

Available online at www.sciencedirect.com**ScienceDirect**

Procedia Economics and Finance 5 (2013) 298 – 307

Procedia
Economics and Finance

www.elsevier.com/locate/procedia

International Conference on Applied Economics (ICOAE) 2013

Stock Selection System: Building Long/Short Portfolios Using Intraday Patterns

Nikitas Goumatianos^{a,b}, Ioannis Christou^{a*} and Peter Lindgren^b^a*Athens Information Technology, 19km Markopoulou Ave. PO Box 68, Paiania, Greece*^b*Aalborg University, Fibigerstræde 16, Aalborg, Denmark, DK-9220*

Abstract

We present the architecture of a complete intraday trading management system using a stock selection algorithm for building long/short portfolios. Our approach for stock selection is based on knowledge discovery in large databases technologies; more specifically, we build techniques that allow one to discover hidden price patterns which are capable of predicting the direction of stock prices. The base of the whole system is a novel pattern mining algorithm from time-series data, which involves highly compute-intensive aggregation calculations as complex but efficient distributed SQL queries on the relational databases that store the time-series. This algorithm follows similar path to Association Rule Mining algorithms (such as the Apriori and FP-Growth algorithms) but with major differences. First, it has one step more at initialization which uses a rule-based generator algorithm that transforms relations and data structures in a binary string format. These patterns contain mixed information of small price patterns (3, 4 or 5 candlesticks) and trading signals/filters produced by technical indicators. As output, instead of generating association rules, it produces probabilities (varies between 60% and 95%) about the future stock price direction such as “within next 5 periods, price will increase (or decrease) more than 2%”. Then, the stock selection system exploits all this information and uses it to find those stocks that appear to form patterns which have the highest accuracy in predicting prices. Each time period (finest interval length being 10 minutes), the stock selector system, having the support of trading systems, examines and decides which stocks should be replaced in current portfolio by other more profitable stocks. The system testing, which produced interested results, is made of portfolio of 10 stocks having open long position and 5 stocks having open short positions.

© 2013 The Authors. Published by Elsevier B.V. Open access under [CC BY-NC-ND license](https://creativecommons.org/licenses/by-nc-nd/4.0/).

Selection and/or peer-review under responsibility of the Organising Committee of ICOAE 2013

Keywords: Portfolio optimization, Portfolio Enhancemt, Data Mining, Price Patterns, Hidden Patterns, Time-Series Forecasting, databases, Stock Selection System, Trading Systems, High Frequency Trading (HFT).

* Corresponding author. Tel.: +30-210-668-2725; fax: +30-210-668-2702.

E-mail address: ichr@ait.edu.gr.

1. Introduction

1.1. Price patterns

The ability to predict the future prices of instruments (stocks, futures, options, etc), based on historical data, is the biggest challenge in investment industry, while from scientific perspective there is a dispute between scientists. Before the 1980s, most researchers were skeptical about the ability to predict prices, especially, when using technical analysis, and concluded that it is not possible to produce as good results as the buy-hold strategy (Alexander, 1961; Jensen and Bennington, 1970; Fama, 1970). However, later studies showed just the opposite (Brock, Lakonishok and Lebaron (1992), Bessembinder and Chan (1995, 1998), Lo, Mamaysky and Wang (2000).

Technical analysis is based on the premise that history repeats itself. Therefore, analyzing the price patterns such as charts formations and candlestick patterns, could be beneficial. An expert system for predicting stock market timing using candlestick charts was proposed by Lee and Jo (1999). Another study made by Gagalp and Laurent (1998), showed that specific candlestick patterns (with names such as “three-white-soldiers”, “three-black-crows” etc.) have predictive capability and indicate a potential profit of 1% or more during a two-day holding period. In contrary, in a paper by Marshall et al. (2006), the authors found candlestick technical analysis has no value on U.S. Dow Jones Industrial Average stocks during the period from 1992–2002.

In the scope of this current work, we are trying to answer the following question: Are there any unknown (hidden) candlestick patterns, that when possibly combined with technical indicators actually carry information about future stock prices? And if there exist any, how can we discover them? A closely related work of discovering profitable candlestick patterns was carried out by Sheng, Hou & Chen (2006). They designed a Knowledge Representation Model which held the information of three (3) successive candlesticks using a bit codification method, called Relative Price Movement (RPM). The training daily data was from January 1, 1994 to December 31, 1998, of 82 stocks, while the testing data was from January 1, 1999 to December 31, 1999. In total, in the test set, the mined patterns occurred less than 100 times, which is very low to make safe conclusions.

