The Application of difference distance measures to cluster financial time series

Research done by Nikita Naychukov

Faculty advisor: Dr. Albert Armisen

Table of contents

Abstract

Financial data, as time series of stock prices and similar, have permanently been assessed as very noisy, disorganized, and non-stationary and, therefore tend to be unpredictable by most conservative models in the light of which financial time series do not contain any patterns to follow. Moreover, most modern researches attempting to reveal some patterns found only weak evidences of patterns existence. Hence, the question whether financial data contain any significant regularities is not resolved yet by using traditional and modern methods which provokes a need in more advanced search instruments. In recent times advanced techniques in time series pattern recognition and image classification become more and more popular among scientists which can also be adapted for research in financial time series. These modern techniques include “shapelets” recognition, Dynamic Programming-based partial matching algorithms, up-to-date similarity-invariant pictures’ descriptors which allows to compare objects invariant to rotation, translation and scale, and fast k-medoids clustering. By reasonably applying them it is possible to test whether these tools could help improve regularities search in financial time series.

In this thesis the main hypothesis tested is whether the model would produce good goodness of fit results comparing different parameters, such as a number of iterations, clusters, objects and different distance functions. The model used is an automated pattern recognition algorithm which employs up-to-date clustering techniques. These techniques encompass machine learning vehicles which allow the algorithm to identify patterns and deliver them into classes without human interruption. The dataset taken encompasses thirty stocks from Dow Jones Industrial Average index (DJIA). To understand how good clustering is performed the goodness of fit approach of Tibshirani & Walther is used.

Chapter 1. Introduction

Section 1.1. General introduction to the topic under study

The issue of pattern recognition in financial data remains a quite important task and even becomes increasingly significant in recent times with regards to practical applications and academia.

With respect to the applied side of the task, it is necessary to appeal to technical analysis (TA) which is forecasting stock further movements relying on past stock prices, volumes and other market data. Generally, this technical approach to financial analysis unfolds an idea that a series of stock prices represents a particular shape determined by investors' attitudes to different economic, social and political factors (Pring, 2002). Technical analysis mostly consists of identifying distinct patterns in financial time series which basically represent some kind of a figure, an object, such as "Double Tops" or "Cup and Handle". Relying on such patterns investors can suppose future price movements which presumably should open a huge potential to make profit.

However, the main problem remains unsolved which is to accurately recognize such a pattern. There is an important research on this topic in which TA is testified using programming algorithms and statistical tests to retrieve specific regularities and prove their existence in financial time series (Lo & Hasanhodzic, 2010). The results of these studies state that there are patterns which can be identified with statistical significance and this motivates for further research in this field. Furthermore, it is necessary to take into considerations possible clustering algorithm optimizations which can be done to improve a pattern recognition process. The other point which makes this issue actual from applied perspective is that current traditional TA methods are not scientific (they are based on a subjective judgment of an investor who relies

on visual tools to assess the data), while there is a need to scientifically prove that financial time series contain statistically significant information, such as previously defined patterns.

From the perspective of academic finance, it is necessary to pay attention to classical tools for financial analysis, such as regressions of various types (linear regression models, non-liner regressions, non-parametric regressions, autoregressive models, such as AR, MA, ARMA, ARIMA, and volatility clustering models, such as ARCH, GARCH, etc.) which provide a framework to account for different factors influencing financial data (Mills & Markellos, 2008; Tsay, 2005). However, these models are not specifically utilized for pattern recognition tasks, they are commonly used in order to account for some determinants of future price movements. Furthermore, these methods are quite sophisticated in practical implementation when tackling with actually huge datasets. They perform poorly in those cases because of their limiting assumptions and inability to capture all possible price factors. Alternative traditional frameworks to analyze market inefficiencies are CAPM and APT. These tools are more accurate than former ones in pattern retrieval because both techniques identify so called “arbitrage patterns” capturing risk-free profits and reflecting market inefficiencies. However, these models appear to be far from reality because of restrictive assumptions and too noisy end-of-day prices which are taken as model inputs (Fama & French, 2004).

Therefore, summing up all the considerations, a new pattern recognition technique for financial time series would have scientific and applied utility. In perspective, this new model might increase efficiency in pattern recognition search algorithm which delivers the opportunity to make profit using those patterns for further predictions. The fields for application of the model are very broad encompassing active traders, hedge funds, mutual funds, asset management companies, pension fund, etc. From the academic viewpoint, the new pattern recognition model, that accurately would identify shape patterns (market inefficiencies), would improve the market efficiency theory by equipping traditional methods, which are not targeted

at pattern recognition, with notion that financial markets contain some statistically significant patterns.

Section 1.2. Literature review of the topic

Technical analysis (TA) took its roots from Japan of XVIII century where it was invented by Munehisa Homma in a form of a candlestick chart representing ongoing prices for rice. Homma identified several regularities in rice prices, such as the prices go down after an increasing market period (bull market) and go up after a decreasing market period (bear market). This founding helped him make profits by utilizing his knowledge and helped him become one of the first financial advisers. Next important period in TA development devoted to Dow's axioms which were released at the end of XIX. Charles H. Dow argued that prices reflect the fundamental value of the underlying and if someone wants to predict the market, it is sufficient to consider current market trends. In XX century, TA disseminated among sophisticated investors who were greatly involved in building the market applying their models and trying forecast the trend. Such popularity triggered the emergence of studies which were aimed at accumulating numerous TA concepts into an ordered structure which gave birth to the first rules of technical trading (Schabacker, 1930). At this point TA reached the pick of its popularity and became a routine instrument for many investors all around the globe. Basically, TA rules were summarized as follows (Bechu & Bertrand, 1999): a traditional approach which treats financial time series as a collection of visual objects and studies shapes of resulting figures; a contemporary approach which relies mostly on quantitative methods, such as different variations of moving average (Taylor & Allen, 1992); a psycho-social approach which aims to describe price movements through identifying behavioral habits of traders. Generally accepted that it is useful to combine all these TA approaches together in order to describe a pattern, a

trend to a maximum extent, and then to use this information for forecasting purposes. Today many practitioners use popular computer applications to extract this information, such as UBS' online service 'KeyInvest", or well-known cloud service "TradingView".

