DOI: 10.1002/ecj.12140

TRANSLATED PAPER

WILEY

Stock price prediction using *k*-medoids clustering with indexing dynamic time warping

Kei Nakagawa[1] | Mitsuyoshi Imamura[1] | Kenichi Yoshida[2]

[1]Nomura Asset Management Co., Ltd. 1-12-1, Nihonbashi, Chuo-ku, Tokyo 103-8260, Japan

[2]Graduate School of Business Sciences, University of Tsukuba, 3-29-1, Otsuka, Bunkyo-ku, Tokyo 112-0012, Japan

Correspondence
Kenichi Yoshida, Graduate School of Business Sciences, University of Tsukuba, 3-29-1, Otsuka, Bunkyo-ku, Tokyo 112-0012, Japan.
Email: yoshida@gssm.otsuka.tsukuba.ac.jp

Translated from Volume 138 Number 8, pages 986–991, DOI: 10.1541/ieejeiss.138.986 of *IEEJ Transactions on Electronics, Information and Systems* (Denki Gakkai Ronbunshi C)

Abstract

Various methods to predict stock prices have been studied. In the field of empirical finance, feature values for prediction include "value" and "momentum". In this research, we use the pattern of stock price fluctuations which has not been fully utilized in the financial market as the input feature of prediction. We extract the representative price fluctuation patterns with *k*-Medoids Clustering with Indexing Dynamic Time Warping method. This method is *k*-medoids clustering on dissimilarity matrix using IDTW which measures DTW distance between indexed time-series. We can visualize and grasp a price fluctuation pattern effective for prediction with the proposed method. To demonstrate the advantages of the proposed method, we analyze its performance using TOPIX. Experimental results show that the proposed method is effective for predicting monthly stock price changes.

KEYWORDS

indexing dynamic time warping, k-medoids clustering, momentum, price fluctuation pattern, reversal, stock price prediction

1 | INTRODUCTION

A number of methods for stock price prediction were proposed so far. Representative methods are those based on time series analysis. Conditional mean models and conditional variance models are used in time series analysis. Typical examples of conditional mean models are AR, MA, and ARMA models. These are used for modeling and prediction of stock price level or return rate. With AR model, future stock prices are predicted based on a linear combination of past prices. With MA model, future stock prices are predicted based on a linear combination of disturbance terms in past prices. ARMA model combines AR and MA models. On the other hand, ARCH model and its generalized version GARCH were proposed as conditional variance models. In previous studies, direct prediction of stock prices by these methods using only past data was considered as difficult; instead, extensive research was done on volatility as a practicable indicator.[1]

On the other hand, there were proposed many methods of price prediction based on so called machine learning with data other than stock price history used for feature values. As distinct from time series analysis, feature values must be properly selected in prediction based on machine learning. In the field of empirical finance, too, many factors (feature values) were found to explain stock price fluctuations.[2,3] Recently, attempts are made to use analyst reports, news, and other text data as feature values. For example, there are studies on text mining of analyst reports suggesting that such reports supply stock markets with non-numerical information,[4] and studies on long-term forecast for Japan's stock market using CPR method composed of three steps: co-occurrence analysis, principal component analysis, and regression analysis.[5] In addition, there is a report on possibility of stock price prediction through deep learning analysis of relationship between information contained in Reuter's news articles and stock prices.[6]

In this study, we extract representative patterns of stock price fluctuations, and use them as feature values for stock price prediction. In order to extract representative patterns of stock price fluctuations, we apply *k*-medoids clustering[9] to dissimilarity matrix using Indexing DTW (IDTW)[8] to measure DTW (Dynamic Time Warping) distance[7] in indexed stock price fluctuations (below referred to as '*k*-medoids clustering with IDTW'). Such time series clustering makes it possible to visualize and grasp price fluctuation patterns useful

for prediction. We conduct empirical analysis of TOPIX to verify effectiveness of the proposed method for prediction of stock price fluctuation patterns that were not yet sufficiently utilized in the finance market. In addition, we demonstrate that clusters representing price fluctuation patterns extracted by the proposed method can be interpreted in terms of *momentum* and *reversal** assumed as useful factors in the field of empirical finance.