1.2. Building Long/Short Equity Portfolio

Long/short equity is an investment strategy generally associated with hedge funds and more recently with certain progressive traditional funds. It involves buying long equities that are expected to increase in value (undervalue) and selling short equities that are expected to decrease in value (overvalue stocks) (Nicholas, 2004). The short portfolio acts both as a hedge against the decline markets (bear markets) and as opportunity to add value by selecting stocks that are expected to underperform. In this strategy, the gross market exposure becomes more than 100% without needed leverage because the income received from selling short can be used for self-financing the long position. Generally, fund managers take positions for Long/Short Equity by using fundamental analysis including industry analysis, forecasting future financial results, valuation analysis, etc. An example of this method is the work done by Morel (2001) who proposed a multi-factor model for selecting stocks. Other fund managers use quantitative and technical analysis methods. Another approach for building long/short portfolios was the development of stock selection system using a rule induction mining algorithm (George H. and Miller, 1996).

Here, the scope of research is first to develop a data mining algorithm to discover patterns which have the highest probability of accurately forecasting the stock price directions, and second to construct a long/ short portfolio strategy.

2. Methods And Application

2.1 Data Setup

The initial setup consists of intraday data (time frame period = 1 minute) of totally 300 stocks belonging to S&P500 index (selecting randomly). The intraday data starts from 2 Jan 2008 until 17 Aug 2012 (4.5 years). The row data of 1 min length were transformed into two different time frames: 10 and 60 min time length. The format of the data is:

Symbol, Date Time <range of 10 min>, Open Price, High Price, Low Price, Close Price, Volume

2.2 System Architecture of the Expert System

The system architecture consists of the following subsystems:

- A Knowledge Representation System.
- A Multi-time frame SQL Rule Analyzer
- A Patterns Validation System
- A Stock Selector System
- A Portfolio Optimizer / Builder for Long/Short Trading

Figure 1 presents the architecture of the whole system:

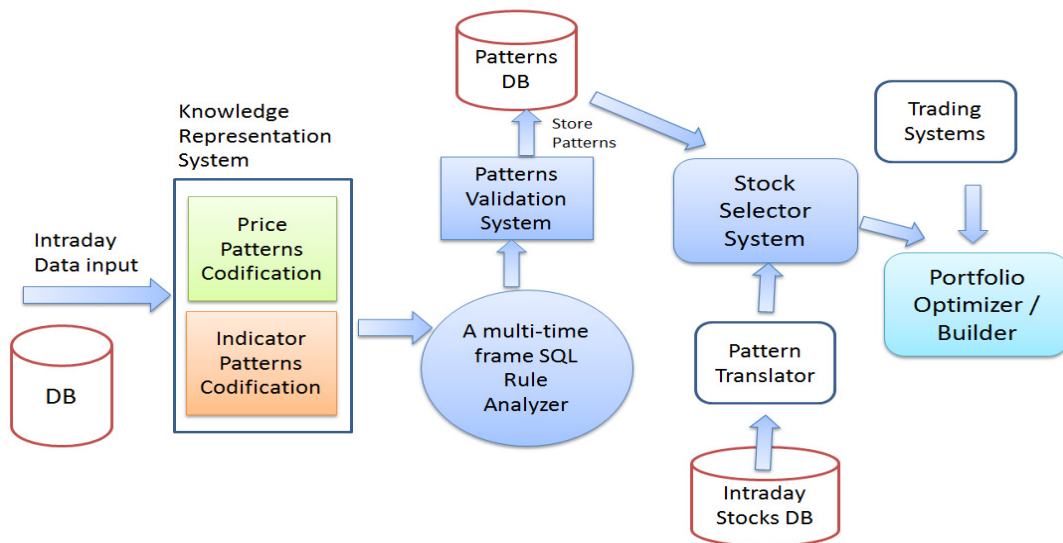


Figure 1: The System Architecture

The Knowledge Representation System is responsible for processing raw intraday data and based on specific rules to translate them in binary format product in such way that in reverse (from binary product) to form the original intraday pattern. Actually, it contains rule-based expressions that record four types of information: bits referred to candlestick itself, bits referred to exact position (relationship) among two or more candlesticks, bits presenting the strength of the price movement and bits referred to technical indicators.

