Chapter 2
Related Work

Abstract This chapter presents background information and reviews the existing literature that is relevant to the development of this project. The first part of this chapter presents a brief description of the two existing approaches to analyze the market, in Sect. 2.1 will be described in detail the fundamental and the technical analysis and its tools. A formal definition of an optimization methodology is given in Sect. 2.2. A review of the existence literature about pattern recognition/detection and its techniques to invest in the market is detailed in Sect. 2.3.

2.1 Market Analysis

The aim of financial market analysis is to predict the behavior of the prices in the market in order to make a better decision (buy/sell) over a financial asset. There are two distinct types of market analysis: Fundamental Analysis and Technical Analysis. The core study of these two distinct types of analysis is different which do not invalidate the fact that the two types can be used simultaneously in order to make the best decision in a financial market. In this work the identification of patterns in the financial markets and the investment rules based on that appeal to the application of technical analysis.

2.1.1 Fundamental Analysis

The Fundamental Analysis [1] is based on a set of financial and economic indicators with the goal of finding the intrinsic value of a company and consequently its stock price. The fundamental analysis studies all the factors, internal or external, that can influence the value of a company. After finding the intrinsic value of a company it is possible to understand if the company is overvalued or undervalued and based on that make a better investment decision. In the case where the stock price of a company is higher than its intrinsic value (overvalued), the better decision is to sell

J. Leitão et al., *Identifying Patterns in Financial Markets*, SpringerBriefs
in Computational Intelligence, https://doi.org/10.1007/978-3-319-70160-8_2

and in the opposite case, where the stock price is lower than its intrinsic value the better decision is to buy because the company is undervalued.

2.1.2 Technical Analysis

The Technical Analysis [2, 3] is based on past market data, such as price and volume, with the goal of predict its behavior. The analysts believe that the stock price in the market already reflects in itself all the fundamental factors that can affect its price, so it is unnecessary to proceed to the Fundamental Analysis [3].

The advantages of this type of analysis are: the data used (price and volume) are easily accessible to anyone and exist in huge quantity, which is very useful to the pattern detection methodologies that will be presented after.

In technical analysis in order to predict correctly the future movement of the financial markets several technical indicators are used [2, 3], which are built based on the price and volume. In addition to technical indicators the analysts also study the formation of chart patterns [3, 4] in the historical prices of financial assets. Some of the most well known and most utilized technical indicators and chart patterns are presented next.

I. **Technical indicators**

A technical indicator is a metric whose value is calculated from the price or the volume of an asset. Its objective is to help predicting the future price, or simply to indicate a general price trend. Some of the most popular technical indicators are presented next. Other technical indicators can be found in [2, 5].

(a) **Simple Moving Average**

This indicator is one of the oldest indicators used by the analysts and represents the mean value of the prices over a certain amount of time (days). Normally, the closing price of each day is used to calculate the value of this indicator.

The Simple Moving Average (SMA) can be calculated for different lengths of time, where the most common are 200, 100, 50, 30, 20, and 10 days. Smaller moving averages are usually used for short-term investments and longer moving averages are used for long-term investments. Naturally, long-term Simple Moving Averages have fewer fluctuations than long-term Simple Moving Averages.

In Fig. 2.1 it is represented the price chart of Apple with a Simple Moving Average of 10 days and other of 20 days for the period 06/06/2014–07/05/2015. It is possible to observe that the line that represents the 10 days average (SMA10) reacts better and faster to the several short-term changes in the price comparing with the line that represents the 20 days average (SMA20).

This indicator can be used by traders to buy a financial asset when the average is in an upward trend and to sell it when the average is in a downward trend. Several

Fig. 2.1 Apple's price chart with 10 and 20 days SMA

and different moving averages can be used simultaneously to determine intersection points between them which originate buy or sell signals.

There are other indicators based on the SMA, like the Exponential Moving Average which attributes more importance (more weight) to the most recent days when calculating the average, in the attempt to better react to sudden changes of price.

(b) **Relative Strength Index (RSI)**

This indicator is one of the most used indicators of the category momentum, which compares the magnitude between the recent profits and losses, with the aim to determine if a financial asset is overbought or oversold. This indicator oscillates between 0 and 100 and it is often calculated for a period of 14 days. In Eq. (2.1) is described the formula to calculate this indicator.

$$\begin{aligned} \text{RSI} &= 100 - \frac{100}{1 + \text{RS}} \\ \text{RS} &= \frac{\text{Average of } x \text{ days' up closes}}{\text{Average of } x \text{ days' down closes}} \end{aligned} \tag{2.1}$$

In Fig. 2.2 it is possible to observe an example with the price chart of Apple and its RSI of 14 days for the period 31/12/2014–07/05/2015. The RSI has several characteristics that can generate different signals:

- If the RSI value of an asset is higher than 50 it means an upward movement of prices and so the asset should be bought. If the RSI value is lower than 50 it means a downward movement of prices and consequently the asset should be sold.

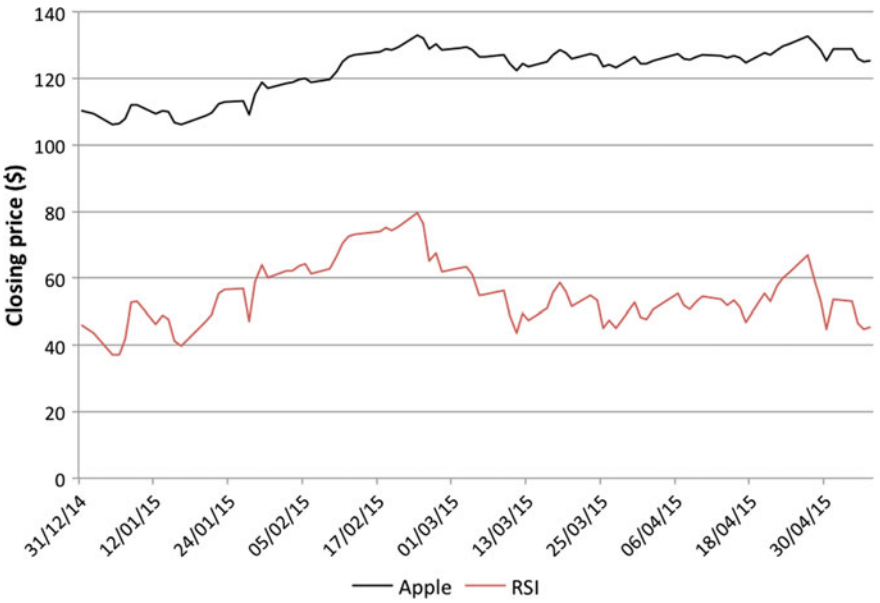


Fig. 2.2 Apple's price chart and its 14 days RSI

- If the RSI value of an asset is higher than 70 it means the asset is overbought and so the asset should be sold. If the RSI value is lower than 30 it means the asset is oversold and so the asset should be bought.

