Dynamic Time Warping Based Clustering for Time Series Analysis

Kun Zhang[1], Shuai Lin[1], Haoxuan Sun[2], Liyao Ma[2(✉)], and Junpeng Xu[3]

[1] Shandong Non-metallic Materials Institute, Jinan 250031, China
[2] School of Electrical Engineering, University of Jinan, Jinan 250022, China
cse_maly@ujn.edu.cn
[3] Shandong Huasheng Pesticide Machinery Co. Ltd., Linyi 276017, China

Abstract. This paper proposes a prediction method based on time series similarity. Based on clustering, Dynamic Time Warping (DTW) algorithm is used to find the influence of similarity and weight on the prediction results. Time series is a structure that records data in time sequence. The characteristics of multiple data at each time point are the same and comparable. According to people's purpose to find the rule of time series, and to the future time forecast. The first chapter introduces the background of the topic. The second chapter mainly introduces the time series, clustering algorithm, similarity, DTW distance and other basic theories involved in this paper. In the third chapter, we study the method of the total forecast data of time series. DTW distance is used for clustering to obtain the similarity with each class and then predict the data.

Keywords: K-MEANS · DTW · Similarity · Dynamic weighting

1 Introduction

With the development of artificial intelligence (AI) and big data, various fields have developed to varying degrees, and more and more artificial intelligence products appear in more industries [1–3]. In the field of transportation, there is also rapid development. Mass time series obtained through bus cards, detectors, cameras, communication equipment, the Internet of things, etc. However, due to periodicity and high noise, how to make use of time series is still a developing problem [4,5].

Although the development and use of AI in China is very strong, the use in the field of intelligent transportation still needs to be improved [6,7]. The penetration rate of intelligent transportation systems in the United States has reached more than 85%, and even cities like New York and Los Angeles have reached 90% [8].

This work is supported by Shandong Key R&D Program grant 2019JZZY021005.

S. Wang et al. (Eds.): IoTCare 2021, LNICST 415, pp. 376–385, 2022.
https://doi.org/10.1007/978-3-030-94182-6_29

2 Background and Related Work

2.1 DTW Distance

In some complicated cases, the expression of the relationship between two time series (or between similar time series) can not effectively use the traditional Euclidean distance measure to express the relationship of similarity degree [9,10]. In general, sequences, taken as a whole, should have very similar shapes, laws, and properties, but they do not align along the X-axis. Therefore, before we can compare their similarities, we need to regularize one or more sequences under the timeline to get a better balance. DTW is an effective method to achieve this regularization distortion. DTW calculates the similarity between two time series to extend the sequence through dynamic programming, find one-to-one or many-to-one points, and calculate the distance between the sequences. Therefore, in 1994, Bemdt and Clifford proposed a dynamic Time Warping algorithm (DTW for short) after research. The DTW distance is mainly to find the shortest distance between sequences after planning [11].

Compared with the Euclidean distance, the length of the two sequences for calculating the DTW distance can be arbitrary and not equal to 0. DTW algorithm can detect similar positions between two sequences and make correspondence, such as points A and C in Fig. 1, rather than points A and B in Euclidean distance. However, the calculation is much more complicated than the Euclidean distance. Two time series $L_1 = \{p_1, p_2, ..., p_i, ..., p_m\}$ and $L_2 = \{q_1, q_2, ..., q_j, ..., q_n\}$ (where i = 1, 2, ..., m, j = 1, 2, ..., n)(m and n may not be equal). To calculate the DTW distance, first construct an m * n matrix, as shown in Fig. 2. Calculate the distance matrix

