Babeş-Bolyai University of Cluj-Napoca

Faculty of Economics and Business Administration

Business Intelligence: Data Mining in Finance

Scientific Adviser: Prof. Dr. Ştefan-Ioan Niţchi

PhD Candidate: Darie Moldovan

December, 2011

Contents

Contents of the thesis

Keywords

Data Mining, Quantitative trading, Time series analysis, Clustering, Genetic Algorithm

1. Introduction

The Business Intelligence enhances the integration of the innovation–creation processes, articulating the initiatives and operations designed for accelerating the business practices. The research and debate in this field allow the identification of contours and offensive or defensive methods of Business Intelligence, promoting the innovation and optimization and control of the technology transfer (geographically, interdisciplinary or cross-cultural).

In a knowledge-based society, the term “intelligence” is becoming more and more important for every level of the business society, from small and medium enterprises to multi-national companies, even if in many areas the concept is still new and somehow vague for executives.

Market globalization, the development of information technologies and Internet had increased the information needs and demanded for more powerful processing and analysis tools. The discipline that emerged in order to satisfy these needs is the Data Mining. It was born due to advances and interactivity in three domains: mathematics / statistics (allowing the creation of new efficient algorithms), the advances in the field of Artificial Intelligence and the database technologies (enhancing the possibility to store large amounts of data and having quick access to them). A broad definition given by (Zanasi, 1998) states that Data Mining is the ensemble of all the techniques that allow us to discover knowledge otherwise hidden in huge databases. The computer algorithms that make this possible are based on sophisticated mathematics and on new, intelligent ways of utilizing computer power.

Throughout the financial industry, data mining is used by the financial companies with the aim of obtaining an advantage over their competitors. Banks, asset management or consulting companies, they all own large amounts of data, collected from their current activity.

Considering these applications of data mining in finance, our interest goes to the trading data analysis. Mining financial data presents some challenges, difficulties and sources of confusion. Considering the efficient market theory a long term trend it is unlikely to be found, but the data miners' task is to determine short term trends, to validate them and be aware when they are not any more useful. Another important use of the financial data mining is the asses portfolios of products (financial products), aiming for optimal combinations.

In this thesis we offer a view of the data mining methods used in the finance industry for analysis and forecasting and their application on the Romanian stock market and on the clustering of the American Dow Jones Indexes. Having an interdisciplinary nature, the main concepts are gathered from finance, statistics, and Machine Learning (Genetic Algorithms).

In two of the chapters (2 and 3) we present the essential background theoretical knowledge gained while researching in these areas. We begin by offering a view on the developments made in the recent years in the field of automated trading on the world's financial markets (Chapter 2), exposing the needs for such processes in the context of market globalization and increasing competition between players, although most of the evidence regarding the trading algorithms and the real dimension of the business remain partially undisclosed due to privacy reasons from the most brokers.

The third chapter presents the methodologies used in the field of Data Mining, describing a brief history of the discipline, and presenting the main process model. Next, we describe in more details the Data Mining methods, showing an insight on the classic methods such as statistical methods, ARIMA, Bayesian classification or nonparametric learning, but also the Evolutionary and Genetic algorithms.

The fourth chapter is dedicated to the quantitative trading, having the aim of describing all the necessary conditions in order to be able to automatically trade on the Romanian stock exchange.

We begin by exposing the general trends in the field of algorithmic trading, then showing in detail the parameters for the performance evaluation of the strategies and the historical simulation environment. We conduct a dedicated statistical research in order to identify the information efficiency level of the Romanian market and we dedicate a distinct subchapter for the risk management issues. Next, we present the business model for the algorithmic trading, describing the specific methodology. We tested the proposed methodology on historical market data gathered from eleven stocks traded at Bucharest Stock Exchange and described our findings in the last part of the chapter.

Combining the nature's efficiency and the computer's speed, the financial possibilities are almost limitless, states (Bauer, 1994), referring to the computational power when a Genetic Algorithm is used in the financial forecasts.

Our approach in the fifth chapter is to obtain a business intelligence system prototype that uses financial data as input and by data mining techniques to group the stocks, considering only their price evolution during the same period of time, with the aim of finding new correlations and interesting, hidden, information. A clustering methodology proposed by Kasabov et al. (Chan & Kasabov, 2004), (Chan, Kasabov, & Collins, 2005) that considers only the evolution in time of the gene values, we found interesting to test in this domain too. Our interest is to investigate if the indirect interactions between stocks could be assimilated to gene interactions, determining their future evolution.

2. The financial markets and the Data Mining techniques

In the last ten years the collection of data has become a normal phenomenon for more and more businesses, especially the data regarding people and their behavior: phone calls made, shopping, visited places or financial transactions made. Back in 1999 it is said that Goldman Sachs, one of the leading investment banks in the US, was tracking more than one million time series, varying from financial instruments such as stocks and bonds to more personal information about their clients such as, for example, the vacations spending. At the same time, the division for proprietary trading at Morgan Stanley was collecting 10 Gigabytes of data every day (Weigend, 1999).

It is difficult to evaluate the dimensions of the data collected and analyzed nowadays by the two companies, but we can estimate it by looking at the growth rate of the e-business sector in the last decade: from a value of 27 billion dollars in 2000 to 176 billion dollars in 2011 (US Census Bureau, 2011) (Mulpuru, 2011).

In the field of financial transactions, the situation is even more impressive. In the last decade the financial market became almost fully electronic. The changes were so massive, that most of the trading is done automatically, based on trading algorithms (Aite Group, 2010).

In our opinion, there are several factors that made this evolution so rapid. Between them we mention:

1. One of the most important is the maturing of the Internet, the fact that in the modern society most of the people are now familiar with it and the rapid growth of the technology led to advances in the trading technology.

2. At the same time with the technology improvements, the investors became more sophisticated, needing quick execution systems, analysis instruments, in order to gain advantage in front of their competitors.

3. The globalization of this area brought, as a consequence, less costly transactions and enhanced access to the global markets for the investors. The exchange mergers and acquisitions determined a more fluent development of the trading systems, the direction being one of compatibility between the various systems around the world. The emerging markets took the business models from the developed markets and created new structures on the same frame.

4. This period was also one of maturity for the business models of the investment banks that evolved rapidly, creating a wide range of products, many of them very exotic, considered to satisfy the needs of the users and align the risk at predetermined levels.

5. The variety of instruments needs powerful tools for computation and quick reactions, many of them being interrelated, acting in a cascade manner. The inter-market analysis became a preoccupation of the financial analysts, the inter-connectivity of the markets giving the possibility for quick rebalance of the portfolios, with multi-asset, multi-market instruments.

3. Computational intelligence methods and tools

3.1 The data mining process

Data mining is the process of employing one or more computer learning techniques to automatically analyze and extract data, according to (Richard J.Roiger, 2002). Another definition, given by (Edelstein, 1999) states that data mining involves the use of sophisticated data analysis tools to discover previously unknown, valid patterns and relationships in large data sets. A simple definition chosen by (Nitchi & Avram-Nitchi, 1997) considers the data mining as the process of extracting predictive information hidden in large datasets.

The data mining has three major roots from which developed. The first is the statistics. Most of the data mining methods are based on a classic statistical approach. The techniques were developed with the aim of improving the usual statistical methods. The concepts of regression, standard deviation, distribution, variance, clustering, classification, confidence intervals are usually used by the data mining technologies and they represent the foundation of many modern algorithms.

