International Journal of Intelligence Science, 2012, 2, 176-180
http://dx.doi.org/10.4236/ijis.2012.224023 Published Online October 2012 (http://www.SciRP.org/journal/ijis)

Using Data Mining with Time Series Data in Short-Term Stocks Prediction: A Literature Review

José Manuel Azevedo[1], Rui Almeida[2], Pedro Almeida[3]
[1]Department of Mathematics, Instituto Politécnico do Porto, Porto, Portugal
[2]Department of Mathematics, Faculdade de Ciências, Universidade da Beira Interior, Covilhã, Portugal
[3]Department of Informatics, Faculdade de Engenharia, Universidade da Beira Interior, Covilhã, Portugal
Email: jazevedo@iscap.ipp.pt, ralmeida@ubi.pt, palmeida@ubi.pt

Received July 1, 2012; revised August 20, 2012; accepted September 1, 2012

ABSTRACT

Data Mining (DM) methods are being increasingly used in prediction with time series data, in addition to traditional statistical approaches. This paper presents a literature review of the use of DM with time series data, focusing on short-time stocks prediction. This is an area that has been attracting a great deal of attention from researchers in the field. The main contribution of this paper is to provide an outline of the use of DM with time series data, using mainly examples related with short-term stocks prediction. This is important to a better understanding of the field. Some of the main trends and open issues will also be introduced.

Keywords: Data Mining; Time Series; Fundamental Data; Data Frequency; Application Domain; Short-Term Stocks Prediction

1. Introduction

Data Mining (DM) is a challenging field for research and has some practical successful application in several different areas. DM methods are being increasingly used in prediction with time series data, in addition to traditional statistical approaches [1-3].

DM can be presented as one of the phases of the Knowledge Discovery in Databases (KDD) process [4-6], and is identified as "the means by which the patterns are extracted from data" [7]. Nowadays, it can be said that the two terms, DM and KDD, are indistinctly used.

The OECD Glossary of statistical terms [8] presents the following definition: "A time series is a set of regular time-ordered observations of a quantitative characteristic of an individual or collective phenomenon taken at successive, in most cases equidistant, periods/points of time". There are several application domains of DM with time series data, being that one important application domain is short-term stocks prediction. This will be the focus of this paper. Short-term stocks prediction is a difficult issue and can be considered as an open research issue [9,10]. Intelligent forecasting models have achieved better results than traditional methods, particularly in short-term forecasts [11]. Although intelligent forecasting methods are better, we can still improve the results in terms of accuracy in addition to other factors.

The main contribution of this paper is to provide an outline of the use of DM with time series data, using mainly examples related with short-term stocks or market indexes predictions. This is important to a better understanding of the field. Some of the main trends and open issues will also be introduced.

The paper is organized as follows: DM with time series data is presented in Section 2, the integration of fundamental data is explored in Section 3, data frequency issues are introduced in Section 4. The paper closes in Section 5, with conclusion and future research directions.

2. Data Mining with Time Series Data

Since the seminal paper of Fayyad in 1996 [4], the Data Mining (DM) area has attracted a great deal of interest and can nowadays be considered as an established field. DM applications can be found in a diversified range of application domains. One important application domain is that of time series data. "A time-series data set consists of sequences of numeric values obtained over repeated measurements of time. The values are typically measured at equal time intervals (e.g., every minute, hour, or day)". [5]. The referred measures can be taken over one variable or several variables—univariate or multivariate time series.

2.1. Data Mining with Time Series Data Applications

DM with time series data is popular and many applications can be found in the literature, for instance, for earthquake forecasting [12], characterization of ozone behavior [13], or flood prediction [14]. Other application example is that of financial decision making. A decision support tool for financial forecasting, named as EDDIE, is presented in [15]. In [16], a new architecture that implements a binary neural network, AURA, to produce discrete probability distribution as forecasts, using high frequency data sets, is presented. The use of support vector machines and back propagation neural networks to predict credit ratings is presented in [17].

One important application concerns short-term stocks prediction, which is the main focus of this paper. In [18], an approach to the paradox of obtaining better results with long-horizon forecasts than with short-horizon forecasts is presented, and it is claimed that the paradox is solved, since the proposed model obtains promising results. Nevertheless, there is a great deal of interest from investors in short-horizon forecasts, thus the authors consider that research focusing on this issue is important, namely in using data mining with time series for short-term stocks prediction.