TA has been very popular since XX century till today, namely enormous amounts of investors (pension funds, mutual funds, hedge funds, individual traders) used TA for trading purposes on numerous markets, such as currency (FOREX), commodity, and stock markets, (Lui & Mole, 1998; Cheung & Chinn, 2001). Although all these TA applications had a place, there were not much research on this topic due to a high portion of skepticism devoted to TA in the academia world. This prejudice to TA has its roots in the concept of market efficiency according to which current prices already reflect the fundamental value of the underlying by containing all the information about the underlying (Malkiel & Fama, 1970). Market efficiency was structured to have three forms: weak form efficiency which is described by the situation when current prices include all historical information, semi-strong form which is described by the situation when prices reflect all current public and historical information, and strong form which is described by the situation when prices reflect all available information including all private sources (Jensen, 1978). Following this logic, if markets are efficient and prices represent the fundamental values of the underlying, then there cannot be any patterns in financial times series which would allow exploit the situation and extract a risk-free profit. Therefore, this discourse makes possible to shed light on the importance of the goal stated in this thesis to produce a pattern recognition algorithm which would allow to elaborate more on the notion of market efficiency. Nevertheless, this research has not the goal to reject the market efficiency hypothesis. The presence of patterns would say only about that at some point in time markets are biased or imperfect which is generally should be true. That may happen due to different reasons, such as failure of regulators to make market cleared out to encompass all possible information.

Although research potential of this topic has recently been discovered, there are already some studies in this field. One of the first notable researches “The Predictive Significance of Five-Point Chart Pattern” (Levy, 1971). Levy testified prediction potential of thirty-two five-point patterns grounded on extrema exploration in financial time series. The set of data for his research encompassed prices of five hundred stocks (from 1964 to 1969) which were traded on New-York Stock Exchange (NYSE). The hypothesis of his research was expected to be true with regards to that identified patterns appeared to be not significant, and there was no opportunity to make arbitrage or predict future movements. Another important study was conducted on prediction graphical models for exchange rates changes (Chang & Osler, 1999). The work produced the result proving that exchange rates movements contain some inefficiencies produced my market imperfections, so that it is possible to identify some regularities and try to forecast future changes. The pattern they identified got the name “Head and Shoulders”, it was explored on the dataset encompassed daily spot exchange rates of six currencies – Canadian dollars, British pounds, German marks, Japanese yens, francs, and Swiss francs – of the time from 1973 to 1994. Chang and Osler applied a bootstrap method (Efron, 1982) to understand if the identified patterns can be treated as statistically significant. This bootstrap tool is a simulation of random data points with the same characteristics (length, mean, variance, skewness and kurtosis) as the original time series, and application of the same exact model which to these data points. Generally, Change and Osler wanted to identify “Head and Shoulders” on the original data points and on the random set, thereafter comparing how often they can find the pattern using a goodness of fit measure. The test showed that there is a significant difference between marks and yens distributions. One more study of the topic was performed where constituents of S&P500 (from 1992 to 1996) had been examined (Caginalp & Laurent, 1998). Caginalp and Laurent identified many reversal regularities (prices revert their direction after a pattern shape is realized) and proved their statistical significance.

The most profound work in the field is “Foundation of Technical Analysis: Computational Algorithms, Statistical Inference, and Empirical Implementation” (Lo, Mamaysky, Wang, 2000). This is the first research exploring technical analysis, which prominently changed academia’s attitude to technical analysis. Lo, Mamaysky and Wang examined a great number of different technical patterns using an automated algorithm which was capable to catch “Head and Shoulders” and its invers version, “Broadening tops/bottoms”, “Triangle tops/bottoms”, “Rectangle tops/bottoms”, and “Double tops/bottoms”. Lo, Mamaysky and Wang devised an ordered system of financial pattern recognition applying a nonparametric kernel regression and an automated technical trends search algorithm on the dataset of U.S. stocks from 1962 to 1996. They used Kolmogorov-Smirnov test to verify if their algorithm can produce statistically significant patterns. They compared an unconditional empirical distribution of daily stock returns with a posterior distribution of technical analysis shapes, such as “Head and Shoulders” or “Double Tops”. The test proved that five (“Head and Shoulders”, “Broadening Bottoms”, “rectangle bottom/tops”, “Double Tops”) of ten preprogrammed technical patterns are embedded in financial time series under consideration. Hence, the main conclusion was that such discovery can provide a significant sign for investors to make profits by exploiting these inefficiencies.

Ultimately, all these studies provided a background to start building further research that could possibly reinforce conclusions of patterns’ significant persistence in financial time series. Furthermore, most researches made a conclusion that financial time series data should contain patterns which means that continuous investigation in this filed apparently can produce fruitful results. However, despite the fact that previous research proved that patterns existed in financial data, each prior work is limited to specific markets, particular shapes, or time horizons. Moreover, previous research conclusions are not strong enough to sum up that patterns investigated are persistent and, therefore exist in financial time series in general.

Hence, this thesis proposes alternative approaches that will be used to identify any patterns in financial times series. These alternative approaches use modern techniques borrowed from areas, such as image classification and general time series pattern recognition. Mentioned tools extensively use machine learning which apparently could increase efficiency of a patterns clustering algorithm. Innovative methods proposed would make possible to solve out the problem of particular shape, markets and time horizons specifications.

Therefore, it is important to shed some light on this filed by introducing relevant works on the topics. The motivating article for this thesis is "Time Series Shapelets: A New Primitive for Data Mining" (Keogh & Ye, 2009). Keogh and Ye described a new methodology for time series clustering, they invented the concept of "shapelet" (the part of a time series which is unique to a specific class of time series). The noticeable example in their work is the problem to cluster leaves of two flower species. The leaves are quite noisy in shape, unstructured and almost identical (their form and color are indistinguishable by a human eye in most cases). Therefore, there is no reason to take into account the entire set of points (which is a restructured reflection of a leave's edge shape), and a reasonable solution to cluster these leaves is to point out some unique piece of a whole leave's edge shape which is different for two species. Consequently, the leaves are depicted into one dimension such that they begin to represent a time series which is an unfolded reflection of a leave's edge shape. Then, a small part of this time series is identified which is the same for all leaves of one species and different for all leaves of the other species. Exactly this part is called a "shapelet" and it represents a unique feature of a particular species. Innovation of this methodology is its unique identifier called "shapelet" which allows to consider only a tiny interval to cluster a whole time series. Keogh and Ye checked their model on different datasets, such as history objects consisting of arm coats, heraldic shields, projectile points (arrowheads), and other clusterable figures. "Shapelets"

technique allowed to reduce clustering time and space, produced more accurate and interpretable classifications against state-of-the-art clustering methods.

"Shapelet's" study is important for this thesis in three respective criteria: Keogh and Ye researched general time series pattern recognition and clustering which is similar to the stated problem of financial time series pattern recognition; authors worked with noised time series which is close to financial time series; Keogh and Ye proposed an idea for recognition of "shapelets" which can be taken into account to this thesis. Nevertheless, the problem of this research is different from what had been proposed by Keogh and Ye: the task for this thesis is not to cluster a whole time series by a distinct subsequence, but to recognize repeating parts (patterns) of time series mix. Therefore, there is no need to fully accustom the proposed methodology, however it would be valuable to acquire some tips for building a clustering algorithm and preparing a dataset from Keogh and Ye.