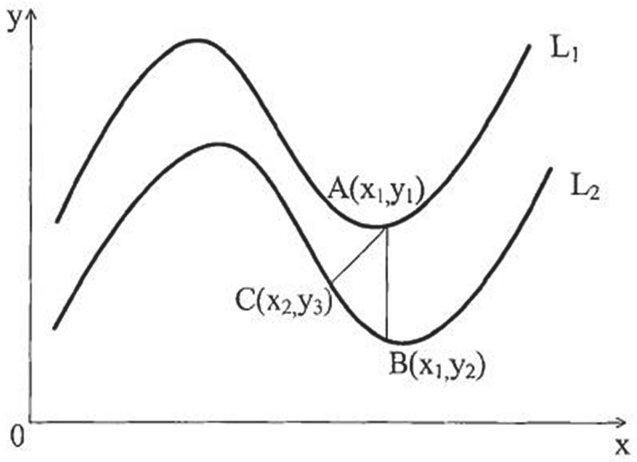

Fig. 1. Dynamic time wraping distance principle

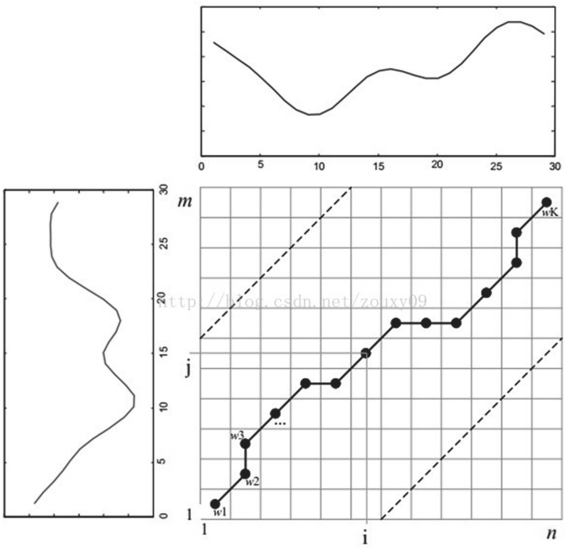

Fig. 2. DTW calculation steps

$$D = \begin{bmatrix} d_{11} & d_{12} & \ldots & d_{1j} & \ldots & d_{1m} \\ d^{21} & \ldots & . & . & . & . \\ . & & . & . & \ldots & . \\ d_{i1} & \ldots & \ldots & d_{ij} & \ldots & d_{im} \\ . & & & & . & \\ d_{n1} & d_{n2} & \ldots & d_{nj} & \ldots & d_{nm} \end{bmatrix} \tag{1}$$

where, the matrix element represents the Euclidean distance from the i-th point in the matrix to the JTH point in the matrix, which can be calculated by using Formula 1, that is, the distance between each point in the matrix and each point in the matrix can be calculated to obtain the similarity. Then you need to find a path through several nodes of the matrix. This path can be called a dynamic time warped path, which is used to represent, and the distance of the path is called the DTW distance. The KTH element is defined as the mapping between reflection and, in fact, the grid point through which the path passes, and this is the process of calculating the nearest path between the two sequences. So that gives us $max(m, n) \leq l < m + n - 1$. The path should meet the following three constraints:

(1) Boundary conditions: $w_1 = (1, 1)$ and $w_k = (m, n)$. The regular path can only start from the lower left and go to the right, up and top right until it reaches the position of (m,n).
(2) Continuity condition: if $w_k = (a, b)$, then $w_{k+1} = (a', b')$ shall satisfy $(a' - a) \leq 1 \& (b' - b) \leq 1$, that is, the dimensions in the time series can only be calculated with adjacent points and proceed in the prescribed order.

This ensures that every point in the sum of the two time series L_1 *and* L_2 can be matched without missing out.

(3) Monotone condition: if $w_k = (a, b)$, then $w_{k+1} = (a', b')$ needs to meet the condition $(a' - a) \geq 0 \& (b' - b) \geq 0$. This ensures that the path goes down in order of direction.

Satisfying the above three constraints, there are many paths, but the DTW distance is the path with the smallest overall normalization loss, i.e.

$$DTW(L_1, L_2) = min(\frac{\sum_{k=1}^{l} w_k}{l}) \tag{2}$$

Here, add up the distances of all the points on the path to get the cumulative distance s(m, n). The cumulative distance has a linear relationship with the cost, and the smaller its value is, the smaller the wrapping cost is. Dynamic cumulative distance, its calculation formula is as follows:

$$s(i, j) = s(p_i, q_j) + min\{s(i-1, j-1), s(i-1, j), s(i, j-1)\} \tag{3}$$

The cumulative distance is the DTW distance we need.

It can be seen that the calculation of distance, after the operation of DTW algorithm, all the data in the sequence will be traversed at least once. However, it can better represent the similarity between the sequences. However, the disadvantage of DTW distance is that both the time complexity and the space complexity are much higher than the calculation of Euclidean distance, which is a problem that cannot be ignored [12].

2.2 Softmax

In artificial neural networks, the most common activation function is the multivariate version of sigmoid [13]. Soft Max can be thought of as an ArgMax operation, primarily to activate functions, which is also a smooth approximation.

