Data mining in finance: From extremes to realism[1]

Boris Kovalerchuk
Professor, Department of Computer Science,
Central Washington University

Evgenii Vityaev
Senior Scientist, Institute of Mathematics,
Russian Academy of Sciences

Abstract

This paper describes data mining in finance by discussing financial tasks and specifics of methodologies and techniques in this data mining area. It includes time dependence, data selection, forecast horizon, measures of success, quality of patterns, hypothesis evaluation, problem ID, method profile, attribute-based, and relational methodologies.

1 This paper is a modified version of authors' chapter 'Data mining for financial applications' from the forthcoming 'Data Mining and Knowledge Discovery Handbook: A Complete Guide for Practitioners and Researchers,' Kluwer Acad. Publ. (Eds. O.Maimon and L.Rokach).

> 'October. This is one of the peculiarly dangerous months to speculate in stocks in. The others are July, January, September, April, November, May, March, June, December, August, and February.'
>
> Mark Twain, 1894

Financial tasks

Forecasting stock market, currency exchange rate, bank bankruptcies, understanding and managing financial risk, trading futures, credit rating, loan management, bank customer profiling, and money laundering analyses are core financial tasks for data mining [Nakhaeizadeh, Steurer & Bartmae (2002)]. Some of these tasks, such as bank customer profiling, have many similarities with data mining for customer profiling in other fields.

Stock market forecasting includes uncovering market trends, planning investment strategies, identifying the best time to purchase the stocks, and what stocks to purchase. Financial institutions produce huge datasets that build a foundation for approaching these enormously complex and dynamic problems with data mining tools. Potential significant benefits of solving these problems motivated extensive research for years.

Almost every computational method has been explored and used for financial modeling. We will name just a few recent studies: Monte-Carlo simulation of option pricing, finite-difference approach to interest rate derivatives, and fast Fourier transform for derivative pricing. New developments augment traditional technical analysis of stock market curves [Murphy (1999)] that have been used extensively by financial institutions. Such stock charting helps to identify buy/sell signals (timing 'flags') using graphical patterns.

Data mining as a process of discovering useful patterns and correlations has its own niche in financial modeling. Similar to other computational methods almost every data mining method and technique has been used in financial modeling. An incomplete list includes a variety of linear and non-linear models, multi-layer neural networks, k-means and hierarchical clustering, k-nearest neighbors, decision tree analysis, regression (logistic regression, general multiple regression), ARIMA, principal component analysis, and Bayesian learning.

Less traditional methods used include rough sets, relational data mining methods (deterministic inductive logic programming and newer probabilistic methods [Muggleton (2002), Kovalerchuk and Vityaev (2000)]), support vector machine, independent component analysis, Markov models, and hidden Markov models.

Bootstrapping and other evaluation techniques have been extensively used for improving data mining results. Specifics of financial time series analyses with ARIMA, neural networks, relational methods, support vector machines, and traditional technical analysis are discussed in Kovalerchuk and Vityaev (2000), Muller, et al. (1997), Murphy (1999), Tsay (2002).

The naive overly optimistic approach to data mining in finance assumes that somebody can provide a cookbook instruction on how to achieve the best result. Some publications continue to foster this unjustified belief, as is evident in commercial data mining software advertisements. Thus, it is not surprising that this overly optimistic cookbook extreme is a fertile ground for an opposite overly pessimistic extreme — nothing can be done.

In fact, the only realistic approach proven to be successful is providing comparisons between different methods conceptually not only based by their performance on a limited dataset. The useful comparison should describe domains of applicability of each method with strengths and weaknesses relative to problem characteristics. In this framework, a user selects a method by matching the data mining problem with these characteristics and task-specific circumstances. In essence, this approach means a clear understanding that data mining in general, and in finance specifically, is still more art than hard science. Fortunately, now there is a growing number of books that discuss issues of matching tasks and methods in a

regular way [Dhar and Stein (1997), Kovalerchuk and Vityaev (2000), Wang (2003)]. Selection of a method is a very complex task.