2.2. Data Mining Techniques Used with Time Series Data for Short-Term Stocks Prediction

Several DM techniques are used with time series data in order to obtain short-term stocks prediction. An interesting approach to portfolio management, using the Gaussian temporal factor analysis technique, is introduced in [19]. Neural networks are one of the most popular techniques for stocks prediction. [20-25] are some examples. In [22] rough sets and classification trees are used, as well. Rough sets are also used in [26]. Support Vector Machines are used in [27].

There were not yet been given strong evidences of some technique being better than other, but nonlinear models are more popular.

2.3. Specific Challenges

Using DM with time series data presents several specific challenges. In [28,29] the authors focus on the issue of representing time series data in order to effectively and efficiently apply DM. In [28], three types of algorithms are presented and compared, namely, the sliding window algorithm, the top-down algorithm, and the bottom-up algorithm, and a new approach, that is claimed to overcome the inconveniences of these three algorithms, is introduced. In [29], a new concept, named as median strings, is presented as a simple and, at the same time, powerful representation for time series data.

Another interesting issue is to find out if different time series, or parts of a time series, have similar behavior. This issue can be approached through the use of similarity measures and indexing techniques. Interesting reviews can be found in [30,31].

Over fitting is a common problem across DM applications and DM with time series data is not an exception. In [32], an approach that intends to overcome this problem is presented.

Other important issue concerns the way to implement each one of the phases of the KDD process, taking into account the specificities of time series data. An application of DM with time series data for short-term stock prediction is presented in [1], analyzing all the phases of the KDD process. Promising results were achieved, but it is referred that the inclusion of fundamental data could help improving the obtained results.

Table 1 presents a resume of the main techniques and challenges.

3. Including Fundamental Data

Concerning short-term stocks prediction, a possible approach is to collect the historical financial data, such as open price, higher price, lower price, close price, and volume. These can be used in a daily basis frequency, or other frequencies considered as appropriate. Several indicators can be derived and used for more adequate analysis. This approach is named as technical analysis. Another possible approach is to use statistical data, such as, macroeconomics indexes, and basic financial indicators of the company. This approach is named as fundamental analysis. **Table 2** resumes some of the technical and fundamental features found in the literature. Other researches, for instance [37-39], present similar indicators.

From the literature review it is clear that one of the main issues in obtaining good predictions is related to the first phase of the KDD process, that is to say, the selec-

Table 1. Data mining with time series data: Main techniques and challenges.

Techniques	Neural networks [20-25]
	Vector machines support [27]
	Rough sets [22,26]
	Classification trees [22]
	Gaussian temporal factor analysis [19]
Challenges	Data representation [28,29]
	Similar behavior [30,31]
	Over fitting [32]
	Implementing all KDD phases [1]

Table 2. Features for technical and fundamental analysis.

Type	Features	References
Fundamental	ROA(A); EBI Gross margin; gross margin growth operating income; operation income growth; net income; net income growth; continued net income; cash flow ratio; sales growth ratio; current ratio; ordinary income growth; continued income growth; total asset growth; return on total asset; quick ratio; liabilities ratio; total asset turnover; account receivable turnover; inventory turnover; fixed asset turnover; days payables outstanding; And several of others: gross national product; real GDP; unemployment rate; real economic growth; monetary supply and amount; gross margin growth; CCI; personal income; industrial production; Taiwan export/import volume; operation income growth liabilities; total asset growth fixed asset turnover; monitoring indicator Export foreign exchange volume; WPI; merchandise trade volume export/import;	Tsai and Hsiao (2010) [33]
Fundamental	Demand index; moving average divergence convergence; relative strength index; positive directional movement index; negative directional movement index; moving average; r-squared; linear regression slope; average true range	Zarandi, Rezaee, Turksen and Neshat (2009) [34]
Technical	Price channel (top); price channel (bottom); price per earning per share; volume; open price; range; changes; close price	
Technical	Average position change; bollinger band %; cutler's relative Strength index; exponential moving average; stochastic oscillator; typical price; volume accumulator; volume weighted RSI-MFI; volume weighted RSI, williams %R; advance decline line; average true range; average position change; chaikin A/D oscillator; on balance volume; stoch. osc.; typical price	Ince and Trafalis (2007) [35]
Fundamental	Money supply (M1B); government consumption level, gross national products, gross domestic products; consumer price index; whole-sale products index; rate of exchange	Cheng, Chen and Lin (2010) [22]
Technical	Moving average convergence/divergence; price rate of change; stochastic %K; stochastic %D; relative strength index; stochastic oscillator and directional indicator	
Technical	On balance volume; moving average; average stock yield	Shen, Guo, Wu and Wu (2011) [36]

tion of the adequate feature combination, since the same methods can yield different results if different features are selected as inputs.

