Referential kNN Regression for Financial Time Series Forecasting

Conference Paper · November 2013
DOI: 10.1007/978-3-642-42054-2_75

5 authors, including:

Tao Ban
National Institute of Information and Communications Technology
73 PUBLICATIONS **207** CITATIONS
SEE PROFILE

Ruibin Zhang
UNITEC Institute of Technology
5 PUBLICATIONS **12** CITATIONS
SEE PROFILE

S. Pang
UNITEC Institute of Technology
65 PUBLICATIONS **1,330** CITATIONS
SEE PROFILE

Abdolhossein Sarrafzadeh
UNITEC Institute of Technology
120 PUBLICATIONS **589** CITATIONS
SEE PROFILE

Referential kNN Regression for Financial Time Series Forecasting

Tao Ban, Ruibin Zhang, Shaoning Pang, Abdolhossein Sarrafzadeh, and Daisuke Inoue

National Institute of Information and Communications Technology
4-2-1 Nukui-Kitamachi, Tokyo, 184-8795, Japan
bantao@nict.go.jp
http://www.nict.go.jp

Abstract. In this paper we propose a new multivariate regression approach for financial time series forecasting based on knowledge shared from referential nearest neighbors. Our approach defines a two-tier architecture. In the top tier, the nearest neighbors that bear referential information for a target time series are identified by exploiting the financial correlation from the historical data. Next, the future status of the target financial time series is inferred from heritage of the time series by using a multivariate k-Nearest-Neighbour (kNN) regression model exploiting the aggregated knowledge from all relevant referential nearest neighbors. The performance of the proposed multivariate kNN approach is assessed by empirical evaluation on the 9-year S&P 500 stock data. The experimental results show that the proposed approach provides enhanced forecasting accuracy than the referred univariate kNN regression.

Keywords: Time series forecasting, correlation analysis, kNN regression, referential kNN regression, S&P 500 Indices.

1 Introduction

Time series forecasting now become a common problem in numerous areas of research (e.g., hydrology, finance, climatology and etc.), In particular financial time series prediction has been drawn substantial attention in both the computer science and financial communities recently and remains a very specialized task. Various studies have shown that financial time series is predictable by using both linear and non-linear models [1]. Actually, recent scholars have revealed that the nonparametric non-linear models tend to outperform linear models in financial time series forecasting [2]. In practice, k-Nearest-Neighbour(kNN) is one of the most commonly employed algorithm in time series forecasting due to its simplicity and intuitiveness in alike instances recovering from large dimensional feature spaces [3], and also the tolerance in high-dimensional and incomplete data [4]. The kNN algorithm assumes that sequences of time series have emerged in the past are likely to have a resemblance to the future sequences and for generating kNN based forecasts, similar patterns of behaviour are masked in

M. Lee et al. (Eds.): ICONIP 2013, Part I, LNCS 8226, pp. 601–608, 2013.

terms of nearest-neighbours, and their development over time is exploited to yield the forecast [5].

Despite of the good general performance of kNN regression in financial time series forecasting, most of the existing researches in kNN regression derived financial time series forecasting are mainly focusing on individual historical data, without referencing any extra associated knowledge. The motivation of this research is, since the fluctuation of financial time series is changing dynamically due to the evolutions of many relevant economic and financial activities [6], the accuracy of the univariate forecasting model will not be optimal when insufficient knowledge has been consulted. The historical knowledges were extracted from stock own are obviously inadequate compare with the knowledges were collected from group of stocks that share similar behaviours. In this study, we attempt to develop a model to forecast the financial time series product, by conducting referencing involved multivariate kNN regression over large scaled historical time series data. The application of referential knowledges facilitate the enhancement of forecasting accuracy since the information used for forecasting will be less biased due to sparse referential consultation.

The beginning of this research is to use numerical distance measurement to extract the interrelationship between all the financial time series based on historical data, and provides a clear picture of the network which allows user to easily identify the neighbors of each time series. Once the cognates of references have been recognized, we will apply the kNN regression algorithm over such findings for each financial time series.