Uncertainty of problem descriptions (problem ID) and method capabilities (method ID) are among the most obvious difficulties in this process (Table 1). Dhar and Stein (1997) introduced and applied a unified vocabulary for business computational intelligence problems and methods that provide a framework for matching problems and methods. A problem is described using a set of desirable values (problem ID profile) and a method is described using its capabilities in the same terms.

Use of unified terms (dimensions) for problems and methods enhances capabilities of comparing alternative methods. Introducing dimensions also accelerates their clarification. Next, users should not be forced to spend time determining a method's capabilities (values of dimensions for the method). This is a task for developers, but users should be able to identify desirable values of dimensions using natural language terms as suggested by Dhar and Stein (1997).

In Table 1, we enhanced such problem description with desirable characteristics by including their necessary characteristics. These characteristics should be taken into account to be

		PROBLEM ID					METHOD CAPABILITIES									
	DIMENSIONS	Stock market forecasting	Trading futures	Portfolio management	Money laundering	Credit rating	Neural networks	ARIMA	Fuzzy logic	Deductive reasoning	Statistical methods	Decision trees	Association rules	Support vector machine	ILP	Probabilistic ILP
	Specifics of financial tasks:															
1	Multidimensional time series	Y	Y	Y			Y	Y		Y	Y			Y		Y
3	Specific efficiency criteria	Y	Y	Y	Y	Y		Y		Y	Y					Y
4	Multiresolution forecast	Y	Y					Y		Y	Y			Y		Y
5	Explained forecast			Y	Y	Y		Y	Y	Y		Y	Y		Y	Y
6	Subtle pattern	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y		Y
7	Include complex relations			Y	Y				Y	Y			Y		Y	Y
8	Use of background knowledge	Y	Y	Y	Y	Y			Y	Y					Y	Y
9	Significant level of noise	Y	Y	Y			Y	Y		Y	Y	Y				Y
10	Control of overfitting	Y	Y	Y				Y			Y			Y		Y
	Data types:															
1	Attribute-based	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y		
2	Relational data types			Y		Y			Y	Y		Y	Y		Y	Y
3	Mixed data	Y	Y	Y	Y	Y			Y	Y	Y					Y
	Hypotheses/models:															
1	Functional	Y	Y	Y	Y	Y	Y	Y	Y		Y			Y		Y
2	Symbolic	Y	Y	Y	Y	Y			Y	Y	Y	Y	Y		Y	Y
	Assumptions:															
1	Probability of events exists	Y	Y					Y			Y					Y
2	Occam's razor	Y	Y	Y	Y	Y			Y	Y		Y	Y		Y	Y

Table 1: Financial problem IDs and data mining methods' capabilities.
In this table, ILP stands for Inductive Logic Programming methods [Muggleton (1999, 2002)]. For explanation of some other terms used in this table, see below.

sure that a method could solve the problem in principle. If a method does not have necessary characteristics, but it has some desirable ones (i.e. user-friendly interface) it would not be sufficient. For instance, for stock market forecasting it is necessary that a method is able to work with multidimensional time series, to use specific financial efficiency criteria, to predict market for different forecast horizon providing a multiresolution forecast, to discover and use subtle patterns, and to work with a high level of noise. These characteristics are presented in Table 1.

Attribute-based learning methods such as neural networks, the nearest neighbors method, and decision trees dominate in financial applications of data mining. These methods are relatively simple, efficient, and can handle noisy data. However, these methods have two serious drawbacks: a limited ability to represent background knowledge and the lack of complex relations. Relational data mining techniques that include Inductive Logic Programming (ILP) [Muggleton (1999), Dzeroski (2002)] intend to overcome these limitations (last two columns in Table 1).

Previously these methods have been relatively computationally inefficient and had rather limited facilities for handling numerical data [Bratko and Muggleton (1995)]. Currently these methods are enhanced in both aspects [Kovalerchuk and Vityaev (2000)] and are especially actively used in bioinformatics [Muggleton (2002), Vityaev et al. (2002)]. We believe that now is the time for applying these methods to financial analyses more intensively, especially to those analyses that deal with probabilistic relational reasoning.