Another aspect that arises from the literature review is that most researchers use only one of the two types of analysis, technical or fundamental. Thus analyzing combinations of both types of indicators is yet under-explored.

In addition, most studies use macroeconomics variables, forgetting the important financial indicators of the companies. Considering the domain application, it is clear that the evolution of stock prices is influenced by both types of variables, so considering it could conduct to good results.

One of the main issues related to the combination of both types of features is that time series data have different frequencies (**Figure 1**). Usually technical features have daily frequencies and fundamental features have monthly, quarterly, and lower frequencies, presenting some integration issues. These integration issues are very important and have several implications.

4. Integrating Features with Different Frequencies

As stated above, interesting results could be obtained through the integration of time series data with different frequencies. With short-term stocks predictions, there is the need to use mainly time series with data collected daily, yielding high frequency time series, opposed to low frequency time series obtained from the collection of fundamental data. Forecasts should be done in a daily basis, thus there are some important issues for research.

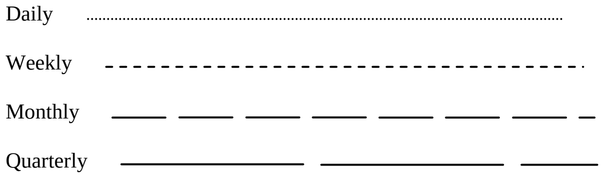

Figure 1. Time series with different frequencies.

Some research can be found in the literature approaching the issue of integrating time series features with different frequencies. Traditional approaches use regression algorithms such as MIDAS [37,38]. Nevertheless, this approach does not use DM.

In the literature review, only a few works, use DM with time series data with different frequencies. [22,34] are two examples. These studies present promising results, but the use of neural networks is somehow a limitation. Neural networks, despite usually yielding good results, functions as a "black box". This way it is difficult to understand the mechanism and the generated model.

From the literature review it can be concluded that these issues needs further research, and it can be useful to test other methods, and to explore the selection of some different features.

The application domain is an important issue to consider when applying DM, thus it should also be considered in this case. Taking into account the application domain will surely bring good insights and will surely yield good results.

5. Conclusions and Future Research Directions

This paper presents a literature review of the use of data mining with time series data. This literature review is very useful, since it brings a better understanding of the field of study, and this is an important contribution of this paper.

From the literature review it can be concluded that this subject attracts a great deal of interest by researchers. Nevertheless, several research issues remain unexplored. One of the ones that were identified during this research is related with the combined use of fundamental and technical indicators. The combined use of both types of indicators reveals also the issue of integrating time series with different frequencies.

Feature selection, corresponding to the first phase of the KDD process, is also an issue that requires more research to be done.

Future research directions include the study of ways to select the best features for DM with time series data. The existence of features with different frequencies is a concern, and methods that will help how to envisage this problem will be planned and implemented.

6. Acknowledgements

This work was partially supported by the research projects: PEst-OE/MAT/UI0212/2011, financed by FEDER through COMPETE—Programa Operacional Factores de Competitividade and by FCT—Fundação para a Ciência e a Tecnologia.

REFERENCES

[1] P. Almeida, “Previsão do Comportamento de Séries Temporais Financeiras com Apoio de Conhecimento Sobre o Domínio,” Ph.D. Thesis, Universidade da Beira Interior, Covilhã, 2003.

[2] L. Breiman, “Statistical Modeling: The Two Cultures,” *Statistical Science*, Vol. 18, No. 3, 2001, pp. 199-231. doi:10.1214/ss/1009213726

[3] M. A. Ruggiero, “Cibernetic Trading Strategies—Developing a Profitable Trading System State-of-the-Art Technologies,” John Wiley & Sons, New York, 1977.

[4] U. M. Fayyad, G. Piatetski-Shapiro and P. Smyth, “From Data Mining to Knowledge Discovery: An Overview,” In: U. M. Fayyad, G. Piatetsky-Shapiro, P. Smyth and R. Uthurusamy, Eds., *Advances in Knowledge Discovery and Data Mining*, The MIT Press, Cambridge, 1996, pp. 1-34.

[5] J. Han, M. Kamber and J. Pei, “Data Mining: Concepts and Techniques,” Morgan Kaufman Publishers, California, 2011.

[6] D. Hand, H. Mannila and P. Smyth, “Principles of Data Mining,” The MIT Press, Cambridge, 2011.

[7] A. Azevedo and M. F. Santos, “KDD, SEMMA, and CRISP-DM: A Parallel Overview,” *Proceedings of the IADIS European Conference on Data Mining*, Amsterdam, 24-26 July 2008, pp. 182-185.