(c) **On-Balance Volume**

This indicator is the oldest and the most well-known volume indicator. The volume and the price are used in the calculation of this indicator, which measures buying and selling pressure. The idea behind this indicator is that using volume to analyze the price chart of an asset, it is possible to predict its behavior or simply confirm its trend. The formula to calculate the OBV is represented in Eq. (2.2).

$$\begin{aligned}
&\text{OBV}\,(x) = \text{OBV}\,(x-1) + \text{Volume}\,(x)\,, \quad \text{if Price}\,(x) > \text{Price}\,(x-1)\\
&\text{OBV}\,(x) = \text{OBV}\,(x-1) - \text{Volume}\,(x)\,, \quad \text{if Price}\,(x) < \text{Price}\,(x-1)\\
&\text{OBV}\,(x) = \text{OBV}\,(x-1)\,, \quad \text{if Price}\,(x) = \text{Price}\,(x-1)
\end{aligned} \tag{2.2}$$

In Fig. 2.3 it is represented as an example of the application of this indicator in the price chart of Apple over the period 09/06/2014–07/05/2015. This indicator has several characteristics that can generate different signals like:

- When the OBV has an upward movement it means that the financial market will start to increase and become a bullish market even if the prices are not rising yet. The performance of a financial market starts to decrease (bearish market) when the OBV has a downward movement even if the prices are not falling.

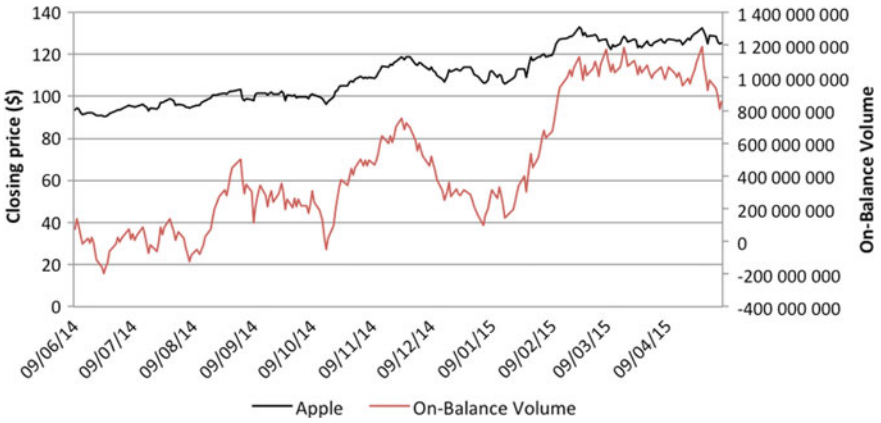


Fig. 2.3 Price chart of Apple with the OBV

- In the case where the OBV and the prices are following an upward movement this means that this trend will be maintained in the future. The same applies to the opposite case.

II. Chart Patterns

Analyzing the historical prices of a financial asset it is possible to observe some similar geometric shapes over the time. Those geometric figures represent the behavior of the traders in the market. Knowing that history repeats, the identification of geometric figures allow the analysts to predict with some confidence the behavior of the traders and consequently the future trend of prices.

The chart patterns, according to [3], can be divided in 2 types: Continuation Patterns and Reversal Patterns. The continuation patterns generally are faster to form than the reversal patterns. In order to be more certain of the future direction of prices, the volume indicator can be used to confirm the formation of chart patterns. In the next sections some of the most used and famous continuation and reversal patterns will be presented. For more information on others chart patterns [4].

(a) Continuation Patterns

This type of chart patterns is characterized by confirming the uptrend or downtrend of the market, despite of the trend of prices become a sideways movement temporarily. When this type of pattern occurs, it can indicate the trend is likely to resume after the pattern completes.

(1) Triangles

There are three types of Triangles: (a) Symmetrical Triangle, (b) Ascending Triangle and (c) Descending Triangle. These patterns have a typical duration of 3 months. In case (a) the prices after breakout the triangle follow the direction of the previous

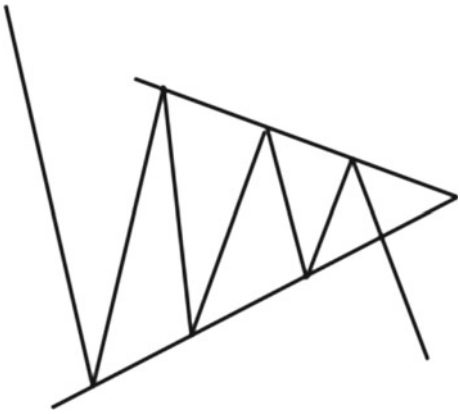


Fig. 2.4 Bullish Symmetrical Triangle (left) and Bearish Symmetrical Triangle (right)

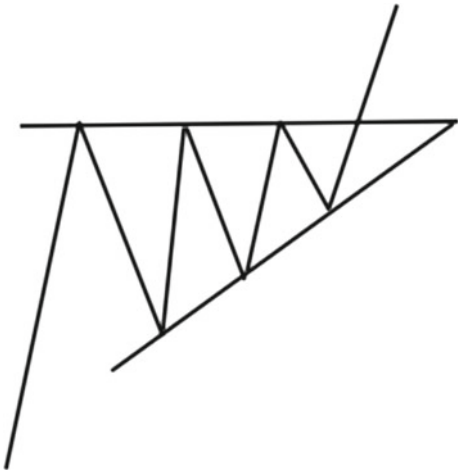


Fig. 2.5 Ascending Triangle

trend. This case applies to bull markets and bear markets as illustrated in Fig. 2.4, respectively.

The case (b) is often a bullish chart pattern, as illustrated in Fig. 2.5, where the prices breakout the triangle with an upward direction thereby confirming the previous trend.

The case (c) is often a bearish chart pattern, where the prices breakout the triangle with a downward direction that confirms the previous trend. This pattern is illustrated in Fig. 2.6.

