Time Series Prediction with Preprocessing and Clustering

Haoxuan Sun[1], Shuai Lin[2], Lin Han[1(✉)], Jidong Feng[1(✉)], and Mingxu Sun[1]

[1] School of EE, University of Jinan, Jinan 250022, China
201921200627@mail.ujn.edu.cn, fjd_844575264@163.com
[2] Shandong Non-metallic Materials Institute, Jinan 250031, China

Abstract. This paper studies the similarity of time series, and studies the influence of weight on prediction results on the basis of clustering. We first introduce the practical significance and research purpose of the selected topic, summarizes the current research situation at home and abroad, and summarizes the research content of this paper. Second, we describe related concepts. Later, based on Dodger data set, we study the flow of total prediction data of time series. First of all, feature extraction of the data, pre-processing work, the original data generation time series. Then the data are processed and divided into training data and test data for the convenience of subsequent processing. Then the clustering algorithm was used to divide the time series into categories, and seven categories were divided according to the characteristics of one week time cycle. The average value of each category is calculated to replace the characteristics of the current category, and then the similarity is compared. Finally, the weight of each category is calculated by similarity degree, and then the data is predicted. MAE, R-squared, MAPE and other indicators were used to analyze and evaluate the forecast data.

Keywords: Time series · Kmeans clustering · Similarity

1 Introduction

With the development of artificial intelligence with big data, various fields have developed to varying degrees [1], and more and more artificial intelligence products appear in more industries [2]. In the field of transportation, there is also rapid development. Mass time series obtained through bus cards, detectors, cameras, communication equipment, the Internet, etc. [3–5]. However, due to periodicity and high noise, how to make use of time series is still an unresolved problem [6].

As an important part of the industry chain of the Internet of Things, the transportation Internet of Things has the characteristics of high market maturity, mature industry sensing technology, and strong government support [7]. In the slogan of building "digital city" and "smart city", intelligent transportation

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

S. Wang et al. (Eds.): IoTCare 2021, LNICST 415, pp. 358–367, 2022.
https://doi.org/10.1007/978-3-030-94182-6_27

system has been applied on a large scale in many cities, with broad market prospects and huge investment opportunities [8,9]. It will become a key area for the development of the Internet of Things industry in the next few years.

2 Related Conception

2.1 Euclidean Distance

The Euclidean metric (most of the time also called the Euclidean distance) is a very familiar distance [10]. It's a calculation of distances between multidimensional vectors usually defined in terms of distances in m-dimensional vector space, you can think of it as a point in a higher dimensional vector space, and the distance between them can be thought of as the distance from this point to the origin, or it can be the actual distance from this point to another point [11] (Fig. 1).

$$distance(X, Y) = \sqrt{\sum (x_{ti} - y_{ti})^2} \quad t = 1, 2, ..., n \tag{1}$$

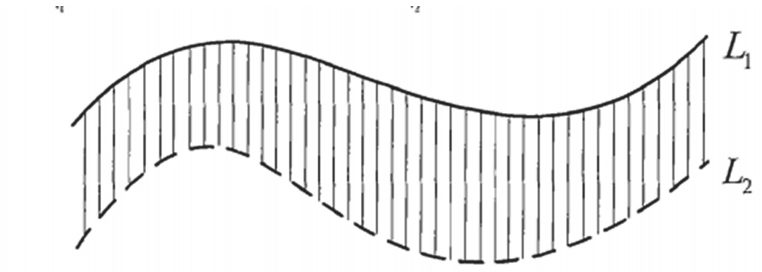

Fig. 1. Time series Euclidean distance schematic diagram

2.2 Softmax

Softmax function in artificial neural network, the activation function used mostly in this, is the sigmoid multivariate versions. Softmax can be thought of as the operation of Arg Max, mainly used for activation functions, which are also approximations of smoothing [12].