Various publications have estimated the use of data mining methods like hybrid architectures of neural networks with genetic algorithms, chaos theory, and fuzzy logic in finance. 'Conservative estimates place about U.S.$5 billion to U.S.$10 billion under the direct management of neural network trading models. This amount is growing steadily as more firms experiment with and gain confidence with neural networks techniques and methods' [Loofbourrow (1995)]. Table 1 contains only base methods, not hybrid methods that are their combinations. However, Table 1 shows that existing methods do not individually meet challenges of financial tasks. Hybrids are really needed to meet these challenges. In more detail, we discuss this issue in Kovalerchuk and Vityaev (2000). Many proprietary financial applications exist and use hybrid data mining methods, but without reporting publicly about methods used as was stated in Von Altrock (1997) and Groth (1998).

Specifics of data mining in finance

The challenge for financial data mining is coming from several difficult to accomplish specific requests:

- Forecast multidimensional time series with high level of noise.
- Accommodate specific efficiency criteria (i.e. the maximum of trading profit) in addition to prediction accuracy, such as R^2.
- Make coordinated multiresolution forecast (minutes, days, weeks, months, and years).
- Incorporate a stream of text signals as input data for forecasting models (i.e. the Enron case, September 11, and others).
- Be able to explain the forecast and the forecasting model (black box models have limited interest and future for significant investment decisions).
- Be able to benefit from very subtle patterns with a short life time.
- Incorporate the impact of market players on market regularities.

These requests are major points for selecting data mining method as Table 1 summarizes.

The current efficient market theory/hypothesis discourages attempts to discover long-term stable trading rules/regularities with significant profit. This theory is based on the idea that if such regularities exist they would be discovered and used by the majority of the market players. This would make

rules less profitable and eventfully useless or even damaging.

Greenstone and Oyer (2000) examine the month by month measures of return for the computer software and computer systems stock indices to determine whether these indices' price movements reflect genuine deviations from random chance using the standard t-test. They concluded that although Wall Street analysts recommended to use the summer swoon rule (sell computer stocks in May and buy them at the end of summer) this rule is not statistically significant. However they were able to confirm several previously known 'calendar effects', such as the 'January effect' noting meanwhile that they are not the first to warn of the dangers of easy data mining and unjustified claims of market inefficiency.

The market efficiency theory does not exclude that hidden short-term local conditional regularities may exist. These regularities can not work forever, and they should be corrected frequently.

It has been shown that the financial data are not random and that the efficient market hypothesis is merely a subset of a larger chaotic market hypothesis [Drake and Kim (1997)]. This hypothesis does not exclude successful short-term forecasting models for prediction of chaotic time series [Casdagli and Eubank (1992)].

Data mining does not try to accept or reject the efficient market theory. Data mining creates tools, which can be useful for discovering subtle short-term conditional patterns and trends (Table 1) in a wide range of financial data. This means that retraining should be a permanent part of data mining in finance and any claim that a silver bullet trading has been found should be treated similarly to claims that a perpetuum mobile has been discovered.

The impact of market players on market regularities stimulated a surge of attempts to use ideas of statistical physics in finance [Bouchaud (2000)]. If an observer is a large marketplace player then such observer can potentially change regularities of the marketplace dynamically. Attempts to forecast in such a dynamic environment with thousands of active agents leads to much more complex models than traditional data mining models were designed for. This is one of the major reasons that such interactions are modeled using ideas from statistical physics rather than from statistical data mining. The physics approach in finance [Voit (2003), Ilinski (2001), and Mandelbrot (1997)] is also known as 'econophysic' and 'physics of finance'. The issue is that the data mining approach is in essence not about developing specific methods for financial tasks, but the physics approach is. It is deeper integrated into the finance subject matter. For instance, Mandelbrot (1997) (known for his famous work on fractals) worked also on proving that the price movement's distribution is scaling invariant.

Data mining approach covers empirical models and regularities derived directly from data and almost only from data with little domain knowledge explicitly involved. Historically, in many domains, deep field-specific theories emerge after the field accumulates enough empirical regularities. We see that the future of data mining in finance would be to generate more empirical regularities and combine them with domain knowledge via generic analytical data mining approach [Mitchel (1997)]. One of the first attempts in this direction is presented in Kovalerchuk and Vityaev (2000).