(2) **Flags and Pennants**

These two types of patterns are very similar due to the fact that they are preceded by a strong increase or decrease movement that is followed by a consolidation period that marks the reset of the initial movement (strong increase or decrease). These patters have a typical duration of one to 4 weeks. In Fig. 2.7 are illustrated the two types

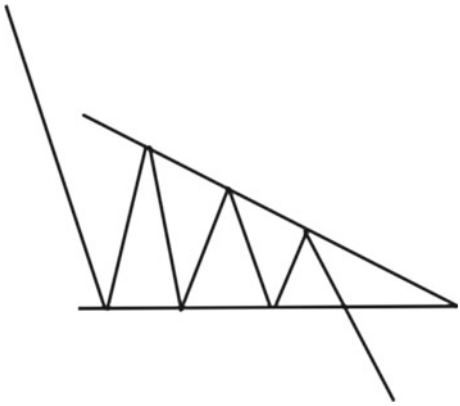


Fig. 2.6 Descending Triangle

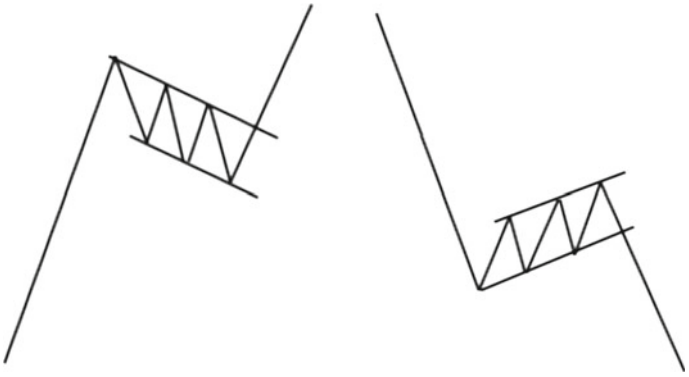


Fig. 2.7 Bull Flag (left) and Bear Flag (right)

of the Flag Pattern: the Bull Flag and the Bear Flag. The two cases of the Pennant Pattern: the Bull Pennant and the Bear Pennant are illustrated in Fig. 2.8.

(3) **Rectangles**

The rectangles represent a period of time where the prices follow a sideways movement delimited by two parallel horizontal lines (resistance and support). During the geometric formation the supply and demand is balanced. In Fig. 2.9 it is possible to observe the Bullish Rectangle and also the Bearish Rectangle.

(b) **Reversal Patterns**

This type of chart patterns, as the name implies, is characterized by a change in the direction of a price trend. An uptrend reverses to a downtrend and a downtrend reverses to an uptrend in this type of patterns. So, in this type of patterns the previous trend is inverted which marks the beginning of the new trend.

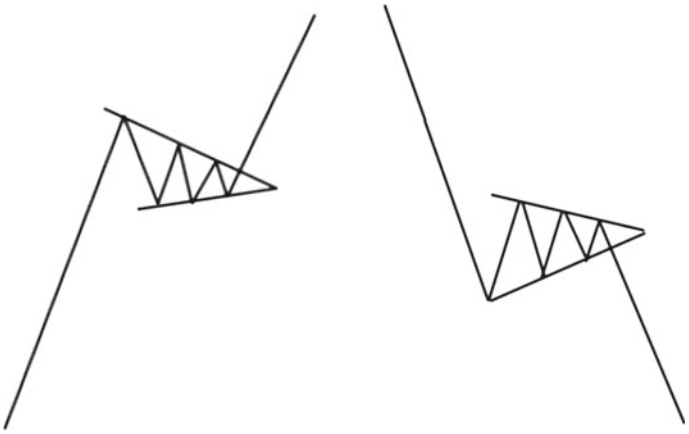


Fig. 2.8 Bull Pennant (left) and Bear Pennant (right)

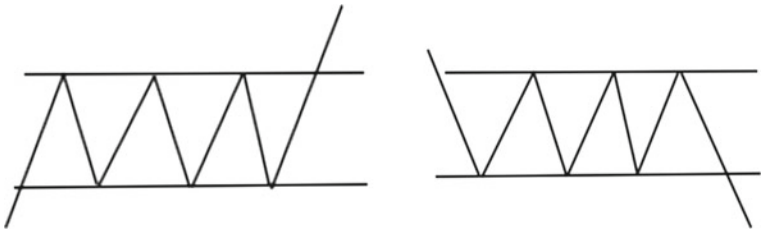


Fig. 2.9 Bullish Rectangle (left) and Bearish Rectangle (right)

Fig. 2.10 Head-and-Shoulders (left) and inverse Head-and-Shoulders (right)

(1) **Head-and-Shoulders**

This pattern is the most reliable and well-known reversal pattern and represents a reversal in the trend, generally the uptrend, where the confirmation of the reversal occurs when the prices breakout the support or the resistance line, named neckline. The two possibilities of this pattern are illustrated in Fig. 2.10.

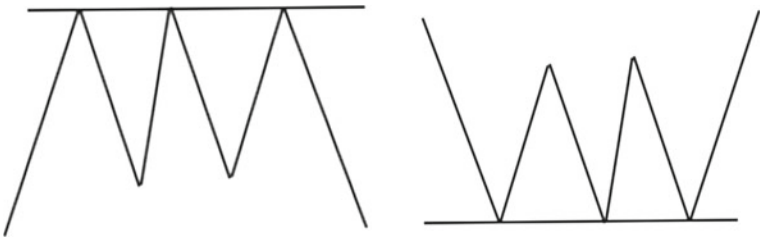


Fig. 2.11 Triple Top (left) and Triple Bottom (right)

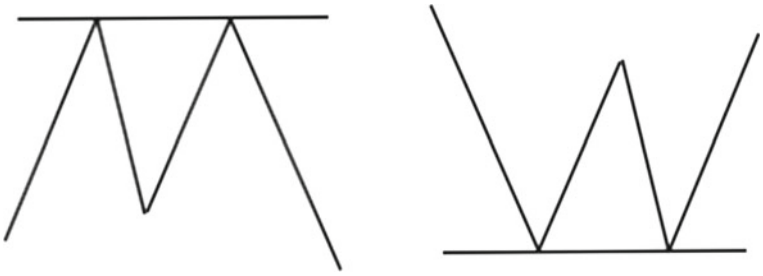


Fig. 2.12 Double Top (left) and Double Bottom (right)

(2) **Triple Top and Triple Bottom**

These patterns, as the name implies, are defined by three peaks (Triple Top) or by three bottoms (Triple Bottom). These patterns are similar to the Head-and-Shoulders pattern except the fact that the three peaks or bottoms, in this case are all at the same amplitude as can be seen in Fig. 2.11.

(3) **Double Top and Double Bottom**

These patterns are frequently seen in price charts and are very similar to the two previous patterns (point 2). The Double Top pattern is represented by 2 peaks with the same amplitude and the Double Bottom pattern is represented by 2 bottoms at the same amplitude too. As can be seen in Fig. 2.12, the Double Top and Double Bottom are often described by the character "M" and the character "W" respectively.

2.2 Optimization Methodologies—Genetic Algorithms

A Genetic Algorithm (GA) [6, 7] is a search heuristic that mimics the process of natural selection. The Genetic Algorithms belong to the class of the evolutionary algorithms, which are used to generate solutions to optimization problems through techniques based on natural evolution. The GA's are widely used in financial markets to find the best combination of parameters of an investment strategy [6, 8].

In the GA a population of potential solutions is used to find the best solution for a problem. Each potential solution is represented by a one-dimensional vector, named chromosome that represents the several parameters to optimize. Each parameter of the chromosome is called gene and for each is assigned a value. For example, some parameters that can be defined as genes can be the Simple Moving Average, the RSI, etc.