The paper is organized in the following way. First, we give an overview of IDTW and *k*-medoids clustering in Section 2. Then in Section 3, we demonstrate effectiveness of the proposed method by example of TOPIX, and discuss relationships between clustered patterns of price fluctuations. In closing, we give a summary of the study.

2 | PROPOSED METHOD: *K*-MEDOIDS CLUSTERING WITH INDEXING DYNAMIC TIME WARPING

As reported in a previous study,[8] daily patterns of stock price fluctuations appropriately screened during a month using IDTW are useful feature values for stock rice prediction. In the present study, we extract underlying fluctuation patterns by time series clustering, and determine relevance of corresponding phenomena. Number of data entries (business days) in daily stock price fluctuations vary from month to month so that clustering methods (such as *k*-means) using Euclidean distance in a simple vector space cannot be applied. Therefore, there is a need to measure similarity among data in a more natural way, without inserting or deleting any values even in case of different data volumes, and to combine such measurement with an appropriate clustering algorithm. In this study, we propose applying *k*-medoids clustering[9] to dissimilarity matrix created by IDTW. The proposed method is advantageous in that natural clustering is possible for financial time series data of different length, as is the case with daily stock prices during a month.

2.1 | Dynamic time warping (DTW)

In order to create dissimilarity matrix for clustering, one must calculate similarity (distance) between time series data. A number of measures were proposed for similarity between time series data. Widely used simple measures include correlation coefficient and Euclidean distance. However, correlation coefficient can reflect only linear relationship between data, while Euclidean distance may produce counter-intuitive results.[10] This is because humans can flexibly recognize

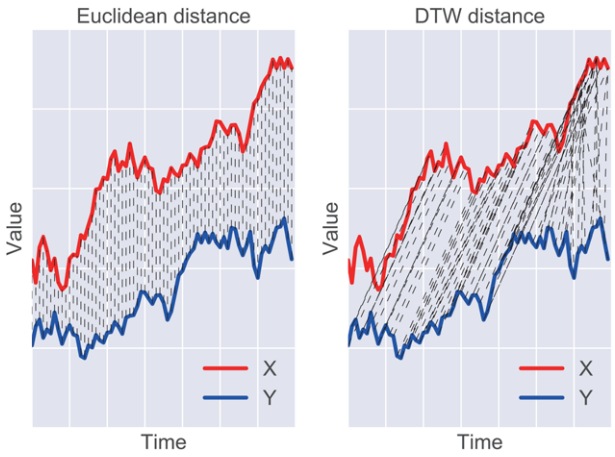


FIGURE 1 Correspondence of time-series data [Color figure can be viewed at wileyonlinelibrary.com]

shapes of time series data, while time direction is fixed with Euclidean distance. Yet another problem with correlation coefficient, Euclidean distance, and the likes is that calculation becomes impossible for two time series of different length.

DTW distance[7] is a measure intended to avoid the problems of Euclidean distance. Particularly, this measure makes possible comparison between time series data of different length; in so doing, optimal match between two time series is performed while warping the time axis (Figure 1). The algorithm to calculate DTW distance is described below (Algorithm 1). Here x and y are time series data with lengths of N and M, respectively; w is a window to constrain warping. As regards the distance function d, we adopted Manhattan distance $d(x, y) = |x - y|$.

ALGORITHM 1 DTW distance

Procedure DTW($x, y, w = 5$)
  ➢ Initialize Matrix D
  Var $D[N, M]$
  $D[1, 1] = 0$
  for $i = 2$ to N **do**
    for $j = 2$ to M **do**
      $D[i, j] = \infty$
    end for
  end for
  ➢ Calculate DTW distance
  for $i = 2$ to N **do**
    for $j = max(1, i - w)$ to $min(M, i + w)$ **do**
      $D[i, j] = d(x[i - 1], y[j - 1])$
        $+ min(D[i, j - 1], D[i - 1, j], D[i - 1, j - 1])$
    end for
  end for
  return $D[N, M]$
end procedure

* Momentum pertains to the phenomenon that rising (falling) prices rise (fall) further; reversal pertains to the phenomenon that rising (falling) prices reverse their trend towards fall (rise).

