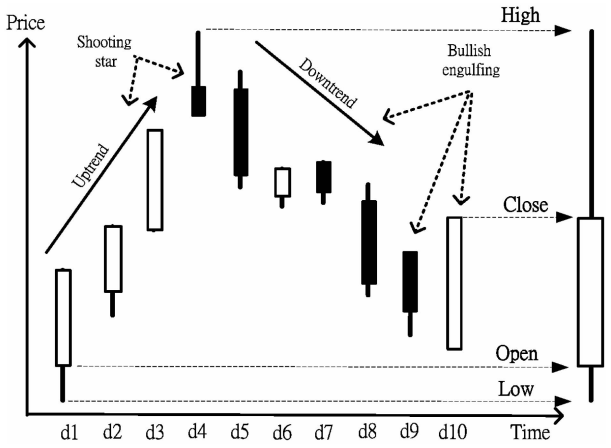

Fig. 2. An example of the candlestick chart.

be interpreted by an experienced investor as an uptrend of the stock price on the days from d1 to d3 because there are many investors who want to buy the stock, making the closing price much higher than the opening price. However, the uptrend might reverse itself on day d4 because there might be investors wanting to sell the stock in the trading period that makes the price close at the lowest price and leaves a long upper shadow. In other words, the candlestick lines at d3 and d4 can be interpreted as the uptrend being returned. At d9, the closing price is lower than the opening price, and at d10, the opening price is much lower than the previous closing price, but closes at the highest price and even higher than the close price on previous day. The lines at d9 and d10 can represent a bouncing back because the downtrend stops at d10.

A candlestick pattern is composed of one or more candlestick lines and the trend before the pattern. By the trading experience, the investor tries to identify the candlestick patterns to help themselves make the investment decisions such as to buy, sell, or hold the stock. There are many existing candlestick patterns which are widely used by the investors [26]. For example, the candlestick line at d4 and the trend formed by d1, d2, and d3 are defined as a pattern which is called Shooting Star to represent the uptrend being returned. Another pattern, called Bullish Engulfing, is also illustrated in Fig. 2 and is composed of a downtrend and the candlestick lines at d9 and d10.

To identify the candlestick patterns from the trading data is simple to the human investors, but identifying the patterns from a large amount of trading data is time consuming. For example, there are more than 1,000 stocks in the Taiwan stock market and most of the stocks contain more than 1,000 historical daily trading data. It takes a lot of time if an investor wants to find which stock has the Bullish Engulfing pattern in a specific day or how many times the Bullish Engulfing pattern appears in the historical data of a specific stock. Moreover, the investors might have different interpretations of the candlestick patterns. Because of the vagueness and impreciseness, effective usage of the candlestick patterns requires many years of investment experiences. For solving the problems above, in our approach, the fuzzy set theory [20] is used to represent the candlestick patterns. The represented candlestick pattern is called the fuzzy candlestick pattern and is described in Section 3.

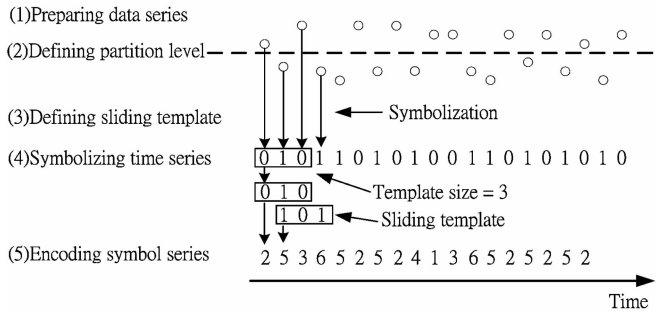

Fig. 3. Processes of symbolizing a time series.

Since the candlestick pattern is represented in computable linguistic variables, finding the candlestick patterns becomes automatic and the query time of candlestick patterns can be reduced.

2.2 Symbolic Time Series

There are many approaches which have been proposed in a large number of papers to index, cluster, classify, and segment the time series. Keogh and Kasetty reimplemented more than two dozens of these papers and depicted the result in [27], and they claimed that it was reasonable if the utility of the approach was only being claimed for a single type of data. We believe that there is no silver bullet for every problem of time series analysis, but carefully preprocessing the data set will reduce the load of numerical computations.

There are great numbers of time series representation methods which have been introduced. Lin and her colleagues survey time series representation approach in [24] and introduced a symbolic representation method for time series to reduce the complexity of time series processes. In [25], the authors also proposed a method to transfer the time series data into symbolic representation and to find specified patterns in time series database.

The advantages of symbolic analysis method for time series are better numerical computation efficiency, less sensitivity to measurement noise, and reduced instrumentation cost and complexity [23]. Although symbolic-based time series analysis is widely used in many areas, the symbols used in many applications are meaningless. The most selected symbols are digits and alphabets, and it is incomprehensible to the users except the system designers. Fig. 3 shows a process for a time series which has been initially converted into a binary symbol values occurring at each point in time, and we can see that the encoded symbol series is difficult to interpret. In this paper, we replace the meaningless symbols with fuzzy linguistic variables to make the encoded symbol series more comprehensible.

2.3 Fuzzy Time Series

The concept of fuzzy time series is proposed by Song and Chissom in [8], [9] and applied in the university enrollment forecasting. We give a brief review of fuzzy time series.

Assume that U is the universe of discourse, where $U = \{x_1, x_2, \ldots, x_n\}$. A fuzzy set A_i of U is defined by

$$A_i = \mu_{A_i}(x_1)/x_1 + \mu_{A_i}(x_2)/x_2 + \ldots + \mu_{A_i}(x_n)/x_n,$$

where μ_{A_i} is the membership function of the fuzzy set A_i, $\mu_{A_i} : U \rightarrow [0, 1]$. x_k is an element of fuzzy set A_i, and $\mu_{A_i}(x_k)$

is the degree of membership of x_k to A_i. $\mu_{A_i}(x_k) \in [0,1]$ and $1 \le k \le n$. Song and Chissom defined the fuzzy time series in [8], [9], [10].

For getting better execution performance and forecasting result, in [11], Chen proposed an arithmetic approach to improve Song's model. Based on Chen's model, there are more methods proposed [12], [13], [14], [15], [16], [17]. Those methods use richer data representation to improve the forecasting results. For example, in [13], Chen and Hwang proposed two-factors fuzzy time series to obtain better forecasting results in temperature prediction.

In many applications, a fuzzy time series $F(t)$ is not only related with $F(t-1)$ but is also caused by $F(t-2), F(t-3), \ldots$ and $F(t-m)$ $(\mathrm{m} = 1, 2, \ldots)$. In this situation, $F(t)$ is called high-order fuzzy time series. The definition of high-order fuzzy time series can be found in [14], [16].

Fuzzy time series provides a good approach to deal with the time series with linguistic variables. However, in order to derive better forecasting results, more factors are needed to fuzzy time series, and considering the causal relationship of fuzzy time series, high-order time series also have to be employed. Both multifactors and high-order of fuzzy time series will cause the computation complexity, and it is difficult to find a procedure to deal with the multifactors high-order fuzzy time series.

We believe that to prepare the fuzzy time series data with domain knowledge is a good way for knowledge discovery. We employ the concept of the Japanese candlestick theory to represent the fuzzy time series data for getting better forecasting results in an acceptable execution performance.

3 Modeling the Candlestick Pattern

Considering the properties of the candlestick chart, a method based on fuzzy logic is proposed to model the candlestick patterns, and we call such a pattern described by fuzzy linguistic variables a fuzzy candlestick pattern. A fuzzy candlestick pattern is composed of related fuzzy candlestick lines in a time period. Each fuzzy candlestick line can be divided into seven parts: sequence, open style, close style, upper shadow, body, body color, and lower shadow.

From the sequence, the location of the candlestick line can be identified. The open style and the close style are used to model the relationship between two continuous candlestick lines, and the rest are used to model the shape of a candlestick line.