Another work which is valuable to consider in this research is "Clustering Time Series using Unsupervised-Shapelets" (Zakaria, Mueenm, Keogh, 2012). The study introduced a novel concept: "u-shapelet". "U-shapelet" is the same unique part as "shapelet", but it can be identified without prior knowledge of a clustering distribution, so "U" means "unsupervised". Unsupervised property is achieved by calculating distance "u-shapelets" to any other time series in a set. That allows to deliver objects into respective clusters measuring their distances to "u-shapelets". In turn, "u-shapelets" initial pick is processed by maximizing the gap between different clusters, such that objects in one cluster are closer to "u-shapelet" of the cluster. This new method produced more accurate clustering results within the group of unsupervised clustering techniques.

The ultimate study on clustering models to introduce is "Scalable Clustering of Time Series with U-Shapelets" (Ulanova, Begum, Keogh, 2015). In this research a hash-based

algorithm “scalable u-shapelet” was introduced which allowed to perform the same unsupervised clustering by “u-shapelets” faster by two orders of magnitude.

It is also important to cover the area of image recognition and classification, which proposes appropriate distance measures for the clustering problem stated in this thesis. Hence, it is necessary to shed light on recent works in this sphere. The study which ideas borrowed in this thesis is “Sketch-based Image Retrieval from Millions of Images under Rotation, Translation and Scale Variations” (Parui & Mittal, 2015). This article is valuable for this research because the main problem is to recognize patterns of shape in stock prices. Consequently, stock prices series can be depicted as sketch-based pictures. Hence, it is need to be acquainted with contemporary image clustering methods. Mittal and Parui produced a distance measure which is rotation, translation, and scale invariant. This distance precisely fits the purpose of this thesis because the task of this work includes to allocate space-different patterns in one cluster. These patterns may be the same in nature but different in a geometric location, such as “Inverted Head and Shoulders” and general “Head and Shoulders”. Mittal and Parui proposed to represent a shape as chains of segments which can be described analytically by similarity-invariant length descriptors. Applying these descriptors, it is possible to evaluate a similarity score among objects even of different lengths by virtue of Dynamic Programming-based partial matching algorithm. Mittal and Parui also proposed to cluster objects using k-medoids clustering. The results of their study implied superiority of their method over rival state-of-the-art techniques.

Section 1.3. Presenting the research gap

The dissemination of advanced pattern recognition tools and their success in many fields convince us to take them into consideration in order to decide which method produces the best

results. Moreover, considering the nature of data sets served for listed studies and financial data sets which are going to be used in this work, it is worth to exploit these tools for pattern recognition in finance. Taking into account that these advanced techniques had not been previously tried with regards to financial time series and that previously proposed automated technical analysis algorithms produced poor results, in this work will be proposed a model which compiles all of these tools and compares them in terms of goodness of fit. However, simply gathering all these into one methodology does not fit the needs. This way, the data sets (which includes just end day share prices) employed in the article of Lo, Mamaysky and Wang are too noisy for pattern recognition problem. Furthermore, their limitation of using just ten predefined figures of technical analysis greatly bounds the investigation of whether financial time series may be inefficient in a way of containing some patterns. The methodology inspected in the studies (Keogh, 2009, 2012, 2015) gives not good enough clustering results if it is applied as a standalone technique. That happens because financial data is over-noised and may even to to contain any patterns we are looking for. Ultimately, proposed picture clustering tools (Mittal, 2015) have not been tested for time series classification. They measured a similarity score (not a distance) which is exponentially defined. That peculiarity makes Mittal's model is quite time consuming. Furthermore, his Dynamic Programming-based partial matching algorithm was purposed for non-continuous objects, which places a need to adapt the algorithm to the problem of continuous shape comparison. Hence, the model used in this thesis is somewhat different from the proposed ones. It takes high and low daily prices instead of close day prices. That makes data more cleared out of excess noise transferring more clean information. The algorithm for this thesis does not limit a search task to some predefined number of patterns, on the contrary it allows to identify patterns based on the quality of clustering. The measure used to identify the quality of clustering represents a goodness of fit approach in clustering tasks which allows to define the best parameters (a number of algorithm's iterations, clusters, objects and a distance

function) for the model (Tibshirani & Walther, 2005). Firstly, the algorithm employed a modified version of Mittal's distance measure, which allowed to have a comparison of different length objects by virtue of Dynamic Programming matching algorithm. However, we identified that our modified distance measure did not satisfy the triangular inequality which creates a problem due to inability to compare distance between objects in this case. To solve out this issue it is necessary to come up with a number of different functions which would allow to compare objects with the best goodness of fit measure. For this we were trying a series of functions proposed to measure the distance between objects like we have (Xu & Xia, 2011). Summing up, the model which was developed for this study allows to identify the best bunch of parameters in order to perform the tasks for time series pattern recognition by employing several machine learning techniques that learns to identify and cluster various patterns in financial time series without human interaction.

Ultimately, in this thesis the main hypothesis tested is whether the model would produce good goodness of fit results comparing different parameters, such as a number of iterations, clusters, objects and different distance functions. The model used is an automated pattern recognition algorithm which employs up-to-date clustering techniques. These techniques encompass machine learning vehicles which allow the algorithm to identify patterns and deliver them into classes without human interruption. The dataset taken encompasses thirty stocks from Dow Jones Industrial Average index (DJIA). To understand how good clustering is performed the goodness of fit approach of Tibshirani & Walther is used.

Chapter 2. Method

Section 2.1 Sample

In this thesis the research is done on 30 U.S. stocks from Dow Jones Industrial Average Index of the timeframe from the 27th June of 2016 to the 27th June of 2017. Yahoo Finance is used as a source of information to get the abovementioned data. This choice is based on a number of R packages which are available to get the data from the internet and to download the dataset right away to a working environment. These 30 U.S. companies are the best choice for the task, because they represent the most profound areas of U.S. economy, and their stocks have different liquidity, such that it is possible to obtain the most reliable results (liquid stocks seldom contain any patterns, while illiquid ones contain them more often).

The original set of data contains daily information on 6 variables: an open price, a high price, a low price, a close price, a volume, and an adjusted close price for the indicated timeframe. Generally, most previous papers employed close prices to investigate patterns, but it would be more accurately to use high and low prices, because these variables allow to distinguish noise from real price movements with more precision. Hence, in this thesis those prices are employed to account more information. Furthermore, this approach allows to adjust price movements for volatility through the normalization process which will be discussed later on.

Section 2.2. Data collecting instruments

The data has been collected through an open-source Yahoo service which is called Yahoo Finance as previously mentioned. This service allows users to search, display and download

any information regarding financial markets including a number of parameters each stock possesses. Namely, “quantmod” R package was employed to acquire all the necessary information and store it into the R environment. This package contains a very useful function “getSymbols” which allows to indicate stock names, a source and a period of time a user wants, and to download the indicated information into R.