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

$$\sigma(x_i) = \frac{e^{x_i}}{\sum_{j=1}^{n} e^{x_j}}, for\ 1, ..., n\ and\ x = [x_1, ..., x_n]^T \in R^n \tag{2}$$

2.3 Clustering

In unsupervised learning, clustering is to train data values without standard classification, reveal the internal rules of the data, and automatically divide the data into similar categories [13,14].

The purpose of clustering is to classify many unrelated subsets in the data set, that is, to divide the data samples into different clusters. Clustering can be a separate process to find the rules of the data set. Clustering is also an essential part of the learning task, providing data for subsequent work.

KMEANS. In 1967, MacQueenE proposed the KMEANS algorithm, which is the most common clustering method and also one of the simplest algorithms. The similarity of KMEANS is represented by the distance between samples. The closer the distance, the higher the similarity will be [15]. Most people will use the reciprocal of distance to express the similarity, so that the two become a positive correlation. And the distance mostly choose European distance or Manhattan distance.

2.4 Evaluation

The evaluation method, just like using academic performance to represent the good or bad learning conditions of students, can be understood in the engineering theory and definition, in order to better optimize the algorithm.

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 [16]. MAE can be obtained by averaging its values.

$$MAE = \frac{1}{N}\sum_{i=1}^{N}|y_i - y_i'| \tag{3}$$

MAPE. Mean Absolute Percentage Error (MAPE) is one of the most popular indicators for evaluating predictive performance [17].

$$M = \frac{1}{n}\sum_{t=1}^{n}\frac{A_t - F_t}{A_t} \tag{4}$$

where A_t represents the actual value and F_t is the predicted value.

3 Clustering with Multiple Similarity Measurements

3.1 Data Preprocessing

The Dodger data set used in this example has a large number of '−1' values, which means that there are many missing values. If missing values are not taken

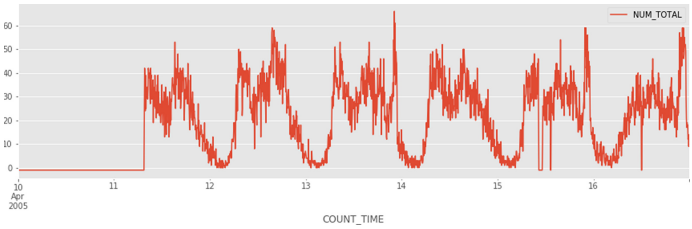

Fig. 2. An overview of several days of raw data

into account, '−1' will have a great impact on the predicted value of the data, making it very difficult to process the subsequent data.

Figure 2 shows a large number of consecutive missing values, with no data for several consecutive days. For example, one day on May 10, 2005. In order to restore the trend of time series as much as possible, all valid data are grouped and the missing data is filled into the average value of the current time point. The corresponding values of 288 time points a day were calculated (data were obtained every 5 min), and the average value of each time point within 24 h was taken.

	time	NUM_TOTAL
0	0.00	8
1	0.05	8
2	0.10	7
3	0.15	7
4	0.20	7
5	0.25	6
6	0.30	6
7	0.35	6
8	0.40	6
9	0.45	5

Fig. 3. Filled data

Rounding the average value of the data to integers to facilitate calculation and reduce the data footprint. As shown in Fig. 3. This shows the trend of the day, as well as the change of the vehicle over the day. In this paper, the characteristics of traffic flow time, using data analysis of the relevant content of data processing. Due to the large amount of data used, Pandas library is used to assist the calculation.