3.1 Modeling the Candlestick Line

In a candlestick chart, the lengths of the shadow and the body play an important role to identify a candlestick pattern and to determine the efficiency of the candlestick pattern. The description of a candlestick line is imprecise and vague like long, middle, or short. There is no crisp value to define the length of body and shadow.

In our approach, four fuzzy linguistic variables EQUAL, SHORT, MIDDLE, and LONG are defined to indicate fuzzy sets of the shadows and the body length. Fig. 4 shows the membership function $\mu(x)$ of the linguistic variables. The ranges of body and shadow length are set to (0, 14) to represent the percentage of the fluctuation of stock price because the varying percentages of the stock prices are limited to 14 percent in the Taiwanese stock market, for example. It should be noted that although we limit the fluctuation of body and shadow length to 14 percent, in other applications, the designer can change the range of the fluctuation length to any number.

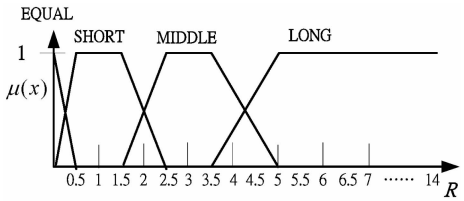

Fig. 4. The membership function of the length of the body and shadow.

In Fig. 4, the footnote of X axis indicates the real length of body or shadow, and the unit of X axis is percentage. The crisp input value of membership function can be calculated by the following equations:

$$\begin{aligned} L_{upper} &= ([high - \max(open, close)]/open) \times 100 \\ L_{lower} &= ([\min(open, close) - low]/open) \times 100 \\ L_{body} &= ([\max(open, close) - \min(open, close)]/open) \times 100. \end{aligned} \tag{1}$$

The character "L" of the equation indicates the length of the upper shadow, lower shadow, or body. The terms with subscripts like "open," "close," "high," and "low" are the prices in an interested time period. The LONG fuzzy set is defined by the following left linear membership function. The parameters (a, b) are equal to (3.5, 5).

$$left\_linear(x : a, b) = \begin{cases} 0 & x < a \\ (x-a)/(b-a) & a \le x \le b \\ 1 & x > b. \end{cases} \tag{2}$$

The membership function of SHORT and MIDDLE is a trapezoid function and the following equation is used.

$$trapezoid \quad (x : a, b, c, d) = \begin{cases} 0 & x < a \\ (x-a)/(b-a) & a \le x < b \\ 1 & b \le x < c \\ (d-x)/(d-c) & c \le x < d \\ 0 & x \ge d. \end{cases} \tag{3}$$

Four parameters (a, b, c, d) of this function to describe the linguistic variables SHORT and MIDDLE are (0, 0.5, 1.5, 2.5) and (1.5, 2.5, 3.5, 5). A right linear membership function is used to model the EQUAL fuzzy set and is defined by the following formula. The parameters (a, b) are equal to (0, 0.5) in this paper.

$$right\_linear(x : a, b) = \begin{cases} 1 & x < a \\ (b-x)/(b-a) & a \le x \le b \\ 0 & x > b. \end{cases} \tag{4}$$

The body color is also an import feature of a candlestick line and can be simply defined by three terms BLACK, WHITE, and CROSS. The situation where open price equals close price has a specific meaning in the candlestick pattern,

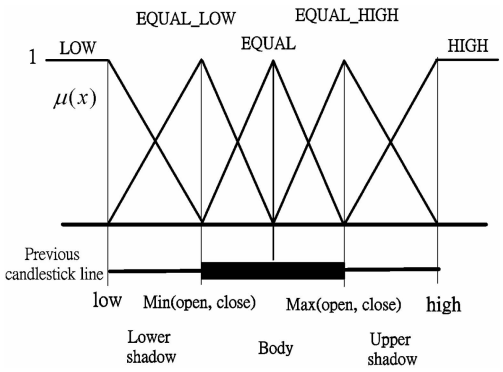

Fig. 5. The membership function of the open and close styles.

so a "CROSS" term is defined to describe this situation. In this case, the height of the body is 0, and the shape is represented with a horizontal bar. The definition of body color is defined as follows:

$$\begin{aligned}&\text{If open-close} > 0 \text{ then the body color is BLACK.}\\&\text{If open-close} < 0 \text{ then the body color is WHITE.}\\&\text{If open-close} = 0 \text{ then the body color is CROSS.}\end{aligned} \quad (5)$$

3.2 Modeling the Candlestick Lines Relationship

Defining the length of body and shadow of a candlestick line is not enough to model the candlestick line. Modeling the length of the body and shadow only determines the shape of a candlestick line. The relationship between a candlestick line and previous candlestick line should be defined to place a candlestick line in a correct position.

Two features are defined to model the candlestick line relationships: the open style and the close style. The related positions of the open and close price to the previous candlestick line are used to model the open style and the close style. Fig. 5 shows the membership function of the linguistic variables of the open style and close style. The candlestick line in the bottom of Fig. 5 is the candlestick line of previous trading day. The unit of X axis is the trading prices of the previous day and the unit of Y axis is the possibility value of the membership function.

Five linguistic variables are defined to represent the open style relationships: OPEN LOW, OPEN EQUAL_LOW, OPEN EQUAL, OPEN EQUAL_HIGH, and OPEN HIGH, and five linguistic variables are defined to represent the close style relationships: CLOSE LOW, CLOSE EQUAL_LOW, CLOSE EQUAL, CLOSE EQUAL_HIGH, and CLOSE HIGH.

The function used to represent OPEN LOW and CLOSE LOW is the right linear function in (4), and the left linear function in (2) is also used to represent the fuzzy sets of OPEN HIGH and CLOSE HIGH. The other fuzzy sets are described by the triangle function in (6).

$$triangle(x:a,b,c) = \begin{cases} 0 & x < a \\ (x-a)/(b-a) & a \le x \le b \\ (c-x)/(c-b) & b < x \le c \\ 0 & x > c. \end{cases} \quad (6)$$

The parameters for the linguistic variables of open and close style are determined by the prices of the previous candlestick line. For example, if the open price in a interested time period is equal to the price of $\mathbf{min}(open, close)$, then the open style is OPEN EQUAL_ LOW and if the close price is equal to the price $\mathbf{max}(open, close)$, then the close style is CLOSE EQUAL_HIGH.

Combining the description of the candlestick line and the relationship between candlestick lines, a candlestick line can be defined completely. The length of body and shadow model the shape of a candlestick line, and the open style and close style model the related position and causal relationship between the continual candlestick lines. Candlestick lines are the basic elements of a candlestick pattern.

3.3 Fuzzy Modifier

The fuzzy modifiers are used to further enhance the flexibility of the linguistic variables in fuzzy candlestick patterns. Using the fuzzy modifier, the system can provide a richer way to help the users define their own fuzzy candlestick patterns or modify the candlestick patterns that are learned by the system.

Modifiers used in phrases such as *VERY LONG* or *ABOVE MIDDLE* change the shape of a fuzzy set in a way that suits the meaning of the word used. For convenience, the modifiers defined in NRC (National Research Council of Canada) Fuzzy Toolkit [28] are used to help the implementation of the system prototype.

Fuzzy modifiers used to modify the fuzzy variables of a candlestick line are ABOVE, BELOW, PLUS, VERY, EXTREMELY, MORE_OR_LESS, SOMEWHAT, and NOT. Assume that x belongs to the fuzzy set of linguistic variables and y indicates the value of the membership function $\mu(x)$. The ABOVE modifier identifies the first x value at which the maximum value is reached. All membership values below this point are set to zero and all membership values above this value are set to $1 - \mathrm{y}$, and the BELOW modifier identifies the first x value at which the maximum value is reached. All membership values above this point are set to zero and all membership values below this value are set to $1 - \mathrm{y}$.