The process of the GA is an iterative process, where the population in each iteration is named generation. In each generation the solutions are evaluated according to a fitness function, i.e. an objective function like maximize profit, and the ones with higher fit value are selected to the next generation. After that, the genetic operators are used in the solutions to create better solutions and to create the next generation. These genetic operators are:

- Crossover—this operation is similar to the human reproduction, where a new solution (children) is created through a combination between the characteristics (parameters) of two solutions (parents).
- Mutation—this operation represents the biologic mutation and it is used to maintain the genetic diversity of populations in the next generations by changing some genes of the chromosomes.

This iterative process is terminated when the stop criteria is reached, that can be defined like: a solution is found that satisfies the minimum criteria, maximum number of generations is reached, lack of improvements of solutions in successive generations, etc.

2.3 Pattern Detection Methodologies

In this section several different techniques to reduce the data dimensionality of time series and its application in the identification of patterns are presented. The first methodology presented is based on matrixes, the second in relevant points and the last in SAX representation.

2.3.1 Heuristic Based on Templates

As the name implies, this method will recognize patterns based on a template pattern approach, where the templates are represented in a matrix format. In [9, 10] is used a method, based on templates, to detect the Bull Flag pattern (Fig. 2.7 left) with the aim of predict a rise in prices in the future. In this approach the goal is to detect the pattern in the historical prices of financial assets, in order to obtain more return than the average return of the financial markets. Through this pattern detection approach were created investment rules that were tested in the NYSE index in [11].

0.5		-1	-1	-1	-1	-1	-1	-1	
1	0.5		-0.5	-1	-1	-1	-1	-0.5	
1	1	0.5		-0.5	-0.5	-0.5	-0.5		0.5
0.5	1	1	0.5		-0.5	-0.5	-0.5		1
	0.5	1	1	0.5				0.5	1
		0.5	1	1	0.5			1	1
-0.5			0.5	1	1	0.5	0.5	1	1
-0.5	-1			0.5	1	1	1	1	
-1	-1	-1	-0.5		0.5	1	1		-2
-1	-1	-1	-1	-0.5		0.5	0.5	-2	-2.5

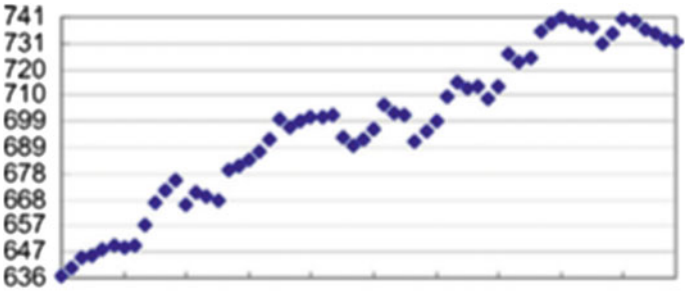
Fig. 2.13 Bull Flag matrix pattern template

This approach aims to detect the Bull Flag pattern based on a template represented by the matrix in Fig. 2.13, which comprises a consolidation area (first seven columns), where prices fluctuate within a channel similar to a parallelogram, which is followed by a strong rise (3 last columns), named breakout, where prices start to increase.

This template illustrated in Fig. 2.13 is represented by a 10×10 matrix named "*T*", where each cell can have a value between -2.5 and $+1.0$ and the cells without value are assigned to 0. Also, the sum of all the values of each column in matrix "*T*", is always equal to 0. Then 10×10 matrixes, named "*I*" are created to represent each time series of the historical prices. The time series that are represented in matrixes "*I*" have a variable number of days (ex: 120, 60, etc.) allowing the identification of patterns with different time lengths. This concept is called sliding window.

In each time series the noise of its data is reduced by replacing the closing prices that exceed a boundary, defined by two standard deviations related to the mean of the time series prices, by its value (boundary value). Then the time series are divided in 10 identical groups, where each group will be mapped in a column of matrix "*I*". As an example, if the time series has a length of 60 days each column of its matrix "*I*" represents 6 days. Using this technique it is possible to represent in matrix "*I*" time series of any length.

After that, the difference between the highest and the lowest price of each time series, which defines the maximum amplitude, is divided by 10 in order to identify the price range of each row in matrix "*I*". As an example if the highest and the lowest price in a time series are 100$ and 50$ respectively, so the first row corresponds to the price range between 95$ and 100$, the second row to the price range 90$–95$ and so on. Then, each cell of matrix "*I*" will have a value between 0.0 and 1.0 dependent on the number of days that are mapped on it. In Fig. 2.14 it is possible to observe an example of a 60 days time series without noise and its matrix "*I*".



0.	0.	0.	0.	0.	0.	0.	0.33	0.83	1.
0.	0.	0.	0.	0.	0.	0.	0.5	0.17	0.
0.	0.	0.	0.	0.	0.	0.5	0.17	0.	0.
0.	0.	0.	0.17	0.5	0.5	0.5	0.	0.	0.
0.	0.	0.	0.5	0.5	0.5	0.	0.	0.	0.
0.	0.	0.33	0.33	0.	0.	0.	0.	0.	0.
0.	0.33	0.33	0.	0.	0.	0.	0.	0.	0.
0.	0.17	0.33	0.	0.	0.	0.	0.	0.	0.
0.33	0.5	0.	0.	0.	0.	0.	0.	0.	0.
0.67	0.	0.	0.	0.	0.	0.	0.	0.	0.

Fig. 2.14 60 days time series and its matrix "*I*". *Source* Ref. [11]

After obtaining the two matrixes, "*I*" and "*T*", it is used a fit function (2.3) that multiplies the two matrixes, the pattern's matrix "*T*" and the time series matrix "*I*" in order to obtain a value which indicates the level of similarity between them and in this case if the matrix of the time series is similar to the Bull Flag matrix pattern template. Thus, highest values will occur when the matrix "*I*" is in highest conformance with matrix "*T*".

$$\text{Fit}_k = \sum_{i=1}^{10} \sum_{j=1}^{10} (T\,(i,\,j) \cdot I_k\,(i,\,j)) \tag{2.3}$$

Then it is calculated, through the Eq. (2.4), the amplitude of the time series where range is the difference between the highest and the lowest price and p_k is the closing price in day k.

$$\text{Height}_k = \frac{\text{range}_k}{p_k} \tag{2.4}$$

The values obtained by the previous formulas Fit and Height are used to create investment rules. So, when those values are higher than a threshold it generates a buy signal, which is followed by a variable holding period (example: 5, 10, 20, 40, 60, 80,

100 days) until the sell order is generated. The results of applying these investment rules are represented in Table 2.2.

