Expert Systems With Applications 66 (2016) 219–233

Contents lists available at ScienceDirect

Expert Systems With Applications

journal homepage: www.elsevier.com/locate/eswa

# Financial time series forecasting using rough sets with time-weighted rule voting

Mariusz Podsiadlo, Henryk Rybinski*

*Institute of Computer Science, Warsaw University of Technology, Poland*

ARTICLE INFO

*Article history:*
Received 3 June 2016
Revised 27 August 2016
Accepted 29 August 2016
Available online 2 September 2016

*Keywords:*
Decision systems
Rough sets
Financial time series prediction

ABSTRACT

In the paper we investigate experimentally the feasibility of rough sets in building profitable trend prediction models for financial time series. In order to improve the decision process for long time series, a novel time-weighted rule voting method, which accounts for information aging, is proposed. The experiments have been performed using market data of multiple stock market indices. The classification efficiency and financial performance of the proposed rough sets models was verified and compared with that of support vector machines models and reference financial indices. The results showed that the rough sets approach with time weighted rule voting outperforms the classical rough sets and support vector machines decision systems and is profitable as compared to the *buy and hold* strategy. In addition, with the use of variable precision rough sets, the effectiveness of generated trading signals was further improved.



## 1. Introduction

Financial ecosystems generate large amounts of noisy and incomplete information and carry inherent uncertainty of any forward looking predictions. Yet, any wrong decision can have potentially catastrophic consequences for the economic well-being of individuals, institutions and nations, as shown by the recent financial crises. It is therefore not surprising that the academia and financial industry have ever since been looking for methods able to analyze the available data and infer reliable observations, applicable to a wide variety of problems, ranging from macroeconomic crisis prevention to stock market movement prediction and optimal composition of investment portfolios.

However, forecasting financial markets based on the available data is considered a very difficult task, as they generate large amounts of noisy and incomplete information, with the further challenge of non-stationarity caused by ongoing evolution of financial markets. Following (Abu-Mostafa & Atiya, 1996) it is worth noting that usually the precision of financial market models is slightly better than 50%. Furthermore, the common interpretation of the *efficient market* hypothesis (Fama, 1970) given by the canonical *random walk* model (Malkiel, 2007) implies that forecasting future stock returns is not possible on the basis of currently available information.

Fortunately, there is a mounting empirical evidence (e.g. Ang and Bekaert, 2007; 1988; Lo & MacKinlay, 2011) that at least a partial predictability based on economic variables linked to a business cycle (e.g. dividend yield, earnings-price ratio, treasury bill rate, inflation, etc.) is possible.[1] The criticism of the mostly in-sample evidence of such a forecastability (Bossaerts & Hillion, 1999; Goyal & Welch, 2003; Welch & Goyal, 2008) was met by an increasing number of studies accommodating model uncertainty and parameter instability, which reported significant positive out-of-sample results (Campbell & Thompson, 2008; Dangl & Halling, 2012; Kelly & Pruitt, 2013; Rapach, Strauss, & Zhou, 2009). On that point, *technical indicators*, being a popular prediction tool among practitioners, were shown as being valuable predictors (Brock, Lakonishok, & LeBaron, 1992; Han, Yang, & Zhou, 2013; Lo, Mamaysky, & Wang, 2000) surpassing the performance of, and complementing economic variables in delivering strong out-of-sample forecasts (Neely, Rapach, Tu, & Zhou, 2014).

Furthermore, the forecast of market direction was shown to be possible and sufficient to generate profitable strategies with remarkable success (Leung, Daouk, & Chen, 2000; Pesaran & Timmermann, 1995; Womack, 1996). Following (Christoffersen & Diebold, 2006), the sign dependency is of a highly nonlinear nature and linked to the well evidenced conditional volatility dependence (Andersen & Bollerslev, 2003; Franses & Van Dijk, 2000; Ghysels, Harvey, & Renault, 1996), which makes any successful use of simple linear sign forecasting models unlikely. In this context, the

* Corresponding author.
*E-mail addresses:* M.Podsiadlo@ii.pw.edu.pl (M. Podsiadlo), h.rybinski@ii.pw.edu.pl (H. Rybinski).

[1] See Rapach and Zhou, 2013 for a survey of related research.

http://dx.doi.org/10.1016/j.eswa.2016.08.066

advantage of capturing possible non-linear relations between the available information and return forecast was shown by Chen and Hong (2010).

Statistics and probability theory traditionally form the basis for formal methods used in economy and finance for time series analysis and prediction. The strong mathematical foundation allows to deliver predictive models being strong enough to capture many aspects of the modeled reality with statistically measurable precision of results. However, the multivariate models, typical for the financial environment driven by multiple causal factors, grow complex together with the number of independent variables, and their non-linear correlations, which makes the analysis difficult to follow by anyone but field experts. The generated knowledge has to be considered within the limits of the underlying model of reality and associated assumptions, be it the type of statistical distribution, a pricing model or stochastic process.

Consequently, attempts to employ data discovery and machine learning models in order to unearth complex data characteristics and relationships have a long track record in economy and finance. The use of neural networks and decision trees goes back to 90s of the last century (see e.g. Azoff, 1994). Also, in parallel with the advances of the Pawlak's *Rough Sets* (RS) theory (Pawlak, 1991), numerous papers addressing problems of using RS to the economic applications appear, with Pawlak and Slowinski (1994) being probably the first. The first thorough review of the papers devoted to RS applications for financial prediction is provided by Tay and Shen (2002).