The multi-time frame SQL Rule Analyzer consists of aggregate SQL views and procedures and processes all binary data to find patterns with capability in predicting futures prices. The Patterns Validation System

selects those binary patterns found by the previous system, and analyzes further using cross validation testing. Patterns to pass the validation (marked as successful) should have prediction accuracy in all tests more than 60%, while obeying at a minimum support and confidence level. Then, those data passed the tests are stored in a patterns database in order to be used by the Stock Selection System. This system takes the raw intraday data, and then, using a pattern translator translates them in binary format. The next step is to make a search in patterns DB to find which patterns are capable of accurately predicting price. The report is sent to the Portfolio Optimizer / Builder which is responsible for building the portfolio of stocks.

2.3 Data Mining Algorithm

The algorithm is distributed in nature, and is implemented as a three step process; each step executes in a different system ran on a different machine. The first step consists of running the Knowledge Representation System.

The rules produced by the knowledge representation system can be classified into two main parts:

- The first part refers to price patterns consists of 3, 4 or 5 candlesticks.
- The second parts represent rules coming from technical analysis (trading signals & filters).

For clarity, we present only samples of bit-condification rules for each category:

a. Rules for price patterns

- (1) *Rules refer to one candlestick itself*: In the following figure 2, rule 1 and 2 are displaying. Note that one candlestick corresponds to a specific time frame (1 min, 5 min, etc). The body is between open and close price. A candlestick could be white (close > open) or black (close < open).

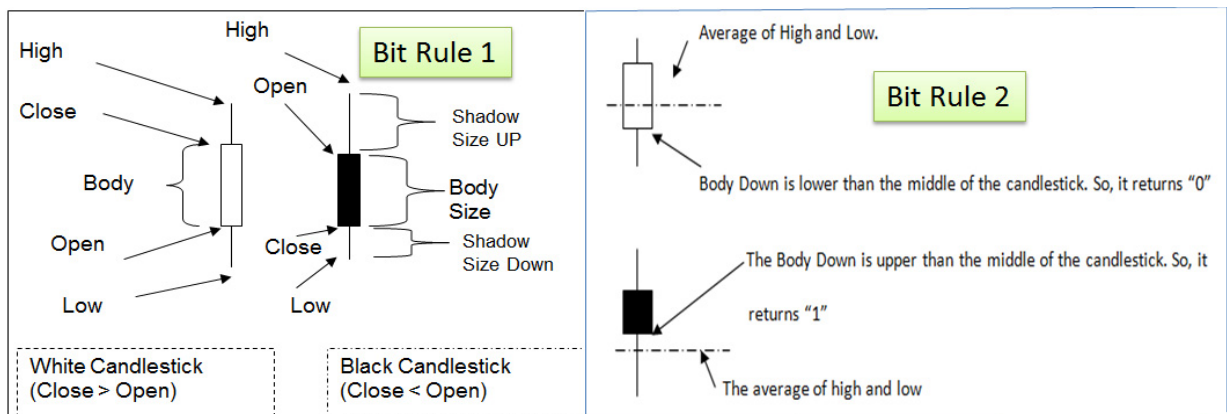


Figure 2: Graphical Representation of two Rules "Candlestick itself"

In the above figure 2, in the left side, rule 1 is displaying. If close price is greater than open price return "1", otherwise return "0". On the right, rule 2 is displaying. If the body down is above of the middle line (point = $(high + low) / 2$), return "1" (it is depicting in right bottom corner), otherwise "0" (it is depicting in right top corner).