Section 2.3. Procedure

All this research is done with help of R programming language using “R studio” and can be structurally separated into the following steps: acquiring data from Yahoo Finance, originating the dataset, cleaning the dataset with the linear approximation process, and applying a clustering algorithm to the dataset. First several steps were already covered in previous sections. They are basically only preliminary stages consisting of preparing the data, while the clustering algorithm is the main idea of this thesis. This algorithm is targeted at searching patterns in financial time series using machine learning techniques.

To start with, we consider that the model of this thesis is the process of clustering enhanced by goodness of fit parameters determination, and these two terms are interchangeable. First of all, it is necessary to provide a clear description of the model in order to being guided through this structure further after. Generally, the model is organized as follows: it takes the dataset we prepared, measures distances, clusters objects from the dataset, and, finally it allows to adjust model’s parameters by applying a goodness of fit measure in order to obtain the best clustering results. More precisely, the clustering algorithm is more complicated and it works in a different way, but for this stage of thesis this description is enough to get a clear sight at the model to follow for further discussion. The sequence of operations in the algorithm is not accidental. In order to cluster very similar objects, it is necessary to define some measure of

their difference which we call distance. This distance must be negligible for similar objects and large enough for different ones.

Having the dataset prepared, the first basic step of the model is to calculate a distance between two objects which is the other way of defining how similar two objects are. Nevertheless, there is no way to determine which distance function to use before having any results. Therefore, we are trying to apply different distance functions which are commonly used in the field, and, then to determine which fits best according to goodness of fit test. The distance functions we used are: Euclidian distance (see Equation 1), Manhattan distance (see Equation 2), Minkowski distance (see Equation 3), Cosine similarity (see Equation 4), Jaccard similarity (see Equation 5), Hellinger distance (see Equation 6), Bhattacharyya distance (see Equation 7), Modified Mittal's distance measure (see Equation 8). However, all other parameters are also needed to be examined according to the goodness of fit test, but we will talk about them later when the clustering process will be described.

$$\sum_{i=1}^{n} \left(x_i - y_j\right)^2, \tag{1}$$

where x_i and y_j - are respective vectors' coordinates.

$$\sum_{i=1}^{n} |x_i - y_j|, \tag{2}$$

where x_i and y_j - are respective vectors' coordinates.

$$\left(\sum_{i=1}^{n} |x_i - y_j|^p\right)^{\frac{1}{p}}, \tag{3}$$

where x_i and y_j – are respective vectors' coordinates, p – being 1 or 2, Manhattan and Euclidian distances respectively.

$$\frac{\sum_{i=1}^{n} x_i y_i}{\sqrt{\sum_{i=1}^{n} {x_i}^2}\sqrt{\sum_{i=1}^{n} {y_i}^2}}, \tag{4}$$

where x_i and y_j – are respective vectors' coordinates.

$$\frac{\sum_{i=1}^{n} \min(x_i, y_i)}{\sum_{i=1}^{n} \max(x_i, y_i)}, \tag{5}$$

where x_i and y_j – are respective vectors' coordinates.

$$(\sum_{i=1}^{n} (\sqrt{x_i} - \sqrt{y_j})^2)^{\frac{1}{2}}, \tag{6}$$

where x_i and y_j – are respective vectors' coordinates.

$$\frac{1}{4}\ln\left(\frac{1}{4}\left(\frac{{\sigma_y}^2}{{\sigma_x}^2} + \frac{{\sigma_x}^2}{{\sigma_y}^2} + 2\right)\right) + \frac{1}{4}\left(\frac{(\mu_x - \mu_y)^2}{{\sigma_y}^2 + {\sigma_x}^2}\right), \text{ where} \tag{7}$$

${\sigma_y}^2$- is the variance of the y-th vector, μ_y- is the mean of the y-th vector,

x, y – are two vectors.

$$D_{jnt}(x, y) = D_{lr}(x, y) + D_{ang}(x, y)$$

$$D_{lr}(x, y) = \lambda_{lr}(\Omega({\gamma_x}^{C_1}, {\gamma_y}^{C_2}) - 1), \text{ where } \Omega(a, b) = \max(\frac{a}{b}, \frac{b}{a})$$

$$D_{ang}(x, y) = \lambda_{ang}|{\theta_x}^{C_1} - {\theta_y}^{C_2}|, \tag{8}$$

where x, y – are two objects described by the following descriptor $\Psi_j =$

$\left\{\gamma_i = \frac{l_{seg_{i+1}}}{l_{seg_i}},\ \theta_i |\ i \in (1, \dots, N-2)\right\}$, where l_{seg_i} – is the length of a segment,

θ_i – is the angle between segments.

Section 2.4. Data analysis

Subsection 2.4.1. Preparing the data

Let us start following the structure of the model by describing the dataset preparation process.

Having the original data described in part 1, it is necessary to normalize all the data and to transform it into the objects which are possible to input for the algorithm of pattern recognition. As have been previously discussed, we take high and low prices from the original dataset (H and L time series).

As a first step, H and L time series need to be adjusted for extreme events which are not of our interest. Those events are all possible corporate actions, such as stock splits, issues, buybacks, etc. The adjusted prices in the original dataset represent exactly such adjusted prices for the case of close ones. Hence, the same procedure needs to be done with respect to high and low prices in order to avoid abnormal price movements. For the next step, it is necessary to extract subsequences of different lengths from H and L time series which will be used further to build final objects. Any potential pattern can be realized during the period of 35 trading days (Lo, Mamaysky, Wang, 2000). However, the length can be easily changed and even be extended to a number of different lengths. Subsequences are retrieved by applying a sliding window of length l_i, where $l_i = \{10, ...\}$ for each corresponding $i = \{1, ...\}$, to H and L time series (Ye, Keogh, 2009). Therefore, the number of extracted subsequences is $\sum_{i=1,2,3}(N - l_i + 1)$, where N – is the number of trading days in price time series. In this thesis we will research patterns of different trading days' length which will be determined according to our goodness of fit model further. We will try different numbers of trading days and the length providing the best goodness of fit result will be the one we are using for the algorithm.