The PLUS, VERY, EXTREMELY, MORE_OR_LESS, and SOMEWHAT modifiers return the expanded fuzzy value passed as its argument. The PLUS modifier raises all the membership values of the fuzzy value by a factor of 1.25, the VERY modifier a factor of 2, the EXTREMELY modifier a factor of 3, the SOMEWHAT modifier a factor of 1/2, and the MORE_OR_LESS modifier a factor of 1/3. The NOT modifier returns the complement of the fuzzy value passed as its argument: $\mathrm{y(x)} = 1 - \mathrm{y(x)}$.

3.4 Modeling the Trends

We define two linguistic variables to model the midterm variation of the trends before and after the candlestick pattern: previous trends and following trends. The weekly candlestick line is used to represent the previous trend of the daily candlestick pattern. Because there are five trading days in the Taiwan stock market every week, the time period of a weekly candlestick line is set to five trading days. Six fuzzy linguistic variables are used to define the previous trend: CROSS, EQUAL, WEAK, NORMAL, STRONG, and EXTREME. The EQUAL fuzzy set is defined by the right linear membership function, EXTREME is defined by left linear membership function, and WEAK,

NORMAL, and STRONG are defined by the trapezoid function.

The parameters of these linguistic variables are (0), (0, 2), (0, 2, 4, 6), (4, 6, 8, 10), (8, 10, 12, 14), and (12, 14). The definition of the body color to reflect the trend is defined as follows, and the variable BEARISH or BULLISH means that the variation of stock price is in downtrend or uptrend.

$$\begin{aligned}&\text{If open-close} > 0 \text{ then the body color is BEARISH.}\\&\text{If open-close} = 0 \text{ then the variable is CROSS.}\\&\text{If open-close} < 0 \text{ then the body color is BULLISH.}\end{aligned} \tag{7}$$

The following trend is derived from the variation of the close price. The variation is calculated by the following equation:

$$\frac{Close_{t+n} - Close_t}{Close_t} \times 100. \tag{8}$$

The terms $Close_{t+n}$ and $Close_t$ mean the close price on day t and n trading days after day t. The unit of the following trend is percentage. The parameter n could be defined by the user. The process to determine the fuzzy sets of the following trend is similar to the fuzzy time series proposed in [11] and is demonstrated in next section.

4 Candlestick Pattern for Financial Time Series Prediction

There are three fundamental problems in pattern recognition [20]. First, the *sensing problem* concerns the acquisition of measured values from recognized objects. Second, the *feature extraction problem* concerns the extraction of characteristic features from the input data. Third, the *patterns classification problem* concerns the determination of optimal decision procedures for the classification of given patterns. In the fuzzy candlestick pattern approach, the measured values are the open, close, high, and low price of trading targets in a specific time period. The features of the trading target price fluctuation are represented by fuzzy candlestick pattern. The classification rules of fuzzy candlestick patterns can be determined by the investors or the computer system.

This section starts with a general description of the candlestick pattern forecasting algorithm (CPFA) and then the details of two financial time series examples: the Taiwan Stock Exchange Capitalization Weighted Stock Index (TAIEX) and the stock of Taiwan Semiconductor Manufacturing Company (2330TSMC) are explained. Because the historical enrollments of the University of Alabama are widely used to illustrate the fuzzy time series forecasting process [9], [10], [11], [12], [15], [16], [17], we also use the data set as an example to show comparative results for the performance evaluation.

4.1 Forecasting Procedure

Using candlestick pattern approach for financial time series prediction consists of the following steps:

Step 1. Preparing the data for candlestick pattern generation. In this step, open, close, high, and low prices should be prepared. The variation percentage between two close prices on time t and time $t + n$ should be calculated by using (8) for deriving the following trend. Based on the variation derived from (8), we can find the minimum increase I_{min} and maximum increase I_{max}. Then, we can define the universe of discourse UoD as $UoD = [I_{min} - D_1, I_{max} + D_2]$, where D_1 and D_2 are suitable positive numbers. For example, if $I_{min} = -5.83$ and $I_{max} = 7.66$. We can set $D_1 = 0.17$ and $D_2 = 0.34$, so UoD can be represented as $UoD = [-6, 8]$.

Step 2. Partitioning the universe of discourse UoD into several intervals $\mathrm{u}_1, \mathrm{u}_2, \ldots, \mathrm{u}_\mathrm{m}$. Taking the UoD in Step 1 as an example, we can partition the UoD into seven intervals, where $\mathrm{u}_1 = [-6, -4], \mathrm{u}_2 = [-4, -2], \ldots, \mathrm{u}_7 = [6, 8]$.

Step 3. Defining fuzzy sets on the UoD. This step determines the linguistic variables represented by fuzzy sets to describe the degree of variation between data of time t and time $t + n$. If we set the number of fuzzy sets to 7, the seven fuzzy sets considered could be:

$$\begin{aligned}A_1 &= (\text{LARGE\_DECREASE}),\\A_2 &= (\text{NORMAL\_DECREASE}),\\A_3 &= (\text{SMALL\_DECREASE}),\\A_4 &= (\text{SMALL\_INCREASE}),\\A_5 &= (\text{NORMAL\_INCREASE}),\\A_6 &= (\text{LARGE\_INCREASE}), \text{ and}\\A_7 &= (\text{EXTREME\_INCREASE}).\end{aligned}$$

Then, the fuzzy sets A_1 to A_7 on the UoD will be as follows:

$$\begin{aligned}A_1 &= 1/u_1 + 0.5/u_2 + 0/u_3 + 0/u_4 + 0/u_5 + 0/u_6 + 0/u_7,\\A_2 &= 0.5/u_1 + 1/u_2 + 0.5/u_3 + 0/u_4 + 0/u_5 + 0/u_6 + 0/u_7,\\A_3 &= 0/u_1 + 0.5/u_2 + 1/u_3 + 0.5/u_4 + 0/u_5 + 0/u_6 + 0/u_7,\\A_4 &= 0/u_1 + 0/u_2 + 0.5/u_3 + 1/u_4 + 0.5/u_5 + 0/u_6 + 0/u_7,\\A_5 &= 0/u_1 + 0/u_2 + 0/u_3 + 0.5/u_4 + 1/u_5 + 0.5/u_6 + 0/u_7,\\A_6 &= 0/u_1 + 0/u_2 + 0/u_3 + 0/u_4 + 0.5/u_5 + 1/u_6 + 0.5/u_7,\\A_7 &= 0/u_1 + 0/u_2 + 0/u_3 + 0/u_4 + 0/u_5 + 0.5/u_6 + 1/u_7.\end{aligned} \tag{9}$$

Step 4. Fuzzifying the values of variation derived in Step 1. The values of variation will be fuzzified in this step. If the number of variation on $\mathrm{time_i}$ is v, where $v \in u_x$, and if there is a value represented by fuzzy set A_y in which the maximum membership value occurs at u_x, then v is translated to A_y.

Step 5. Calculating the candlestick patterns. In this step, we will use the definitions of the candlestick pattern proposed in Section 3 to calculate the parameters of candlestick patterns. The data of open, close, high, and low prices will be transformed into candlestick lines.

Step 6. Refining the extracted patterns. Since the raw data have been transformed into candlestick patterns, we want to find more important attributes to affect the following trend of the candlestick pattern. This is a classification problem. Because the data is represented by symbolic patterns, based on the following trend, the ID3 classification algorithm [29] can be used to classify the extracted fuzzy candlestick patterns. The reasons for using the ID3 algorithm are because it is a method for approximating discrete-valued functions, robust to noisy data, and capable of learning disjunctive expressions. We use the algorithm to filter the attributes which are less important to the following trend.