- (2) *Rules refer to the relationship between candlesticks*, as in the following figure:

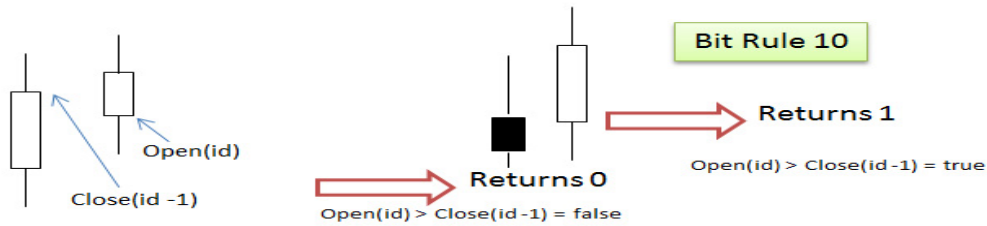


Figure 3: Graphical Representation of an Example of Relationship Rule

Here, the rule compares if the current open price ($id = \text{current}$) is greater than previous close price ($\text{previous} = id - 1$). If it is true returns “1”, otherwise returns “0”.

Finally, after applying all rules, the product of price pattern, regarding a specific position (time) is formed by combining all produced bits. For a pattern consists of three candlesticks, the formula is as following:

$$\text{Product} = \text{Bits of itself Candlestick}(id-2) + \text{Bits of itself Candlestick}(id-1) + \text{Bits of itself Candlestick}(id) + \text{Bits of Relationship Between Candlesticks}(id) \text{ and } (id-1) + \text{Bits of Relationship Between Candlesticks}(id) \text{ and } (id-2) + \text{Bits of Relationship Between Candlesticks}(id-1) \text{ and } (id-2).$$

Indicator	Condition	Is true?	Is false?
Moving Average of 20 periods	Is below current price close value?	Return “1”	Return “0”
Moving Average of 50 periods	Is below current price close value?	Return “1”	Return “0”
Highest - Lowest Value of 40 previous periods	Is current price > above 70% of the current?	Return “1”	Return “0”
Relative Strength Index (RSI)	Is value above 70? Is value below 30?	Return “0” Return “1”	Return “1” Return “0”
Parabolic SAR Indicator	Is greater than close value?	Return “0”	Return “1”

Figure 4: Sample of list of indicators (rules) used for coding

The second part of the product is formed by involving technical indicators filters and trading signals. In above figure 4, the table includes all technical indicators as well as their rules (conditions true/false). For each type of price pattern (length of 3, 4, or 5 candlesticks), we add a part of information of the indicator rules. There is an algorithm which selects those rules that add capability in predicting future prices. First, it executes a greedy local search to find the most value-adding subset of rules for the system. With those subsets passing the validation, the second step is to execute an optimization algorithm, where the objective function of this optimization combines two factors, namely (i) that the total patterns appearing must have a minimum support level, and (ii) that the selected rules should increase the prediction accuracy of the system. This algorithm is executing in the Multi-time frame SQL Rule Analyzer. It processes all binary data patterns, for each training data set, and for each type of prediction. An example of pattern prediction accuracy is the following:

Pattern “A” prediction = (times that price climbed to more than 2% within next 60 minutes) / (Total times Pattern “A” has appeared).

Next, the Patterns Validation System will examine all training and testing sets to examine if the following rules are satisfied:

- In Each Training Data Set: Prediction Accuracy $\geq 65\%$, then in Testing Set Data: Prediction Accuracy should exceed 60%.
- The total occurrence of each pattern should be more than 30 times.

Finally, the Patterns Validation System, for those patterns that have passed the validation, it will mark them as successful and store them in the patterns database. The format of the produced information is as follows: [Info of Pattern], [Type of Prediction], [Time Frame], [Pattern Binary Code], [Occurrences], [Probability].

The prediction variables selected in the algorithm are displaying in the following figure 5:

Time frame 60 min	Time frame 10 min
Price will increase more than 0.6% within next 2 periods (2 hours)	Price will increase more than 0.3% within next 2 periods (20 min)
Price will decrease less than -0.6% within next 2 periods (2 hours)	Price will decrease less than -0.3% within next 2 periods (20 min)
Price will increase more than 1% within next 5 periods (5 hours)	Price will increase more than 0.6% within next 5 periods (50 min)
Price will decrease less than -1% within next 5 periods (2 hours)	Price will decrease less than -0.6% within next 5 periods (20 min)
Price will increase more than 1.5% within next 10 periods (10 hours)	Price will increase more than 1% within next 10 periods (100 min)
Price will decrease less than -1.5 % within next 10 periods (10 hours)	Price will decrease less than -1 % within next 10 periods (100 min)

Figure 5: List of Predictive Variables used by Rule Analyzer

2.4 Methodology of Portfolio Construction

The Stock Selector System is able to search information for any type of pattern from the patterns database, which have been updated before by the Patterns Validation System. The system, during a testing process, constructs all type of patterns for all stocks (total 300 in DB) for current time frame. It is responsible for collecting those stocks having the highest accuracy value in prediction including both long and short positions.