The outline of the paper is organized as follows: In Section 2, we briefly introduces the background of financial correlation analysis and kNN regression algorithm, plus the review of related works in the financial time series forecasting. Then, we present the proposed referential financial time series forecasting method in Section 3. Next, the experimental design regarding to proposed method, and discussion of comparison in results have been presented in Section 4. Finally, we have drawn the conclusion in Section 5.

2 Related Work

2.1 Correlation Analysis in Financial Time Series

In fundamental stock correlation analysis, the focus is on the investigation of stocks' fundamental attributes [7], regardless of any numerical financial calculation. In [8], Clarke stated that economic intuition supports the idea that firms in the same industry share high return correlations compared to firms in different industries.

Also, as suggested in [7], stocks can be categorised into homogeneous groups using criteria other than industry affiliation and in the literature, academic researchers and investment practitioners have figured a variety of approaches to construct homogeneous stock groups. Notable works include the heuristics approaches proposed by Farrell[9], Elton and Gruber[10], the clustering method proposed by Brown and Goetzmann [11].

In addition, stocks are commonly grouped on the basis of primary economic attributes such as market capitalization or operating performance.

2.2 k-Nearest-Neighbor Regression

Because of its simplicity and intuitiveness, k-nearest-neighbour (kNN) algorithm is widely adopted for classification and regression [12,13].

The application of kNN to time series forecasting under nonparametric locally weighted regression condition was presented independently by Yakowitz [14] and Cleveland [15] within the community of statistics. In financial time series forecasting field, kNN has also drawn pretty much of attention because of the work by Meade [16], Fernando et al. [3], and its performance have been validated through numerical studies.

The underlying intuition to apply kNN to univariate time series is that consistent data-generating processes often produces observations of repeated patterns of behavior. Therefore, if a previous pattern can be identified as similar to the current behavior of the time series, the subsequent behavior of previous pattern can provide valuable information to predict the behavior in the immediate future. In the kNN regression algorithm introduced by Meade [16], the target variable of a time series forecasting problem is presented as a sequence of interval scaled values. Given a pattern whose future value is to be predicted, the algorithm identifies the k most similar past patterns and combines their future values to make the prediction on future value.

3 The Proposed Model

As aforementioned, stock prices in the market fluctuates with the evolution of related economical and financial factors. Intuitively, stocks share the common characteristics more or less, as proved by related economical studies. Our correlation analysis among related stocks indicated that, stocks from the same industry are more close to each other in terms of correlation coefficients due to their consensus to the same external factors. Therefore, we propose to improve the accuracy of previous studies on kNN regression by incorporating associated knowledge from closely related stock indices.

The proposed method is implemented in two steps. In the first step, we try to identify a set of referential nearest neighbors (RNN) which are likely to provide relevant information for the prediction. In the second step, pattern searching and prediction is done by respecting all the historical information in RNN. The details of the two steps are addressed in the following two subsections.

3.1 Search for Referential Nearest Neighbors

Let $\mathbb{X} = \{\boldsymbol{x}_i | i = 1, ..., N\}$ be a collection of stock indices, where $\boldsymbol{x}_i$ is a set of observations x_{it}, each being recorded at the closing of n consecutive working

days. To retrieve the referential nearest neighbors (RNN) of a target time series $\boldsymbol{x}_i$, we make use of the Pearson's correlation coefficient,

$$\rho(\boldsymbol{x}_i, \boldsymbol{x}_j) = \frac{\sum_{t=1}^{n}(x_{it} - \bar{\boldsymbol{x}}_i)(x_{jt} - \bar{\boldsymbol{x}}_j)}{\sqrt{\sum_{t=1}^{n}(x_{it} - \bar{\boldsymbol{x}}_i)^2}\sqrt{\sum_{t=1}^{n}(x_{jt} - \bar{\boldsymbol{x}}_j)^2}}, \tag{1}$$

where $\bar{\boldsymbol{x}}_{\cdot}$ is the mean of the vector. Then the following dissimilarity metric [17] is adopted to compute the proximity between $\boldsymbol{x}_i$ and $\boldsymbol{x}_j$:

$$D(\boldsymbol{x}_i, \boldsymbol{x}_j) = 1 - \rho^2(\boldsymbol{x}_i, \boldsymbol{x}_j). \tag{2}$$

Based on this distance metric, we can identify a set $\mathbb{K}_i$ of k nearest neighbors, namely, the RNNs, which is comprised of stock indices with strong resemblance to characteristics of $\boldsymbol{x}_i$ and therefore can provide reference for predicting future values of $\boldsymbol{x}_i$. Note that $\boldsymbol{x}_i$ is included in $\mathbb{K}_i$ by default because the distance between $\boldsymbol{x}_i$ and itself is always 0.