Time series analysis

A temporal dataset T, called a time series, is modeled in an attempt to discover its main components, such as long-term trend, L(T), cyclic variation, C(T), seasonal variation, S(T) and irregular movements, I(T). Assume that T is a time series, such as daily closing price of a share or S&P 500 index, from moment 0 to current moment k, then the next value of the time series T(k+n) is modeled by formula (1):

$$T(k+n)= L(T)+ C(T)+ S(T)+ I(T) \quad (1)$$

Traditionally classical ARIMA models occupy this area for finding parameters of functions used in formula (1). ARIMA models

are well developed but are difficult to use for highly non-stationary stochastic processes that is typical in finance. Potentially, data mining methods can be used to build such models to overcome ARIMA limitations. The advantage of this four-component model in comparison with black box models, such as neural networks, is that components in formula (1) have an interpretation.

Data selection and forecast horizon

Data mining in finance has the same challenges as general data mining has in data selection for building models. In finance, this question is tightly connected to the selection of the target variable. There are several options for target variable y: y=T(k+1), y=T(k+2),...,y=T(k+n), where y=T(k+1) represents forecast for the next time moment, and y=T(k+n) represents forecast for n moments ahead. Selection of dataset T and its size for a specific desired forecast horizon n is a significant challenge.

For stationary stochastic processes the answer is well-known, that a better model can be built for longer training duration. For financial time series, such as S&P 500 index, this is not the case [Mehta and Bhattacharyya (2004)]. Longer training duration may produce many and contradictory profit patterns that reflect bear and bull market periods. Models built using too short durations may suffer from overfitting and hardly applicable to the situations where market is moving from the bull period to the bear period. Also in finance, the long-horizon returns could be forecasted better than short-horizon returns depending on the training data used and model parameters [Krolzig and Toro (2004)].

In standard data mining it is typically assumed that the quality of the model does not depend on frequency of its use. In financial application the frequency of trading is one of the parameters that impact the quality of the model. This happens because in finance the criterion of the model quality is not limited to the accuracy of prediction, but is driven by profitability of the model. It is obvious that frequency of trading impacts the profit as well as the trading rules and strategy.

Measures of success

Traditionally, the quality of financial data mining forecasting models has been measured by the standard deviation between forecast and actual values on training and testing data. This approach works well in many domains, but this assumption should be revisited for trading tasks. Two models can have the same standard deviation but may provide very different trading returns. The small R^2 is not sufficient to judge that the forecasting model will correctly forecast stock change direction (sign and magnitude). More appropriate measures of success in financial data mining are measures, such as average monthly excess return (AMER) and potential trading profits (PTP) [Greenstone and Oyer (2000)]:

$$AMER_j = R_{ij} - \beta_i R_{500j} - \left(\sum_{j=1}^{12} (R_{ij} - \beta_i R_{500j})/12\right),$$

Where R_{ij} is the average return for the S&P 500 index in industry i and month j and R_{500j} is the average return of the S&P 500 in month j. The β_i values adjust the AMER for the index's sensitivity to the overall market. A second measure of return is potential trading profits (PTP):

$$PTP_{ij}=R_{ij}- R_{500j}$$

PTP shows investor's trading profit versus the alternative investment based on the broader S&P 500 index.

Quality of patterns and hypothesis evaluation

An important issue in data mining in general and in finance in particular is the evaluation of quality of discovered pattern P measured by its statistical significance. A typical approach assumes the testing of the null hypothesis H that pattern P is not statistically significant at level a. A meaningful statistical test requires that pattern parameters, such as the month(s) of the year and the relevant sectoral index in a trading rule pattern P, have been chosen randomly [Greenstone and Oyer (2000)]. In many tasks, this is not the case.

Greenstone and Oyer argue that in the 'summer swoon' trading rule mentioned above, the parameters are not selected

randomly, but are produced by data snooping — checking combinations of industry sectors and months of return and then reporting only a few significant combinations. This means that rigorous tests would require testing a different null hypothesis not only about one significant combination, but also about the family of combinations. Each combination is about an individual industry sector by month's return. In this setting, the return for the family is tested versus the overall market return.