The algorithm for the Stock Selector System is as follows:

[For each time frame of data]:

Construct six (6) types of composite patterns (price & indicators) [for each stock] (totally 6 X 300 = 1800 patterns for current selected time frame).

Apply database search for each pattern (repeat research for each pattern of total 1800)

Classify them into two categories: patterns for long and patterns for short positions

Sort them by the prediction accuracy.

Return a list of the top 10 stocks for long position (having highest accuracy)

Return a list of the top 5 stocks for short position (having highest accuracy)

Here is one of the available methods of estimating the highest accuracy (60 min time frame) for long position:

$Estimated\ Accuracy = prediction_next2GreaterThan0.6\% + prediction_next5GreaterThan1\% \times (1 + weighting\ Factor) + prediction_next10GreaterThan1.5\% \times (1 + 2 \times weighting\ Factor)$

where the *weighting Factor* gives either more value to shorter time prediction (when negative) or to longer prediction (when positive).

Next, the Portfolio Optimizer / Builder for Long/Short Trading will take above information and with the assistance of five(5) trading systems will choose which stocks will close position of the total portfolio and create new positions by those proposed by the Stock Selector System. During the simulation some parameters were optimized such as the scope of trading (select those having highest prediction in next two periods or those having the highest prediction within next 10 periods?), how many periods is the buy and hold strategy.

3. Testing the System

3.1 Results from Data Mining Algorithm

We tested the produced patterns using 4-fold cross validation and across six (6) variables displayed in figure 5. In the following figures are displayed the results of running our mining algorithm. Testing is done on years 2008 to 2011 and does not include data from year 2012, which is left for testing the long/short portfolio. Figure 6 corresponds to 60 min intraday patterns. The first bar (blue), which is higher than red bar, depicts the total support for each category of patterns (3, 4 and 5) which at least one predictive variable has accuracy more than 65% (within next 2 periods price increase more than 0.6% => prediction > 65% **OR** within next 5 periods price increase more than 1% => prediction > 65% **OR**...). The red bar is for accuracy in prediction more that 75% for at least one predictive variable.

