Expert Systems With Applications 57 (2016) 37–48

Contents lists available at ScienceDirect

Expert Systems With Applications

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

Forecasting of stock return prices with sparse representation of financial time series over redundant dictionaries

Roberto Rosas-Romero*, Alejandro Díaz-Torres, Gibran Etcheverry

Department of Electrical & Computer Engineering, Universidad de las Américas-Puebla, Mexico

ARTICLE INFO

Article history:
Received 4 October 2015
Revised 4 March 2016
Accepted 5 March 2016
Available online 19 March 2016

Keywords:
Financial time series
Artificial financial predictors
Sparse representation
Learned over-redundant dictionaries
Temporal feature extraction
Time-domain pattern recognition

ABSTRACT

This paper presents the theory, methodology and application of a new *predictive model* for time series within the financial sector, specifically data from 20 companies listed on the U.S. stock exchange market. The main impact of this article is (1) the proposal of a recommender system for financial investment to increase the cumulative gain; (2) an artificial predictor that beats the market in most cases; and (3) the fact that, to the best of our knowledge, this is the first effort to predict time series by *learning redundant dictionaries* to *sparsely reconstruct* these signals. The methodology is conducted by finding the optimal set of predicting model *atoms* through two directions for *dictionaries* generation: the first one by extracting *atoms* from past *daily return price* values in order to build *untrained dictionaries;* and the second one, by atom extraction followed by *training of dictionaries* though K-SVD. Prediction of financial time series is a periodic process where each cycle consists of two stages: (1) *training of the model* to learn the dictionary that maximizes the probability of occurrence of an *observation sequence* of return values, (2) *prediction of the return value* for the next coming trading day. The motivation for such research is the fact that a tool, which might generate confidence of the potential benefits obtained from using formal financial services, would encourage more participation in a formal system such as the stock market. Theory, issues, challenges and results related to the application of sparse representation to the prediction of financial time series, as well as the performance of the method, are presented.

1. Introduction

This paper presents a method, based on sparse representation of time series over redundant dictionaries, which is implemented and applied as a predictive model for time series within financial contexts, specifically the return of the price of an individual stock from some companies that are listed on the U.S. stock exchange market. The problem is approached using the framework of dictionary learning, early addressed by Olshausen and Field (1997), with the motivation that there are no reports about forecasting of stock return prices using the framework of sparse representation of time series over learned dictionaries. This research field focuses on developing algorithms to learn dictionaries with elements, called atoms, so that a signal of interest can be reconstructed as a linear combination of a very few atoms.

There are three categories of dictionary leaning algorithms. In this application, the learning algorithm follows the direction of the clustering methods (Tosic & Frossard, 2011). We use dictionaries which are constructed in two ways by (1) extraction of an observed sequence of stock price return values from historical information about a certain company, and (2) by extraction of an observation sequence followed by dictionary adaptation which consists in minimizing a cost function through optimization methods such as K-SVD (Aharon, Elad, & Bruckstein, 2006). The application of untrained dictionaries has been successful in some applications such as face recognition (Wright, Yang, Ganesh, Sastry, & Ma, 2009).

The data to evaluate are the shares corresponding to 20 companies such as Bank of America, Intel, Nanosphere, Haliburton, Nokia, Office Depot, Twitter, Netflix, General Electric Company, Oracle Corporation, Rite Aid Corporation and the Coca-Cola Company. The price of an individual stock is analyzed to predict changes that will happen in the immediate future so that a decision is taken with such information. To perform an adequate forecasting analysis, historical information, about companies listed in the stock exchange market, is required. There are tools that allow us to obtain this information such as *Yahoo! Finance*, which provides historical prices of companies that trade on various exchange markets around the world.

The stock market is an institution that people attend to protect and enhance their financial savings and the resources obtained, allow companies and government to finance productive projects that

* Corresponding author. Tel.: +52 2222292677.
E-mail addresses: roberto.rosas@udlap.mx (R. Rosas-Romero), alejandro.diazts@udlap.mx (A. Díaz-Torres), gibran.etcheverry@udlap.mx (G. Etcheverry).

http://dx.doi.org/10.1016/j.eswa.2016.03.021
0957-4174/

generate development. It is important to bring formal saving systems to the population since this is what drives an active economy. This is possible by working on the development of systems that provide attractive and safe yields and, above all, accurate performance forecasting.

The rest of the paper is organized as follows. Section 2 reviews relevant and recent related work that combines machine learning with financial engineering. Section 3 presents a description of basic concepts of financial time series and how a predictive system is used to analyze and predict these signals. Section 4 gives a detailed overview of the framework of sparse representation and dictionary learning. Description of the proposed method for prediction of financial time series based on sparse representation over dictionaries is described in Section 5. Section 6 provides the experimental results obtained by using the aforementioned method. Conclusions are presented in Section 7.

2. Review of related work

Stock market prediction is a challenging research area where different methods have been developed with the aim of predicting the future of return gain values (Guresen, Kayakutlu, & Daim, 2011). The aim is to design models with the potential to predict stock price behavioral movements in the stock exchange market; however, predicting such trends is very challenging due to the fact that stock market data are noisy and time varying in nature (Atsalakis & Valavanis, 2009). To address the topic of future stock price predictions, different techniques from artificial intelligence, fuzzy systems, and machine learning areas have been applied to predict the future stock prices movements and trends.