[8] OECD, “Time Series,” 2006. http://stats.oecd.org/glossary/detail.asp?ID=2708

[9] M. A. Ferreira and P. Santa-Clara, “Forecasting Stock Market Returns: The Sum of the Parts Is More than the Whole,” *Journal of Financial Economics*, Vol. 100, No. 3, 2011, pp. 514-537. doi:10.1016/j.jfineco.2011.02.003

[10] T. Fu, “A Review on Time Series Data Mining,” *Engineering Applications of Artificial Intelligence*, Vol. 21, No. 1, 2011, pp. 164-181. doi:10.1016/j.engappai.2010.09.007

[11] T. O. Hill, M. Connor and W. Remus, “Neural Network Models for Time Series Forecasts,” *Management Science*, Vol. 42, No. 7, 1996, pp. 1082-1092. doi:10.1287/mnsc.42.7.1082

[12] S. Fong and Z. Nannan, “Towards an Adaptive Forecasting of Earthquake Time Series from Decomposable and Salient Characteritics,” *Proceedings of the 3rd International Conference on Pervasive Patterns and Applications*, Rome, 25 September 2011, pp. 53-60.

[13] K. J. Walsh, M. Milligan, M. Woodman and J. Sherwell, “Data Mining to Characterize Ozone Behavior in Baltimore and Washington DC,” *Journal of Atmospheric Environment*, Vol. 42, No. 18, 2008, pp. 4280-4292. doi:10.1016/j.atmosenv.2008.01.012

[14] C. Damle and A. Yalcin, “Flood Prediction Using Time Series Data Mining,” *Journal of Hidrology*, Vol. 333, No. 2-4, 2007, pp. 305-316. doi:10.1016/j.jhydrol.2006.09.001

[15] E. Tsang, P. Yung and J. Li, “EDDIE-Automation, a Decision Support Tool for Financial Forecasting,” *Decision Support Systems*, Vol. 37, No. 4, 2004, pp. 559-565. doi:10.1016/S0167-9236(03)00087-3

[16] A. Pasley and J. Austin, “Distribution Forecasting of High Frequency Time Series,” *Decision Support Systems*, Vol. 37, No. 4, 2004, pp. 501-513. doi:10.1016/S0167-9236(03)00083-6

[17] Z. Huang, H. Chen, C. J. Hsu, W. H. Chen and S. Wu, “Credit Ratings Analysis with Support Vector Machines and Neural Networks. A Market Comparative Study,” *Decision Support Systems*, Vol. 37, No. 4, 2004, pp. 542-558. doi:10.1016/S0167-9236(03)00086-1

[18] H. M. Krolzig and J. Toro, “Multiperiod Forecasting in Stock Market: A Paradox Solved,” *Decision Support Systems*, Vol. 37, No. 4, 2004, pp. 531-542. doi:10.1016/S0167-9236(03)00085-X

[19] K. C. Chiu and L. Xu, “Arbitrage Pricing Theory-Based Gaussian Temporal Factor Analysis for Adaptive Portfolio Management,” *Decision Support Systems*, Vol. 37, No. 4, 2004, pp. 485-500. doi:10.1016/S0167-9236(03)00082-4

[20] O. Coupelon, "Nneural Network Modeling for Stock Movement Prediction: A State of the Art," 2007. http://olivier.coupelon.free.fr/Neural_network_modeling_for_stock_movemen_prediction.pdf

[21] M. Kordos and A. Cwiok, "A New Approach to Neural Network Based Stock Trading Strategy," *Proceedings of the* 12*th International Conference on Intelligent Data Engineering and Automated Learning*, Norwich, 7-9 September 2011, pp. 429-436.

[22] J. H. Cheng, H. P. Chen and Y. M. Lin, "A Hybrid Forecast Marketing Timing Model Based on Probabilistic Neural Network, Rough Set and C 4.5," *Expert Systems with Applications*, Vol. 37, No. 4, 2010, pp. 1814-1820. doi:10.1016/j.eswa.2009.07.019

[23] Z. Yudong and W. Lenan, "Stock Market Prediction of S & P 500 via Combination of Improved BCO Approach and BP Neural Network," *Expert Systems with Applications*, Vol. 36, No. 5, 2009, pp. 8849-8854. doi:10.1016/j.eswa.2008.11.028

[24] X. Lin, Z. Yang and Y. Song, "Short-Term Stock Price Based on Echo State Networks," *Expert Systems with Applications*, Vol. 36, No. 3, 2009, pp. 7313-7317. doi:10.1016/j.eswa.2008.09.049