The second component that influenced the data mining is the Artificial Intelligence approach. The AI attempts to model processes specific to the human mind, opposed in a way to the statistics. The AI concepts became more and more used along with the technologic capabilities increasing. They require high computational power, which always represented an issue, as the dimension of the problems increased with the same speed as the storage capacities and processing speed increased.

The third discipline that helped the development of the data mining is the databases. The databases concentrate on large-scale data. While for a statistician data mining is the inference of the models, for a database person data mining is a form of analytic processing (queries that examine large amounts of data) (Anand Rajaraman, 2011). The results of the same problem will be a query answer for the person involved in databases and the parameters of a model for the statistician.

The data mining can be defined as the union of the developments in statistics, AI and databases. The success of the data mining consists in the fact that managed to obtain (by applying its methods) hidden knowledge from the analysis of large amounts of data, probably impossible to find otherwise. It shows the difference between data and information: data mining transforms data into information.

For the success of a data mining implementation, systematization is needed on the form of a process model. In the literature we found a few models developed by the software producers or other interested organizations. They described the necessary steps of the process, leading the user from gathering the data until collecting the results. The most well-known models, according to (Edelstein, 1999), are 5A, developed by SPSS, SEMMA, proposed by SAS, but the one that caught most of the attention in the field was CRISP-DM, launched by a consortium of companies: NCR, Daimler-Benz, SPSS and OHRA.

3.2 The data mining methods

A variety of methods for data mining exists, depending on the goals proposed. There is a need for understanding them, their interconnectivity and classification. Two large categories can be defined: the exploration-oriented and discovery-oriented methods. (Sayad, 2011) The exploration methods are used to explain the past and bring into attention important aspects of the data, while the discovery methods try to build models.

The explorative methods study the value of a hypothesis, using generally the statistical methods (variance analysis, T-test, ANOVA). While, by principle, Data Mining tries to find new, useful knowledge about the data, the explorative methods remain less associated with the concept of Data Mining, while the discovery methods focus on its main objectives: models creation.

The discovery methods are following two principal ways to achieve their task. One is based on prediction and the other on description. The description methods interpret the data, studying the relations between data. The prediction methods focus on the behavior of data, creating models and based on them to predict the future values of the studied variables, but also arrange the data in a way that becomes understandable and useful for the end-user. These techniques are usually based on induction, the model learning some rules from a sample dataset and then tested on fresh data, until a certain acceptable level of accuracy is obtained.

A different approach in classifying the methods comes from the machine learning area: supervised and unsupervised learning. The supervised learning aims to define a function from supervised data that should predict the output value for a given input. This type of learning is based on classification or regression, which of them being used depending on the type of the problem in study. The unsupervised learning works with unlabeled data and has the target to discover a structure of the data. (Bishop, 2006) (Witten & Frank, 2005)

The Data Mining methods are under the influence of some parameters like: the data, the data types and the algorithms applied.

A data mining system must be properly integrated to be able to use it in a continuous manner. The most important issue is the relation with the database management system. Either is an offline or online analysis data mining system, the access to the database must be excellent designed.

3.3 Evolutionary methods

The growing rate of the data stored in databases led to a high need of quick analysis of the data, as valuable information resides in it. As the number of analysts is limited and also their capacity, the need of automatic processing and analysis has arisen. Besides the methods presented in the previous chapter, we propose the use of evolutionary methods for data mining, and in particular the genetic algorithms (GA).

The use of evolutionary algorithms is represented by stochastic search techniques, based on the abstraction of biological evolution (reproduction, mutation, recombination, selection). Each of the individuals of a certain studied population is considered a candidate solution. A fitness function evaluates them and calculates the quality of the solution. In this way, using natural selection, the individuals evolve through a selection procedure. Operators based on genetics are applied with pre-defined probabilities in this procedure so the stronger the “gene” of an individual, the higher the possibility for parts of its candidate solution to be transmitted to later generations of individuals. The mutations could result in genes that do not exist in the individuals of the first population. In opposition, if the operators are stochastic crossover operators, the gene values will not change, but they will be swapped between individuals.

The evolutionary algorithms can be used to solve problems in many areas of research, the two main issues being to choose the individual representation (which candidate solution an individual represents) and the fitness function to evaluate the individuals.

In problems of rule-discovery the individual is represented by a candidate rule or rules and the fitness function is represented by the measurement of the rules quality. The best rules of a generation will be selected and a genetic operator will transform the candidate rules into a new set of rules. In contrast with other algorithms, the evolutionary algorithms, by using stochastic operators, perform a global search of the rule space, a single operator being able to change a large number of the rule set. They also perform a complete evaluation on a candidate rule set, not leaving partial candidate rules. At the same time they have a high computational power, working with populations of candidates rules at a time.

In practice, it is recommended to use a combination between an induction algorithm and an evolutionary one, in order to gain improved results in the data mining tasks (Freitas A. A., 2002).

4 Quantitative trading

4.1 Overview on the Algorithmic Trading

The quantitative trading represents the process of securities trading automatically by a computer algorithm, without the direct interaction of a human, or according to Chan (Chan E. P., 2008) the trading based strictly on the buy/sell decisions of computer algorithms.

A report published by Aite Group (Aite Group, 2009) shows that in the past three years the algorithmic trading became dominant in the financial markets, with an impressive year to year growth. In 2009 the estimation made was that 70% of the daily traded volumes in US equities were realized in this automated way. This expansion in the recent years was stimulated by the large profitability the algorithms are providing. According to a FixProtocol

(Donefer, 2008) report the total profit resulted annually from the quantitative automated trading is around 20 billion dollars.

The algorithms are developed considering strategies used by traders, based on the historical data available, tested and improved. In this way, competitive strategies were created, obtaining automated solutions that will react quickly at the market changing conditions.

Most of the algorithms are high-frequency trading algorithms. The difference between high-frequency and low-frequency strategies is that in the first case a very low profit per trade is the target of the investment, a large capital is involved and of course the number of trades in a strategy is very high, trying to exploit most of the market movements. Positions are usually closed in the same day with very small profit, but the number of profitable trades should lead to comfortable gains.

Even if it is based historical data, it should not be confounded with the technical analysis. The technical analysis can be a part of the quantitative strategy if it can be input using a programming language. Also fundamental data can be incorporated in a strategy, news or comments about a certain company. The computational power of the machines can be used to make comparisons for fundamentals of thousands of companies or to interpret recent news much quicker than actually a person could read and understand it.

One important consequence of the development of the algorithmic trading is that the markets in which it activates become more efficient. The trading algorithms tend to exploit more and more the market inefficiencies, the new information being absorbed in the price more quickly. Another consequence is the increased liquidity generated by the high frequency strategies, a great benefit for the markets. The liquidity benefit in the markets where the trading algorithms are present consist also in lower costs for all the investors because of the smaller bid/ask spread and a lower risk for the investments, specifically the counterparty risk.

Nevertheless the trading algorithms lead to advances in computational power research, determining progress in computer science technology regarding the efficiency of the trading systems. In the recent years the execution time for the trades declined tremendously, because of the increased need of rapidity, concerning the trading algorithms. The competition between the algorithms developers, institutions or individuals that use them takes place not only on the market but also on the infrastructure built to support and increase the efficiency of the algorithms. A direct consequence of the technology advance is the dramatic reduce of the minimum period for an investment. The holding period can be reduced to seconds now.