3.2 Reconstruct Time Series

Since the resulting data is not a time series but a timestamp, permutation and combination are required to generate the time series. Since the time period of each day represented by each time series is different, the similar time period between different days is selected as the data clustering. Through the experiment, it is found that the length of half a day as the search step is appropriate. 1248 is because more than 8 steps (40 min) is considered to be a medium term forecast rather than a short term forecast. Half a day's traffic flow data (144 pieces of data in 5 min for each group) were selected as the search step of the time series, and expressed as

$$x_i = \{t_i, t_{i+1}, \ldots, t_{i+n}\} \quad n = 143 \tag{5}$$

In addition, the next period of time series is selected for prediction, and the data with the step length of 1, 2, 4, and 8 are selected for prediction, which is expressed as

$$y_i = \{t_{i+n+1}, t_{i+n+2}, t_{i+n+3}, t_{i+n+4}\} \quad n = 143 \tag{6}$$

With the help of Python's Numpy and Pandas library, the sequence of datasets for the production experiment is shown in Fig. 4 below

var1(t-143)	var1(t-142)	var1(t-141)	var1(t-140)	var1(t-139)	var1(t-138)	var1(t-137)	var1(t-136)	var1(t-135)	var1(t-134)	...	var1(t-5)	var1(t-4)	var1(t-3)	var1(t-2)	var1(t-1)	var1(t)	var1(t+1)	var1(t+2)	var1(t+4)	var1(t+8)
8	8	7	7	7	6	6	6	6	5	...	24	25	26	26	26	25	24	24	25	25
8	7	7	7	6	6	6	6	5	5	...	25	26	26	26	25	24	24	25	25	26
7	7	7	6	6	6	6	5	5	5	...	26	26	26	25	24	24	25	25	25	25
7	7	6	6	6	6	5	5	5	5	...	26	26	25	24	24	25	25	25	25	25
7	6	6	6	6	5	5	5	5	5	...	26	25	24	24	25	25	25	25	25	24
6	6	6	6	5	5	5	5	5	4	...	25	24	24	25	25	25	25	25	26	24
6	6	6	5	5	5	5	5	4	4	...	24	24	25	25	25	25	25	25	25	24
6	6	5	5	5	5	5	4	4	4	...	24	25	25	25	25	25	25	26	25	26
6	5	5	5	5	5	4	4	4	5	...	25	25	25	25	25	25	26	25	24	25
5	5	5	5	5	4	4	4	5	4	...	25	25	25	25	25	26	25	25	24	25

Fig. 4. A sequence of total data sets

In order to have a better generation effect, the data were divided into test set and training set, which were different from each other, and the order was random. The data was broken up to simulate the real data prediction results. The train_test_split method was used to take 25% of the whole data as the predicted data.

3.3 Clustering Similarity

Clustering with Euclidean Distance. The Kmeans algorithm used in this section is data clustering, which is a partition-based clustering method. Since you are using a time series as your data, you need to find similar sequences. Clustering analysis can be used to cluster time series to a certain extent. In the

case of unsupervised learning, the algorithm will pay more attention to sequence similarity for sequence classification, and then roughly obtain the distribution of time series. Therefore, the method to express the similarity measure between samples is very important. In the Kmeans algorithm, the Euclidean distance and Manhattan distance used by most people represent the similarity measure of time series. But the Euclidean distance is more common. But the limitation of Euclidean distance is that one dimension is much bigger than the others, and the distance is excessively affected by this dimension.

Python's Sklearn library can use the Kmeans algorithm, but the default Euclidean distance is not easy to modify.

Classification. It can be seen that the clustering methods based on these distances have some differences in the distribution of classification of prediction data, as shown in the figure below

```
array([0.12571429, 0.10857143, 0.10857143, 0.18285714, 0.17142857,
       0.21142857, 0.09142857])
```

Fig. 5. Cluster number distribution of two distances

As can be seen from the Fig. 5, the main purpose of using the test set is to simulate the effect of the algorithm in the real scenario and show whether the algorithm has a good generalization effect, so as to better apply the improvement and optimization to the real scenario.

It can be clearly seen from the clustering results that the clustering data situation. Considering the morphological characteristics of traffic flow time series and the purpose of clustering for 7 days, all kinds of values should tend to average. The result of clustering is even.