[25] T. Chang, "A Comparative Study of Artificial Neural Networks, and Decision Trees for Digital Game Content Stocks Price Prediction," *Expert Systems with Applications*, Vol. 38, No. 12, 2011, pp. 14846-14851. doi:10.1016/j.eswa.2011.05.063

[26] L. Shen and H. T. Loh, "Applying Rough Set to Market Timing Decisions," *Decision Support System*, Vol. 37, No. 4, 2004, pp. 583-597. doi:10.1016/S0167-9236(03)00089-7

[27] Q. Wen, Z. Yang, Y. Song and P. Jia, "Automatic Stock Decision Support System Based on Box Theory and SVM Algorithm," *Expert Systems with Applications*, Vol. 37, No. 2, 2010, pp. 1015-1022. doi:10.1016/j.eswa.2009.05.093

[28] E. Keogh, S. Chu, D. Hart and M. Pazzani, "Segmenting Time Series: A Survey and Novel Approach," In: M. Last, A. Kandel and H. Bunke, Eds., *Data Mining in Time Series Databases—Series in Machine Perception Artificial Intelligence*, World Scientific, Singapore, 2004, pp. 1-21.

[29] X. Jiang, H. Bunke and J. Csirik, "Median Strings: A Review," In: M. Last, A. Kandel and H. Bunke, Eds., *Data Mining in Time Series Databases—Series in Machine Perception Artificial Intelligence*, World Scientific, Singapore, 2004, pp. 173-192.

[30] G. Das and D. Gunopulos, "Time Series Similarity and Indexing," In: N. Ye, Ed., *The Handbook of Data Mining*, Lawrence Erlbaum Associates, London, 2003, pp. 279-304.

[31] M. L. Hetland, "A Survey of Recent Methods for Efficient Retrieval of Similar Time Sequences," In: M. Last, A. Kandel and H. Bunke, Eds., *Data Mining in Time Series Databases—Series in Machine Perception Artificial Intelligence*, World Scientific, Singapore, 2004, pp. 23-42.

[32] K. Mehta and S. Bhattacharya, "Adequacy of Training Data for Evolutionary Mining of Trading Rules," *Decision Support Systems*, Vol. 37, No. 4, 2004, pp. 461-474. doi:10.1016/S0167-9236(03)00091-5

[33] C. F. Tsai and Y. C. Hsiao, "Combining Multiple Feature Selection Methods for Stock Prediction: Union, Intersection, and Multi-Intersection Approaches," *Decision Support Systems*, Vol. 50, No. 1, 2010, pp. 258-269. doi:10.1016/j.dss.2010.08.028

[34] M. H. F. Zarandi, B. Rezaee, I. B. Turksen and E. Neshat, "A Type-2 Fuzzy Rule-Based Expert System Model for Stock Price Analysis," *Decision Support Systems*, Vol. 36, No. 1, 2009, pp. 139-154.

[35] H. Ince and T. Trafalis, "Kernel Principal Component Analysis and Support Vector Machines for Stock Price Prediction," *IIE Transactions*, Vol. 39, No. 6, 2007, pp. 629-637. doi:10.1080/07408170600897486

[36] W. Shen, X. Guo, C. Wu and D. Wu, "Forecasting Stock Indices Using Radial Basis Function Neural Networks Optimized by Artificial Fish Swarm Algorithm," *Knowledge-Based Systems*, Vol. 24, No. 3, 2011, pp. 378-385. doi:10.1016/j.knosys.2010.11.001

[37] B. C. O. Tas, "Private Information of the Fed and Predictability of Stock Returns," *Applied Economics*, Vol. 43, No. 19, 2011, pp. 2381-2398. doi:10.1080/00036840903194220

[38] P. M. Dechow, A. P. Hutton, L. Meulbroek and R. G. Sloan, "Short-Sellers, Fundamental Analysis, and Stock Returns," *Journal of Financial Economics*, Vol. 61, No. 1, 2001, pp. 77-106. doi:10.1016/S0304-405X(01)00056-3

[39] M. Lam, "Neural Network Techniques for Financial Performance Prediction: Integrating Fundamental and Technical Analysis," *Decision Support Systems*, Vol. 37, No. 4, 2004, pp. 567-581. doi:10.1016/S0167-9236(03)00088-5

[40] K. Wohlrabe, "Forecasting with Mixed-Frequency Time Series Models," Ph.D. Thesis, Ludwig Maximilians Universitat, Munchen, 2008.

[41] E. Andreou, E. Ghysels and A. Kourtellos, "Forecasting with Mixed-Frequency Data," *Journal of Econometrics*, Vol. 158, No. 2, 2010, pp. 246-261. doi:10.1016/j.jeconom.2010.01.004