An overview of the most common neural networks applied to forecast spatio-temporal patterns from feed forward to recurrent neural networks to model non-linear dependencies in spatio-temporal patterns, is presented by Dorffner (1996). Prediction models based on Artificial Neural Networks (ANN), Support Vector Machines (SVM), Random Forest and naïve-Bayes have been applied to predict the stock movement direction and the stock price index for Indian stock markets according to Patel, Shah, Thakkar, and Kotecha (2015). Three classical learning based forecasting methods: Neural Networks (NN), Adaptive Neuro-Fuzzy Inference Systems (ANFIS), and Least-Squares Support Vector Machines (LS-SVM), have been used (Wu & Lee, 2015) for time series prediction through a local modeling strategy. Artificial Neural Networks (ANNs) were applied by Vanstone and Finnie (2010) to financial time series by outlining an empirical methodology for creating and testing ANNs for stock market trading systems. Chua, Suardy, and Tsiaplias (2013) examine the forecasting performance of Bayesian Model Averaging (BMA) for a set of single factor models of short-term rates. Kara, Boyacioglu, and Baykan (2011) provided a comparable pattern whereby Neural Networks and plain SVM were compared for stock price movement prediction. Zbikowski (2015) applied a modified Support Vector Machine (SVM) and the Fisher method for feature selection to create a stock trading strategy based on stock market short-term trends forecasting.

Sparse representation of signals has received considerable attention as a tool to solve different problems. An extensive survey of the challenges, motivation, approaches and applications of the main algorithms in the field of dictionary learning for sparse representation is presented by Tosic and Frossard (2011) and Elad (2010). Aharon et al. (2006) proposed the groundbreaking *K-SVD* algorithm for dictionaries adaptation in order to achieve the best sparse reconstruction for each member from a set of signals under sparsity conditions, which is an iterative method that alternates between sparse coding of signals over the current dictionary and a dictionary updating process to better fit the signals set. Mairal, Bach, Ponce, Sapiro, and Zisserman (2008) introduced two different algorithms to perform an efficient optimization of an energy function to learn dictionaries, which are explicitly optimized to be both *reconstructive* and *discriminative*. Modeling of signals through sparse representation has been a very useful for different applications. The solution to the problem of dictionary learning for sparse representation has proven to be successful in different applications such as classification for EEG based computer-brain interfaces (Shin et al., 2015), short text classification (Gao, Zhou, & Guan, 2015), face recognition with occlusion (Zhao and Hu, 2015), Alzheimer disease classification (Xu, Wu, Chen, & Yao, 2016), image super-resolution (Zhao, Chen, Sui, & Gu, 2015), image denoising (Aharon et al., 2006; Elad & Aharon, 2006), compression (Marcellin, Gormish, Bilgin, & Boliek, 2000), and color restoration (Mairal et al., 2008).

3. Financial time series and forecasting of future returns

A *financial time series* is a sequence of *observations* (*feature vectors*), $\boldsymbol{O} = \boldsymbol{o}_1, \ \boldsymbol{o}_2, \ldots, \boldsymbol{o}_N$, with features of some financial asset (such as price P_t, return R_t, gain G_t, etc.) defined over discrete time, where the time index ($t = 1, \ \ldots, \ N$) corresponds to an hour, or a trading day, or a week, or a month, etc. The *price of an asset, P_t*, is an essential feature and given two consecutive daily prices, P_{t-1} and P_t, another financial feature, called *return*, is defined

$$R_t = \frac{P_t - P_{t-1}}{P_{t-1}}. \tag{1}$$

where $-1 \ \leq \ R_t$. In our experiments, the *daily stock price return* R_t, rather than *price*, is used as a temporal feature because of the fact that it requires a smaller dynamic range which makes it easier to quantize. A *positive return, $R_t > 0$*, represents an *increase* in the asset price while a *negative return, $R_t < 0$*, means a *price drop*. Another asset feature, related to the daily return, is the *one-period gain G_t* which is given by $G_t = \frac{P_t}{P_{t-1}} = 1 + \ R_t$. A *positive return* or price *increase* implies a *gain* greater than one while a *negative return* or *price drop* introduces a *gain* less than one. *Prediction* of financial time series consists in estimating a future *price return*, R_{t+n}, with n trading days in advance by processing a set of past return observations $R_t, \ R_{t-1}, \ldots, R_{t-\ell+1}$. The cumulative return price P_c, after a period of n trading days, is given by the product of the *initial price* of the asset P and the *cumulative gain* according to,

$$P_c = \ P(1+R_1)(1+R_2)\ldots(1+R_n). \tag{2}$$

The *return* is not only used as a feature, but it is also used to measure the *performance* of an *artificial financial predictor*. A financial predictor is efficient if it provides an investor with a higher *cumulative gain* than that obtained by not using it. For instance, let us assume that a financial predictor has estimated four consecutive price return values according to the four-day sequence {*price increase, price drop, price increase, price drop*}. The consequence of this prediction is that the investor decides to sell his/her shares at those days when a *price drop* is expected, so that the obtained return at those days is zero $R_t = 0$ instead of negative, with a gain $1 + R_t = 1$. Therefore, the *cumulative gain* over the four-day period is $(1+R_2)(1+R_4)$ which is expected to be an accurate prediction and higher than the original index $(1+R_1)(1+R_2)(1+R_3)(1+R_4)$. A *return gain curve* shows the evolution of the return gain as time increases and the final goal of an artificial financial predictor is to generate a return gain curve higher than the index curve. The higher the prediction accuracy, the higher the obtained return. Fig. 1 shows a set of financial time series, which includes curves for stock price, return value, gain at each training day, and the cumulative gain. All curves are shown over an interval of 1765 trading days.

An artificial financial *predictor* is a dynamic system that evolves over time to learn the changes of financial time series, which are