3.2 Referential Nearest Neighbor Regression

Now we move on to forecast future values of $\boldsymbol{x}_i$ based on the historical knowledge stored in $\mathbb{K}_i$. The current state of $\boldsymbol{x}_i$ is represented by the latest z consecutive observation of $\boldsymbol{x}_i$ at time n, i.e., $\boldsymbol{p}_i = [x_{n-z+1}, x_{n-z+2}, \cdots, x_n]$, where z is the *pattern length* parameter determined later. Let $\mathbb{T}_i$ be the set of all consecutive patterns of length z that could be extracted from time series in $\mathbb{K}_i$. Based on the distance metric in (2), we select M most resembling patterns to form a referential pattern set $\mathbb{R}_i$. Then prediction of $x_{i(n+h)}$, where h denotes the *forecasting horizon*, is obtained by

$$\hat{x}_{i(n+h)} = \frac{1}{M}\sum_{j=1}^{M} p_{j(n'_j+h)}, \tag{3}$$

where n'_j is the last time index of referential pattern $\boldsymbol{p}_j \in \mathbb{R}_i$.

3.3 Evaluation Metrics

There are many criteria that can be used to evaluate the performances of forecasting models in the numerical study [4]. In this paper, the prediction performance of the proposed forecasting model is evaluated through two typical statistical metrics: Mean Absolute Percentage Error (MAPE) and Normalised Mean Squared Error (NMSE). The usage of MAPE and NMSE is to measure the derivation between the predicted and actual values, the smaller values of MAPE and NMSE means the closer between predicted value and actual value.

These metrics are determined by $x_{i(n+1)}$ and $\hat{x}_{i(n+1)}$: the current observation of time series $\boldsymbol{x}_i$ at time $n+1$, and the prediction for that observation. Then the two evaluation metrics are defined as,

$$MAPE = \frac{1}{n}\sum_{j=n+1}^{n+N}\left|\frac{\hat{x}_{ij} - x_{ij}}{x_{ij}}\right|, \tag{4}$$

$$NMSE = \frac{N-1}{N} \sum_{j=n+1}^{n+N} \frac{(\hat{x}_{ij} - x_{ij})^2}{(x_{ij} - \bar{\boldsymbol{x}}_i)^2}, \tag{5}$$

where forecast errors are evaluated over a period of N consecutive days.

4 Experiment and Discussions

The performance of proposed algorithm in forecasting stocks daily closing price has been assessed in this experiment, with the comparison between conventional k-NN regression, Empirical Mode Decomposition, and benchmark algorithms. In order to avoid biases from the training samples, we use the 'unprocessed' time series data, daily closing price, and 'processed' time series data, daily return on both traditional and modified kNN regressions.

Like most typical time series forecasting studies represented, using real financial data will draw our attention into actual problem analysis, and would be useful to an economist studying the effect of various indicators on the market. In this case, we form the data pool by entire S&P 500 stocks. The S&P 500 has been widely regarded as the best single gauge of the large cap U.S. equities market since the index was first published in 1957.

4.1 Data Selection

We examine our model over 2500 observations of 121 stocks from S&P 500, by using daily closing price between January 3, 2001 and February 29, 2011. The entire stock data was downloaded from Yahoo Finance. Those 121 stocks have been selected from the following 4 sectors: Energy, Consumer Discretionary, Health Care, and Information Technology.

4.2 Parameters Modulation

On the purpose of demonstrating robust empirical results, the determination of the most crucial parameters: length of the reference pattern z and number of referential neighbors k has been recovered prior to the evaluation stage since parameter settings are dominating the performance of proposed algorithm. During our research, one of the statistical evaluating methods, the hold-out cross-validation has been implemented for parameters' tuning, and the benefits of this cross-validation technique has been sourced by Refaeilzadeh [18].