2.2 | Indexing dynamic time warping (IDTW)

Fluctuation range of stock prices varies with the observation period (day, week, month), while the price level greatly varies over time. Therefore, to apply DTW to stock prices, one needs (1) an appropriate observation period and (2) a normalization method. Seasonality of stock prices is widely recognized among investors; particularly, 'Sell in May' is a well-known strategy. The effect of 'Sell in May' has been confirmed in world's stock markets.[11] Besides, monthly return is a fundamental evaluation unit for funds and factor returns. From the above, one can say that investors recognize stock price fluctuations per month. Therefore, as regards (1), daily closing stock prices during a month are defined as observation period. When practically comparing stock prices between different periods, indexing with respect to the beginning period taken as 1 is more natural than comparison via standardization or return rate. This indicates that investors recognize fluctuation patterns of stock prices over a certain period rather than change of the prices. Therefore, as regards (2), we take end-of-previous-month value as 1, and apply indexing to express daily stock prices in a month by comparison to previous day.

In the above context, IDTW was proposed as a measure of similarity between stock prices in a previous report.[8] That report is distinct in that future stock prices are predicted only from historic data. However, accurate extraction of similar stock price fluctuation patterns and utilization of the extracted patterns for price prediction constitute a different issue. Thus we employed k-NN algorithm and its modification k^*-NN[12] to verify that price fluctuation patterns extracted by IDTW are useful as feature values for prediction. On the other hand, both k-NN and k^*-NN are lazy learning algorithms, and relationship between feature values and learning labels are hard to comprehend. Thus in this study, we perform time series clustering with k-medoids algorithm to visualize and identify price fluctuation patterns useful for prediction.

The algorithm for calculation of IDTW distance is described below (Algorithm 2).

ALGORITHM 2 IDTW distance

procedure IDTW(x, y)
  ➢ Scaling Data
  Var Ix, Iy
  $Ix[1] = 1, Iy[1] = 1$
  for $i = 2$ to N **do**
    $Ix[i] = Ix[i-1]\frac{x[i]}{x[i-1]}$
  end for
  for $j = 2$ to M **do**
    $Iy[j] = Iy[j-1]\frac{y[j]}{y[j-1]}$
  end for
  ➢ Apply DTW
  return $DTW(Ix, Iy)$
end procedure

2.3 | k-medoids clustering with Indexing dynamic time warping

k-medoids algorithm is a method of partitional optimization clustering related to k-means. The difference from k-means is that clusters are formed around medoids rather than centroids. Medoid is a data point in a cluster with the minimal total dissimilarity to all other points in the cluster. Intuitively, it is the point closest to the center of a cluster. Therefore, medoids always exist within the data subject to clustering. For this reason, k-medoids algorithm can be applied as long as dissimilarity matrix is specified for classified data, and clustering can be realized for arbitrary dissimilarity measures (distances). That is, this algorithm can be applied if dissimilarity is defined even for data not expressed by vectors, such as time series of different length. Besides, errors are evaluated by squared distance in k-means but by distance in k-medoids, which provides robustness against noise and outliers. With financial time series, one can expect for thick-tailed distributions so that outliers have a strong effect; therefore, k-medoids algorithm is a favorable method to analyze financial time series.

In this paper, we propose applying k-medoids algorithm with IDTW used as a dissimilarity measure to financial time series of different length such as monthly stock prices. The use of IDTW as a measure of similarity makes possible natural clustering of financial time series data. The algorithm is described below (Algorithm 3). Here $x_1, \ldots, x_N$ are time series data of different length, and k is the number pf clusters. The stopping criterion is applied when cluster allocation remained unchanged, or when the iteration count reached its limit (100 times).