Similarity Weights. After clustering all sequences, the grouping situation is obtained, and the average value of feature and label of each cluster is calculated to represent the current cluster. The similarity between the predicted sequence and the 7 clusters can be calculated. The distance matrix of the formula 1 can be generated. In the process of calculating the distance, we use the above mentioned distance metrics to calculate the distance. And through the reciprocal, can be directly converted to generate A Update the weight matrix W, dist related to A, and obtain the current final weight matrix after the modification of Formula 2.

```
array([[198.38870561, 312.30204145, 147.5663145 , 221.7634799 ,
        132.98371663, 179.37599818, 275.94923531]])
```

Fig. 6. Distance between sequence and each cluster

```
array([[0.00504061, 0.00320203, 0.00677661, 0.00450931, 0.00751972,
        0.00557488, 0.00362385]])
```

Fig. 7. Similarity between sequence and each cluster

Meanwhile, set the weights w1, w2, ..., w7, W = sofetmax (A), and finally the data can be predicted (Figs. 6 and 7)

$$predict = W \times Y_m = [24.35835927, 24.12020417, 21.55549108, 21.44721639] \quad (7)$$

Y_m is the label of each cluster.

3.4 Calculation of Total Data Value

A set of data obtained at this point. Iterate the remaining test set data. The data of all test sets are completed and the algorithm is finished. So now I have my prediction for the next time, and I do the same thing and I get my prediction after time 1, 2, 4, 8 (Fig. 8).

	tn+1	tn+2	tn+4	tn+8
0	12.107919	12.167536	11.333130	12.605538
1	30.857143	31.500000	32.214286	31.571429
2	18.315708	19.278737	19.036072	21.510601
3	23.068570	22.870936	24.329254	21.738762
4	17.198980	18.747107	16.647945	17.276340
...	...	...	...	...
12558	14.408070	17.267959	17.929496	22.149490
12559	13.604976	14.614037	14.665070	16.120974
12560	28.903030	29.597521	29.757800	27.339606
12561	19.573403	20.247104	22.313147	21.755453
12562	32.181818	34.909091	34.727273	36.363636

Fig. 8. Total data of prediction

3.5 Analysis

In the Euclidean distance prediction data, compare the predicted data at four times with the real data in the test set, and draw the following figure

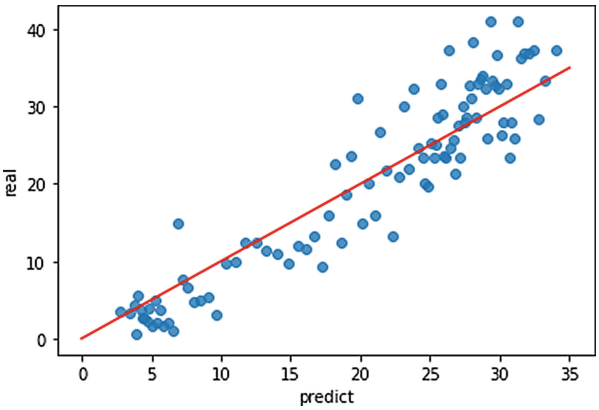

Fig. 9. The Q-Q graph of the method

As can be seen from the picture, the weight vector can reflect the data characteristics to some extent, but we can see that there is a part of the data close to the position of horizontal axis 0. Since the data set was recorded 15 years ago, the conditions at that time could not be compared with now, so the record was caused by many missing data, power outages and other problems, which could not be avoided. The domestic traffic environment is improved to ensure people's personal safety, and protective measures are gradually in place. However, it can be seen that on the oblique line of 45°, a large number of data sets indicate the rules of most fitting sequences of data.

The clustering prediction of Euclidean distance takes about 21 s to calculate 100 data. Therefore, the Kmeans clustering algorithm based on the Sklearn library is feasible and effective, and it is easier to use in actual scenes.

As can be seen from Fig. 9, it is obvious to see the gap between the predicted data of each category and the real data.

Looking at the values of MAE, R2 and MAPE after the completion of the prediction, we can clearly see that the actual difference between the evaluation values of the data at the four moments is not very big.