Several testing options are available. Sullivan et al. (1999) use a bootstrapping method to evaluate statistical significance of such hypotheses adjusted for the effects of data snooping in trading rules and calendar anomalies. Greenstone and Oyer (2000) suggest a simple computational method — combining individual t-test results by using the Bonferroni inequality. Another option would be to test whether the statements are jointly true using the traditional F-test. However if the null hypothesis about a joint statement is rejected it does not identify the profitable trading strategies.

The sequential semantic probabilistic reasoning that uses statistical F-test addresses this issue. Kovalerchuk and Vityaev (2000) were able to identify profitable and statistically significant patterns for the S&P 500 index using this method. These types of methods are identified in Table 1 as Probabilistic ILP, where ILP stands for Inductive Logic Programming. We also were able to demonstrate that this approach can be beneficial for uncovering money laundering schemes in forensic accounting. [Klösgen and Zytkow (2002), Kovalerchuk and Vityaev (2003)]. This technique that combines first-order logic and probabilistic semantic inference is discussed in the next section.

Relational data mining in finance

Decision tree methods are very popular in data mining applications in general and in finance specifically. They provide a set of human readable, consistent rules, but discovering small trees for complex problems can be a significant challenge in finance. In addition, rules extracted from decision trees fail to compare two attribute values, as it is possible with relational methods.

It seems that relational data mining methods also known as relational knowledge discovery methods are gaining momentum in different fields [Muggleton (2002), Dzeroski (2002), Vityaev et al. (2002), Cowan (2002)]. Below we provide some major concepts from relational data mining.

Data can be represented by attributes A_1, A_2,...,A_n of objects, that is each object x is given by a set of values $A_1(x)$, $A_2(x)$,...,$A_n(x)$. The common data mining methodology assume this type of data. Such data form the base of an attribute-based or attribute-value methodology. It covers a wide range of statistical and connectionist (neural network) methods. Examples of the most popular attributes in financial time series are index value at open, index value at close, highest index value, lowest index value, and trading volume and lagged returns from the time series of interest. Fundamental factors include the price of gold, retail sales index, industrial production indices, and foreign currency exchange rates. Technical attributes include variables that are derived from time series such as moving averages.

The relational data type is a second data type, where objects are represented by their relations with other objects, for instance, $x > y$, $y < z$, $x > z$. In this example we may not know that x=3, y=1 and z=2, but we know relations between x, y and z. Objects may have different attribute values (i.e., x=5, y=2, and z=4), but still have the same relations. A less traditional relational methodology is based on the relational data type.

Many data mining methods assume a functional form of the relationship. For instance, the linear discriminant analysis assumes linearity of the border that discriminates between two classes in the space of attributes. Often it is hard to justify such functional form in advance. Relational data mining methodology intends to learn symbolic relations on numerical data. The following technical analysis rule is in this category. To derive a conclusion it compares values of two attributes

such, as 5 and 15 day moving averages (ME5 and ME15) and derivatives of moving averages for 10 and 30 days (DerivativeME10, DerivativeME30):

If ME5(t)=ME15(t) & DerivativeME10(t)>0 DerivativeME30(t)>0, then buy stock at moment (t+1).

This rule can be read as 'If moving averages for 5 and 15 days are equal and derivatives for moving averages for 10 and 30 days are positive then buy stock on the next day.'

Data mining in finance leads the application of relational data mining for multidimensional time series, such as stock market time series. A. Cowan, a senior financial economist from U.S. Department of the Treasury, noticed that examples and arguments available in Kovalerchuk and Vityaev (2000) for the application of relational data mining methods, such as Machine Method for Discovering Regularities (MMDR), to financial problems produce expectations of great advancements in this field in the near future for financial applications [Cowan (2002)].

It was strengthened in several publications by suggestions that relational data mining area is moving toward probabilistic first-order rules to avoid the limitations of deterministic systems [Muggleton (2002)]. Relational methods in finance, such as MMDR, are equipped with probabilistic mechanism that is necessary for time series with high level of noise. MMDR is well suited to financial applications given its ability to handle numerical data with high levels of noise [Cowan (2002)].