The algorithmic trading is applied to foreign exchange markets, equities, futures and other derivatives. In the recent years more and more exotic products were developed to feed the need of the investors for risk, from hedging to extreme leverage. The dissemination of news, the speed and quality of financial analysis, the possibility of a very quick reaction to news led to more efficient markets, transparent and trustworthy for the investors.

The trading process has simplified along the years. The automation of quotations, the possibility to visualize live quotations and place orders in the market without the direct help of a broker, mean less costs for the clients. The main errors that were reduced due to technology improvements were the risk that the quotations transmitted to the client to be outdated (a trader was needed before, to check and communicate the quotations via phone or other slower methods) or the possibility not to understand right the quotations told.

According to a report released by Aite Group (Aite Group, 2010) in Europe the trades executed via algorithms or using Direct Market Acces increased to more than 50%, while in US the percentage is above 70%. The sophistication of the traditional customers and the new preferences for the algorithmic trading made the brokers to diversify their offer. The trend of electronic trading gaining more and more from the market is considered to continue in the following years. The report estimates that almost 100% of US equities to be traded in this way in 2010.

But as trading using algorithms became more and more popular, the clients gained more knowledge in this domain and request more and more advanced technologies. This evolution is leading to a change in the types of algorithms used. The high volatility in the markets during the last 2 years changed the preferences for more adaptive algorithms, the classic volume data based, becoming less popular. Real time analysis, risk management and news processing are requested in order to adapt quickly to the changing market conditions. They must incorporate real time variables and adjust dynamically the trading strategy.

4.2 An automatic trading strategy designed for Bucharest Stock Exchange stocks

4.2.1 Motivation

The use of technical analysis indicators in decision making for stock investments stays a controversial subject, being appreciated by some investors, but rejected by others (Edwards, Magee, & Bassetti, 2007). While professionals and researchers from the academic world developed new methods and indicators, live or simulated tests are needed to validate them (Silaghi & Robu, 2005).

The price prediction is a very complex issue, and selecting the right technical indicators for the analysis of a particular stock is one of the first preoccupations of the investors that use the technical analysis. One difficulty is the tuning of the parameters of these indicators in a way that makes their signals correct in a percentage as high as possible (Bodas-Sagi, Fernández, Hidalgo, Soltero, & Risco-Martín, 2009). While the behavior of the stocks is different from one to another and changes during time, the choice of parameters' values becomes a difficult task without the help of an advanced computational method.

The data mining methods are considered to be a smart choice for selecting the right technical indicators, allowing tests on very large datasets (an essential condition, regarding the large volume of financial data available) and many combinations of parameters' values, combining daily, weekly or monthly prices for tests (Bodas-Sagi, Fernández, Hidalgo, Soltero, & Risco-Martín, 2009) (Silaghi & Robu, 2005).

Our objective is to propose a methodology that combines different technical indicators, based on tests conducted over data sets gathered from the international or local stock markets, and obtaining buy or sell signals with an improved accuracy, compared to the results obtained using the use of a singular indicator, comparing the results with other research conducted.

4.2.2 Proposed methodology

The signal indicator considered in our model proposal is a combination of three momentum indicators used by the technical analysis and the benchmark will be if a signal aggregation will lead to better results than the signals gathered individually from the indicators.

The three indicators are MACD (Moving Average Convergence-Divergence), ROC (Price Rate of Change) and STS (Stochastic Oscillator).

- MACD is a widely used indicator and tracks the changes in strength, direction, momentum and direction of a trend. It is calculated considering the Exponential Moving Average (EMA) for two different periods and compares them (Gerald, 1999).

The formula for calculating an EMA at a certain point is the following:

$$EMA_t = EMA_{t-1} + \propto * (price_t - EMA_{t-1}), \tag{4.35}$$

where $\propto$ is a constant smoothing factor expressed like a percent or as the number of periods

.

Generally,

$$\text{EMA} = \propto * (p_1 + (1-\propto)p_2 + (1-\propto)^2 p_3 + (1-\propto)^3 p_4 + \cdots). \tag{4.36}$$

The weighting factor in each data point (p) is decreasing exponentially, so the older the data point, the less influence will have in the result.

Next, the MACD formula is the following:

$$\text{MACD} = EMA_a - EMA_b \text{ , where a < b.} \tag{4.37}$$

The trade signals are given when the EMA for the shorter period increases at a higher value than the longer EMA ($EMA_a > EMA_a$) - a buy signal - or the shorter EMA becomes smaller than the longer EMA($EMA_a < EMA_a$).

- The Price Rate of Change (ROC) indicator is an oscillator that calculates the movement of the price between the current time price and the one of n periods of time before.

The calculation formula is the following:

$$ROC = Price_t - Price_{t-n} \tag{4.38}$$

or the relative value:

$$ROC\% = \frac{P_t - P_{t-n}}{P_{t-n}} * 100 \tag{4.39}$$

where t is the current time and n is the number of periods of time in the back.

The ROC signals when a certain stock is overbought or oversold, the trading signals occurring when a divergence appear against the current price evolution.

- The Stochastic Oscillator measures the momentum of the market by considering the trading range from a certain period of time (Lane, 1984).

For the calculation we use the following formulas:

$$\%\mathrm{K} = 100\ \frac{\mathrm{C-L}}{\mathrm{H-L}} \tag{4.40}$$

where C represents the closing price of the stock, L the lowest price for the period, while H is the maximum price.

$$\%D = 3 \text{ period } MA(\%K), \tag{4.41}$$

where MA(%K) is the moving average of %K.

The algorithm gets in touch with two independent entities and an efficient communication between them must be assured. First the database from which the data for analysis is gathered in real time. This is a one way communication, from the database to the system. An important issue is the consistency of the data, missing or unformatted data is not acceptable, the technical indicators being very sensitive.

The other entity is the market (e.g. Stock Exchange). A mutual communication is needed in this case. The system sends trading orders to the market and the market sends responses whether they were executed or not. A very stable and fast connection is needed between the two, the execution speed being a very important factor for the success of a trading algorithm, sometimes the precision must be of milliseconds (Chaboud, Chiquoine, Hjalmarsson, & Vega, 2009).

Inside the system, the trading algorithm is the core. All other processes send their signals to the algorithm, which reacts depending on the calculations made, deciding whether to send orders to the market, close the open positions or not to react at all, just wait for the data to change so new information becomes available (Moldovan, Moca, & Nitchi, 2011).

The system will perform additional tasks, in addition to the trading signal aggregation. These tasks are the open positions management and risk management.

Figure 1 shows the integration of the algorithm in the whole system.