During the parameter modulation, we record score for all parameter combinations, and at the end, we adopt the parameter configuration in accordance with the best cross-validation scores. Since it impossible to go through all the possible

Table 1. Notation of parameters and grid values

Meta-Parameters	Definitions	Grid Values
k	Number of referential neighbors for each TS	$\{1, 3, \ldots, 27, 31\}$
M	Number of neighbors for reference pattern among each TS	$\{1, 3, 5, 7, 9, 11\}$
z	Length of the referential pattern	$\{5, 10, \ldots, 25, 30\}$

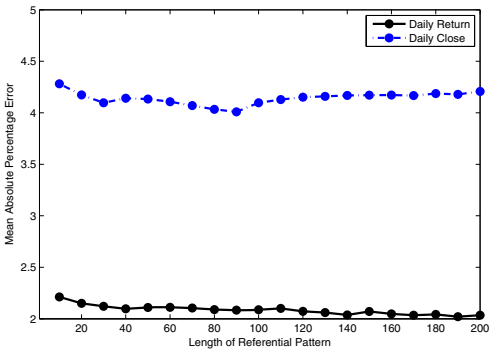

Fig. 1. Tuning parameter z using hold-out cross-validation

parameter combinations, we carefully selected sparse grid values for parameter settings, in order to provides reasonable coverage of parameter settings with optimal performance.

We have listed all the meta-parameters have been involved during the experiments, and range of grid values have been examined in Table 1, and as an instance of the parameter tuning process, Figure 1 shows how the length of referential pattern z is determined for the latter forecasting stage.

4.3 Experiment Results and Discussions

Due to the limitation of the paragraph, we only present the results of 10 randomly selected stocks across all the test data in Table 2, and the lowest error has been highlighted in bold. Despite the occasional better performance from the conventional kNN regression method on the table, the average result at the bottom of the table is concluding that, the proposed referential kNN regression provides the highest forecast accuracy that outperforms conventional kNN regression along with the benchmark algorithms by 34.69% (1.62% percentage points) and 33.7% (1.55% percentage points) respectively in MAPE, 51.90% (1.5% percentage points) and 55.31%(1.72% percentage points) respectively in NMSE.

Table 2. Random selected evaluation results among various forecasting techniques

Stock ID		NN		kNN		EMD	AR
		Price(%)	Return(%)	Price(%)	Return(%)	Price(%)	Return(%)
Stock 1	MAPE:	**3.76**	5.56	**3.76**	3.89	6.62	5.56
	NMSE:	**1.12**	2.09	**1.12**	1.16	7.22	2.41
Stock 2	MAPE:	**3.01**	4.60	**3.01**	3.07	4.72	4.33
	NMSE:	**0.69**	1.54	**0.69**	0.71	1.43	1.38
Stock 3	MAPE:	**2.84**	4.30	**2.84**	2.91	3.54	4.05
	NMSE:	**1.98**	4.23	**1.98**	2.05	3.06	4.06
Stock 4	MAPE:	2.71	4.14	2.71	**2.70**	3.67	3.91
	NMSE:	2.11	4.39	2.11	**2.01**	3.65	4.62
Stock 5	MAPE:	**4.23**	5.91	**4.23**	4.27	6.04	5.88
	NMSE:	**1.30**	2.34	**1.30**	1.31	2.29	2.52
Stock 6	MAPE:	**3.86**	5.45	**3.86**	3.93	5.43	5.40
	NMSE:	**1.22**	2.36	**1.22**	1.28	2.31	2.44
Stock 7	MAPE:	2.71	3.94	2.71	**2.70**	3.66	3.97
	NMSE:	1.60	2.75	1.60	**1.51**	2.75	3.45
Stock 8	MAPE:	2.82	4.07	2.82	**2.78**	3.74	4.08
	NMSE:	1.96	3.84	1.96	**1.93**	3.33	4.38
Stock 9	MAPE:	2.71	3.86	2.71	**2.67**	3.58	3.92
	NMSE:	1.68	3.20	1.68	**1.66**	2.58	3.63
Stock 10	MAPE	3.37	4.83	3.37	**1.60**	4.31	4.89
	NMSE:	1.09	2.20	1.09	**0.29**	1.78	2.22
Average	MAPE:	3.20	4.67	3.20	**3.05**	4.53	4.60
	NMSE:	1.48	2.89	1.48	**1.39**	3.04	3.11

