Clustering-XGB Based Dynamic Time Series Prediction

Haoxuan Sun[1], Kun Zhang[2], Tingting Wang[1(✉)], Wanfeng Ma[1(✉)], and Qinjun Zhao[1]

[1] School of EE, University of Jinan, Jinan 250022, China
202021100395@mail.ujn.edu.cn, elephantfff@163.com
[2] Shandong Non-metallic Materials Institute, Jinan 250031, China

Abstract. This work analyzes time series and find the rules and statistical characteristics from the numerous data. According to the purpose of the time series analysis, we find the rules and conduct the future time forecast. This paper is mainly based on the similarity of time series. Based on clustering results, XGB is used to reflect the relationship between similarity and clusters' weights and to predict the value. Overall, it is a time series prediction model based on clustering and XGB regulated weights. The process of model prediction is realized by using instances in dataset, and the relationship between similarity and weights is obtained by using XGB.

Keywords: Time series · KMEANS clustering · XGBoost

1 Introduction

As a highly integrated and integrated application of the new generation of information technology, the Internet has strong penetrating power, strong driving effect and good comprehensive benefits [1–3]. It is the computer, Internet and mobile communication networks after the development of the information industry is another driving force. Logistics industry, as an important part of the internet of things industry chain, with its characteristics of high market maturity, wide market prospects and big investment opportunities, will become a key area of logistics networking industry development in the next few years [4,5].

With the development of big data of artificial intelligence, various fields have been developed to different extent, artificial intelligent products appear more and more in more industries [6]. There have also been rapid developments in the area of transport. Quality time series obtained through bus cards, detectors, cameras, communication equipment, Internet, etc. However, due to the periodicity of time and the influence of noise, how to utilize time series is still a problem to be solved [7,8].

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

S. Wang et al. (Eds.): IoTCare 2021, LNICST 415, pp. 341–351, 2022.
https://doi.org/10.1007/978-3-030-94182-6_25

2 Related Conception

2.1 Euclidean Distance

Much of the time of Euclidean equations is also known as euclid distance, which is a familiar distance [9,10]. It is a calculation of the distance between multidimensional vectors, usually defined in terms of distances in m-dimensional vector spaces, and you can think of it as a point in a vector space with a higher dimension. The distance perception between them can be seen as one from this point to the origin, or as the actual distance from one to another [11].

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

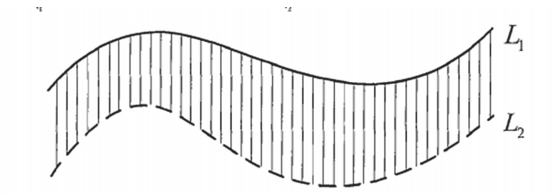

Fig. 1. Time series Euclidean distance schematic diagram

2.2 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 clusters [12].

KMEANS. In 1967, MacQueene proposed the KMEANS algorithm, one of the simplest and most common clustering methods [13,14]. 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 Europe or Manhattan.

2.3 Evaluation

Evaluation methods can be understood in engineering theory and definitions, just as learning achievement is used to represent a student' s learning performance.

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$C = \{C_1, C_2, ..., C_k\}$

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 [15].

$$MAE = \frac{1}{N} \sum_{i=1}^{N} |y_i - y_i'| \tag{2}$$

$\boldsymbol{R^2}$. R-squaredis a relative measure that measures the sum of residuals of existing models far from the standard of the sum of residuals of the benchmark model [16].

$$SSR = \sum_{i=1}^{N} w_i (\overline{y_i} - \widehat{y_i})^2 \tag{3}$$

$$SST = \sum_{i=1}^{N} w_i (\overline{y_i} - y_i)^2 \tag{4}$$

so,

$$R^2 = 1 - \frac{SSR}{SST} \tag{5}$$

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

$$M = \frac{1}{n}\sum_{t=1}^{n}\frac{A_t - F_t}{A_t} \tag{6}$$

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

2.4 Grid Search

Grid Search, a common method, is mainly used for tuning parameters. The model is tested by using different parameters in turn in order to find the parameters corresponding to the best effect. Is to use exhaustive search for hyperparameters, for each case to solve. The super parameters are all the cases of the pre-specified parameters, select from the inside to find the best performance of the set of parameters [18].