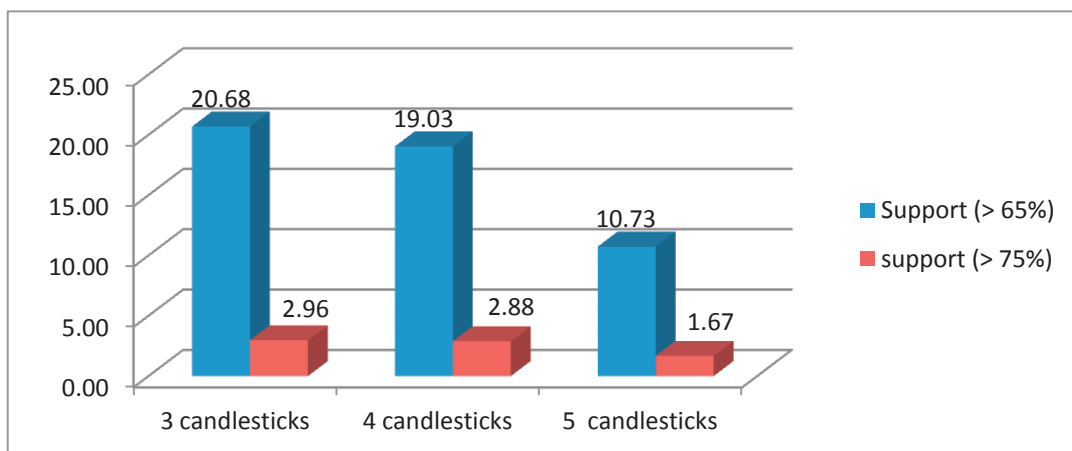


Figure 6: Pattern Results of 60 Min Time Frame

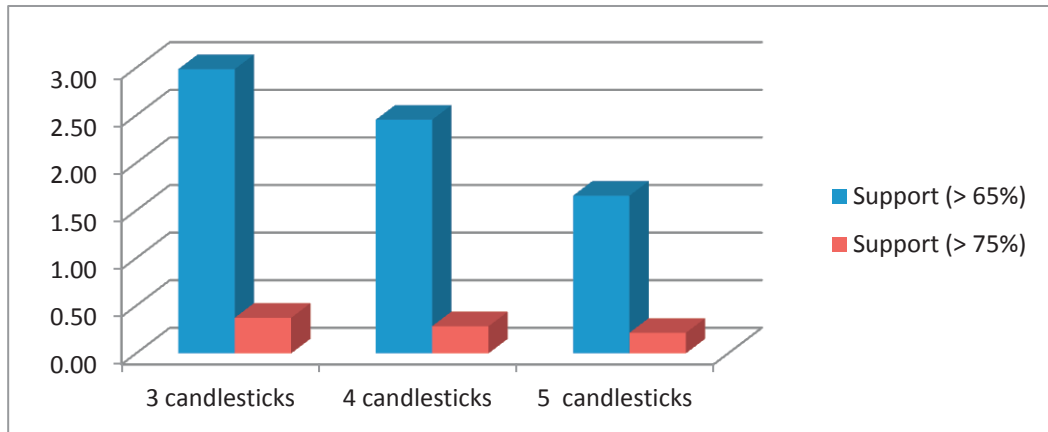


Figure 7: Pattern Results of 10 Min Time Frame

Figure 7 corresponds to 10 min intraday patterns, in which each bar, similarly to the previous figure, represents the support (%) of two predictions: accuracy greater than 65% (blue bar) and 75% (red bar). Comparing the above two figures, the smaller time frame has less support in predicting the future prices.

Totally, for the “60 min” it has been found 17,200 different patterns that correspond to 715,826 occurrences while for the “10 min” found 16,009 different patterns that correspond to 612,501 occurrences. It should be noted that the total estimation (by simulation) support for the 60 min patterns is less than 25% (not the some of all because of repeating patterns), while for 10 min patterns is low not more than totally 4%.

3.2 Portfolio Profitability Testing

The testing period was specified as 15 January 2012 to 15 August 2012 and contained 281,134 rows for the 60 min time frame and 1,808,061 rows of 10 min time. The initial investment is set to USD \$100,000, initially distributed equally among each position (totally 10 stocks for long and 5 stocks for short). The commission (fee) per transaction by default is set to be \$ 4.95 per transaction (which is the lowest in the market). Each trade costs about 2x\$4.95. The results of applying the portfolio trading algorithm for 7 month-testing periods are shown in Figure 8.

	<i>Net Profit \$ - %</i>	<i>Commissions in \$</i>	<i># of trades</i>
<i>60 min Time Frame (Long Positions) with /or not of reinvest from sell short</i>	5,245 (=5.245%) <u>12,659 (=12.659%)</u>	10,375	1,048
<i>60 min Time Frame (Short positions)</i>	2,753 (=2.753%)	3,277	331
<i>10 min Time Frame Long Positions) with /or not of reinvest from sell short</i>	1,531 (=1.531 %) <u>8,384 (=8.384%)</u>	13,088	1,322
<i>10 min Time Frame(Short</i>	520 (=0.520%)	4,079	412

positions)			
------------	--	--	--

Figure 8: Portfolio Performance Results

In Fig. 8, the total portfolio performance for the directional Long/Short strategy of 100/50 based on 60 min intraday patterns, is about 8% (with no reinvesting money from the short selling), while at the same time the S&P 500 was increased by 8,8 % (15 Jan 2012 was 1290.22 & 15 Aug 2012 was 1403.89). If we follow the Long/Short strategy of 150/50 by reinvesting the money received from shorting, then, the total net profit from long positions increases at **12.659%**. This can be explained because for each trade and for each stock we invest additional 5,000 \$ (totally 15,000 instead of 10,000 for each stock of long position) and additionally the commission is independent of the investing amount, but only dependent on the number of trades. Of course, this is true only for small amounts of each trade such as in our setup. The net profit per transaction for long positions is equal to:

$$\text{Net Profit} = (\text{Sell Price} - \text{Buy Price}) \times \text{Number of shares} - \text{commission (for Buy + for Sell)}$$

In the above equation, the net profit for specific trades (produced by a system) is dependent only on the number of shares. Therefore, the net profit percentage could be increased by investing more money in each stock or reducing the transaction costs by reducing the number of permitting opened positions (supposing we are applying the same strategy).