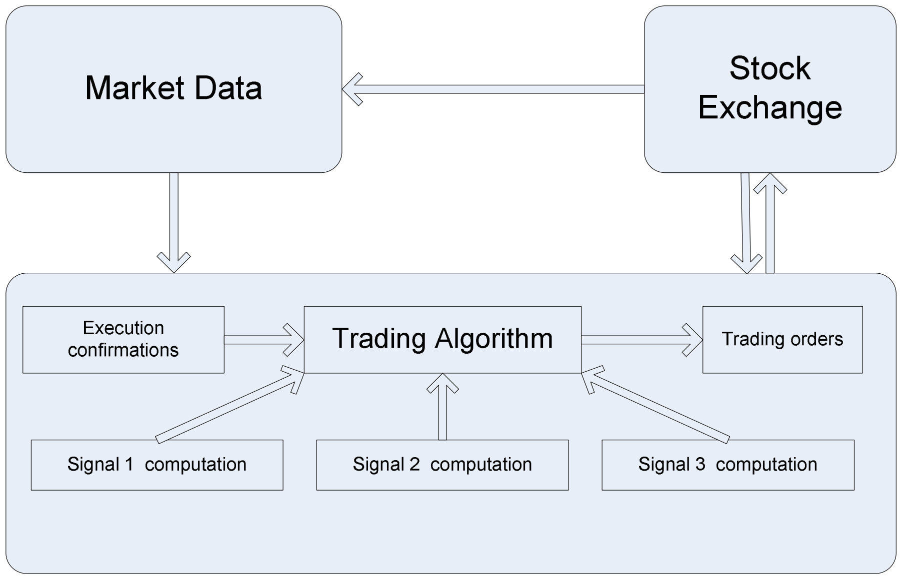


Figure 1. Integration

The functionality of the system depends not only on the optimal design, but also on the programming language used. (Russell & Yoon, 2005) consider the .Net framework the most appropriate due to its superior flexibility, scalability compatibility and interoperability.

Before entering a trading order, the system must do some validations regarding the management of the positions owned and the risk management.

A very often verification of the trading signal must be performed (this interval can be settled depending on the trading strategy, whether it will be a high frequency trading or not). If a signal was issued the status of the current open positions must be checked. An exposure limit that was reached will not allow an order to get in the market.

4.2.3 Experimental results

For implementing the methodology we used the AFL – Amibroker Formula Language (Amibroker), a programming language used to develop custom indicators, setup the risk management and back-testing.

For testing the proposed methodology we chose a group of eleven stocks from the Bucharest Stock Exchange, components of the BET and BET-Fi indexes. They are among the most liquid stocks on the market; therefore we considered them as the most relevant for testing the strategy. The time series containing the companies' evolutions start from January 2007, ending July 2011, with a number of approximately 50000 records each. The price history is recorded with a frequency of five minutes during the period mentioned above.

To define the benchmarks for the tests we considered several approaches. First, we tested the system against strategies using only one indicator, in accordance with our declared aim, of obtaining better results in trading by composing the signals of several indicators in order to perform better than the strategies using only one of the indicators.

Secondly, we compared the performances of the two indexes, BET and BET-FI (the „Buy and Hold" strategy). We compared the results in different time frames, one for the whole period and others for each year, in order to identify the behavior and performances of the system during different phases of the market. The „Buy and Hold" strategy assumes that the index is bought on the first day of the trading period and is sold in the last day, having only two trades for the whole period.

Another useful benchmark is the risk-free interest, which even if it belongs to a different category of risk it can be considered a base reference for every type of investment. More, it is used in the calculation of several fitness or performance indicators.

In the next table we provide the interest value for every year considered in our analysis and the benchmark indexes values (year to year percentage evolution). For the risk-free interest rate we used the policy rate determined by the National Bank of Romania.

Table 1. Benchmark values

	2007	2008	2009	2010	2011	Overall
Interest rate	7.42%	9.75%	9.06%	6.26%	6.25%	41.15%
BET	16.29%	-69.68%	57.2%	10.89%	-15.15%	-44.36%
BET-FI	14.95%	-83.62%	83.33%	-10.09%	-21.02%	-74.19%
MACD	19.73%	-53.88%	73.85%	-30.75	-19.16%	-45.13%
ROC	-2.74%	-8.35%	14.43%	-34.94%	-35.4%	-60.3%
St. Oscillator	18.9%	-21.88%	86%	-3.93%	-7.03%	53.13%
Proposed methodology	13.98%	8.27%	81.81%	3.45%	14.26%	146%

We can note that overall, the benchmark indexes had a negative performance, the investment in the stocks composing them generating significant losses in the case of a strategy of „Buy and Hold”. One can identify though periods with high volatility and positive returns.

From the strategies using the technical indicators, only the Stochastic Oscillator proved to be winning overall.

Since the market where we tested our proposed methodology allows only long positions (no short selling allowed) we tuned the parameters of the technical indicators in order to identify and take profit of the up trends and to avoid the as much as possible the falls of the market. Therefore the system has limitations concerning its use on the markets where short selling is permitted since it has not been tested for opening short positions. The only sells allowed were for closing the long positions.

The virtual initial equity for our tests was defined at 100000 RON, considering the reduced liquidity of the market. The indicators were calculated using hourly time series, generating in this way a sufficient number of trades to be conclusive for the tests.

Regarding the risk management, we determined the minimum position value to be of 10000 RON, and the maximum drawdown for a trade at 5%.

The tests were conducted in two directions, first considering a year to year approach, in this way resetting the portfolios at the beginning of each year to 100000 RON, not including the performance obtained during the past year; in the second approach we performed a test for the whole period, assuming that all the profits are reinvested, no cash withdraw being made and also no cash input. A commission per trade of 0.3% was taken into account when we evaluated the results.

4.3 Conclusion

The use of technical analysis indicators in decision making for stock investments stays a controversial subject, being appreciated by some investors, but rejected by others. While professionals and researchers from the academic world developed new methods and indicators, live or simulated tests are needed to validate them.

Following the statistical tests we applied on the daily returns of the Romanian stock index Bet-Fi, linear and non-linear correlations were found, the price of the stocks being mostly influenced by the new information that arrives in the market, the random walk hypothesis being rejected. We cannot sustain the existence of a weak form of information efficiency in this case, the usefulness of technical analysis not being rejected.

The combination of the three indicators in getting trading signals was successfully tested using the AFL –Amibroker Formula Language (Amibroker). We conducted the tests on time series for the period of 2007-2011 gathered from stocks traded on the Bucharest Stock Exchange and consistent results were obtained. Even the conditions imposed led to a prudent approach, the system, generally, performed better than the benchmarks and showing a very accurate control of the losses. Even in some cases the strategy using signals from an individual indicator performed better than system proposed by us, the overall results showed the approach of aggregating signals from the three indicators as being more effective.

For optimizing the system, we consider an integration with a Genetic Algorithm or other adaptive method is a highly desirable solution in order to tune up the parameters of the indicators in a quick and reliable way, but also they can be used in the discovery of new trading rules.

A combination between automatically discovered trading rules and others discovered by experts' observations could be a performing way of setting up the automated trading system.

Considering that every individual has different risk averseness, the risk-return balance must be controlled, which may lead in a development of an evolutionary trading system.

A limitation of the state of the art knowledge about developments in this field is a fact that could be an obstacle in the study, in most cases the developers of the algorithms do not make public the results and the methodology used.

5. Learning the stock market sectors

5.1 Overview

The behavior in time of a single stock can't describe the evolution of the entire market, but studied alongside with other ones, weighting their importance, one can tell the main direction of the group. For this reason stock indexes were created. In reverse, it is easier to forecast the price evolution of a single stock, taken away from a group where most of the stocks have a similar behavior.

A stock market index is a method for measuring a section of the market. In the last few decades, indexing has been a strong preoccupation for every fund manager, raising the performance expectations (Burr, 2005). Created by financial services companies or news providers, the indexes are the first benchmark for the performance of a portfolio.