With the extensions of the RS theory, provided by Pawlak and Skowron (2007), the use of RS approaches to the financial prediction has got even more attention. In particular, (Yao & Herbert, 2009) proposes using RS to analyzing financial time series, (Teoh, Cheng, Chu, & Chen, 2008) constructs a model combining fuzzy time series approach with the probabilistic and RS approach. In Lee, Ahn, Oh, and Kim (2010) RS were shown to be applicable to any time scale, down to intraday trading.

A newer analysis of RS based approaches to financial market data has been made recently by Podsiadło and Rybiński (2014). The employed RS models were usually tested by providing a mixture of (lagged) time series price and volume attributes of the predicted asset (most commonly a futures index), other time series objects and macroeconomic variables believed to be correlated, as well as derived technical indicators. The experimental results reported mostly excellent classification accuracy but often used relatively small data sets divided into outright training and testing samples (see for example Kim & Han, 2001; Yao & Herbert, 2009), thus being vulnerable to the sample bias. In order to avoid the bias, some authors (Kumar, Agrawal, & Joshi, 2004; Nair, Mohandas, & Sakthivel, 2010) used cross-validation.[2] However, the use of cross-validation without considering the data sample time order (i.e. random sampling) was common, in spite of the fact that it could cause a look-ahead bias. The application of the rolling window validation, which respects the time order as recommended by Rapach and Zhou (2013), was seldom (Kim & Enke, 2016; Teoh et al., 2008).

The main motivation for using RS was that they neatly combined possibilities to successfully mine from vague data with interpretability of results, which are just decision rules. Especially the ability to interpret results in terms of decision rules is very important in finance, where little is left to chance. RS properties of attribute reduction (so called *reducts*) and ability to produce strong decision rules from incomplete and noisy data (Grzymala-Busse, 1992; Skowron, 1993), deliver flexibility but also interpretability, lacked by black-box models based on neural networks and other machine learning methods.

Noteworthy, feature selection is an inherent value and quality of RS, which means that given data in form of a decision table we can obtain a reduced table and a set of decision rules, equivalent to the original table. There are numerous studies comparing the RS approach with other methods. Already Wong, Ziarko, and Ye (1986) have compared the RS approach with statistical methods in machine learning, showing that the RS-based concept of approximate classification is closely related to the statistical approach. Zhong, Dong, and Ohsuga (2001) discusses some disadvantages of two feature selection methods for selecting relevant attributes, namely the filter approach and the wrapper approach, comparing them to the RS approach. Dimitras, Slowinski, Susmaga, and Zopounidis (1999) show superiority of the RS approach versus discriminant analysis for the business failure prediction problem, stating that the RS approach is especially useful for the cases with large number of attributes. Therefore, RS are also frequently used as the feature selection component of hybrid classifiers (Kruczyk et al., 2013; Salamó & López-Sánchez, 2011), where another model, e.g. neural networks, genetic algorithms or *support vector machines* (SVM), is used to perform the actual classification. Zhang, Sai, and Yuan (2008) analyses the combination of RS + SVM for stock index prediction, using RS approach to reduce the number of attributes, and then for the reduced model applying the SVM model, showing increase of the precision by some 3%. In this case the interpretability of results is limited only to the feature selection step, as the ability of RS to generate human-readable decision rules and perform classification is not used. Therefore, in our paper the full capacity of RS, i.e. feature selection and interpretable classification, is tested.

As shown by Leitch and Tanner (1991), forecast profitability is a more relevant metric for assessing forecasts, while the relationship between MSFE and forecast profitability is weak, with direction-of-change being the only conventional forecast error measure showing significant correlation with the forecast profitability. However, RS-based experiments in this area, displaying both, the classification accuracy and financial profitability of the generated models, were relatively few (Ang & Quek, 2005; Huang, 2009; Ruggiero, 1994), as was the market direction forecast (Kim & Han, 2001; Kim & Enke, 2016; Shen & Loh, 2004).

Consequently, our work attempts to verify the feasibility of RS and their *Variable Precision Rough Sets* (VPRS) extension (Ziarko, 1993) to financial time series forecasting by applying the generated models to a large real life data set of multiple stock market indices (U.S. S&P500, German DAX, and Hong Kong HSI), using the rollover time window validation of results, which considered the time order of data samples, thus alleviating the likelihood of the sample or look-ahead bias. This enabled generation of realistic classification and financial performance benchmarks for the RS models.

In order to verify the quality of the proposed RS models we have decided to compare them with other models. SVM is increasingly popular in financial time series prediction (Cao & Tay, 2003; Huang, Nakamori, & Wang, 2005; Kim, 2003), and compares favorably to other neural network models (Tay & Cao, 2001; Zhu & Yi, 2012). To this end we use SVM as a reference soft computing model for performance comparison with the investigated RS models. The financial performance of the generated models was also tested in terms of the profit generated using a simple long only market timing strategy vs. an index *buy and hold*.

Inherent data inconsistency may cause RS models to create *non-deterministic* rules. It results from the fact that financial market data are subject of evolving in time, so rules from various periods may be inconsistent, implying multiple possible classifica-

[2] Cross-validation is a classifier validation technique, where the available data sample is randomly divided into $k$ mutually exclusive subsets (folds). The classifier is tested $k$-times, whereas each fold acts in turn as the validation sample against the remaining folds being the training sample (Kohavi, 1995).