5 Conclusions

In this paper, we develop a new multivariate kNN regression approach for financial time series forecasting in regards of referential nearest neighbors. We determine the referential knowledge of the target time series by conducting the financial correlation analysis among historical time series data. Then, we apply the aggregated knowledge that extracted from related rNNs with k-Nearest Neighbour(kNN) regression model to forecast the future status of the time series. The effectiveness of the proposed hybrid approach is assessed by a robust empirical evaluation over 9 years S&P 500 stock data. The experiment results demonstrate that the proposed multivariate kNN approach provides enhanced forecasting accuracy beyond classical univariate kNN regression.

References

1. Campbell, J.Y., Shiller, R.J.: The dividend-price ratio and expectations of future dividends and discount factors. Review of Financial Studies 1(3), 195–228 (1988)
2. Kanas, A., Yannopoulos, A.: Comparing linear and nonlinear forecasts for stock returns. International Review of Economics & Finance 10(4), 383–398 (2001)
3. Andrada-Félix, J., Fernadez-Rodriguez, F., Garcia-Artiles, M.-D., Sosvilla-Rivero, S.: An empirical evaluation of non-linear trading rules. Studies in Nonlinear Dynamics & Econometrics 7(3) (2003)
4. Lin, A., Shang, P., Feng, G., Zhong, B.: Application of empirical mode decomposition combined with k-nearest neighbors approach in financial time series forecasting. Fluctuation and Noise Letters 11(2) (2012)
5. Kanas, A.: Non-linear forecasts of stock returns. Journal of Forecasting 22(4), 299–315 (2003)
6. Ciora, C., Munteanu, S.M., Hrinca, I.G., Ciobanu, R.: Uncertainty and fluctuations on the stock markets. In: International Conference on Financial Management and Economics. (Singapore), International Proceedings of Economics Development and Research, vol. 11. IACSIT Press (2011)
7. Chan, L.K., Lakonishok, J., Swaminathan, B.: Industry classifications and return comovement. Financial Analysts Journal, 56–70 (2007)
8. Clarke, R.N.: Sics as delineators of economic markets. Journal of Business, 17–31 (1989)
9. Farrell, J.L.: Analyzing covariation of returns to determine homogeneous stock groupings. The Journal of Business 47(2), 186–207 (1974)
10. Elton, E.J., Gruber, M.J.: Homogeneous groups and the testing of economic hypotheses. Journal of Financial and Quantitative Analysis, 581–602 (1970)
11. Brown, S.J., Goetzmann, W.N.: Mutual fund styles. Journal of financial Economics 43(3), 373–399 (1997)
12. Wu, X., Kumar, V., Quinlan, J.R., Ghosh, J., Yang, Q., Motoda, H., McLachlan, G.J., Ng, A., Liu, B., Philip, S.Y., et al.: Top 10 algorithms in data mining. Knowledge and Information Systems 14(1), 1–37 (2008)
13. Brandsma, T., Buishand, T.A., et al.: Simulation of extreme precipitation in the rhine basin by nearest-neighbour resampling. Hydrology and Earth System Sciences Discussions 2(2/3), 195–209 (1998)
14. Yakowitz, S.: Nearest-neighbour methods for time series analysis. Journal of Time Series Analysis 8(2), 235–247 (1987)
15. Cleveland, W.S.: Robust locally weighted regression and smoothing scatterplots. Journal of the American Statistical Association 74(368), 829–836 (1979)
16. Meade, N.: A comparison of the accuracy of short term foreign exchange forecasting methods. International Journal of Forecasting 18(1), 67–83 (2002)
17. Mantegna, R.N.: Hierarchical structure in financial markets. The European Physical Journal B-Condensed Matter and Complex Systems 11(1), 193–197 (1999)
18. Refaeilzadeh, P., Tang, L., Liu, H.: Cross-validation. Encyclopedia of Database Systems 5 (2009)