There are many types of indexes, based on the size, specific sector, type of management or other criteria considered useful by their creators. Our approach is to obtain a business intelligence system prototype that uses financial data as input and by data mining techniques to group the stocks, considering only their price evolution during the same period of time, with the aim of finding new correlations and interesting, hidden, information. A clustering methodology proposed by prof. Kasabov et al. (Chan & Kasabov, 2004), (Chan, Kasabov, & Collins, 2005) that considers only the evolution in time of the gene values, we found interesting to test in this domain too. Our aim is to investigate if the indirect interactions between stocks could be assimilated to gene interactions, determining their future evolution. For experimentation we chose the 65 companies (traded on the New York Stock Exchange and NASDAQ) composing the DJA index. Since the number of cases is relatively small, the results will be easier to understand and verify.

5.2 Financial data clustering

A stock time-series clustering becomes valid if the price fluctuations within a group are correlated, and the price fluctuations between groups are uncorrelated, or less correlated.

Clustering statistics will tell how closely the stocks in a group resemble to each other. Differentiation between groups must be seen, in order to argue the need of the clusterization, which will lead us to a possible application of the financial clustering: by analyzing a stock's evolution we will determine which group the most behaves like, which is not necessarily the group the Dow Jones categorized it as.

As discussed in (Chan, Kasabov, & Collins, 2005), the possibilities of choosing a clustering algorithm vary from the classic K-means clustering, hierarchical clustering, Tree-based algorithm to the newer Autoregressive models, B-splines and the Multiple Linear

Regression Models (MLR).The last choice was chosen for the gene trajectory clustering. To solve the problem of the local optimization method used by the clustering method, the Genetic Algorithm was applied, obtaining a hybrid algorithm.

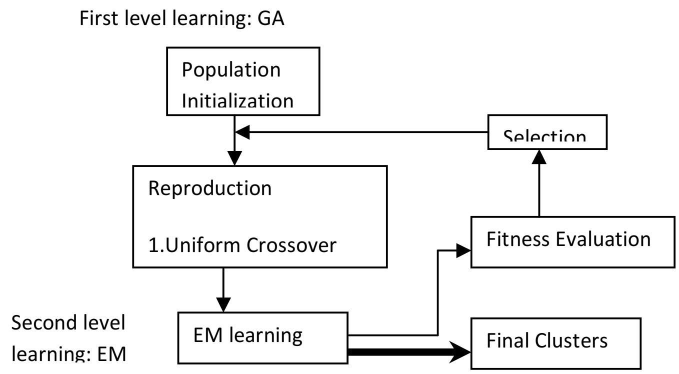


Figure 2. The clustering algorithm design (Chan, Kasabov, & Collins, 2005)

For the financial data to fit the GNetXP rigors we needed to make a preprocessing. For our tests we considered the daily adjusted close price for the stocks found in DJA's composition between the years 2000 and 2007. It is divided in three sub indexes (Dow Jones Industrial Average, Dow Jones Transportation Average and Dow Jones Utility Average) by the main field of activity of the companies composing it.

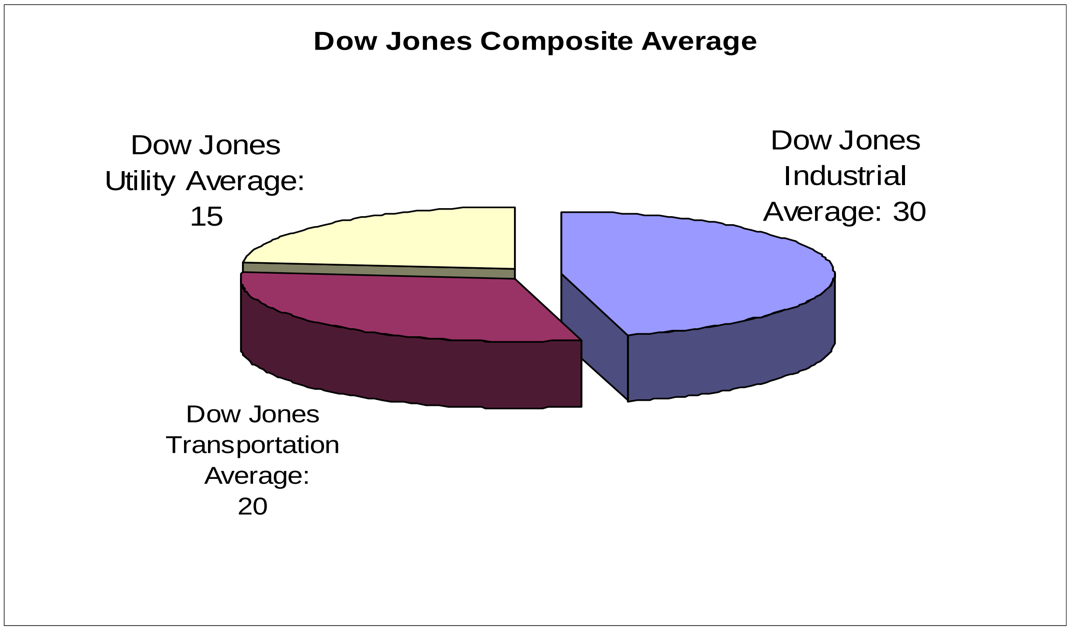


Figure 3. Dow Jones Composite Average components

We downloaded the historic data available on Yahoo! Finance[1] for each stock. The price itself doesn't say much about a stock's value among the others, a unique measure was needed to benchmark the performances. We started by calculating the daily logarithmic return of each stock, to obtain a homogeneous image of their evolution during the periods. To be able to project the data as trajectories we scaled it, considering a start point for every stock at 100 points at the beginning of every year, and applied the logarithmic return obtained at the previous step. Figure 4 shows the year 2000 data, ready for being tested.