Informally, the idea of semantic probabilistic reasoning used in MMDR method is coming from the principle of Occam's razor (a law of simplicity) in science and philosophy. For trading it was written as follows in Occam's razor! (2004):

- When you have two competing trading theories which make exactly the same predictions, the one that is simpler is the better and more profitable one.
- If you have two trading/investing theories which both explain the observed facts then you should use the simplest one until more evidence comes along.
- The simplest explanation for a commodity or stock price movement phenomenon is more likely to be accurate than more complicated explanations.
- If you have two equally likely solutions to a trading or day trading problem, pick the simplest.
- The price movement explanation requiring the fewest assumptions is most likely to be correct.

Conclusion

To be successful a data-mining project should be driven by the application needs and results should be tested quickly. Financial applications provide a unique environment where efficiency of the methods can be tested almost instantly, not only by using traditional training and testing data but making real stock forecast and testing it the same day. This process can be repeated daily for several months collecting quality estimates.

The relational data mining methods outlined in this paper advances pattern discovery methods that deal with complex numeric and non-numeric data, involve structured objects, text, and data in a variety of discrete and continuous scales (nominal, order, absolute, and so on).

Currently the success of data mining exercises has been reported in literature extensively. Typically it is done by comparing simulated trading and forecasting results with results of other methods and real gain/loss and stock. For instance, recently Huang et al. (2004) claimed that data mining methods achieved better performance than traditional statistical methods in predicting credit ratings. Much less has been reported publicly on success of data mining in real trading by financial institutions. It seems that the market efficiency theory is applicable to reporting success. If real success is reported then competitors can apply the same methods and the advantage will disappear because in essence all fundamental data mining methods are not proprietary.

The next step is for the development of practical decision support software tools that make it easier to operate in a data mining environment specific for financial tasks, where hundreds and thousands of models, such as neural networks and decision trees, need to be analyzed and adjusted every day with a new data stream coming every minute.

Inside of the field of data mining in finance we expect an extensive growth of hybrid methods that combine different models and provide a better performance than can be achieved by individuals. In such an integrative approach individual models are interpreted as trained artificial experts. Therefore their combinations can be organized similar to a consultation of real human experts. Moreover, these artificial experts can be effectively combined with real experts. It is expected that these artificial experts will be built as autonomous intelligent software agents. Thus, experts to be combined can be data mining models, real financial experts, trader, and virtual experts that runs trading rules extracted from real experts. A virtual expert is a software intelligent agent that is in essence an expert system. We coined a new term 'expert mining' as an umbrella term for extracting knowledge from real human experts that is needed to populate virtual experts.

We expect that in coming years data mining in finance will be shaped as a distinct field that blends knowledge from finance and data mining, similar to what we see now in bioinformatics, where integration of field specifics and data mining is close to maturity. We also expect that the blending with ideas from the theory of dynamic systems, chaos theory, and physics of finance will deepen.