As can be seen from Fig. 10, the further back the predicted data is, the worse the accuracy of the prediction will be, and the higher the prediction needs to be after a long period of time. Therefore, the closer the moment of the predicted data is to the time series, the higher the accuracy of the prediction will be.

```
第t n+ 1 时刻的预测值
5.541900932250141       0.3957427310493161      0.49763984017579377
第t n+ 2 时刻的预测值
5.587645373660939       0.377720614522142       0.4969103345177101
第t n+ 4 时刻的预测值
5.647849973231797       0.3556265976099703      0.5171025866886609
第t n+ 8 时刻的预测值
5.7866137881857425      0.2963791757598844      0.5277307158591259
```

Fig. 10. Estimates of the three distances

4 Conclusion and Prospect

Through the analysis of the characteristics of the time series of vehicle flow, combined with the actual scene application, this paper studies a method of combining the various similarity distances to forecast data, and analyzes the different results of the application of several similarity measures. The core idea is as follows: the improved DTW algorithm is used to measure the similarity distance, the Kmeans clustering algorithm is used to obtain the centroid of multiple clustering results of the total data, the classification of clustering results is obtained, and the prediction of the final data is completed by combining with the similarity, and the evaluation of the validity index of clustering results is completed.

In 2019, the total investment scale of smart transportation in China exceeded 227.8 billion yuan. Even under the influence of the epidemic, the transportation investment of 10 million yuan project in the first quarter of 2020 increased by 15% year on year. In February 2021, Shandong Provincial Department of Public Security issued the Guiding Opinions on Strengthening the Application of Urban Road Traffic Signal Control, which required the integration of big data, intelligent new thinking and new technology to further improve the application ability of urban road traffic signal control. The algorithm used in this paper is efficient, accurate and economical. In particular, with the improvement of machine learning and deep learning algorithm accuracy, the improvement of computing power and the decrease of cost, the cost of intelligent transportation is constantly reduced and the advantages of scale are constantly emerging, which greatly promote the implementation and application of the algorithms listed in this paper.

This work is supported by Shandong Non-metallic Materials Institute under grant WSJL20206C069.

References

1. Wang, W., Shan, X.: Study on regular pattern of railway passenger flow in three-day holiday based on clustering method of time series. Railw. Comput. Appl. **04**, 23–27 (2015)

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. Ji, M., Xiao, L.: A dynamic k-means clustering algorithm for time series data. Comput. Digit. Eng. **48**(8), 1852–1857 (2020). https://doi.org/10.3969/j.issn.1672-9722.2020.08.007
5. Lin, Q.: Research on Feature Screening and Clustering Analysis of Time Series Data - A Case Study of the CSI 300 Index. Southwestern University of Finance and Economics (2017)
6. 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/
7. 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), pp. 2283–2289. IEEE, Cambridge, United Kingdom (2018)
8. Zheng, C.Z.L.: Shape clustering on time series data. In: Proceedings of Information Technology and Environmental System Sciences, ITESS 2008, vol. 3, pp. 1249–1253 (2008)
9. 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 (209)
10. 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)
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. Li, F., Tan, L., et al.: On the data-mining oriented methods for clustering time series. Comput. Sci. **027**(012), 76–80 (2000)
13. Zijian, T.: Time Series Forecast via Similar Fluctuate Pattern. Hefei University of Technology (2016)
14. 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)
15. Liu, C.: Research on Interactive Prediction of Airport Noise Monitoring Points Based on Time Series Similarity Measure. Nanjing University of Aeronautics and Astronautics
16. Sun, B., Ma, L., Shen, T., et al.: A robust data-driven method for muti-seasonal and heteroscedastic IoT time series preprocessing. Wirel. Commun. Mob. Comput. (WCMC) **2021**, 1–11 (2021). Article ID 6692390
17. Lai, Y.: Study on Real-Time Prediction of Arrival Time for Floating Transit Vehicle. Chongqing University (2011)