Step 7. Selecting the patterns for forecasting. In this step, we are interested in when a pattern appears and what probability the real following trend falls into the predicted fuzzy sets. This can be described by the Bay's rule as follows:

$$p[A_x/p_y] = \frac{p[p_y/A_x]p[A_x]}{p[p_y/A_x]p[A_x] + p[p_y/A_{\overline{x}}]p[A_{\overline{x}}]}, \quad (10)$$

where p is the probability of events, A_x is the event of the following trend that falls into a fuzzy set of specific following trend, p_y is the event of a specific pattern which is identified, and $A_{\overline{x}}$ is the event of the following trend not falling into the fuzzy set of specific following trend. The prior probabilities $p[A_x]$ can be derived by previous experiments and $[A_{\overline{x}}] = 1 - p[A_x]$. The posterior probability $p[A_x/p_y]$ can be estimated more precisely by acquiring more experiments data sets. In this paper, for simplifying the problem, we use the original data set as a sample to estimate $p[A_x/p_y]$. The equation to select the pattern can be reduced as follows:

$$T = \frac{count(p_y \cap A_x)}{count(p_y \cap A_x) + count(p_y \cap A_{\overline{x}})} = \frac{count(p_y \cap A_x)}{count(p_y)}, \quad (11)$$

where T represents a threshold to select the pattern, the term *count* indicates a function to count the appearing times of refined pattern p_y, p_y with following trend A_x, and p_y with following trend $A_{\overline{x}}$.

For example, assume that the selection rule is "if $T > 0.5$, then selects the pattern." We want to calculate the T of a refined pattern p_a to determine whether the pattern should be selected. The following trend of p_a is A_4. By looking up all of the extracted patterns in Step 5, assume that there are three patterns p_x, p_y, and p_z retrieved by p_a. The following trend of p_x and p_y are A_4, and the following trend of p_z is A_2, then we can derive T as follows:

$$T = \frac{2}{1+2} = 0.67 > 0.5.$$

The pattern p_a will be selected for forecasting.

Step 8. Forecasting the trend to follow. We use the process introduced in Step 5 to transform the testing data set into testing candlestick patterns, match the test patterns with the selected patterns, and apply the following rules to forecast the following trend of the test patterns.

Rule 1. If the test pattern cannot be found by matching the selected patterns, then set the variation of following trend to 0.

Rule 2. If there is exactly one selected pattern matched with the test pattern and the following trend of selected pattern is A_k, use the midpoint m_k of the fuzzy set A_k as the following variation of the test pattern.

Rule 3. If there are more than one selected patterns matched with the test pattern, use the arithmetic average of midpoints as the following variation of the test pattern.

$$v_{following} = \frac{\sum_{i=1}^{k} m_i}{k}. \quad (12)$$

TABLE 1
TAIEX Data, Data Variation, and Fuzzy Sets

Years	Open	Close	High	Low	One day variations	Fuzzy variations A_i
2004/1/2	5907.15	6041.56	6043.3	5907.15	1.39	A_7
2004/1/5	6080.18	6125.42	6137.91	6061.86	0.30	A_5
2004/1/6	6170.02	6144.01	6170.02	6110.69	-0.04	A_4
2004/1/7	6174.95	6141.25	6215.45	6130.35	0.45	A_5
2004/1/8	6180.36	6169.17	6189.6	6142.24	0.94	A_6
2004/1/9	6241.36	6226.98	6257.89	6207.69	-0.12	A_4
2004/1/12	6225.23	6219.71	6246.62	6196.41	-0.15	A_4
2004/1/13	6239.16	6210.22	6255.94	6195.64	1.04	A_7
2004/1/14	6195.09	6274.97	6298.72	6190.35	-0.17	A_4
2004/1/15	6297.59	6264.37	6306.95	6253.64	0.09	A_5
2004/1/16	6300.76	6269.71	6311.46	6264.04	1.83	A_8
…	…	…	…	…	…	…
2004/2/12	6505.53	6436.95	6578.82	6397.34	1.74	A_8
2004/2/13	6467.15	6549.18	6554.78	6459.14	0.25	A_5
2004/2/16	6566.76	6565.37	6590.25	6518.1	0.53	A_6
2004/2/17	6586.49	6600.47	6616.04	6576.22	0.08	A_5
2004/2/18	6656.1	6605.85	6661.13	6586.74	1.15	A_7
2004/2/19	6647.86	6681.52	6688.14	6626.47	-0.24	A_4
2004/2/20	6692.03	6665.54	6708.04	6656.06	0.01	A_5
2004/2/23	6686.67	6665.89	6695.45	6607.36	-1.15	A_2
2004/2/24	6650.16	6589.23	6686.85	6589.23	0.84	A_6
2004/2/25	6619.69	6644.28	6672.24	6595.24	0.74	A_6
2004/2/26	6709.43	6693.25	6730.96	6660.52	0.86	A_6
2004/2/27	6732.63	6750.54	6756.76	6713.67		

Finally, the forecasting data can be derived as follows:

$$Forecast = close + close \times v_{following}. \quad (13)$$

Step 9. Evaluating the forecasting results. In order to evaluate the forecasting results and compare them with other works, the forecasting results are evaluated by the mean square error (MSE) defined as follows:

$$MSE = \frac{\sum_{i=1}^{n} (Forecasted_i - Actual_i)^2}{n}. \quad (14)$$

In (14), $Forecasted_i$ is the forecasted value at time i and $Actual_i$ is actual value at time i.

4.2 TAIEX Forecasting

For demonstrating the effectiveness of the fuzzy candlestick pattern approach, we use daily TAIEX open, close, high, and low process covering the period from 2004-01-02 to 2005-01-31. We choose 2-month data as training data and the following one month as testing data. For example, the data from January 2004 to February 2004 are chosen as the first training data set, and the data of 2004 March is chosen as the testing data set. We take the data from 2004-01-02 to 2004-2-27 as examples to explain the CPFA and demonstrate the experiment results to compare with other forecasting methods. The process is described as follows.

After performing steps 1 to 4 of CPFA, the activities in TAIEX from 2004-01-02 to 2004-2-27, related data variation, and related fuzzy sets are shown in Table 1. In this example, the variations interval of TAIEX is one day. By (8) and the training data, we can find that $I_{min} = -1.15$ and $I_{max} = 1.83$. For example, the close prices of 2004-01-02 and 2004-01-05 are 6041.56 and 6125.42; the variation is calculated by performing (8):

$$\frac{Close_{2004\text{-}01\text{-}05} - Close_{2004\text{-}01\text{-}02}}{Close_{2004\text{-}01\text{-}02}} \times 100 = \frac{6125.42 - 6041.56}{6041.56} \times 100 = 1.39.$$

We set $D_1 = 0.85$ and $D_2 = 0.17$, so the UoD can be represented as $UoD = [-2, 2]$. We partition the UoD into eight intervals, where

$$u_1 = [-2, -1.5], u_2 = [-1.5, -1], \ldots, u_8 = [1.5, 2].$$

The eight fuzzy sets considered are:

$$A_1 = (\text{EXTREME\_DECREASE}),$$
$$A_2 = (\text{LARGE\_DECREASE}),$$
$$A_3 = (\text{NORMAL\_DECREASE}),$$
$$A_4 = (\text{SMALL\_DECREASE}),$$
$$A_5 = (\text{SMALL\_INCREASE}), \ldots, \text{ and}$$
$$A_8 = (\text{EXTREME\_INCREASE}).$$

The definition of fuzzy sets A_1 to A_8 on the UoD is similar to (9).