ALGORITHM 3 IDTW based k-medoids clustering

procedure IDTW BASED K-CLUSTERING ($\{x_1, \cdots, x_N\}, k$)
  Randomize $m_1, \cdots, m_k$
  while stopping criterion has not been met **do** ➢ Cluster Assignment
    for $i = 1$ to k **do**
      $c_i := \{x_i | IDTW(x_i, m_k) \leq IDTW(x_i, m_l)\}$
    end for ➢ Update Medoids
    for $j = 1$ to k **do**
      $m_i := min_{x_i \in c_j} \sum_{i=1}^{N} IDTW(x_l, x_j)$
    end for
  end while
  ➢ Return the medoids
  return $m_1, \cdots, m_k$
end procedure

3 | EMPIRICAL ANALYSIS

3.1 | Analysis procedure

In this section, we confirm clustering of similar price fluctuation patterns and effectivity of extracted clusters as

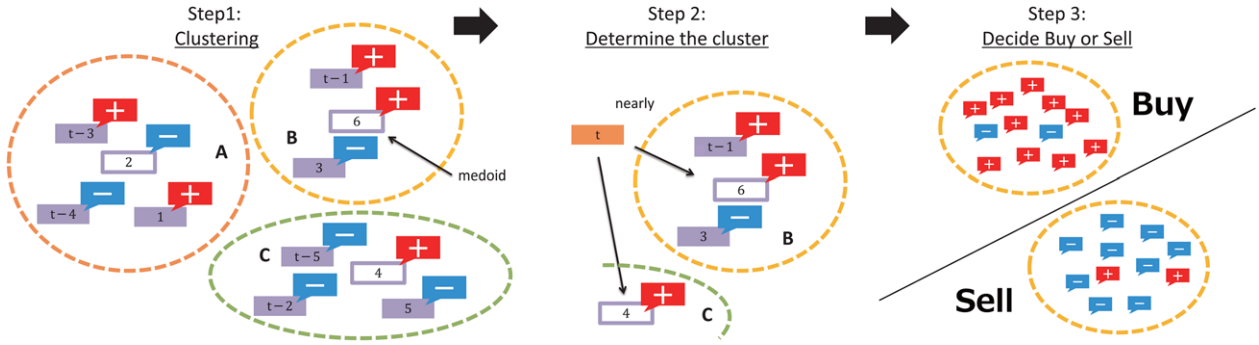


FIGURE 2 Stock price prediction framework [Color figure can be viewed at wileyonlinelibrary.com]

feature values through empirical analysis using TOPIX Dividend Index. The index data were acquired from Bloomberg Terminal. As regards clustering, in order to verify effectivity of the proposed method, we compared (1) prediction with AR (below referred to as AR) model as the simplest time series model, (2) k-medoids clustering with DTW (below referred to as DTW) as an appropriate benchmark according to previous research,[13] (3) IDTW + k^*-NN (below referred to as k^*-NN)[8] that combines IDTW and k^*-NN as improvement of k-NN,[12] and (4) k-medoids clustering with IDTW (below referred to as IDTW) proposed in this study. As regards the number k of clusters, we vary it at $2 \leq k \leq 12$, and extract as many clusters as required to achieve the best prediction accuracy, that is, to explain price fluctuations in the best way. Prediction accuracy is evaluated via both return rate and correctness. Specific analysis procedures are described below (with reference to Figure 2). As regards AR, parameters were re-estimated using monthly return for all usable monthly data. Besides, lag order was chosen every month as an optimal value up to 10 using AIC criterion. With k^*-NN, prediction was performed under the same conditions as in previous studies.[8,12]

The data period was set from January 1989 through March 2017, and the verification period was set from January 200 through March 2017.

Step 1

Based on past month-long daily price fluctuations up to month t – 1, the number k of clusters is fixed, and clustering is performed. For every month, next-month return rise or fall is kept as a label.

Step 2

Clusters containing price fluctuations during month t are identified.

Step 3

Cluster to which month t belongs is found; if there are more rise (fall) labels, then profit is calculated for sell (buy) at end-of-month price. In case of same number of both labels, average return is calculated for rise and fall labels, and the greater is adopted.

Step 4

Analysis returns to Step 1 at t = t + 1.

TABLE 1 Average accuracy of all years and total return for DTW and IDTW with k-medoids clustering (out-of-sample period is from January 2007 to March 2017; bold values are best measurements of each column)

	Total Return[%]		Accuracy[%]	
	IDTW	DTW	IDTW	DTW
$k = 2$	113	23	58.87	53.23
$k = 3$	98	−17	57.26	51.61
$k = 4$	148	66	61.29	53.23
$k = 5$	**162**	64	63.71	54.03
$k = 6$	151	50	**64.52**	54.03
$k = 7$	131	5	61.29	51.61
$k = 8$	132	6	62.1	53.23
$k = 9$	147	45	62.9	54.84
$k = 10$	114	41	58.06	55.65
$k = 11$	104	**79**	58.87	**59.68**
$k = 12$	90	56	59.68	58.06

3.2 | Analysis results

A summary of results obtained using DTW and IDTW at varied number of clusters is given in Table 1.