The authors of the previous studies used this pattern identification approach with neural networks and genetic algorithms as optimization methodologies with the goal of create an investment decision model in [12, 13]. The neural network contained 22 input nodes, where 20 of them correspond to the fit values of each column (10 to price and 10 to volume) and the last 2 correspond to the height of price and height of volume. The output of the neural network in one case is a prediction of the future price and in the other case are two confidence values that allow, through a "thresholding" technique, avoiding false buy signals and also reducing the number of wrong entries/exits operations in the market. The genetic algorithms are used in [12] to optimize and identify the subset of 22 input data nodes that should be used in the neural network. The results of these studies are represented in Table 2.2.

The advantages of this methodology are the representation of patterns through matrixes templates is very visually intuitive, also is very efficient to identify simple patterns and the implementation of its method do not required a great complexity.

A disadvantage of this methodology is that using this method it is not possible to represent complex patterns like the Head-and-Shoulders (Fig. 2.10) due to the lack of space in the matrix template. It would be possible to increase the size of the matrix, but the sliding window would also have to increase more. Other disadvantage is the lack of explanation about the technique used to build the template of Fig. 2.13 therefore the template building is a black box for the users. To solve this problem in [14] is described a simple technique to build the matrix template for several types of patterns, where the user only need to put values equal to 1 in each column and the rest of the values of each column are automatic generated because its sum must be equal to 0.

Other disadvantage of this method is related with the weights assigned to the cells of the matrix because, as can be seen in Fig. 2.15, the matrix on top is much more similar to the Bull Flag pattern than the matrix on the bottom, however its fit value (6.5) is lower than the fit value of the bottom matrix (7.5), which can cause the wrong identification of the pattern. To resolve this problem a new Bull Flag template was created in [15] with new weights in each cell of its matrix. The results of this study are represented in Table 2.2.

In [16] the authors used the template pattern approach to represent several patterns in order to be able to identify more and different cases in the historical prices. With the combination of different patterns it is possible to identify more entry and exit points, which creates more complete and robust investment strategies. The Genetic Algorithm was used in this study to optimize important parameters of the process like: sliding window size, noise reduction rate, FitBuy which is the minimum value to generate a buy signal and FitSell which is the minimum value to generate a sell signal. The results of this study outperformed the Buy&Hold strategy and are represented in Table 2.2.

In a recently approach, described in [17], the creation of the Bull Flag template followed a new methodology in order to mitigate the problem identified in Fig. 2.15. Unlike the previous studies [9, 10, 14, 16] that defined the Bull Flag pattern through

0.5	0	-1	-1	-1	-1	-1	-1	-1	0
1	0.5	0	-0.5	-1	-1	-1	-1	-0.5	0
1	1	0.5	0	-0.5	-0.5	-0.5	-0.5	0	0.5
0.5	1	1	0.5	0	-0.5	-0.5	-0.5	0	1
0	0.5	1	1	0.5	0	0	0	0.5	1
0	0	0.5	1	1	0.5	0	0	1	1
-0.5	0	0	0.5	1	1	0.5	0.5	1	1
-0.5	-1	0	0	0.5	1	1	1	1	0
-1	-1	-1	-0.5	0	0.5	1	1	0	-2
-1	-1	-1	-1	-0.5	0	0.5	0.5	-2	-2.5

0.5	0	-1	-1	-1	-1	-1	-1	-1	0
1	0.5	0	-0.5	-1	-1	-1	-1	-0.5	0
1	1	0.5	0	-0.5	-0.5	-0.5	-0.5	0	0.5
0.5	1	1	0.5	0	-0.5	-0.5	-0.5	0	1
0	0.5	1	1	0.5	0	0	0	0.5	1
0	0	0.5	1	1	0.5	0	0	1	1
-0.5	0	0	0.5	1	1	0.5	0.5	1	1
-0.5	-1	0	0	0.5	1	1	1	1	0
-1	-1	-1	-0.5	0	0.5	1	1	0	-2
-1	-1	-1	-1	-0.5	0	0.5	0.5	-2	-2.5

Fig. 2.15 Matrix with fit value of 6.5 (top) and matrix with fit value of 7.5 (bottom)

a consolidation period followed by a strong rise, this approach defines the Bull Flag pattern as a strong rise (first four columns) followed by a consolidation period, illustrated in Fig. 2.16.

Also the values assignment to the matrix is completely different in this case, because there is only one cell with a positive value (first column, last row), which ensures that in order to obtain a positive fit value it is necessary that the prices of a time series pass through this cell. Cells with negative values are areas where prices should not pass through and cells with value equal to 0 are areas where prices can pass because do not affect the fit value.

This method is similar to an if-then rule because, as an example, *if* only the time series with fit value equal or higher than 4 are considered as Bull Flag pattern *then* the prices of the time series have to pass mandatorily through the cell with value 5 and can only pass through one cell with value -1, thereby constraining the values of the eight remaining columns, that must be 0.

0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0
0	0	0	0	-1	-1	-1	-1	-1	-1
0	0	0	-1	-2	-2	-2	-2	-2	-2
0	0	-1	-3	-3	-3	-3	-3	-3	-3
0	-1	-3	-5	-5	-5	-5	-5	-5	-5
0	-1	-5	-5	-5	-5	-5	-5	-5	-5
0	-1	-5	-5	-5	-5	-5	-5	-5	-5
5	-1	-5	-5	-5	-5	-5	-5	-5	-5

Fig. 2.16 New Bull Flag matrix pattern template

In this study, in order to evaluate the return of each operation in the investment strategies, a different sell/exit method was adopted instead of defining a holding period of time to generate the sell order. The sell/exit method is a dynamic method, where the exit point is defined by the evolution of the price and not by time. To do that two variables were defined, *take profit* and *stop loss*, which are related with the maximum amplitude of the time series identified as a pattern that limit the profit and the loss of each operation, respectively. Thus, whenever the price reaches the *take profit* value or the *stop loss* value the operation is closed. The gain at the take profit level is often greater than the loss at the stop loss level so that the total profit of a strategy depends on the success rate of operations. The results of this study are represented in Table 2.2.

2.3.2 Perceptually Important Points (PIPs)

In this approach, as the name implies, the time series are represented by a set of Perceptually Important Points (PIPs) which are the most relevant points because are the ones who characterize the time series and the patterns. Patterns are characterized by a set of critical points, as an example the Head-and-Shoulders pattern can be defined by a head point, two points for the shoulders and two more points for the neckline. These points are the most relevant points because are the ones that define the shape of this pattern. So, in order to identify PIPs in time series, a technique based on distance measures was used in [18]. The algorithm to identify PIPs is described as:

- The sequence P is the set of time series data points. The first and the last point of the sequence P are the first two PIPs identified. The next PIP is the point of sequence P with maximum distance to the first two PIPs. Then, the fourth PIP will then be the point in P with maximum distance to its two adjacent PIPs, i.e., in between the first and the second PIPs or the second and the last PIPs. This process ends when the number of PIPs identified is equal to the number of PIPs of the pattern.