The fuzzy candlestick pattern can be derived by performing step 5 of CPFA. We take the candlestick pattern on 2004-01-12 as an example to explain the creation of candlestick patterns. In this example, the number of candlestick lines in each candlestick pattern is 2. The calculation of the previous trend is introduced in Section 3.4. In 2004-01-02 and 2004-01-12, the close prices are 6219.71 and 6041.56. The variation of previous trend on 2004-01-12 $v_{2004\text{-}01\text{-}12}$ is calculated as follows:

$$v_{2004\text{-}01\text{-}12} = \frac{close_{2004\text{-}01\text{-}12} - close_{2004\text{-}01\text{-}02}}{close_{2004\text{-}01\text{-}02}} \times 100 = \frac{6219.71 - 6041.56}{6041.56} \times 100 = 2.95.$$

We find that the membership function of fuzzy set WEAK (0, 2, 4, 6) has the maximum membership value, then $v_{2004\text{-}01\text{-}12}$ belong to WEAK, and because the close price of 2004-01-12 is higher than the close price of 2004-01-02, by (7), the previous trend is translated to WEAK_BULLISH.

The open price of 2004-01-12 is 6225.23 which are within the low price and the middle of body on 2004-01-09, by the definition of the open style, the fuzzy set OPEN_EQUAL_LOW has the maximum membership value, and then the open style of 2004-01-12 is translated to OPEN_EQUAL_LOW. The close price of 2004-01-12 is 6219.71 which are similar to the open price within the low price and the middle of body on 2004-01-09, and then the close style of 2004-01-12 will be interpreted as CLOSE_EQUAL_LOW. The length of upper shadow of candlestick line on 2004-01-12 can be calculated as follows:

$$L_{upper_{(2004\text{-}01\text{-}12)}} = \frac{high_{2004\text{-}01\text{-}12} - \max(open_{2004\text{-}01\text{-}12}, close_{2004\text{-}01\text{-}12})}{open_{2004\text{-}01\text{-}12}} \times 100 = \frac{6246.62 - 6225.23}{6225.23} \times 100 = 0.34.$$

Because $L_{upper_{(2004\text{-}01\text{-}12)}}$ equals 0.34, by the definition of the upper shadow, the fuzzy set SHORT has the maximum membership value, and then the upper shadow of 2004-01-12 is translated to SHORT. The lower shadow is treated in similar way and translated to SHORT. The length of body can be calculated as follows:

$$L_{body_{(2004\text{-}01\text{-}12)}} = ((\max(open_{2004\text{-}01\text{-}12}, close_{2004\text{-}01\text{-}12}) - \min(open_{2004\text{-}01\text{-}12}, close_{2004\text{-}01\text{-}12})) \div (open_{2004\text{-}01\text{-}12})) \times 100 = \frac{6225.23 - 6219.71}{6225.23} \times 100 = 0.09.$$

By the definition of the body, the fuzzy set EQUAL has the maximum membership value and the close price is higher than the open price on 2004-01-12, then the body of 2004-01-12 is translated to EQUAL_BLACK.

The fuzzy candlestick line on 2004-01-09 can be calculated in the same way. Combining the previous trend and the candlestick lines on 2004-01-12 and 2004-01-09, the candlestick pattern on 2004-01-12 can be derived. For calculation convenience, we set the candlestick line of 2004-01-12 to line 1 and 2004-01-09 to line 2. Table 2 shows the fuzzy candlestick pattern transformed from the TAIEX data. The line number of the candlestick patterns is 2.

After performing the action of Step 6 and Step 7 of CPFA, a set of patterns can be derived. Using the patterns shown in Table 2 as input, the refined and selected patterns from TAIEX 2004-01-02 to 2004-2-27 are shown in Table 3. The symbol "null" in Table 3 indicates that the related attributes is a "don't care" and the refined patterns can be regarded as a set of fuzzy rules. For example, the first pattern in Table 3 can be represented by the following fuzzy rule:

IF previous-trend is NORMAL_BULLISH and open style is OPEN_EQUAL, THEN the following-trend is SMALL_INCREASE.

In this example, the pattern selection rule is set to "if $T > 0.5$, then selects the pattern." In our approach, the threshold of T is not restricted and the user can set different T in different applications. If T is set to 0, all of the refined patterns in Step 6 will be used as selected patterns.

After performing Step 8 of CPFA, the forecasting results are shown in Table 4. We use a different number of candlestick lines in this example for comparing the effect of the number of candlestick lines. Because the most popular numbers of candlestick lines used in the candlestick theory are 1 to 3, we set the number of candlestick lines from 1 to 3. After performing the CPFA, the experiment results are shown in Table 5 and compared with the methods proposed in [11], [12], [16]. The uppercase alphabet is used to represent the training and testing data sets. For example, the training data set A is the data from January 2004 to February 2004 and the testing data set A' is the data set of 2004 March. The training data set B is the data from March 2004 to April 2004 and the testing data set B' is the data set of May 2004.

The simulation result shows that our method gets better forecasting results than the proposed method in [11], [12], [16], not only in training data forecasting, but also in testing data prediction. When applying Chen's method [11], [16] to forecast testing data B', the values of MSE were very great. This is because the maximum and minimum training data from 2004-03-01 to 2004-04-30 is 7,034 and 6,118, but the maximum and minimum testing data from 2004-05-03 to 2004-05-31 is 6,188 and 5,483. In other words, when applying Chen's model, if the testing data is far from the interval of training data, using the trained model to forecast testing data will derive a bad result. Because Hwang's

TABLE 2
Extracted Candlestick Patterns of TAIEX Data (Candlestick Lines = 2)

Date	Previous trend, line1 Open style, line1 Close style, line1 Upper shadow, line1 Body, line1 Lower Shadow, line2 Open style, line2 Close style, line2 Upper shadow, line2 Body, line2 Lower Shadow,	Following trend
2004-01-12	WEAK_BULLISH, OPEN_EQUAL_LOW, CLOSE_EQUAL_LOW, SHORT, EQUAL_BLACK, SHORT, OPEN_HIGH, CLOSE_HIGH, SHORT, EQUAL_BLACK, SHORT	A_4
2004-01-13	EQUAL_BULLISH, OPEN_HIGH, CLOSE_EQUAL_LOW, SHORT, SHORT_BLACK, EQUAL, OPEN_EQUAL_LOW, CLOSE_EQUAL_LOW, SHORT, EQUAL_BLACK, SHORT	A_7
2004-01-14	WEAK_BULLISH, OPEN_LOW, CLOSE_HIGH, SHORT, SHORT_WHITE, EQUAL, OPEN_HIGH, CLOSE_EQUAL_LOW, SHORT, SHORT_BLACK, EQUAL	A_4
2004-01-15	WEAK_BULLISH, OPEN_HIGH, CLOSE_EQUAL_HIGH, EQUAL, SHORT_BLACK, EQUAL, OPEN_LOW, CLOSE_HIGH, SHORT, SHORT_WHITE, EQUAL	A_5
2004-01-16	WEAK_BULLISH, OPEN_EQUAL_HIGH, CLOSE_EQUAL_LOW, EQUAL, SHORT_BLACK, EQUAL, OPEN_HIGH, CLOSE_EQUAL_HIGH, EQUAL, SHORT_BLACK, EQUAL	A_8
…	…	…
2004-02-20	WEAK_BULLISH, OPEN_HIGH, CLOSE_EQUAL, EQUAL, SHORT_BLACK, EQUAL, OPEN_EQUAL_HIGH, CLOSE_HIGH, EQUAL, SHORT_WHITE, SHORT	A_5
2004-02-23	WEAK_BULLISH, OPEN_EQUAL_HIGH, CLOSE_EQUAL_LOW, EQUAL, SHORT_BLACK, SHORT, OPEN_HIGH, CLOSE_EQUAL, EQUAL, SHORT_BLACK, EQUAL	A_2
2004-02-24	EQUAL_BULLISH, OPEN_EQUAL_LOW, CLOSE_LOW, SHORT, SHORT_BLACK, EQUAL, OPEN_EQUAL_HIGH, CLOSE_EQUAL_LOW, EQUAL, SHORT_BLACK, SHORT	A_6
2004-02-25	EQUAL_BULLISH, OPEN_EQUAL, CLOSE_EQUAL_HIGH, SHORT, SHORT_WHITE, SHORT, OPEN_EQUAL_LOW, CLOSE_LOW, SHORT, SHORT_BLACK, EQUAL	A_6
2004-02-26	EQUAL_BULLISH, OPEN_HIGH, CLOSE_HIGH, SHORT, EQUAL_BLACK, SHORT, OPEN_EQUAL, CLOSE_EQUAL_HIGH, SHORT, SHORT_WHITE, SHORT	A_6