Given a one-dimensional vector, Softmax function maps it to a probability distribution. The softmax function $R^n \rightarrow R^n$ is defined by the following formula

$$\sigma(x_i) = \frac{e^{x_i}}{\sum_{j=1}^{n} e^{x_j}}, for\ 1, ..., n\ and\ x = [x_1, ..., x_n]^T \in R^n \tag{4}$$

2.3 Clustering

In unsupervised learning, cluster analysis is the training of unstandardized classification of data, revealing the inherent laws of the data and automatically classifying data into similar categories.

The purpose of clustering is to classify many irrelevant subsets of a dataset, that is, to divide the data samples into different clusters. A cluster can be a separate procedure for finding data set properties. Internal indicators of cluster

results do not require external data. The distance between the sample point and the cluster center is used to express the quality of the clustering analysis results. Clustering is also an important part of learning tasks, providing data for future work.

And a good clustering, there are many characteristics, specific performance in practical applications, but also

(1) Good scalability.
(2) The ability to process different data.
(3) Noise processing.
(4) Insensitive to sequential samples.
(5) Constraints.
(6) Interpretation and ease of use.

KMEANS. In 1967, MacQueene proposed the KMEANS algorithm, one of the simplest and most common clustering methods. The similarity of KMEANS is reflected in the distance between samples. The closer you are, the more likeness you have. The degree of similarity directly affects the classification criteria. However, most people use the countdown of distance to express similarities, making the two positively related. Most of the distance is from Euclidean distance or Manhattan distance [14].

Algorithm 1. Kmeans algorithm

Input: $D = \{x_1, x_2, ..., x_m\}$;
Cluster number k.

Randomly select k samples from D as the initial mean vector $\{\mu_1, \mu_2, ..., \mu_k\}$
repeat
  $C_i = \phi(1 \leq i \leq k)$
  for $j = 1, 2, .., m$ **do**
    Calculate the distance between sample x_j and each mean vector $\mu_i(1 \leq j \leq k)$: $d_{ji} = \|x_j - \mu_i\|_2$;
    The cluster marker of x_j is determined according to the nearest mean vector :$\lambda_j = arg\ min_{i \in \{1,2,...,k\}} d_{ji}$;
    Assign the sample x_j to the corresponding cluster: $C_{\lambda_j} = C_{\lambda_j} \bigcup \{x_j\}$;
  end for
  for i=1,2,...,k **do**
    Calculate the new mean vector$\mu_i' = \frac{1}{|C_i|} \sum_{x \in C_i} \boldsymbol{x}$;
    if $\mu_i' \neq \mu_i$ **then**
      Update the current mean vector μ_i to μ_i'
    else
      Leave the current vector unchanged
    end if
  end for
until None of the current mean vectors have been updated

Output: Cluster partition$C = \{C_1, C_2, ..., C_k\}$

2.4 Evaluation

Evaluation methods can be understood in engineering theory and definitions, just as learning achievement is used to represent a student's learning performance [15]. Evaluate the rationality of existing models and algorithms and compare with better models.

MAE. Mean Absolute Error is referred to as MAE for the purpose of finding the difference between the predicted value and the real value and the Absolute Error. MAE can be obtained by averaging its values.

$$MAE = \frac{1}{N}\sum_{i=1}^{N}|y_i - y_i'| \tag{5}$$

3 Methods

3.1 Clustering of Time Series

After compiling time data, the scale of the time series used for comparison was selected. Experiments show that half a day's length is suitable for the search step. 1248 is considered to be a medium-term forecast rather than a short-term forecast as it goes beyond eight steps. The generated sample sequence is characterized by half-day data sequences and the size of the search step 1248 is tabulated. Generates a time series as a data set. The clustering algorithm compares the size of sequences by comparing the similarity between sequences, and classifies them precisely so as to prepare for the traffic flow prediction at the next traffic intersections.

$$x_i = \{t_i, t_{i+1}, \ldots, t_{i+n}\} \tag{6}$$

$$y_i = \{t_{i+n+1}, t_{i+n+2}, t_{i+n+3}, t_{i+n+4}\} \tag{7}$$

n is length of search step.