IDTW tends to outperform DTW in both return rate and correctness. As regards the number of clusters in IDTW, the peak is reached around $k = 4, 5, 6$; with more clusters, both return rate and correctness decline. Thus one can say that the optimal number of clusters of TOPIX price fluctuation patterns is about 5 from the viewpoint of prediction. In this case, cumulative return rate was as high as 162% at correctness of 64%. This prediction performs is comparable to a study that predicted monthly TOPIC index for 10 years, same as in this study (though a different period was considered), based on text information.[5] On the other hand, in case of clustering using DTW, the peak was reached with more clusters as compared to IDTW, namely, around $k = 11, 12, 13$.

Table 2 compares return rate and correctness among AR (benchmark), k^*-NN, and best cases of DTW ($k = 11$) and IDTW ($k = 5$). As can be seen from the table, clustering with IDTW outperforms AR and k^*-NN in both return rate and correctness.

TABLE 2 Average accuracy of all years and total return for each method (out-of-sample period is from January 2007 to March 2017; bold values are the measurements of each row)

	AR	k^*-NN	DTW ($k = 11$)	IDTW ($k = 5$)
Accuracy[%]	56.45	58.87	59.68	**64.52**
Total Return[%]	65	101	79	**162**

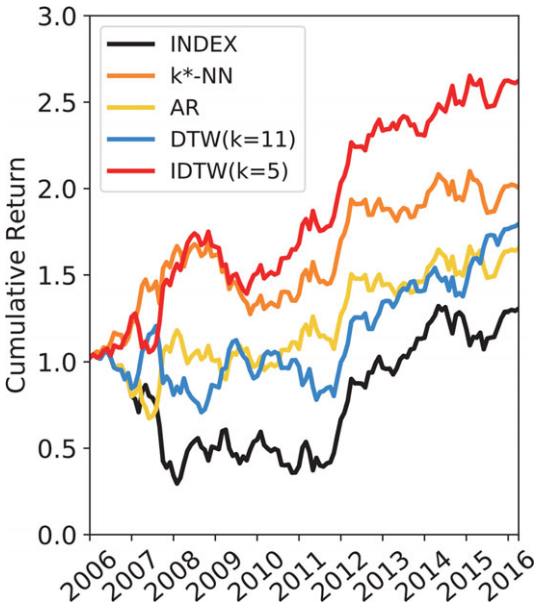


FIGURE 3 Cumulative return for each prediction method [Color figure can be viewed at wileyonlinelibrary.com]

In addition, Figure 3 shows trends of return rate for TOPIX as compared to the best results of DTW and IDTW. IDTW shows increasing return regardless of the time period, thus exceeding TOPIX and DTW. Therefore, we confirmed that representative price fluctuation patterns extracted by the proposed method work for prediction.

Next, we explain about extracted clusters. Clustering results obtained using IDTW with $k = 5$ at Step 1 for March 2017 are presented in Figure 4. Here black lines show price fluctuations selected as central (medoids); red lines show price fluctuations belonging to respective clusters. The diagram title format is *Sample: number of samples in cluster (probability of rise in next-month's return)*. We comment on clusters from left to right. In the top left diagram, the central black line remains quite flat regardless of fluctuations, and the parenthesized number suggests high probability of fall in next month. In the upper right diagram, probability of rise is high despite of the downward price trend, which is indicative of a reversal from fall to rise. In the middle left diagram, a strong rise in prices is observed, while probability of rise in next month is high; that is prices continue to rise, which is indicative of a rising momentum. In the middle right diagram, fluctuations pattern is almost flat, and this would last into next month. In the bottom left diagram, a strong fall in prices is observed, while probability of fall in next month is