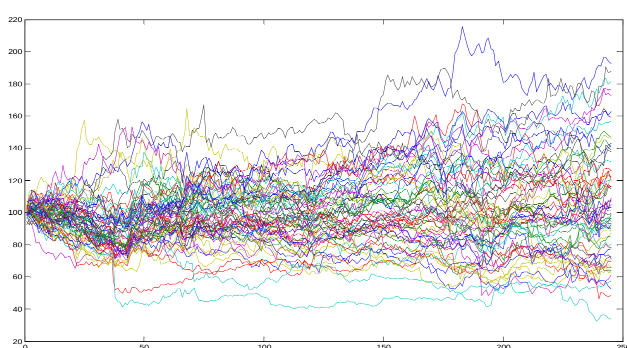


Figure 4. Data ready for testing

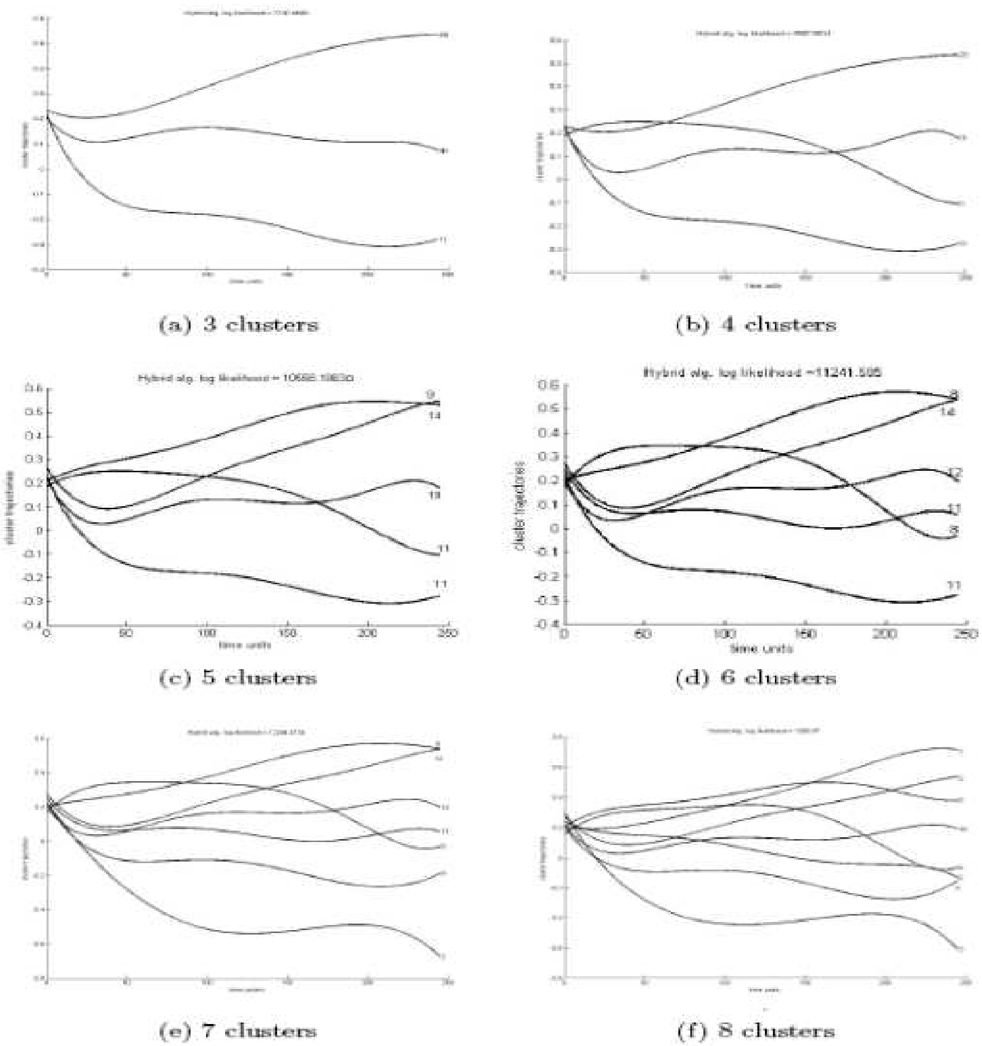
[1] http://finance.yahoo.com

5.3 Experiments results

We applied the procedure presented above for each of the eight datasets, from year 2000 to 2007. To obtain accurate results we ran the algorithm 30 times for each year. The hybrid algorithm performed as planned, just a few of the clusters being different from run to run.

Our interest in analyzing the results was on three directions:

- to observe the performance of the algorithm and to make sure the log likelihoods of the clustering are at an acceptable level;
- to see how the stocks are grouped, compared to the natural classification, based on industry;
- to find new, interesting correlations between stocks, from an economical point of view (Moldovan & Silaghi, 2009).



(a) 3 clusters (b) 4 clusters (c) 5 clusters (d) 6 clusters (e) 7 clusters (f) 8 clusters

Figure 5. Cluster trajectories obtained for different number of clusters

5.4 Conclusion

Our interests were in creating a business intelligence system prototype with use in financial data analysis, by help of the data mining processes.

After preprocessing the data, the algorithm was tested and the results were encouraging to continue the research. Considering the cluster analysis we concluded that the algorithm fitted well the data and it was appropriate for clustering the financial data.

The results showed that in many cases the natural division of the stocks is not the most adequate, and there are many correlations between stocks, that should be considered when investing.

We also found that building clusters using the proposed methodology is time consuming and for a more complex research or a system that needs almost real time results update there is a need for more powerful computational resources, such as the parallel computing.

6 Concluding remarks and future development

We organized the thesis into four main parts, from which the first two (Chapter 2 and 3) represented the theoretical support for the research, providing the necessary background in order to obtain an essential part of the knowledge needed before proceeding with the practical investigations. The two chapters provide both the business information and the methodologies for the research.

Our aims were to offer a business model for automated trading on the Romanian financial market, work detailed in chapter 4 and to investigate the possibility of application in finance of a methodology first used in the gene trajectory in medicine by (Chan & Kasabov, 2004) (Chan, Kasabov, & Collins, 2005), which we presented in chapter 5.

From the bibliographic research we found some insights in the domain of Financial Data Mining, we discovered the parameters of development of the financial sector, emphasizing the main interest and trends in the area of intelligent computational methods applied in finance. The data mining techniques can be approached to analyze financial time series, to find patterns and detect anomalies in the price behavior or to determine scenarios with a high-probability of success or risk. It may also be able to predict price movements, improve indicators and techniques, by the use of powerful intelligent methods such as neural networks or Genetic Algorithm and to combine the methods in order to obtain high accuracy predictions.

In the first part of Chapter 3 we described the general model of the Data Mining process, presenting the CRISP-DM model in particular, being also the model adopted by us for the research.

Our findings after the study of the opportunity of using intelligent methods to analyze the financial data were that they are recommended by the professionals in the industry as very powerful tools for forecasting and there is a large inter-operability of the methods between very different domains.

Chapter 4 was dedicated to the study of quantitative trading, describing the steps in building a completely automatic trading strategy, defining the parameters of performance evaluation and risk management.

The combination of the three indicators in getting trading signals was successfully tested. The tests were conducted on time series for the period of 2007-2011 gathered from stocks traded on the Bucharest Stock Exchange and consistent results were obtained. Even the conditions imposed led to a prudent approach, the system, generally, performed better than the benchmarks and showing a very accurate control of the losses. Even in some cases the a strategy using signals from an individual indicator performed better than system proposed by us, the overall results showed the approach of aggregating signals from the three indicators as being more effective.

We consider the integration with a Genetic Algorithm as a highly desirable solution in order to tune up the parameters of the indicators in a quick and reliable way, but also can be used in the discovery of new trading rules.

A combination between automatically discovered trading rules and others discovered by experts' observations could be a performing way of setting up the automated trading system. Considering that every individual has different risk averseness, the risk-return balance must be controlled, which may lead in a development of an evolutionary trading system.

Our interests shown in the fifth chapter were in creating a business intelligence system prototype with use in financial data analysis, by help of the data mining processes.

The approach was to try to apply in finance a methodology first used in a different domain (medicine): the gene trajectory study. Our aim was to find if natural division of the stocks, grouped in indexes, by natural reasons, such as their activity profile is the best solution. More, some correlations between stocks were expected and a wish for economic and logical explanations for these facts.

Considering the cluster analysis we concluded that the algorithm fitted well the data and it was appropriate for clustering the financial data.

The results showed that in many cases the natural division of the stocks is not the most adequate, and there are many correlations between stocks, that should be considered when investing.

Considering the business cluster analysis, we conclude that the GTC algorithm applied was appropriate for clustering the financial data and that there are many cases when the natural division of the stocks by the company profile is not a solution for grouping them, finding the technology companies uncorrelated with each other, and the banks correlated only in the last three years.

We also identified an area for improvement in use for the method tested: while the processing time was found to be long (between 4 to 5 minutes) and this could cause problems when used for active investing, in a real time environment, a distributed processing solution could be the answer.

Bibliography

Aite Group. (2009). *New World Order: The High Frequency Trading Community and Its Impact on Market Structure.* Boston.