To measure the maximum distance between one point and its two adjacent PIPs, in [18] are presented 3 methods:

1. **Euclidean distance (ED)**

Calculates the sum of the ED (2.5) of the test point p_3 to its adjacent PIPs p_1 e p_2

$$\text{ED}\,(p_3,\, p_2,\, p_1) = \sqrt{(x_2 - x_3)^2 + (y_2 - y_3)^2} + \sqrt{(x_1 - x_3)^2 + (y_1 - y_3)^2} \tag{2.5}$$

2. **Perpendicular Distance (PD)**

Calculates the PD (2.9) between the test point p_3 and the line connecting the two adjacent PIPs p_1 e p_2.

$$\text{Slope}\,(p_1,\, p_2) = s = \frac{y_2 - y_1}{x_2 - x_1} \tag{2.6}$$

$$x_c = \frac{x_3 + s y_3 + s y_2 - s^2 x_2}{1 + s^2} \tag{2.7}$$

$$y_c = s x_c - s x_2 + y_2 \tag{2.8}$$

$$\text{PD}\,(p_3,\, p_c) = \sqrt{(x_c - x_3)^2 + (y_c - y_3)^2} \tag{2.9}$$

3. **Vertical Distance (VD)**

Calculates the VD (2.10) between the test point p_3 and the line connecting the two adjacent PIPs p_1 e p_2.

$$\text{VD}\,(p_3,\, p_c) = |y_c - y_3| = \left| \left(y_1 + (y_2 - y_1)\, \frac{x_c - x_1}{x_2 - x_1} \right) - y_3 \right| \tag{2.10}$$

These three distance methods were tested in [18], using 2500 points of data from Hang Seng Index (HSI), and the Vertical Distance (VD) method proved to be the best in capturing the shapes of patterns.

After the identification of PIPs in time series, the next step is to detect pattern based on this representation. In order to do that two distinct methodologies were used in [19]. The first is based on templates and the second is based on rules.

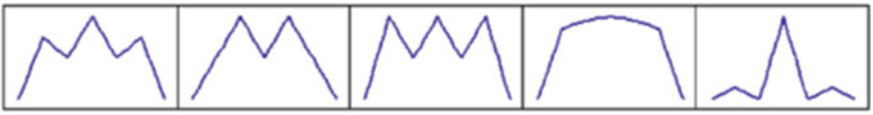


Fig. 2.17 Five typical patterns represented by 7 PIPs. *Source* Ref. [19]

I. Pattern detection based on templates

In this approach, the structure of the patterns is defined visually which allows comparison point-to-point between the time series and the patterns. In Fig. 2.17 it is possible to observe a set of well-known patterns with length equal to 7 PIPs.

As different time series may have different amplitudes, it is necessary to normalize the PIPs identified in the time series in order to facilitate the comparison between the different time series (e.g. range 0–1). After that, the Amplitude Distance (AD) between the pattern's template and time series is calculated through point-to-point direct comparison, Eq. (2.11).

$$\mathrm{AD}\,(\mathrm{SP},\,Q) = \sqrt{\frac{1}{n}\sum_{k=1}^{n}\left(\mathrm{sp}_k - q_k\right)^2} \tag{2.11}$$

The variables SP and sp_k denote the PIPs identified in the time series P and the variables Q and q_k denote the PIPs of the pattern template. It is also necessary to consider the horizontal distortion (time dimension) of the time series against the pattern templates. The Temporal Distance (TD) between P (time series) and Q (pattern template) is defined in Eq. (2.12).

$$\mathrm{TD}\,(\mathrm{SP},\,Q) = \sqrt{\frac{1}{n-1}\sum_{k=2}^{n}\left(\mathrm{sp}_k^t - q_k^t\right)^2} \tag{2.12}$$

where sp_k^t and q_k^t denote the time coordinate of the sequence points sp_k and q_k, respectively. In order to take both horizontal and vertical distortion into consideration in the similarity measure, the formula of this measure is defined as:

$$D\,(\mathrm{SP},\,Q) = w_1 \times \mathrm{AD}\,(\mathrm{SP},\,Q) + (1 - w_1) \times \mathrm{TD}\,(\mathrm{SP},\,Q) \tag{2.13}$$

where w_1 represents the weight of AD and TD that is specified by the users. The results of this methodology are represented in Table 2.2.

II. Pattern detection based on rules

One disadvantage of the template-based methodology is the difficulty of defining the relationship between the relevant points. In this approach, a set of rules between

PIPs is created to describe the shape of the patterns. For example, in the Head-and-Shoulders pattern, the two shoulders must be lower than the point that defines the head and must have a similar degree of amplitude.

Using the patterns from Fig. 2.17 and assuming that all of them have a length of 7 PIPs, sp1 until sp7, a set of rules can be defined for each pattern. The set of rules that define the Head-and-Shoulders pattern are the following:

- sp4 > sp2 e sp6
- sp2 > sp1 e sp3
- sp6 > sp5 e sp7
- sp3 > sp1
- sp5 > sp7
- diff(sp2, sp6) < 15%
- diff(sp3, sp5) < 15%

The set of rules that define the other four patterns of Fig. 2.17 can be found in [19].

After the definition of a set of rules for each pattern, the time series are represented by its PIPs, in this case by 7 PIPs, and those who comply with all the rules of a pattern are identified as one. The results of this approach are represented in Table 2.2 and in general, were lower than the results of the previous approach (template-based). However, this approach obtained excellent results in the distinction between the Head-and-Shoulders pattern, Triple Top pattern and Double Top pattern. The advantages of this new approach, i.e. PIPs representation and detection of patterns based on templates or rules are:

- High complexity reduction of time series and patterns because only a small set of points are used to represent time series and identify patterns.
- Possibility to detect complex and detailed patterns.

The main disadvantage of this approach is related with the detection of patterns, where the number of PIPs that define the patterns and the time series must always be equal in order to enable the comparison point-to-point in the template-based methodology or the validation of rules in the rule-based methodology. For example, the Head-and-Shoulders pattern of Fig. 2.17 is represented by 7 PIPs, which force the time series to be represented by the same number of PIPs to be possible to compare them with the pattern. To resolve this problem a DTW (Dynamic Time Warping) algorithm was used in [20] to find an optimized alignment between two sequences combined with the three distance measures described previously (ED, PD and VD). Using this algorithm it is possible to measure the similarity between a pattern and a time series with different lengths of PIPs.

2.3.3 *Symbolic Aggregate approXimation (SAX) Representation*

Traditionally, the representation of time series and its dimensionality reduction was made through numeric methods like Wavelet Discrete Transform [21]. The SAX representation approach [22] allows defining metrics between data representation, which is related with the real distance between time series. The SAX representation solves the problem related with the distance between the real data and its representation because it is possible to obtain a lower bounding approximation for the distance measures and also this representation allows a significant reduction of the data dimensionality of time series.