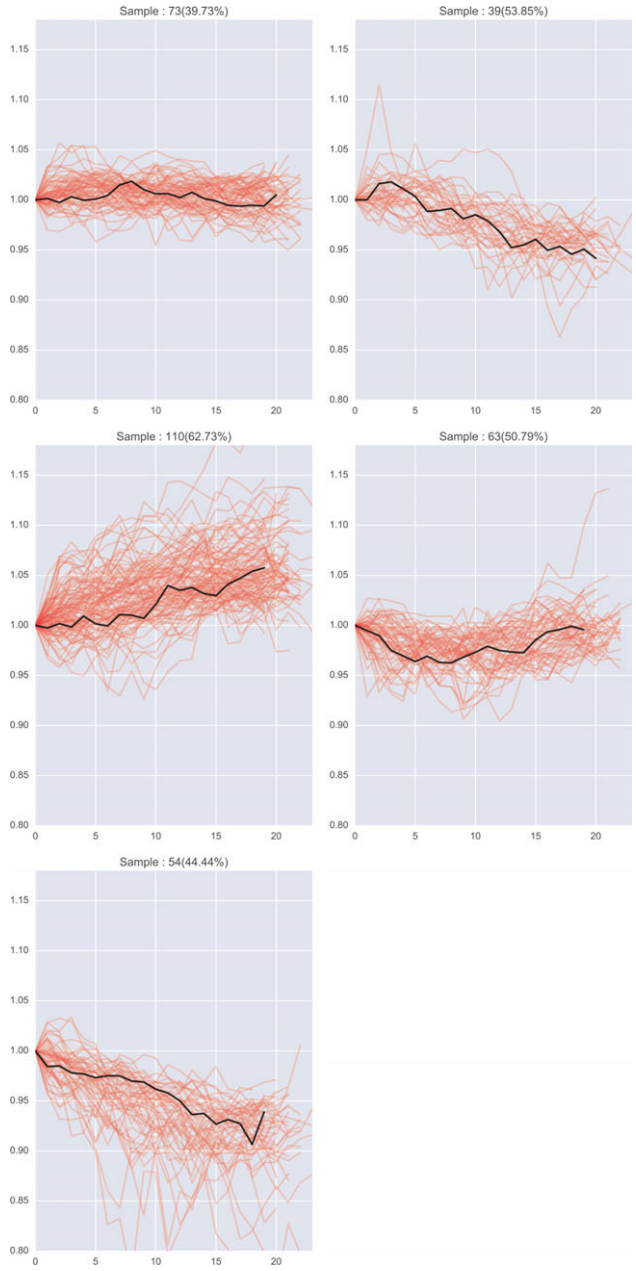


FIGURE 4 IDTW based k-medoids clustering as of 2017/3 [Color figure can be viewed at wileyonlinelibrary.com]

high; that is, prices continue to fall, which is indicative of a falling momentum.

In Japan's stock market, effectivity of momentum cannot be confirmed in many cases.[14] However, in this section we demonstrated momentum has a strong effect on both rise and fall when prices heavy fluctuate in a month of interest provided that clusters are properly defined based on similarity between price fluctuations. On the other hand, effect of reversal was only pronounced in case of the trend from fall to rise.

Such interpretation of every cluster by price fluctuation patterns is advantageous in that the concepts of momentum and reversal in empirical analysis can be extended from simple rise or fall with respect to reference month to price fluctuation patterns.

4 | CONCLUSION

In this study, we conducted clustering of stock price fluctuation patterns, and extracted representative fluctuation patterns as feature values for prediction. Particularly, we proposed a method called '*k*-medoids clustering with IDTW' in which *k*-medoids clustering is applied to dissimilarity matrix using IDTW to measure DTW distance in indexed price fluctuations. Time series clustering by this method makes it possible to visualize and grasp price fluctuation patterns useful for prediction.

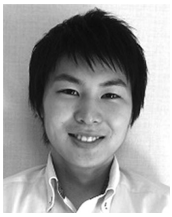
As a result of empirical analysis using TOPIX index and cluster visualization, we confirmed the following.

- The proposed *k*-medoids clustering with IDTW shoed high prediction accuracy, and outperformed TOPIX index and both benchmark methods in both return rate and correctness.
- Due to using *k*-medoids clustering with IDTW, the number of price fluctuation clusters required from the viewpoint of prediction accuracy is about five.
- The effect of momentum not recognized in Japan's market was confirmed for both rising and falling prices in case of heavy price fluctuations in a month of interest.