TABLE 3
Selected Patterns of TAIEX Data (Candlestick Lines = 2)

No.	Previous trend, line1 Open style, line1 Close style, line1 Upper shadow, line1 Body, line1 Lower Shadow, line2 Open style, line2 Close style, line2 Upper shadow, line2 Body, line2 Lower Shadow,	Following trend
1	NORMAL_BULLISH OPEN_EQUAL null null null null null null null null null	A_5
2	WEAK_BULLISH OPEN_EQUAL null null null null null null null null null	A_3
3	EQUAL_BULLISH OPEN_EQUAL null null null null null null null null null	A_6
4	null OPEN_EQUAL_HIGH CLOSE_EQUAL_LOW null null null null CLOSE_EQUAL_HIGH null null null	A_8
5	null OPEN_EQUAL_HIGH CLOSE_HIGH null null null null CLOSE_EQUAL_HIGH null null null	A_4
6	null OPEN_EQUAL_HIGH null null null null null CLOSE_EQUAL null null null	A_2
7	null OPEN_EQUAL_HIGH null null null null null CLOSE_HIGH null null null	A_2
8	EQUAL_BEARISH OPEN_EQUAL null null null null null null null null null	A_5
...	…	…
18	EQUAL_BEARISH OPEN_HIGH null null null null OPEN_EQUAL_LOW null null null null	A_8
19	null OPEN_HIGH null null null null OPEN_LOW null null null null	A_5
20	WEAK_BULLISH OPEN_EQUAL_LOW null null null null null null null null null	A_2
21	EQUAL_BULLISH OPEN_EQUAL_LOW null null null null null null null null null	A_6
22	WEAK_BEARISH OPEN_EQUAL_LOW null null null null null null null null null	A_7
23	null OPEN_LOW null null null null null null null null null	A_4

approach and Candlestick pattern approach both use the sum of variation and close price on the day t to determine the forecasting results on day $t+1$, this shortage can be avoided.

4.3 Stock Forecasting

There are several ways to use the fuzzy candlestick patterns. In the TAIEX example, the forecasting model generates forecasting results day by day. An investor has to determine how to use the daily forecasting result to buy or sell their stock. However, in many situations, the investor prefers to know when the uptrend is returned or the downtrend is bouncing back rather than to forecast the variation of stock price every time, because buying or selling stocks too frequently will cost much. In other words, the problem is how to find the point when the trend will return or bounce back.

The stock of 2330 (TSMC) in the Taiwan market are used as an example to explain the proposed method. The data from 1997-10-23 to 2002-12-25 are used as the training data, and we apply the patterns learned from the training data to invest the stock from 2003-01-02 to 2003-12-31. The algorithm is similar to CPFA and is described as follows:

Step 1. Determine the parameters of the fuzzy candlestick patterns. In this example, the variations interval of 2330 (TSMC) is set to five days. In other words, we want to

TABLE 4
TAIEX Forecasted Results

Date	(1) Actual index	(2) Forecasted index	Errors = |(1) – (2)|
2004-01-12	6210.22	6204.16	6.06
2004-01-13	6274.97	6287.85	12.88
2004-01-14	6264.37	6259.28	5.09
2004-01-15	6269.71	6280.03	10.32
2004-01-16	6384.63	6379.43	5.20
…	…	…	…
2004-02-20	6665.89	6682.20	16.31
2004-02-23	6589.23	6582.57	6.66
2004-02-24	6644.28	6638.65	5.63
2004-02-25	6693.25	6694.11	0.86
2004-02-26	6750.54	6743.45	7.09

TABLE 5
Mean Square Error of TAIEX Forecasted Results

	Chen1996 [11]	Hwang1998 [12] (window=5)	Chen2002 [16] (order=6)	Candlestick pattern (line =1)	Candlestick pattern (line =2)	Candlestick pattern (line =3)
Training data A	3958	5115	1235	313	82	84
Testing data A'	29904	31855	43171	26083	18202	20363
Training data B	15932	19217	2452	612	335	343
Testing data B'	187383	34469	69183	22877	26769	31312
Training data C	16474	22295	1101	3919	216	215
Testing data C'	22078	8246	10579	16239	15660	15340
Training data D	3828	7520	534	221	214	218
Testing data D'	13567	4525	3469	2003	3400	3745
Training data E	3579	4191	629	754	174	179
Testing data E'	5102	5187	6966	6298	6573	4244
Training data F	3715	3938	619	498	68	68
Testing data F'	3028	2980	3235	5278	4771	2747

forecast a five-day variation of the stock price. The UoD is represented as $UoD = [-20, 20]$ and partitioned into 10 intervals. The UoD and the data variation distribution are shown in Table 6. The fuzzy sets considered are:

$$A_1 = (\text{EXTREME\_DECREASE}),$$
$$A_2 = (\text{LARGE\_DECREASE}),$$
$$A_3 = (\text{NORMAL\_DECREASE}),$$
$$A_4 = (\text{SMALL\_DECREASE}),$$
$$A_5 = (\text{TINY\_DECREASE}), \ldots, \text{ and}$$
$$A_{10} = (\text{EXTREME\_INCREASE}).$$

The definition of fuzzy sets A_1 to A_{10} on the UoD is similar to (9). The candlestick line number of a candlestick pattern is set to 2.

Step 2. Extract the patterns from the stock open, close, high, and low prices. This step is the same as Step 5 of CPFA. There are 1,342 patterns extracted from the stock data.

Step 3. Refine the extracted patterns. This step is the same as Step 6 of CPFA. There are 1,178 refined patterns.

Step 4. Select the refined patterns. The pattern selection rule is defined as follows: In this example, we assume that the user wants to select the patterns with a following uptrend. There are 79 patterns selected for forecasting.

IF $T > 0.5$ and the following trend of the pattern equals to "STRONG_INCREASE" or "EXTREME_INCREASE" THEN select the pattern.

TABLE 6
UoD Data Distribution of 2330 (TSMC)

UoD	Interval	Data distribution
u_1	-20.0 -> -16.0	26
u_2	-16.0 -> -12.0	35
u_3	-12.0 -> -8.0	108
u_4	-8.0 -> -4.0	196
u_5	-4.0 -> 0.0	371
u_6	0.0 -> 4.0	270
u_7	4.0 -> 8.0	188
u_8	8.0 -> 12.0	79
u_9	12.0 -> 16.0	42
u_{10}	16.0 -> 20.0	34

Step 5. Use the selected pattern to predict the stock trend. The meaning of the selected patterns is when the selected patterns appear then the following trend will be an uptrend. Table 7 shows the date of any selected pattern matched and the close price variance v. The variance v is defined as follows:

$$v = \frac{close_{t+5} - close_t}{close_t} \times 100, \tag{15}$$

where $close_t$ is the close price when any selected pattern is matched on date t and $close_{t+5}$ is the close price of date $t + 5$.

4.4 Enrollment Forecasting

This section describes how to use the candlestick pattern approach for enrollment forecasting. To generate, a candlestick line needs four data: open, close, high, and low. Because there is only yearly data within the historical enrollment data shown in [9], we choose four years as a window to derive the candlestick patterns. For example, the enrollment data from 1977 to 1980 is 15,603, 15,861, 16,807, and 16,919. The open, close, high, and low data for 1980 are 15,603, 16,919, 16,919, and 15,603, respectively. We use (8) and set the parameter n in (8) to 1 for deriving the variation of two continue years. In this example, we can derive $I_{min} = -5.83$ and $I_{max} = 7.66$.