Regarding the strategy of using the 10 min intraday patterns, the performance of the portfolio was less than that of using a higher time frame (in our case 60 min). It seems, that this is due to the increased number of trades (higher transaction costs) and the lesser support of intraday patterns. It is noted that the algorithm of the 10 min intraday patterns for the portfolio construction was assisted by involving technical indicator trading signals comparing to the 60 min intraday patterns which was mainly based on intraday patterns.

4. Discussion & Conclusion

The results show that the proposed system can discover patterns that can assist intraday trading to select the best stocks whose value is expected to increase (take long positions) or to decrease (take short positions). This research has shown that it is possible to design pattern mining algorithms that are able to discover hitherto unknown useful sequences of candlesticks (3, 4 or 5), because the only patterns known to the community of traders and professionals are those from technical analysis which are no more than a hundred (100). These patterns cannot be visually detected by a trader not only due to the large number (thousands) of patterns but also due to the complexity of information they present.

The application of intraday patterns could be used additionally in various ways such as in producing an alerting mechanism to aid a trader in their decisions regarding trading at particular times (filter system), or as input in neural network systems.

It is expected that the profitability of the current portfolio trading could produce much better results than those presented above. We are in progress of applying an optimization method of all involved parameters contained in this algorithm. Some of the parameters we refer to are the following:

- Which types of predictions should count more: those having more accuracy for the next two periods or those having for the next 10 periods,
- The time period of buy and hold strategies.
- Resolving conflicts regarding predictions.
- When to involve technical indicator signals
- Money management (to minimize the cost of transactions discussed in previous paragraph).

Finally, it appears likely that higher time frames (e.g. daily candlesticks) may provide better accuracy in future time length prediction. Therefore, an investigation in discovering daily patterns may be more useful in long/short strategies.

References

- Alexander, S.S. 1961. Price movements in speculative markets: trends or random walks, *Ind. Mgmt. Rev.*, vol.2, pp. 7-26.
- Jensen M.C., Bennington, G.A., 1970. Random walks and technical theories: Some additional evidence, *J. of Fin.*, 25(2):469-482.
- Fama, E.F. 1970. Efficient capital markets: A review of theory and empirical work, *J. of Fin.*, 25:383-417.
- Brock, W., Lakonishok, J., LeBaron, B., 1992. Simple technical trading rules and the stochastic properties of stock returns, *J. of Fin.*, 47:1731-1764.
- Bessembinder, H., Chan, K., 1998. Market efficiency and the returns to technical analysis, *Fin. Mgmt.* 27:5-17.
- Lo, A.W., Mamaysky, H., Wang, J., 2000. Foundations of technical analysis: Computational Algorithms, Statistical Inference, and Empirical Implementation, *J. of Fin.*, 55:1705 – 1770.
- Lee, K.H., Jo, G.S., 1999. Expert system for predicting stock market timing using a candlestick chart. *Exp. Sys. Appl.* 16:257-364.
- Gagilalp, G., Laurent, H., 1998. The predictive power of price patterns. *App. Math. Fin.* 5:181-205
- Marshall, B.R., Young, M.R., Rose, L.C., 2006. Candlestick technical trading strategies: can they create value for investors? *J. Bank. Fin.* 30:2303-2323.
- Sheng, Y.P., Hou W.C., Chen, Z. 2006. Mining for profitable patterns in the stock market. In: *Encyclopaedia of Data Warehousing and Mining*, Idea Group Inc, 2006
- Cristophe Morel, , 2000, Stock selection using a multi-factor model - empirical evidence from the French stock market, *The European Journal of Finance*, 7(4):312-334
- George H. John, Peter Miller, 1996, *Building Long/Short Portfolios Using Rule Induction*, The IEE Conference on Computational Intelligence in Financial Engineering, New York 1996