In [23] was used a method based on SAX to represent time series with the aim of identifying patterns, which begins by dividing larger time series in smaller time series windows. The data of each smaller time series is then normalized, according to Eq. (2.14), to guarantee that the time series can be compared between each other. In Eq. (2.14), x_i corresponds to a point of the time series, μ_x and σ_x correspond to the mean and standard deviation of the time series, respectively.

$$x_i' = \frac{x_i - \mu_x}{\sigma_x} \tag{2.14}$$

After the normalization of data it is necessary to reduce its dimensionality in each time series and to do that the Piecewise Aggregate Approximation (PAA) method was used [24]. With PAA a time series is divided in equal segments, where each of them is represented by its arithmetic mean (2.15).

$$\bar{x}_j = \frac{w}{n} \sum_{i=\frac{n}{w}(j-1)+1}^{\frac{n}{w}j} x_i' \tag{2.15}$$

where w represents the number of segments, n represents the time series size and x_i' is the data point in the window. As can be seen in Fig. 2.18, this process allows the representation of a time series by the arithmetic mean of each segment, which makes a set of data points to be now represented only by its mean (red line).

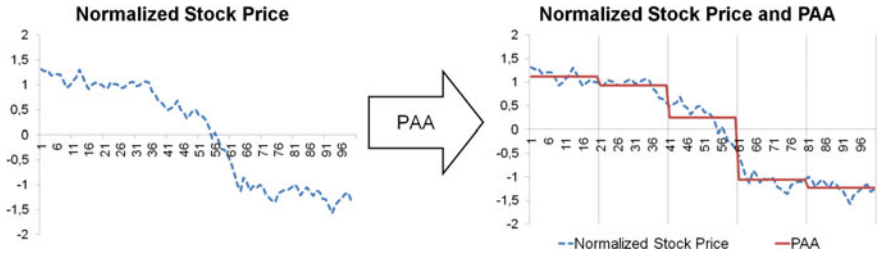


Fig. 2.18 PAA representation. *Source* Ref. [23]

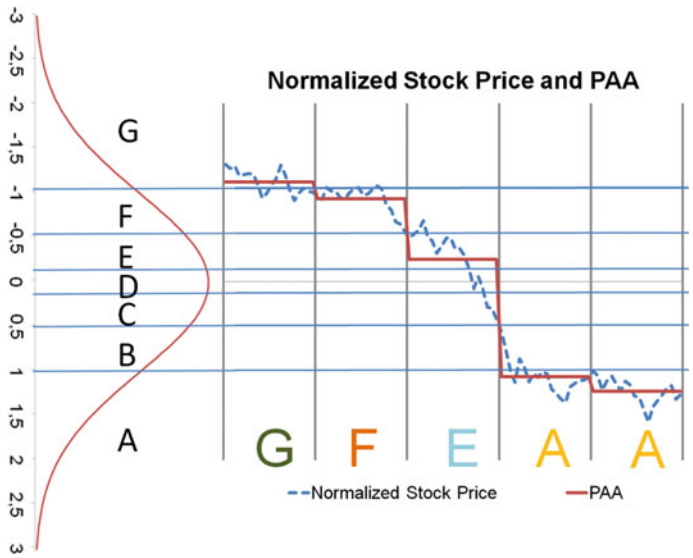


Fig. 2.19 SAX representation. *Source* Ref. [23]

Table 2.1 Breakpoints for a intervals of the normal distribution curve

	$a = 3$	$a = 4$	$a = 5$	$a = 6$
β_1	−0.43	−0.67	−0.84	−0.97
β_2	0.43	0	−0.25	−0.43
β_3		0.67	0.25	0
β_4			0.84	0.43
β_5				0.97

Source Ref. [23]

After the application of PAA, the amplitude of each time series is divided in equiprobable intervals, using a normal distribution curve over the vertical axis, where breakpoints are calculated to produce equal areas for each interval under the curve. Thus, it is possible to assign a different symbol to each interval and consequently assign a symbol to each segment by determining to which interval the segment belongs, as illustrated in Fig. 2.19. Applying this method to all segments of a time series allows the representation of the time series by a sequence of SAX symbols (string).

To find patterns with this approach, the SAX sequences of symbols must be compared with each other or compared with well-known SAX sequence that defines some wanted pattern. To measure the similarity between SAX sequences two distance measures were used: MINDIST (2.17) and ALPHAB.DIST (2.18). The ALPHAB.DIST method proved to be faster than the MINDIST due to the fact that does not need to identify the breakpoints of Table 2.1.

Fig. 2.20 Chromosome used in GA. *Source* Ref. [23]

$$\text{dist}\,(p_i, q_i) = \begin{cases} 0 & \text{se } |i-j| \leq 1 \\ \beta_{j-1} - \beta_i & \text{se } i < j-1 \\ \beta_{i-1} - \beta_j & \text{se } i > j+1 \end{cases} \tag{2.16}$$

The β's are the breakpoints defined in Table 2.1.

$$\text{MINDIST}\,(P, Q) = \sqrt{\frac{n}{w}} \sqrt{\sum_{i=1}^{w} (\text{dist}\,(p_i, q_i))^2} \tag{2.17}$$

$$\text{ALPHAB.DIST}(T, P) = \sqrt{\sum_{i=1}^{W} (T_i - P_i)^2} \tag{2.18}$$

where Ti e P_i are the symbols i of the sequence T and P, respectively.

An advantage of SAX representation is the simplicity to identify patterns because in this approach the identification is simply a comparison between two sequences of symbols, i.e. two strings. Other advantage is the simple implementation of this methodology and also the transformation of time series in SAX sequences of symbols is fast. The main advantage is that this approach allows a huge reduction of dimensionality of data and at the same time maintains the main characteristics of time series and patterns.

The authors of this study [23] also used genetic algorithms to optimize the investment strategies based on the total return. In Fig. 2.20 it is possible to observe the chromosome used and its genes. This chromosome is divided in 2 parts, in the first (first four genes) are the parameters that support the buy/sell decisions and in the second ($P_1 - P_w$) is the sequence of symbols that define the pattern, where each gene defines a symbol. The first two genes define the distances between the time series and the pattern that allow to identify if the pattern is presented and a buy order should be generated (Distance to Buy) or if it is not present and a sell order should be generated (Distance to Sell). The third gene (Days to Sell) defines the holding period to maintain the financial asset until it is sold. The fourth gene (Measure Type) defines which distance measure (MINDIST and ALPHAB.DIST) should be used to find the similarity between sequences. The results of this study are presented in Table 2.2.