TABLE 7
Stock Prediction Results

Date	Variance
2003-12-19	3.3
2003-11-21	-3.1
2003-11-20	-0.7
2003-11-04	-0.7
2003-10-01	5.3
2003-09-30	6.8
2003-09-17	-4.9
2003-09-12	3.6
2003-08-15	4.9
2003-08-13	5.9
2003-07-28	3.4
2003-07-22	2.7
2003-07-02	3.3
2003-06-17	-0.9
2003-05-30	9.3
2003-03-13	8.6
Total variance: 46.8	

TABLE 8
Enrollment Forecasted Results

Years	(1) Actual enrollments	(2) Forecasted enrollments	Errors = |(1) – (2)|
1979	16807		
1980	16919	16975.07	56.07
1981	16388	16411.43	23.43
1982	15433	15568.6	135.60
1983	15497	15587.33	90.33
1984	15145	15032.09	112.91
1985	15163	15296.45	133.45
1986	15984	15921.15	62.85
1987	16859	16783.2	75.8
1988	18150	18039.13	110.87
1989	18970	19057.5	87.5
1990	19328	19159.7	168.3
1991	19337	19521.28	184.28
1992	18867	19530.37	654.37

We set $D_1 = 0.17$ and $D_2 = 0.34$, so that UoD can be represented as $UoD = [-6, 8]$. The UoD is partitioned into seven intervals.

The number of candlestick lines in the enrollment candlestick pattern is 1. The selection rule is "if $T > 0.5$, then selects the pattern." Finally, the forecasting result of the enrollment data set is shown in Table 8.

We carefully reimplemented the methods proposed in [11], [12], [15], [16] to obtain the MSE values for comparison. The MSEs of our method and other methods are shown in Table 9.

In our survey, the method proposed in [17] received the smallest MSE. In the paper, the authors assume that the forecasted fuzzy set A_f can be exactly derived and use the data difference of differences between years n-1 and n-2 and between years n-2 and n-3 to fine tune the forecasting results. However, how to exactly derive the fuzzy set is not mentioned in the paper. Thus, we can see the result as an ideal benchmark.

4.5 Discussion

The time complexity of decision tree with n instances and m attributes induction is $O(mn \log n) + (n \log n)^2$ [29] and is the most important factor of time complexity when construct the forecasting model in our approach. The time complexity of forecasting step is $O(kn)$, where n is the number of selected patterns and k is the attributes of the patterns.

Although the proposed method takes a larger time complexity than other proposed methods, we argue that it is acceptable for two reasons. First, the forecasting model only needs to be constructed once and the forecasting step takes similar time complexity in [16]. Second, the time spent is still acceptable with the computational power of the current personal computers. We implemented our method in Java language and executed it in our PC with Pentium1.6G, 512M RAM, Windows 2000 operation system. When entering 2,722 instances with 12 attributes (two candlestick lines), the model construction time is 0.3 second. If there are 300 patterns selected as the forecasting model and 500 daily stock data need to be identified, the execution time of total forecasting process is less than five seconds.

TABLE 9
Enrollment Forecasting Comparison Results

Methods	MSE
Song1993[9]	423027
Song1994[10]	775687
Chen1996[11]	407521
Hwang1998[12](window=5)	284144
Hurang2001[15]	226611
Chen2002[16](order=6)	98215
Chen2004[17]	5353
Candlestick pattern	44686

The investors can use the method proposed in the stock forecasting example to find patterns which they are interested in. Furthermore, the investors can use fuzzy modifier proposed in Section 3 to fine tune the pattern and apply the pattern to other target.

For example, assume that the following patterns are selected by the pattern selection process in the stock example.

P1. IF previous-trend = STRONG_BEARISH and open style = OPEN_LOW and upper shadow = EQUAL and body = CROSS and lower shadow = MIDDLE, then following trend = STRONG_INCREASE.

P2. IF previous trend = EXTREME_ BEARISH and open style = OPEN_LOW and upper shadow = EQUAL and body = CROSS and lower shadow = LONG, then following trend = EXTREME_INCREASE.

The investor can merge P1 and P2 by using the ABOVE fuzzy modifier as follows:

P3. IF previous trend = ABOVE STRONG_ BEARISH and open style = OPEN_LOW and upper shadow = EQUAL and body = CROSS and lower shadow = ABOVE MIDDLE, then following trend = ABOVE STRONG_INCREASE.

The investors can further reuse the fine tuned patterns as the selected patterns to help themselves making the investment decision. How to fine tune the candlestick patterns needs investment experience, and our approach provides a good way for the users in representing and validating their investment experience. The patterns can also be used in different places. For example, the user can apply the pattern found in stock 2330 (TSMC) to TAIEX forecasting or share the selected pattern to other investors.

In candlestick theory, the fluctuation of the stock price in the same time period can be represented in the length of shadows and body. Moreover, the causal relationship between continual candlestick lines can be described in the open style and the close style. The fuzzy candlestick pattern contains richer information than the symbolic time series and the fuzzy time series, and the added information can be used to increase the effectiveness and correctness of the forecasting results.

Another advantage of modeling the financial time series with fuzzy candlestick pattern is that the time series data become visual, interpretable, and editable to the investor. In other words, the investors can understand the reasons of the investment decision and participate in the forecasting process by defining and modifying the patterns. We have implemented a system prototype to demonstrate this idea. The fuzzy candlestick pattern not only can be used as the preprocessed data set for the data mining and machine

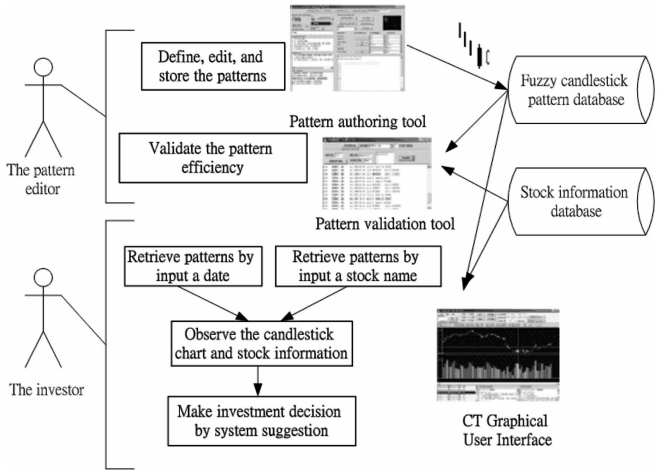

Fig. 6. Operations of candlestick tutor.

learning methods, but also can be used directly for pattern recognition.

5 Implementation

A system prototype, named Candlestick Tutor (CT) which was proposed in [30], is adapted to realize the idea of this paper. For the user to learn and share the knowledge about the candlestick patterns, the CT includes a graphical user interface (GUI) for displaying candlestick charts, a pattern authoring and acquiring tool for mining patterns, a pattern recognition module for recognizing patterns, and a pattern validation tool for checking pattern efficiency. The CPFA is implemented in the pattern authoring and acquiring tool to facilitate the pattern editing process.

An information agent, a stock information database, and a fuzzy candlestick pattern database are also designed to support the system. The information agent connects to the Web site which provides the daily stock information for acquiring the stock information. The system operation is illustrated in Fig. 6. For programming convenience, the system is coded in Java language. The NRC Fuzzy Toolkit provides a set of Java API which enhances the system with the capabilities of handling fuzzy concepts and reasoning.