The most common K-mean algorithm is chosen for clustering time series. Taking into account the periodicity of the time series, the 7-day cycle per week will be selected for cluster prediction. Because each time series represents a different time period, similar time periods between different dates are selected as data for the data cluster. First, choose the value of the k class, k is 7 elements, K points are randomly selected. When calculating Kage, if the data skew of some samples in the dataset is too large, it will make operation difficult and not easy to converge. However, it is better suited to analyze large amounts of data than hierarchical clustering.

DTW distance represents the similarity between sequences, which is better than Euclidean distance, etc., to find sequences that are similar between sequences. In the process of Kmeans clustering, DTW distance can be used to calculate the distance between the two to classify the classes. But you can choose the Tslearn library to carry out the Kmeans algorithm to calculate the DTW

distance. However, due to the high time complexity, the calculation of DTW distance is a process of dynamic programming, so the termination condition cannot be the same as the Euclidean distance when each mean is iterated to no update. The clustering can only be ended when the number of iterations of the better clustering is better. In the case of sacrificing certain accuracy, the running time of the algorithm is reduced and the overall operation efficiency of the model is improved.

After clustering all time series, the following table is given,

Table 1. Number of categories

Clusters	Amount
1	Number1
2	Number2
3	Number3
4	Number4
5	Number5
6	Number6
7	Number7

Seeing 7 clusters from 1 to 7, and the number of sequences in each category. The number of categories should be relatively equal.

3.2 Mean Value of Each Cluster

All the sequences of each category make up a group so that it is easy to calculate the average value for each attribute of the entire group and represent a sequence of current clusters. Other clusters did the same and obtained a sequence corresponding to the average of the seven clusters. The data from the sequence can be used as predictive data for each galaxy cluster. The following

$$mean\_clusters^{(i)} = \{t_1^{(i)}, t_2^{(i)}, ..., t_{143}^{(i)}, t_{144}^{(i)}, ..., t_{147}^{(i)}\} \quad i = 1, 2, ..., 7 \tag{8}$$

i is the cluster.

3.3 Similarity Metric

Since the data is a time series and there are no specific classification criteria, clustering analysis can be used to cluster such sequences to some extent [?]. Under the condition of unsupervised learning, the algorithm pays more attention to the similarity of sequences, and obtains the time series distribution roughly.

The predicted sequence is

$$sequence_of_feature = \{t_1, t_2, ..., t_{144}\} \tag{9}$$

There is no vehicle data in the last four moments of Formula 8, because the last four data are the data we need, so the predicted time series is only 144 features. What needs to be predicted is

$$prediction_of_label = \{t_{145}, t_{146}, t_{147}, t_{148}\} \tag{10}$$

And is used to calculate the similarity of 7 clusters, that is, the DTW distance is obtained by Formula 3. The distance matrix with seven clusters can be obtained

$$dist = [d_1, d_2, ..., d_7] \tag{11}$$

3.4 Similarity Weights and Data Prediction

By understanding the similarity, we can calculate the W-weighted value by using the relationship between similarity degrees and predict the value of the data. In other words, there is a mapping relationship between similarity and weight. The distance used here represents W's weight, and the greater the distance, the smaller the W_i value of the class.

From Formula 11, we can know which clusters of the data to be predicted belongs to, and the weight is needed to calculate the data to be predicted for Formula 9 and Formula 10. The predicted value is closely related to the similarity matrix A, as shown in the following formula

$$A = \frac{1}{dist} = \{\frac{1}{d_1}, \frac{1}{d_2}, ..., \frac{1}{d_7}\} \tag{12}$$

Initialize a weight matrix W,

$$W = |w_1, w_2, ..., w_7| \tag{13}$$

Since the predicted values are related to the values of the 7 clusters, only the degree of correlation is different, so the values of the row need to be added up to one, so as to show the relationship between the predicted data and the prediction values of the 7 clusters. So we need to sum each row of W to 1.

Using the Formula (4), we get

$$W = softmax(A) \tag{14}$$

So we know that the total number is

$$Forecast\_data = W \times mean\_clusters[:, 144:] = \{pred_1, pred_2, pred_4, pred_8\} \tag{15}$$

This is the current data that needs to be predicted. Perform the same operation to complete the prediction after the future time 1, 2, 4 and 8.

Put the predicted value into Formula 4 to get the value of the evaluation coefficient.

4 Conclusion and Prospect

Based on the analysis of the characteristics of vehicle flow time series and the practical application in the field, the method of combining the similar distance with the forecast data is studied. The core idea is to measure the similarity distance by using the improved DTW algorithm, to obtain the center of multiple clustering results of the total data by K-mean clustering analysis, and to get the result classification.