Following the abovementioned procedure, we need to conduct z-score normalization in order to make the final objects comparable. The process goes is as follows: $z = \frac{x-\bar{y}}{\hat{\sigma}}$, where x – is a price subsequence (for example, high prices), $\bar{y}$ – is the subsequence of mean adjusted close prices for a corresponding time period of the price subsequence, $\hat{\sigma}$ – is Parkinson volatility (see Equation 9). The subsequence of mean adjusted close prices is employed in order to make sure we are deducting the same exact value from both high and low prices. This ensures that our normalized values are centered around mean adjusted close prices and high prices are always higher that low ones. Parkinson estimator for volatility, that assumes high and low prices, is more appropriate for our task to acquire less noise data due to it is approximately five times more efficient than close-to-close estimators (Alizadeh, Brandt, Diebold, 2002).

$$\sqrt{\frac{1}{4Nln(2)}\sum_{i=1}^{N}\ln\left(\frac{h_i}{l_i}\right)^2}, \tag{9}$$

where h_i – is the high price and l_i – is the low price.

Following the normalization procedure, it is possible to comprise high and low subsequences into the final objects that are of our interest. Analytically, an object is basically a vector of numbers (high and low prices), which looks like that: $v_1 = (60.09, 58.23 \dots 57.85, 56.91 \dots)$. Alternatively, an object can also be represented by a number of points with two coordinates: a price and day. Graphically, the same object is shown on figure 1, where a black line is the subsequence of high prices, and a red line is the subsequence of low prices:

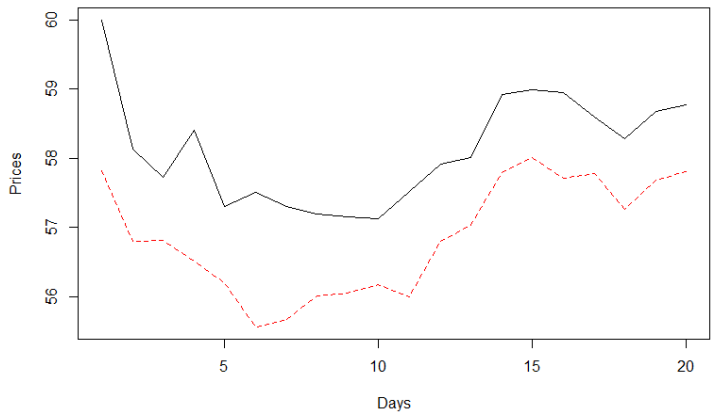

Figure 1: The example of a final object.
Source: Own elaboration based on Yahoo Finance data.

This way, a number of such objects of lengths 10 has been collected and stored in a dataset. Ultimately, as our interest lies in the field of sophisticated patterns like “Head and Shoulders”, it is important to clean the dataset out of simple trends which are generally straight lines of numerous variations:

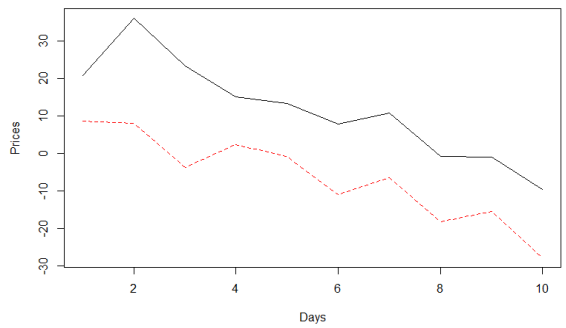

Figure 2: The example of a simple trend
Source: Own elaboration based on Yahoo Finance data

To do that, we apply a linear approximation that describes an object according to precision defined as R^2. So, we use this linear approximation to every object in the dataset with $R^2 = 0.95$, which practically showed decent results in identifying patterns of interest from simple

trends. As a next step, we look at how many data points remain after the approximation process. An empirical criterion for objects sophistication is that there are 7 points remained after the procedure. Although many simple objects (like the trend on Figure 2) can depicted by four points, there are numerous objects which have more than four points after linearization. This founding led us to the empirical criterion we derived. Graphically, the linear approximation can be represented as follows:

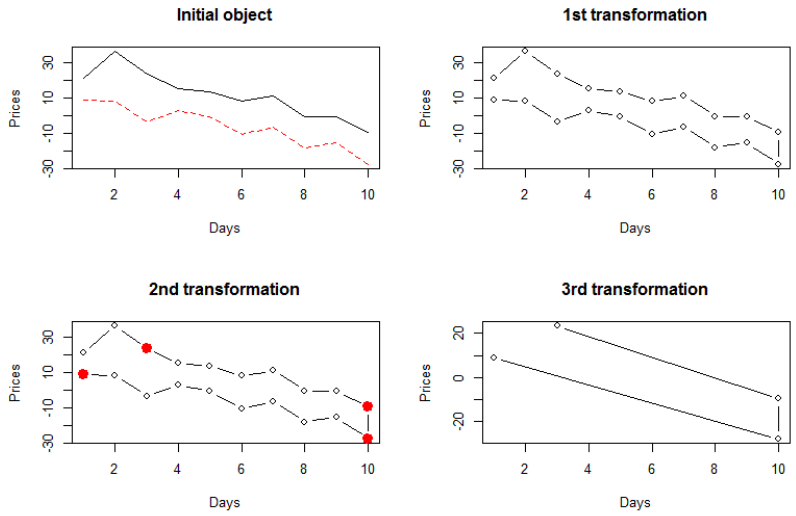

Figure 3: A linear approximation process
Source: Own elaboration based on Yahoo Finance data.

The final dataset after linearization contains 7365 objects, such as depicted on Figure 1, while 82 objects has been considered not sophisticated enough and deleted from our dataset. Finally, we have the dataset of objects with a complex shape, their length is 4 trading days, they represent high and low prices of 30 DJIA stocks for the period of time from the 27th June of 2016 to the 27th June of 2017.

Subsection 2.4.2. Applying the algorithm

Moving forward to the step of measuring the distance, it is necessary to say that our distance choice and the model itself will make possible to retrieve graphically similar patterns as

technical analysts do so. Any trader who is following technical approach follows commonly the same process as our algorithm does when trying to find figure similarities in stock charts on a computer screen. Professionals in place just search for patterns by looking and analysing all similar and different parts of subsequences from financial time series. In this regard we are creating a machine learning application, because we try to make the algorithm to behave like a human. Nevertheless, many analysts base their conclusions on subjectivity of their views in recognising if those subsequences are real patterns or just noisy outbreaks. Therefore, there is a reasonable question why there is a need to teach a computer to see similar objects in financial data, if practitioners already can do it. And the answer would be yes - when algorithms can recognize similarities, it can aggregate them and cluster in a rigorous way, such that there will be a mathematical prove of the existence of patterns in financial data. Hence, at this point we need to introduce a clustering method we are using.

The foundation of the clustering model is classic k-means algorithm which is a machine learning technique to classify objects. However, it still requires a user-defined number of clusters. The primary idea of k-means algorithm is as follows: suppose $X = \{x_1, \dots, x_n\}$, which is a set of objects; and consequently k-means procedure minimizes an objective function in order to classify n objects out of X into k clusters $K_1, \dots, K_k$, where μ_i is a center of cluster K_i.:

$$F = \sum_{i=1}^{k} \sum_{x \in K_i} ||x - \mu_i||^2, \tag{9}$$

where x – is a cluster object and μ_i – is a center of the cluster.