Aite Group. (2010). *The European Equity Electronic Trading Landscape: How Deep Is Your Pool?* London.

Aldridge, I. (2009). *High-Frequency Trading: A Practical Guide to Algorithmic Strategies and Trading Systems.* Wiley.

Amibroker. (n.d.). Retrieved 2011, from http://www.amibroker.com/

Anand Rajaraman, J. U. (2011). *Mining of Massive Datasets.* Retrieved 2011, from http://infolab.stanford.edu/~ullman/mmds.html

Babu, G. M. (1994). Clustering with evolution strategies. *Pattern Recognition , 27* (2), 3210-330.

Banzhaf, W., Nordin, P., Keller, R. E., & Francone, F. D. (1997). *Genetic Programming: An Introduction.* Morgan Kaufman.

Bauer, R. J. (1994). *Genetic Algorithms and Investment Strategies.* Wiley Finance.

Berthold, M., & Hand, D. J. (2007). *Intelligent Data Analysis.* Springer.

Bishop, C. M. (2006). *Pattern Recognition and Machine Learning.* Springer.

Bodas-Sagi, D. J., Fernández, P., Hidalgo, J. I., Soltero, F. J., & Risco-Martín, J. L. (2009). Multiobjective optimization of technical market indicators. *Proceedings of the 11th Annual Conference Companion on Genetic and Evolutionary Computation Conference: Late Breaking Papers.* (pp. 1994-2004). New York: ACM.

Bollerslev, T. E. (1988). A Capital Asset Pricing Model with Time. *Journal of Political Economy , 96*, 116-131.

Bollerslev, T. (1986). Generalised Auto-Regressive Conditional Heteroscedasticity. *Journal of Econometrics* , 307-327.

Bollerslev, T. (2008). *Glossary to ARCH (GARCH).* School of Economics and Management, University of Aarhus.

Box, G. E. (1994). *Time Series Analysis: Forecasting and Control.* Prentice–Hall.

Burke, G. (1994). A Sharper Sharpe Ratio. *Futures* .

Burr, B. B. (2005). Essential book of indexing. *Gale Group* .

Carlin, P. S. (1992). A Monte Carlo approach to nonnormal and nonlinear state space modeling. *Journal of the American Statistical Association* , 493-500.

Chaboud, A., Chiquoine, B., Hjalmarsson, E., & Vega, C. (2009). *Rise of the Machines: Algorithmic Trading in the Foreign Exchange Market.* Federal Reserve Board - International Finance Discussion Papers.

Chan, E. P. (2008). *Quantitative Trading: How to Build Your Own Algorithmic Trading Business.* Wiley.

Chan, Z., & Kasabov, N. (2004). Gene trajectory clustering with a hybrid genetic algorithm and expectation maximization method. *IEEE International Conference on Neural Networks* (pp. 1669 - 1674). IEEE Computer Society.

Chan, Z., Kasabov, N., & Collins, L. (2005). A hybrid genetic algorithm and expectation maximization method for global gene trajectory clustering . *Bioinformatics and Computational Biology 3(5)* , 1227-1242.

Chao Jin, V. (2008). An extension of MapReduce for Parallelizing Genetic Algorithms. *eScience '08. IEEE Fourth International Conference on e-Science and Grid Computing*, (pp. 214-221).

Chen, R. T. (1993). Functional-Coefficient Autoregressive Models. *Journal of the American Statistical Association* , 298-308 .

Chen, R. T. (1993). Nonlinear Additive ARX Models. *Journal of the American Statistical Association* , 955-967 .

Cherkaue, K., & Shavlik, J. (1996). Growing simpler decision trees to facilitate knowledge discovery. *Proceedings of the 2nd International Conference of Knowledge Discovery & Data Mining* (pp. 315-318). AAAI Press.

De Jong, V. (1975). Analysis of the behavior of a class of genetic adaptive systems. *Ph.D. Disertation* . University of Michigan.

Deza, M. M. (2009). *Encyclopedia of Distances.* Springer.

Donefer, B. S. (2008). *Risk Management and Electronic Trading.* FIXProtocol.

Edelstein, H. A. (1999). *Introduction to Data Mining and Knowledge Discovery.* Two Crows Corporation.

Edwards, R. D., Magee, J., & Bassetti, W. (2007). *Technical Analysis of Stock Trends, 9th Edition (Hardcover).* American Management Association.

Embrechts, P. (1999). Extreme Value Theory as a Risk Management Tool. North American. *Actuarial Journal* .

Engle, R. L. (1987). Estimating Time Varying Risk Premia in the Term. *Econometrica , 55*, 391-407.

Engle, R. (2001). The use of ARCH/GARCH Models in Applied Econometrics. *Journal of Economic Perspectives* , 23-33.

Fama, E. (1965). The Behaviour of Stock Market Prices. *Journal of Business* , 34-105.

Financial Services Authority. (2009). *The turner review: A regulatory response to the global banking crisis*. Retrieved from http://www.fsa.gov.uk/pubs/other/turner_review.pdf

Fisher, R. A. (1928). Limiting forms of the frequency distribution of the largest or smallest member of a sample. *Proceeding of Cambridge Philoshophical Society*, (pp. 180-190).

Freitas, A. (2003). A survey of evolutionary algorithms for data mining and knowledge discovery. In *Advances in evolutionary computing* (pp. 819 - 845). Springer-Verlag.

Freitas, A. A. (2002). *Data Mining and Knowledge Discovery with Evolutionary Algorithms.* Springer.

Fuller, W. A. (1996). *Introduction to statistical time series.* Wiley Series in Probability and Statistics.

Gavrilov, M., Anguelov, D., Indyk, P., & Matwani, R. (2000). Mining the stock market: which measure is best? *Proceedings of the sixth ACM SIGKDD international conference on knowledge discovery and data mining* (pp. 487-496). ACM.

Gerald, A. (1999). *Technical Analysis Power Tools for Active Investors.* Financial Times Prentice Hall.

Glosten, L. J. (1993). On the Relation Between the Expected Value and. *Journal of Finance , 48*, 1779-1801.

Goldeberg, D. (1989). *Genetic Algorithms in Search, Optimization, and Machine Learning.* Addison-Wesley Professional.

Granger, C. W. (1978). *An Introduction to Bilinear Time Series Models.* Vandenheur and Ruprecht.

Hamilton, J. (1989). A New Approach to the Economic Analysis of Non-Stationary Time. *Econometrica* , 357-384.

Haupt, R. H. (2004). *Practical Genetic Algorithms.* Wiley Interscience.

Hoffmeister, F., & Bäck, T. (1991). Genetic self learning. *Towards a Practice on Autonomous Systems: Proceedings of the First European Conference on Artificial Life* (pp. 227-235). Paris: MIT Press.

Holland, J. (1975). *Adaptation in Natural and Artifical Systems.* Ann Arbor: University of Michigan Press.

J. Arneric, E. J. (2007). Theoretical Distributions in Risk Measuring on Stock Market. *8th WSEAS Int. Conference on Mathematics and Computers in Business and Economics.*

Jobson, J., & Korkie, B. M. (1981). Performance Hypothesis Testing with the Sharpe and Treynor Measures. *Journal of Finance* , 889-908.

John Y. Campbell, A. W. (1996). *The Econometrics of Financial Markets.* Princeton University Press.

Johnson, B. (2010). *Algorithmic Trading and DMA: An introduction to direct access trading strategies.* 4Myeloma Press.