REFERENCES

1. Hamilton JD. Time series analysis. Princeton: Princeton University Press; 1994.
2. Fama EF, French KR. The cross-section of expected stock returns. *J Finance*. 1992;47:427–465.
3. Harvey CR, Liu Y, Zhu H. … and the cross-section of expected returns. *Rev Financ Stud*. 2016;29:5–68.
4. Takahashi S, Takahashi H, Tsuda K, Terano T. "Analyzing asset management knowledge from analyst's reports through text mining", International IPSI-2004, 2004;11.
5. Kuramoto T, Izumi K, Toshimura S, et al. Analysis of Long-term market trend by Text-mining of news articles. *Trans JSAI*. 2013;28:291–296. (in Japanese)
6. Goshima K, Takahashi H, Terano T. Estimating news articles' negative-positive by deep learning", 29th Annual Conf. *JSAI*. 2015. (in Japanese)
7. Itakura F. Minimum prediction residual principle applied to speech recognition. *IEEE Trans Acoust Speech Signal Process*. 1975;23:67–72.
8. Nakagawa K, Imamura M, Yoshida K, "Stock price prediction using k*-nearest neighbors and indexing dynamic time warping", in Artificial Intelligence of and for Business (AI-Biz 2017) (2017)
9. Kaufman L, Rousseeuw P. *Clustering by means of medoids*. North-Holland. 1987.
10. Keogh EJ, Pazzani MJ, "Scaling up dynamic time warping for datamining applications", in Proceedings of the sixth ACM SIGKDD International conference on Knowledge discovery and data mining, pp. 285–289, ACM 2000.
11. Bouman S, Jacobsen B. The halloween indicator, "Sell in May and go away": another puzzle. *Am Econ Rev*. 2002;92:1618–1635.
12. Anava O, Levy K, "k*-nearest neighbors: From global to local", in Advances in Neural Information Processing Systems, pp. 4916–4924, 2016.
13. Niennattrakul V, Ratanamahatana CA, "On clustering multimedia time series data using k-means and dynamic time warping", in Multimedia and Ubiquitous Engineering, 2007 MUE'07. International Conference, pp. 733–738, IEEE 2007.
14. Asness CS, Moskowitz TJ, Pedersen LH. Value and momentum everywhere. *J Finance*. 2013;68:929–985.

AUTHORS' BIOGRAPHIES



Kei Nakagawa, non-member, In 2012 graduated from Kyoto University (Fac. of Econ.), 2015 completed master's course at University of Tsukuba (Grad. School of Business Sci.), now 2nd term of doctorate at University of Tsukuba (Grad. School of Business Sci.). Was employed by Nissei Asset Management Co., Ltd, 2014 quants fund manager at Mitsui-Sumitomo Asset Management Co, Ltd., since 2018 Nomura Asset Management Co, Ltd. (Innovative Lab). Research in finance market analysis using methods of financial engineering and machine learning.

Mitsuyoshi Imamura, non-member, In 2014 completed master's course at Japan Advanced Institute of Science and Technology (Information Sci.), and was employed by Hewlett-Packard Japan (now Hewlett-Packard Enterprise), as network engineer, then Microsoft Japan, now Nomura Asset Management Co, Ltd. (Innovative Lab), 2nd term of doctorate at University of Tsukuba (Grad. School of Systems and Information Eng.).

Kenichi Yoshida, non-member, In 1980 graduated from Tokyo Institute of Technology (Fac. of Sci., Information Sci.), and was employed by Hitachi, Ltd. 1992 earned doctor's degree (Doctor of Eng., Osaka University). Since 2002 professor at University of Tsukuba (Grad. School of Business Sci.). Research in machine learning for analysis of various data on the Internet. Membership: IPSJ, JSAI.

How to cite this article: Nakagawa K, Imamura M, Yoshida K. Stock price prediction using *k*-medoids clustering with indexing dynamic time warping. *Electron Comm Jpn*. 2019;102:3–8. https://doi.org/10.1002/ecj.12140