The algorithm repeats this optimization following a number of iterations by reassigning centers through finding an average among elements within a cluster until it converges or, in other words, stops changing centers, clusters. However, we are not borrowing exact k-means for our

task, because it assumes only Euclidian distance and redefines centers by creating a new center and not just repacking among existing objects. With our aim of objects clustering this classic approach is not appropriate, because we are not allowed to create any new objects as this will produce bad results which would be a complete mess of never existent objects. Hence, we are adapting that classic k-means by employing some modifications proposed in the method called k-medoids. The difference of these approaches is that in k-medoids we do not create a center by calculating within cluster average, but choose from existing objects within a cluster. Moreover, k-medoids works with various distance measures. Our advanced k-medoids model works conceptually the same as a former one does except for recalculation of centers. K-medoids proposed by us rearranging centers by identifying if it is on average closer to a previous center than all other objects. As inputs, the model takes objects form the dataset which are represented as vectors, such as $v_1 = (60.09, 58.23 \ldots 57.85, 56.91 \ldots)$ which is shown on Figure 1. As output, it produces a number for every object which determines a cluster of an object. The step-by-step description of the algorithm is as follows:

Step 1: Takes objects from the dataset and randomly assigns centers to k of them, where k is a user-defined number of clusters.

Step 2: Calculates a distance between every object and every cluster center, and divides the dataset into clusters.

Step 3: Redefines cluster centers by calculating an average distance of objects to a center within a cluster. A new center is defined as an object which is on average closer to a cluster center than others. Formally, it chooses such an object which has a minimum distance to a cluster center after an average distance is subtracted.

Step 4: Reiterates from step 1.

As initial centers are assigned randomly, we propose to repeat the algorithm several times with different initial centers. We call these repeats global iterations. These initial random

centers may have significant influence on a resulting clustering, so that it might be a good idea to have some insurance against mistakes in a form of global iterations. As our algorithm may not have natural convergence, we also propose to restrict by the number of iterations from in the form step 1 to step 3 by some number. Those iterations we call local ones. In order to define those numbers, we are applying our goodness of fit criteria. The final numbers will be picked with the highest goodness of fit score.

Ultimately, we need to identify the best parameters for the model by using our goodness of fit algorithm (Tibshirani & Walther, 2005) which is embedded into the clustering model. This algorithm works as follows:

Step 1: Divides the dataset into 5 subsets: 4 training subsets and 1 test subset.

Step 2: Applies the clustering algorithm to training sets of data.

Step 3: Applies the clustering algorithm to the test set of data.

Step 4: Assigns each observation of the test set to a cluster center defined for training sets according to a minimum distance.

Step 5: Calculate how many pairs of observations are remained in the same clusters by comparing clusters from a test clustering to step 4.

Step 6: Calculates a ratio of remained pairs to all pairs.

Step 7: Reiterates 5 times from step 1.

Step 8: Takes mean of the ratio in step 6 as a goodness of fit measure.

This algorithm allows us to try a different number of global iterations, local iterations, clusters, objects and all previously discussed distance functions.

Chapter 3. Results and Discussion

The main hypothesis of this thesis is to test if we can produce a clustering model which would give us good goodness of fit results. Namely, we wanted to identify which parameters to adjust and how in order to obtain the best goodness of fit results according to the algorithm described in the previous section. So, we tested different variations for a number of objects, algorithm's iterations (global and local ones), clusters, and a distance function.

With regard to the number of objects we mean a length of a subsequence. We tried different lengths from $l = 2$ to $l = 30$ and obtained the following goodness of fit distribution:

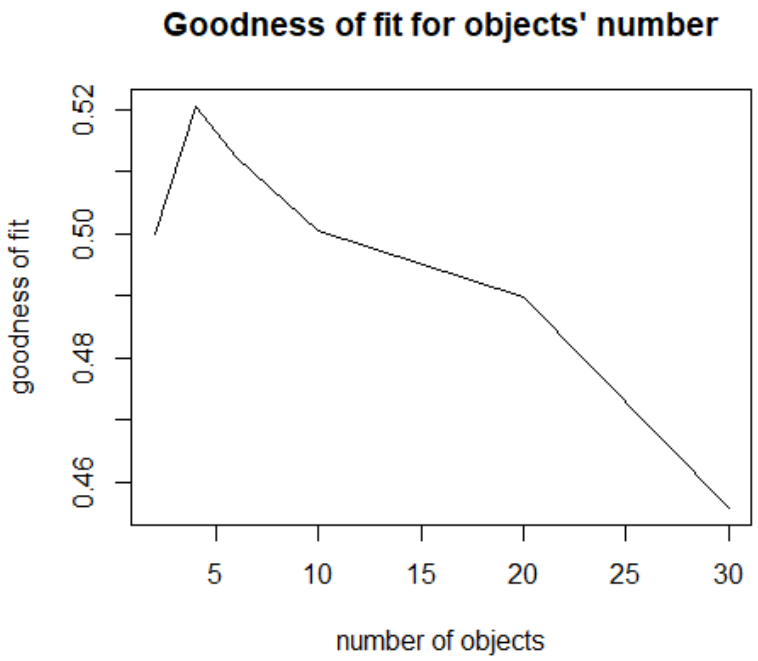

Figure 4: Goodness of fit distribution for the number of objects
Source: Own elaboration based on Yahoo Finance data.

As it can be seen, the best goodness of fit measure is produced by the number of 4. Hence, this parameter for the number of objects is considered the best and will be used for all consequent clustering trials. This number indicates that the most robust clustering distribution of objects is produced, when we consider 4 trading days. That can be a sign that patterns in financial markets

are realised exactly during those amount of days. This observation in some sense approves the following statement: any potential pattern can be realized during the period of 35 trading days (Lo, Mamaysky, Wang, 2000). 4 trading days is included in 35-day period and, therefore there is no contradiction to previous research. However, this period is quite short which may rise new discussion towards the optimum amount of days for pattern realization. This shortage from 35 trading to 4 trading days may be explained by the data taken into consideration. The dataset for this thesis includes most recent data for the period of last year, while that past research (Lo, Mamaysky, Wang, 2000) considers the period from 1962 to 1996. Hence, we may reasonably suppose that patterns in recent financial data more frequently appear, change and disappear. This assumption moves along the facts that financial markets are more regulated, and there are more instruments used including high-frequency trading which greatly affects the possibility to extract profits from inefficiencies.