Table 2.2 Results comparison of some studies presented [25]

Ref.	Year	Method	Used Data	Period		Financial Market	Algorithm Performance		Buy-and-Hold Performance	
[11]	2008	Bull Flag w/Matrix Template	Stock price	04/08/1967–12/05/2003		NYSE Composite Index	4.59% (Transaction average over the period)		1.83% (Transaction average over the period)	
[15]	2007	Bull Flag w/Matrix Template	Stock price	NASDAQ	TWI	NASDAQ and TWI	NASDAQ	TWI	NASDAQ	TWI
				03/04/1985–20/03/2004	01/06/1971–24/03/2004		4.38% (Transaction average over the period)	6.48% (Transaction average over the period)	3.27% (Transaction average over the period)	4.03% (Transaction average over the period)
[13]	2002	Hybrid Neural Network w/Pattern detection	Stock price and Vol.	24/07/1984–11/06/1998		NYSE Composite Index	66% (Days market goes up after buying order)		60% (Days market goes up after buying order)	
[17]	2015	Bull Flag w/Matrix Template	Stock price	22/05/2000–29/11/2013		Dow Jones Industrial Average Index	13% (Average return)		N/A	

(continued)

Table 2.2 (continued)

Ref.	Year	Method	Used Data	Period	Financial Market	Algorithm Performance	Buy-and-Hold Performance
[16]	2011	Uptrend pattern w/Matrix Template + GA	Stock price	1998–2010	S&P500 Index	36.92% (Total return)	−4.69% (Total return)
[19]	2007	Template-Based	Stock price	N/A	Several	96% (Hits on pattern identification)	N/A
[19]	2007	Rule-Based	Stock price	N/A	Several	38% (Hits on pattern identification)	N/A
[19]	2007	PAA	Stock price	N/A	Several	82% (Hits on pattern identification)	N/A
[23]	2013	SAX + GA	Stock price	1998–2010	S&P500 Index	16.28% (Average annual return)	7.79% (Average annual return)

References

1. Helfert, E.: Financial Analysis Tools and Techniques: A Guide for Managers. McGraw-Hill Education (2001)
2. Kirkpatrick II, C.D., Dahlquist, J.R.: Technical Analysis: The Complete Resource for Financial Market Technicians, 2nd edn. (2010)
3. Murphy, J.J.: Technical Analysis of the Financial Markets: A Comprehensive Guide to Trading Methods and Applications (1999)
4. Bulkowski, T.N.: Encyclopedia of Chart Patterns, 2nd edn. Wiley (2005)
5. Colby, R.W.: The Encyclopedia of Technical Market Indicators. McGraw-Hill (2003)
6. Lin, L., Cao, L., Wang, J., Zhang, C.: The Applications of Genetic Algorithms in Stock Market Data Mining Optimization (2000)
7. Chen, S.H.: Genetic Algorithms and Genetic Programming in Computational Finance. Springer, Boston (2002)
8. Pinto, J., Neves, R., Horta, N.: Fitness function evaluation for MA trading strategies based on genetic algorithms. In: Proceedings of the 13th Annual Conference Companion on Genetic and Evolutionary Computation, pp. 819–820 (2011)
9. Leigh, W., Modani, N., Purvis, R., Roberts, T.: Stock market trading rule discovery using technical charting heuristics. Expert Syst. Appl. **23**(2), 155–159 (2002)
10. Leigh, W., Paz, N., Purvis, R.: Market timing: a test of a charting heuristic. Econ. Lett. **77**(1), 55–63 (2002)
11. Leigh, W., Frohlich, C.J., Hornik, S., Purvis, R., Roberts, T.: Trading with a stock chart heuristic. Syst. Man Cybern. Part A Syst. Hum. **38**(1), 93–104 (2008)
12. Leigh, W., Purvis, R., Ragusa, J.M.: Forecasting the NYSE composite index with technical analysis, pattern recognizer, neural network and genetic algorithm: a case study in romantic decision support. Decis. Support Syst. **32**(4), 361–377 (2002)
13. Leigh, W., Paz, M., Purvis, R.: An analysis of a hybrid neural network and pattern recognition technique for predicting short-term increases in the NYSE composite index. Omega **30**(2), 69–76 (2002)
14. Wang, J., Chan, S.: Trading rule discovery in the US stock market: an empirical study. Expert Syst. Appl. **36**(2), 5450–5455 (2009)
15. Wang, J., Chan, S.: Stock market trading rule discovery using pattern recognition and technical analysis. Expert Syst. Appl. **33**(2), 304–315 (2007)
16. Parracho, P., Neves, R., Horta, N.: Trading with optimized uptrend and downtrend pattern templates using a genetic algorithm kernel. In: IEEE Congress on Evolutionary Computation, pp. 1895–1901 (2011)
17. Cervelló-Royo, R., Guijarro, F., Michniuk, K.: Stock market trading rule based on pattern recognition and technical analysis: forecasting the DJIA index with intraday data. Expert Syst. Appl. **42**(14), 5963–5975 (2015)
18. Fu, T., Chung, F., Luk, R., Ng, C.: Representing financial time series based on data point importance. Eng. Appl. Artif. Intell. **21**(2), 277–300 (2008)
19. Fu, T., Chung, F., Luk, R., Ng, C.: Stock time series pattern matching: template-based vs. rule-based approaches. Eng. Appl. Artif. Intell. **20**(3), 347–364 (2007)
20. Tsinaslanidis, P.E., Kugiumtzis, D.: A prediction scheme using perceptually important points and dynamic time warping. Expert Syst. Appl. **41**(15), 6848–6860 (2014)
21. Ni, H.: Profitability of technical chart pattern trading on FX rates: analyzed by wavelet transform. In: Third International Symposium on Intelligent Information Technology Application, pp. 138–141. Nanchang (2009)
22. Lin, J., Keogh, E., Wei, L., Lonardi, S.: Experiencing SAX: a novel symbolic representation of time series. Data Min. Knowl. Disc. **15**(2), 107–144 (2007)
23. Canelas, A., Neves, R., Horta, N.: A SAX-GA approach to evolve investment strategies on financial markets based on pattern discovery techniques. Expert Syst. Appl. **40**(5), 1579–1590 (2013)

24. Keogh, E., Chakrabarti, K., Pazzani, M., Mehrotra, S.: Dimensionality reduction for fast similarity search in large time series databases. J. Knowl. Inf. Syst. **3**(3), 263–286 (2000)
25. Leitão, J., Neves, R.F., Horta, N.: Combining rules between PIPs and SAX to identify patterns in financial markets. Expert Syst. Appl. **65**, 242–254 (2016) (Reprinted with permission from Elsevier)

Springer

http://www.springer.com/978-3-319-70159-2

Identifying Patterns in Financial Markets

New Approach Combining Rules Between PIPs and SAX

Leitão, J.; Neves, R.F.; Horta, N.C.G.

2018, XVII, 66 p. 69 illus., Softcover

ISBN: 978-3-319-70159-2