References

- Bouchaud, J., and M. Potters, 2000, Theory of Financial Risks: From Statistical Physics to Risk Management, Cambridge Univ. Press, Cambridge, UK.
- Bratko, I., and S. Muggleton, 1995, "Applications of Inductive Logic Programming. Communications of ACM," 38:11, 65-70
- Casdagli, M., and S. Eubank, 1992, (Eds). Nonlinear modeling and forecasting, Addison Wesley
- Cowan, A., 2002, "Book review: Data Mining in Finance," International Journal of Forecasting, 18:1, 155-156
- Dhar, V., and R. Stein, 1997, Intelligent decision support methods, Prentice Hall
- Dzeroski S., 2002, "Inductive logic programming approaches," In: Klösgen W., and J. Zytkow, Handbook of data mining and knowledge discovery, Oxford University Press, 348-353
- Drake, K., and Y. Kim, 1997, "Abductive information modeling applied to financial time series forecasting," In: Nonlinear financial forecasting, finance and technology, 95-109
- Groth, R., 1998, Data Mining, Prentice Hall
- Greenstone, M., and P. Oyer, 2000, "Are there sectoral anomalies too? The pitfalls of unreported multiple hypothesis testing and a simple solution," Review of Quantitative Finance and Accounting, 15, 37-55
- Huang, Z., H. Chen, C. J. Hsu, W. H. Chen, and S. Wu, 2004, "Credit rating analysis with support vector machines and neural networks: a market comparative study," Decision support systems, 37:4, 543-558
- Ilinski, K., 2001, Physics of finance: Gauge modeling in non-equilibrium pricing, Wiley
- Klösgen W., and J. Zytkow, 2002, Handbook of data mining and knowledge discovery, Oxford University Press, Oxford
- Kovalerchuk, B., and E. Vityaev, 2000, Data mining in finance: Advances in relational and hybrid methods, Kluwer, 2000.
- Kovalerchuk, B., E. Vityaev, and J. F. Ruiz, 2001, "Consistent and complete data and 'mining' in medicine," In: Medical data mining and knowledge discovery, Springer, 238-280
- Kovalerchuk, B., and E. Vityaev, 2003, "Detecting patterns of fraudulent behavior in forensic accounting," In Proceedings of the Seventh International Conference "Knowledge-based Intelligent Information and Engineering on Systems," Oxford, UK, Sept, Part 1, 502-509
- Krolzig, M., and J. Toro, 2004, "Multiperiod forecasting in stock markets: A paradox solved," Decision Support Systems, 37:4, 531-542
- Loofbourrow, J., and T. Loofbourrow, 1995, "What AI brings to trading and portfolio management," In: Freedman R., R. Klein, and J. Lederman, Artificial intelligence in the capital markets, Irwin, Chicago, 3-28
- Mandelbrot, B., 1997, Fractals and scaling in finance, Springer
- Mehta, K., and S. Bhattacharyya, 2004, "Adequacy of training data for evolutionary mining of trading rules," Decision Support Systems, 37:4, 461-474
- Mitchell, T., 1997, Machine learning, McGraw Hill
- Muller, K. R., A. Smola, G. Rtsch, B. Schlkopf, J. Kohlmorgen, and V. Vapnik, 1997, "Using support vector machines for time series prediction," In: Advances in Kernel methods support vector learning, MIT Press
- Murphy, J., 1999, Technical analysis of the financial markets: A comprehensive guide to trading methods and applications, Prentice Hall
- Muggleton, S., 2002, "Learning structure and parameters of stochastic logic programs," 12th International Conference, ILP 2002, Sydney, Australia, July 9-11. Lecture notes in Computer Science 2583, Springer 2003, 198-206
- Muggleton S., 1999, "Scientific Knowledge Discovery Using Inductive Logic Programming," Communications of ACM, 42:11, 42-46
- Nakhaeizadeh, G., E. Steurer, and K. Bartmae, 2002, "Banking and Finance," In: Klösgen W., and J. Zytkow, Handbook of data mining and knowledge discovery, Oxford University Press, 771-780
- Occam's Razor!, 2003, http://www.commoditytradingadvisor.com/occams-razor.htm
- Quinlan J.R., 1993, "C4.5: programs for machine learning," Morgan Kaufmann Publishers Inc., San Francisco, CA
- Sullivan, R., A. Timmermann, and H. White, 1999, "Data-snooping, technical trading rule performance, and the bootstrap," Journal of Finance, 54, 1647-1691
- Tsang, E., P. Yung, and J. Li, 2004, "EDDIE-Automation, a decision support tool for financial forecasting," Journal of Decision Support Systems, Special Issue on Data mining for financial decision making, 37:4, 559-565
- Tsay, R, 2002, Analysis of financial time series, Wiley
- Vityaev E.E., L. Orlov Yu, O. V. Vishnevsky, B. Ya. Kovalerchuk, A. S. Belenok, N. L. Podkolodnii, and N. A. Kolchanov, 2002, "Knowledge discovery for gene regulatory regions analysis," In: Knowledge-based intelligent information engineering systems and allied technologies, KES, (Eds.) Damiani, E., R. Howlett, L. Jain, and N. Ichalkaranje, IOS Press, Amsterdam, Part 1, 487-491
- Voit, J., 2003, The statistical mechanics of financial markets, Vol. 2, Springer
- Von Altrock C., 1997, Fuzzy logic and NeuroFuzzy applications in business and finance, Prentice Hall
- Wang J., 2003, Data mining; opportunities and challenges, Idea Group, London