Jorion, P. (2000). *Value at Risk: the new benchmark for managing financial risk.* McGraw-Hill.

Kalyanmoy, D. (2001). *Multi-Objective Optimization Using Evolutionary Algorithms.* John Wiley & Sons.

KDnuggets. (2007). Retrieved from KDnuggets.com: http://www.kdnuggets.com/polls/2007/data_mining_methodology.htm

Keating, S. (2002). A Universal Performance Measure. *Journal of Performance Measurement* .

Kestner, L. (1996). Getting a Handle on True Performance. *Futures* .

KIM, K. (2007). *Electronic and Algorithmic Trading Technology.* Burlington: Elsevier.

Kovalerchuk, B., & Vityaev, E. (2000). *Data Mining in Finance: Advances in Relational and Hybrid Methods.* New York: Kluwer Academic Publishers.

Lance, G. W. (1966). A General Theory of Classifactory Sorting Strategies. *Computer Journal* , 373-380.

Lane, G. (1984). Lane's Stochastics. *Technical Analysis of Stocks and Commodities* (2), pp. 87-90.

Larose, D. T. (2006). *Data Mining Methods and Models.* New Jersey: John Wiley & Sons.

Lewis, A. S. (1991). Nonlinear Modeling of Time Series Using Multivariate Adaptive Regression Splines . *Journal of the American Statistical Association* .

Lin, L., & Cao, L. (2008). Mining in-depth patterns in stock market. *International Journal of Intelligent Systems Technologies and Applications* .

Longerstaey, J., & Spencer, M. (1996). *RiskMetrics—Technical Document.* New York: Morgan Guaranty Trust.

Maggini, M., Giles, C. L., & Horne, B. G. (1997). *Financial Time Series Forecasting Using K-Nearest Neighbors Prediction.* Finance & Technology Publishers.

Makulowich, J. (1999). Government Data Mining Systems Defy Definition. *Washington Technology* .

Mardia, K., Kent, J., & Bibby, J. (1979). *Multivariate Analysis.* Academic Press.

McLachlan, G. K. (1997). *The EM Algorithm and Extensions.* John Wiley and Sons.

Mills, T. (1990). *Time series techniques for economists .* Cambridge University Press.

Mitchell, J. (1999). *An Introduction to Genethic Algorithms* (Fifth Edition ed.). MIT Press.

Moldovan, D. (2010). Testing the efficiency of the Romanian stock market. *International Conference on Development, Energy, Environment, Economics*, (pp. 378-381). Tenerife, Spain.

Moldovan, D., & Silaghi, G. C. (2009). A clustering of DJA stocks – an application in finance of a method first used in gene trajectory study. *Analele Universităţii Oradea , 5* (1), 1006-1011.

Moldovan, D., & Silaghi, G. (2009). Gene Trajectory Clustering for Learning the Stock Market Sectors. *Lecture Notes in Computer Science vol. 5494* , 559-569.

Moldovan, D., Moca, M., & Nitchi, S. (2011). A Stock Trading Algorithm Model Proposal, based on Technical Indicators Signals. *Informatica Economica , 15* (1), 183-188.

Mulpuru, S. (2011). *US Online Retail Forecast, 2010 To 2015 .* Forrester Research.

Mutu, S., Balogh, P., & Moldovan, D. (2011). The Efficiency of Value at Risk Models on Central and Eastern European Stock Markets. *International Journal of Mathematics and Computers in Simulation , 5* (2), 110-117.

Nelson, D. (1991). Conditional Heteroskedasticity in Asset Returns: A new Approach. *Econometrica , 59* (2), 347-370.

Nitchi, S., & Avram-Nitchi, R. (1997, Feb). Data mining, o noua era in informatica. *Byte Romania* .

Pardo, R. (2008). *The Evaluation and Optimization of Trading Strategies.* Wiley.

Priestley, M. B. (1980). STATE-DEPENDENT MODELS: A GENERAL APPROACH TO NON-LINEAR TIME SERIES ANALYSIS. *Journal of Time Series Analysis* .

Rechenberg, I. (1989). Evolution Strategy: Nature's Way of Optimization. *Optimization: Methods and Applications, Possibilities and Limitations* , 106-126.

Richard J.Roiger, M. G. (2002). *Data Mining: A Tutorial Based Primer.* Addison Wesley.

Russel, S., & Norvig, P. (2003). *Artificial Intelligence: A Modern Approach (2nd Edition) .* New Jersey: Pearson Education.

Russell, S., & Yoon, V. (2005). Heterogeneous Agent Development: A Multi-Agent System for Testing Stock Trading Algorithms. *AMCIS 2005 Proceedings.*

Sayad, S. (2011). *Real Time Data Mining.* Self-Help Publishers.

Sharpe, W. (1966). Mutual Fund Performance. *Journal of Business* , 119-138.

Sharpe, W. (1994). The Sharpe Ratio. *The Journal of Portfolio Management , 21* (1), 49-58.

Shearer, C. (2000). The Crisp-DM Model: The New Blueprint for Data Mining. *Journal of Data Wharehousing* , 13-22.

Silaghi, G. C., & Robu, V. (2005). An agent strategy for automated stock market trading combining price and order book information. *2005 ICSC Congress on Computational Intelligence Methods and Applications.* IEEE Computer Society.

Sortino, F., & van der Meer, R. (1991). Downside Risk. *Journal of Portfolio Management* .

Subramanian, H., Ramamoorthy, S., Stone, P., & Kuipers, B. (2006). Designing Safe, Profitable Automated Stock Trading Agents Using Evolutionary Algorithms. *Proceedings of the Genetic and Evolutionary Computation Conference.*

Tan, P.-N., Steinbach, M., & Kumar, V. (2005). *Introduction to Data Mining.* Addison-Wesley.

Thompson, S. (1999). Pruning boosted classifiers with a real valued genetic algorithm. *Knowledge-Based Systems , 12* (5-6), 277-284.

Todea, A. (2005). *Eficiența informațională a piețelor de capital. Studii empirice pe piața românească.* Cluj-Napoca: Casa Cărții de Stiință.

Tong, H. (1983). *Threshold Models in Non-linear Time Series Analysis.* Springer-Verlag.

Treynor, J. (1965). How to Rate Management of Investment Funds. *Harvard Business Review* .

Tsay, R. S. (2002). *Analysis of Financial Time Series.* Wiley-Interscience.

US Census Bureau. (2011, May 26). *E-stats.* Retrieved 2011, from US Census Bureau: http://www.census.gov/econ/estats/index.html

Vlaar, P. (2000). Value at risk models for Dutch bond portfolios. *Journal of Banking & Finance* , 1131-1154.

Ward, J. H. (1963). *Hierachical grouping to optimize an objective function.* Am. Statist. Assoc.

Weigend, A. S. (1999, January 20). *Data mining in finance - Course notes.* Retrieved 2011, from http://www-psych.stanford.edu/~andreas/Teaching/DataMiningFinance/S99/Notes/01Description.html

Witten, I. H., & Frank, E. (2005). *Data Mining: Practical Machine Learning Tools and Techniques.* Elsevier.

Wold, H. (1938). *A Study in the Analysis of Stationary Time Series.* Almqvist & Wiksells.

Young, T. (1991). Calmar Ratio: A Smoother Tool. *Futures* .

Zanasi, A. (1998). Competitive intelligence through data mining public sources. *Competitive Intelligence Review, 9* , 44-54.