Speaking about iterations it is necessary to clarify again what is meant by that. Global iterations are those which assign initial centers randomly. Local iterations are our artificial restriction due to the lack of natural convergence in the algorithm. With regard to those iterations we obtained the following results:

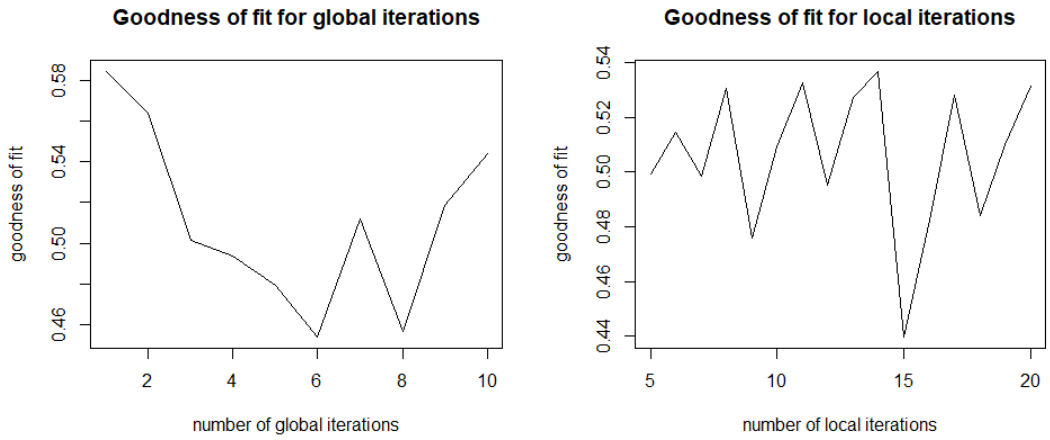

Figure 5: Goodness of fit distribution for the iterations
Source: Own elaboration based on Yahoo Finance data.

With respect to these results, we can conclude that 1 global iteration and 14 local iterations produce the best goodness of fit. However, the result for global iterations may seem surprising. As it was mentioned, in global iterations, cluster centers are assigned randomly, so it is not really clear why several repeats of this process decreases goodness of fit. In our perception, any random centers' initial allocation should be eventually converged to some reasonable centers' allocation which is assumed by the information contained in time series. Hence, the more we repeat this random allocation, the more possibility we get that convergence. According to such reasoning, the graph should increase and not to decrease. With regard to local iterations, the graph seems to be expected. In our algorithm we have the function to exit the loop if cluster centers start to repeat, so this goodness of fit allows us to detect the optimal amount of iterations to preserve the best goodness of fit without any loss of information about possible centers.

In terms of the number of clusters, we obtained very unexpected results which may cross the main hypothesis. The results are as follows:

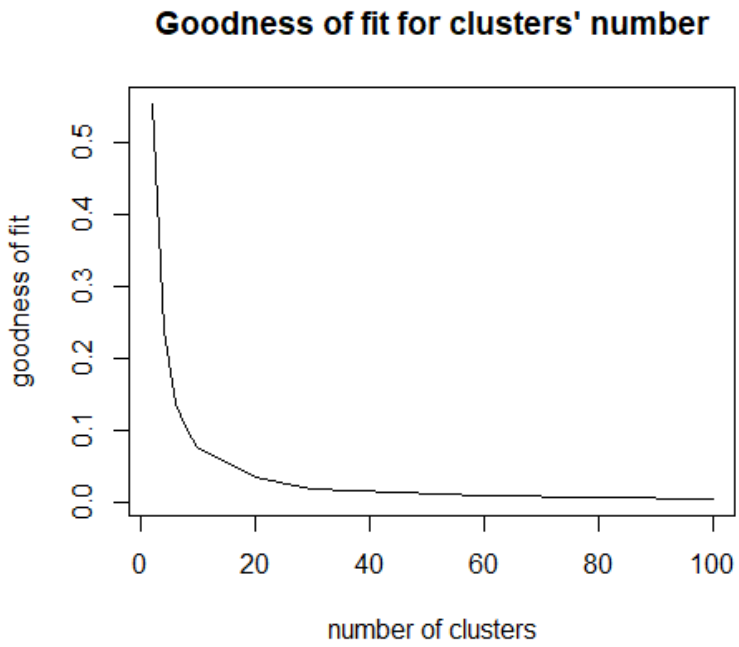

Figure 6: Goodness of fit distribution for the number of clusters
Source: Own elaboration based on Yahoo Finance data.

As it can be seen the best goodness of fit results are obtained, when the number of clusters is equal to zero or one, which is equivalent. Hence, that means that the best clustering distribution is obtained if the dataset is not clustered at all. That may mean that there are no patterns embedded in the data, or that we cannot resolve the pattern recognition task using the methods we proposed in this thesis. This result is very controversial and it does not support previous research (Caginalp & Laurent, 1998; Chang & Osler, 1999; Lo, Mamaysky, Wang, 2000) which stated that patterns can be identified in financial time series and there is at least a weak presence of regularities which can be clustered. Therefore, there is more evidence towards that there is a problem with the set of methods we are using to identify patterns and cluster them than in the dataset itself. However, there may be a chance that the nature of financial data has been significantly changed during recent times due to increased regulations, new instruments and institutions which may increase efficiency of markets and fade out all patterns from the data.

The function which produces the best goodness of fit results is Jaccard similarity. For this type of objects this function seems to give the best measure of similarity in the clustering task. We have discussed that initially we tried a modified version of Mittal's distance measure, but we decided to come up with a variety of different functions due to the modified one does not satisfy a triangular inequality. However, we decided to compare goodness of fit for all of them in order to have a complete view on all possibilities. The results are presented as follows: Euclidian distance (1), Manhattan distance (2), Minkowski distance (3), Cosine similarity (4), Jaccard similarity (5), Hellinger distance (6), Bhattacharyya distance (7), Modified Mittal's distance measure (8):

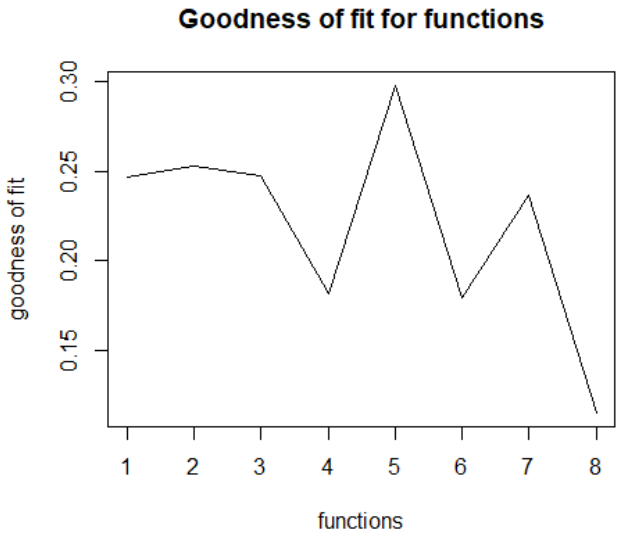

Figure 7: Goodness of fit distribution for the number of clusters
Source: Own elaboration based on Yahoo Finance data.