Although in theory, DTW distance can better represent the similarity of sequences, due to the high time complexity, it can not complete the clustering. The effect is good, but the time is too long, not suitable for the real environment. The completion of clustering will also take a long time to calculate the data.

Although the theoretical complexity of DTW algorithm is $O(n^2)$, with the cheap hardware, and because a desktop with GPU can support more than 200 terminals, it can support more than 50 intersections according to the calculation of 4 terminals at each intersection, which fully meets the needs of the main roads in general cities.

By 2019, China's total investment in smart transport will exceed 227.8 billion yuan. Investment in transport for 10 million projects will increase by 15% in the first quarter of 2020, even under the impact of the epidemic. In February 2021, the Shandong provincial public security bureau issued guidelines for strengthening the application of urban road traffic signal control, calling for the integration of big data, intelligent new thinking and new technologies to further improve the ability to apply urban traffic signals control. The algorithm adopted in this paper is efficient, accurate, economical and reasonable, especially with the improvement of the precision of machine learning and depth learning algorithm, the enhancement of computational power and the reduction of cost, and intelligent transportation costs continue to drop, scale advantage continues to appear, which greatly promotes the landing and application of these algorithms.

References

1. Zhang, G.: Research and Application on Interval Time Series Clustering Based on DTW. Northwest Normal University (2020)
2. Geng, R., Sun, B., Ma, L., Zhao, Q., Shen, T.: Anomaly-aware in sequence data based on MSM-H with EXPoSE. In: 40th Chinese Control Conference, CCC 2021, Shanghai, China (2021)
3. Sun, B., Cheng, W., Goswami, P., Bai, G.: Short-term traffic forecasting using self-adjusting k-nearest neighbours. IET Intel. Transp. Syst. **12**(1), 41–48 (2018)
4. Li, S.: Clustering for Time Series Based on the Feature. Guangxi Normal University (2014)
5. Ma, L., Sun, B., Ziyi, L.: Bagging likelihood-based belief decision trees. In: 20th International Conference on Information Fusion (FUSION), Xi-An, China, pp. 1–6 (2017). http://ieeexplore.ieee.org/abstract/document/8009664/
6. Lafuente-Rego, B., Vllar, J.A.: Clustering of time series using quantite autocovariances. Adv. Data Anal. Classif. **10**(3), 391–415 (2016)

7. Sun, B., Cheng, W., Bai, G., Goswami, P.: Correcting and complementing freeway traffic accident data using Mahalanobis distance based outlier detection. Tehnicki Vjesnik (Tech. Gaz.) **24**(5), 1597–1607 (2017)
8. Plant, C., Wohlschhiger, A.M., Zherdin, A.: Interaction-based clustering of multivariate time series. In: The 9th IEEE International Conference on Data Mining, ICDM 2009, Miami, Florida, USA, 6–9 December 2009, pp. 914–919 (2009)
9. Sun, B., Wei, C., Liyao, M., Prashant, G.: Anomaly-aware traffic prediction based on automated conditional information fusion. In: International Conference on Information Fusion (FUSION), Cambridge, United Kingdom, pp. 2283–2289. IEEE (2018)
10. Ying, L.: Research on Clustering Methods for Time Series. Liaoning Normal University (2012)
11. Ma, L., Sun, B., Han, C.: Learning decision forest from evidential data: the random training set sampling approach. In: 4th International Conference on Systems and Informatics (ICSAI), Hangzhou, China (2017)
12. Sun, B., Cheng, W., Goswami, P., Bai, G.: An overview of parameter and data strategies for k-nearest neighbours based short-term traffic prediction. In: 2017 ACM International Conference Proceeding Series, pp. 68–74. ACM (2017)
13. Jianle, S.: Stock Price Trend Prediction Research Based on Time Series Similarity. Chongqing Jiaotong University (2014)
14. Ashish, S., Dale, E.: Clustering for multivariate time series data. In: Proceedings of the American Control Conference, Anchorage, May 2002, pp. 586–591 (2002)
15. Sun, B., Ma, L., Shen, T., et al.: A robust data-driven method for multi-seasonal and heteroscedastic IoT time series preprocessing. Wirel. Commun. Mob. Comput. (WCMC) **2021**, 1–11 (2021). Article ID 6692390