6 Conclusions

In this paper, we proposed a knowledge-based method to represent the financial time series and to facilitate the knowledge discovery process of the time series. The variance of the stock price is represented in fuzzy candlestick patterns which make the imprecise and vague investment knowledge comprehensible, computable, visual, and editable. The fuzzy candlestick patterns carry rich information and can be used to increase the efficiency of the knowledge discovery process of financial time series. Pattern construction and the recognition process are introduced and implemented in a system prototype to illustrate the usage of the fuzzy candlestick patterns. Moreover, investors can save and share their investment experience. By reusing and modifying the stored candlestick pattern information, the investor can also increase the efficiency of their investing strategies.

Acknowledgment

The authors would like to thank Kao-Shing Hwang for helpful discussions on the fuzzy sets and the anonymous reviewers for their valuable comments. This work was supported in part by the Department of Industrial Technology, Ministry of Economic Affairs (Taiwan) under Grant 94-EC-17-A-02-S1-029 and by the National Science Council under grant NSC-93-2213-E-194-039.

References

[1] J.C.P. Shieh, *Contemporary Investments-Analysis And Management.* Taipei, Taiwan: Best-Wise, pp. 317-431, 1998.
[2] K. Mehmed, *Data Mining: Concepts, Models, Methods, and Algorithm,* pp. 19-38. John Wiley & Sons, 2003.
[3] E.W. Saad, D.V. Prokhorov, and D.C. II Wunsch, "Comparative Study of Stock Trend Prediction Using Time Delay, Recurrent and Probabilistic Neural Networks," *IEEE Trans. Neural Networks,* vol. 9, no. 6, pp. 1456-1470, Nov. 1998.
[4] A.P. N. Refenes and W.T. Holt, "Forecasting Volatility with Neural Regression: A Contribution to Model Adequacy," *IEEE Trans. Neural Networks,* vol. 12, no. 4, pp. 850-864, July 2001.
[5] K.N. Pantazopoulos, L.H. Tsoukalas, N.G. Bourbakis, M.J. Brun, and E.N. Houstis, "Financial Prediction and Trading Strategies Using Neurofuzzy Approaches," *IEEE Trans. Systems, Man, and Cybernetics, Part B,* vol. 28, no. 4, pp. 520-531, Aug. 1998.
[6] V. Dhar and D. Chou, "A Comparison of Nonlinear Methods for Predicting Earnings Surprises and Returns," *IEEE Trans. Neural Networks,* vol. 12, no. 4, pp. 907-921, July 2001.
[7] L.J. Cao and F.E. H. Tay, "Support Vector Machine with Adaptive Parameters in Financial Time Series Forecasting," *IEEE Trans. Neural Networks,* vol. 14, no. 6, pp. 1506-1518, Nov. 2003.
[8] Q. Song and B.S. Chissom, "Fuzzy Time Series and Its Models," *Fuzzy Sets and Systems,* vol. 54, pp. 269-277, 1993.
[9] Q. Song and B.S. Chissom, "Forecasting Enrollments with Fuzzy Time Series—Part 1," *Fuzzy Sets and Systems,* vol. 54, pp. 1-9, 1993.
[10] Q. Song and B.S. Chissom, "Forecasting Enrollments with Fuzzy Time Series—Part 2," *Fuzzy Sets and Systems,* vol. 62, pp. 1-8, 1994.
[11] S.M. Chen, "Forecasting Enrollments Based on Fuzzy Time Series," *Fuzzy Sets and Systems,* vol. 81, pp. 311-319, 1996.
[12] J.R. Hwang, S.M. Chen, and C.H. Lee, "Handling Forecasting Problems Using Fuzzy Time Series," *Fuzzy Sets and Systems,* vol. 100, pp. 217-228, 1998.
[13] S.M. Chen and J.R. Hwang, "Temperature Predictin Using Fuzzy Time Series," *IEEE Trans. Systems, Man, and Cybernetics, Part B,* vol. 30, no. 2, pp. 263-275, 2000.
[14] C.C. Tsai and S.J. Wu, "A Study for Second-order Modeling of Fuzzy Time Series," *Proc. IEEE Fuzzy System Conf.,* pp. 719-725, 1999.
[15] K. Huarng, "Heuristic Models of Fuzzy Time Series for Forecasting," *Fuzzy Sets and System,"* vol. 123, pp. 369-386, 2001.
[16] S.M. Chen, "Forecasting Enrollments Based on High-Order Fuzzy Time Series," *Cybernetics and Systems: An Int'l J.,* vol. 33, pp. 1-16, 2002.
[17] S.M. Chen and C.C. Hsu, "A New Method to Forecast Enrollments Using Fuzzy Time Series," *Int'l J. Applied Science and Eng.,* vol. 2, no. 3, pp. 234-244, 2004.
[18] Taiwan Futures Exchange, http://www.taifex.com.tw/eng/eng_home.htm, 2006.
[19] NASDAQ, New York, http://www.nasdaq.com/, 2006.
[20] G.J. Klir and B. Yuan, *Fuzzy Sets and Fuzzy Logic Theory and Application.* Prentice Hall, pp. 357-378, 1995.
[21] J. Giarratano and G. Riley, *Expert System—Principles and Programming,* third ed. Boston: PWS, pp. 57-91, 1998.
[22] S.K. Pal and S.C. K. Shiu, *Foundations of Soft Case-Based Reasoning.* John Wiley & Sons, pp. 3-25, 2004.
[23] C.S. Daw and C.E. A. Finney, "A Review of Symbolic Analysis of Experimental Data," http://www-chaos.engr.utk.edu/pap/crg-rsi2002.pdf, 2006.

[24] J. Lin, E. Keogh, S. Lonardi, and B. Chiu, "A Symbolic Representation of Time Series, with Implications for Streaming Algorithms," *Proc. SIGMOD Workshop Research Issues on Data Mining and Knowledge Discovery,* pp. 2-11, 2003.
[25] E. Keogh, S. Lonardi, and B. Chiu, "Finding Surprising Patterns in a Time Series Database in Linear Time and Space," *Proc. SIGKDD Conf.,* pp. 550-556, 2002.
[26] G.L. Morris, *Candlestick Charting Explained: Timeless Techniques for Trading Stocks and Futures,* second ed. McGraw-Hill Trade, pp. 8-139, 1995.
[27] E. Keogh and S. Kasetty, "On the Need for Time Series Data Mining Benchmarks: A Survey and Empirical Demonstratin," *Proc. SIGKDD Conf.,* pp. 102-111, 2002.
[28] NRC, Fuzzy Toolkit, http://www.iit.nrc.ca/IR_public/fuzzy/fuzzyJDocs/index.html, 2006.
[29] H.W. Ian and F. Eide, *Data Mining-Practical Machine Learning Tools and Techniques with Java Implementations.* San Francisco: Morgan Kaufmann, 2000.
[30] C.H.L Lee, W. Chen, and A. Liu, "An Implementation of Knowledge Based Pattern Recognition for Finicial Prediction," *Proc. 2004 IEEE Conf. Cybernetics and Intelligent Systems (CIS '04),* pp. 218-223, 2004.

Chiung-Hon Leon Lee received the BS degree in electronic engineering from the National Taiwan Institute of Technology in 1996 and the MS degree in electrical engineering from the National Chung Cheng University in Taiwan in 1998. He is a PhD candidate and lecturer in the Electrical Engineering Department at National Chung Cheng University. His research interests are in agent-based software engineering, Web services, knowledge representation, and fuzzy time series.

Alan Liu received the PhD degree in electrical engineering and computer science from the University of Illinois at Chicago in 1994. He is an associate professor in the Department of Electrical Engineering, National Chung Cheng University in Taiwan. His research interests in artificial intelligence and software engineering include knowledge acquisition, requirements analysis, intelligent agents, and applications in embedded systems, and robotic systems. He is also a member of the IEEE, ACM, and TAAI.

Wen-Sung Chen received the BS degree in electrical engineering from National Kaohsiung University of Applied Sciences in Taiwan in 2003. He is an MS student in the Department of Electrical Engineering at National Cheng Kung University, Taiwan. His research interests include expert systems and fuzzy system applications.