Here we can conclude that the similarity measure proposed (Parui & Mittal, 2015) and modified by us is not appropriate in the context of our model. It produces the lowest goodness of fit results with respect to our test and data under consideration. In this sense we can see that image recognition tools were not suitable for our financial time series or these tools need more developments and our objects need more transformation to be closer to pictures.

Chapter 4. Conclusion

Section 4.1. Main contributions of this paper

In this paper we presented the automated algorithm enhanced with goodness of fit parameter determination for searching patterns in financial time series data using machine learning techniques. We applied our algorithm on the dataset of 7365 from the original dataset described in sample section, and adjusted the parameters of the model to obtain the best goodness of fit results. The hypothesis of whether we can structure a model which would produce decent clustering results is not confirmed with respect to instruments of our research.

The main contribution of this research is a trial to combine different machine learning tools and construct a model (an objective measure) to understand how good these tools are. In principle, this work is inspired by previous research of pattern recognition in financial time series using an automated algorithm (Lo, Mamaysky, Wang, 2000). In this thesis we tried different tools to construct an automated algorithm and combine them with goodness of fit measure to tune the model to acquire the best results. We extended the principle of previous research of pattern recognition automated algorithm to modern data and methods from pattern and image recognition which promised outstanding results. However, the clustering task was badly performed by those methods and there is still much room to improve.

We believe that our model could be significantly extended by considering the suggestions at the end of this thesis, and it would significantly increase the efficiency of patterns retrieval from financial time series data.

Sections 4.2 - 4.3. Limitations and Prospective

Eventually, it is important to state directions for further work which are implied by limitations of our model. Firstly, it is necessary to investigate the issue why goodness of fit graphs is not monotonic. Because of their heterogeneity it is actually not precise to infer information about the best parameter according to goodness of fit. This happens because any time further – if we add more variations of a parameter on x-axis – we can find even a better parameter. Hence, it is important to understand what causes them to be not monotonic and fix this issue.

The other limitation which creates a room for improvement is to understand why we obtained the best goodness measure for one-size cluster which underlies that there are no patterns in the dataset. This may happen because our dataset is not prepared in a way that makes it possible to reveal any patterns, or because our algorithm – or tools which we employ - does not allow to detect patterns in the dataset, or because our goodness of fit measure cannot capture enough information about the quality of clustering. Ultimately, it may be the case that data itself does not contain any regularities and financial markets are perfectly structured such that it is impossible to see any regularities. In other words, the markets are efficient enough to incorporate any patterns occurring, so that those patterns disappear. Hence, it is necessary further investigate this issue.

Consequently, the other significant space for further research is clustering algorithm which we use. Essentially, it is not logical to use clustering algorithm with user-defined number of clusters when the task is to classify unlabeled data, even if we produce a goodness of fit measure to determine the best number of clusters. Moreover, we are using a rectangular slicing window which limits our search for patterns to specific sets of cases. Because of these drawbacks, our final clusters contain excess noisy objects which disturb the attention when

looking at other normal cluster objects. But the main reason we use our k-medoids algorithm is that it saves a significant amount of time in comparison to brute-force-like algorithms and produces comparatively good results when we define the number of clusters that is closer to the real one. The good extension of this work would be to try DBSCAN clustering algorithm which is the most cited in scientific literature till then.

Another direction for further work might be to proceed further from building the model which would allow to determine best parameters for clustering to the statistical tests and models which would allow to determine if those clusters which we detect are significant and distinct from noise data. This way, it would be possible to judge if we have any possibility to forecast stock movements.

It is also important to extent our work by developing the trading algorithm which would use the retrieved patterns for price forecasting. Such algorithm would help in research of the topic of how possible to earn excess returns using our pattern recognition algorithm.

The last obvious direction of work is to research larger or different datasets. Moreover, it would be the good idea to add the stock label and the time label to the dataset in order to determine to which company and period of time patterns belong.

References

Alizadeh, Brandt. “Range-based estimation of stochastic volatility models.” The Journal of Finance, 2002.

Bechu, Bertrand. “Laanalyse technique: pratiques et methods.” 1999.

Caginalp, Laurent. “The predictive power of price patterns.” Applied Mathematical Finance, 1998.

Chang, Carol L. Osler. “Methodical Madness: Technical Analysis and the Irrationality of Exchange-rate Forecasts.” The Economic Journal1, 1999.

Cheung, Chinn. “Currency traders and exchange rate dynamics: a survey of the US market.” Journal of international money and finance, 2001.

Efron. “The jackknife, the bootstrap and other resampling plans.” 1982.

Fama, French. “The capital asset pricing model: Theory and evidence.” Journal of Economic Perspectives, 2004.

Fama, Malkiel. “Efficient capital markets: A review of theory and empirical work.” The journal of Finance, 1970.

Jensen. “Some anomalous evidence regarding market efficiency.” Journal of financial economics, 1978.

Lo, Hadzic. “The heretics of finance: Conversations with leading practitioners of technical analysis.” 2010.

Lo, Mamaysky, Wang. “Foundations of technical analysis: Computational algorithms, statistical inference, and empirical implementation.” The journal of finance, 2000.

Lui, Mole. “The use of fundamental and technical analyses by foreign exchange dealers: Hong Kong evidence.” Journal of International Money and Finance, 1998.

Mills, Markellos. “The econometric modelling of financial time series.” 2008.

Parui, Mittal. “Sketch-based Image Retrieval from Millions of Images under Rotation, Translation and Scale Variations.” 2015.

Pring. “Technical analysis explained: The successful investor's guide to spotting investment trends and turning points.” 2002.

Schabacker. “Stock market theory and practice.” 1930.

Taylor, Allen. “The use of technical analysis in the foreign exchange market.” Journal of international Money and Finance, 1992.

Tibshirani, Walther. “Cluster validation by prediction strength.” Journal of Computational and Graphical Statistics, 2005.

Tsay. “Analysis of financial time series.” 2005.

Ulanova, Begum, Keogh. “Scalable Clustering of Time Series with U-Shapelets.” Conference on Data Mining (SDM 2015), 2015.

Xu, Xia. “Distance and similarity measures for hesitant fuzzy sets.” Information Sciences, 2011.

Ye, Keogh. “Time series shapelets: a new primitive for data mining.” Proceedings of the 15th ACM SIGKDD international, 2009.

Zakaria, Mueen, Keogh. “Clustering time series using unsupervised-shapelets.” Data Mining (ICDM), 2012.