2.5 XGBOOST

It can be referred to as XGB, which was developed by Dr. Tianqi Chen from the University of Washington in the United States [19]. The idea of the algorithm is to grow a tree by continuously adding trees and continuously splitting features. Each addition of a tree is actually learning a new function to fit the residual error predicted last time. When we complete the training and get k trees, we need to predict the score of a sample. In fact, according to the characteristics of the sample, it will fall to a corresponding leaf node in each tree, and each leaf node corresponds to a score. Finally, we only need to add up the corresponding scores of each tree to get the predicted value of the sample.

$$\begin{gathered}\widehat{y} = \Phi(x_i) = \sum_{k=1}^{K} f_k(x_i) \\ where\ F = \{f(x) = w_{q(x)}\}(q : R^m \to T, w \in R^T)\end{gathered} \tag{7}$$

$w_{q(x)}$ is the fraction of leaf node q, and f(x) is one of the regression trees. Objective function is

$$\begin{gathered}Obj = \sum_{i=1}^{n} l(y_i, \widehat{y_i}) + \sum_{k=1}^{K} \Omega(f_k) \\ where\ \Omega(f) = \gamma T + \frac{1}{2}\lambda\|w\|^2\end{gathered} \tag{8}$$

The objective function above has two parts, the former is the difference between the real and the predicted, the latter is the regularization term, T is the number of leaf nodes, w is the fraction of leaf nodes. γ controls the number of leaf nodes, λ controls the fraction of leaf nodes will not be too large, can prevent overfitting.

As shown in the following example, two decision trees are trained

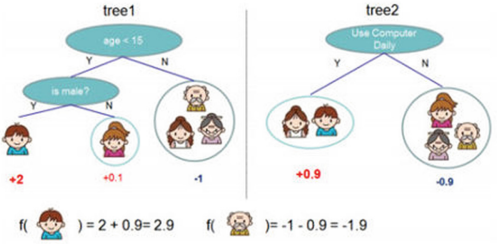

Fig. 2. A simple example of XGB

3 Clustering Prediction and Comparison Based on Multiple Similarity Degrees

3.1 Data Preprocessing

The dodge dataset used in this example has a large number of "−1" values, which means there are many defaults. If missing values are not taken into account, −1 will have a significant impact on the predicted values of the data, making it difficult to process subsequent data.

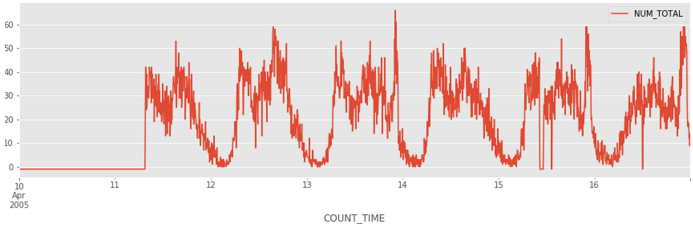

Fig. 3. An overview of several days of raw data

As shown in Fig. 3, in the absence of data for several consecutive days, a large number of consecutive missing values are found in data and a significant number continue to be missing. For example, one day on 10 May 2005. In order to eliminate the disturbance of missing values, combined with the fact that traffic data at traffic junctions is on a one-day cycle, all day data are used to fill in the missing value. In order to minimize changes in the sequence, the time points of missing data are added to the average of the current points and all valid data is grouped. The corresponding values of 288 points of time in the daily HT data are calculated every five minutes to fill the data with the average values in each point in time over a 24-h period.

The average data are rounded to integers to facilitate operation and reduce the data space occupied.

3.2 Generating Time Series

Because time series represent different time periods, a similar time period between different dates is selected as data for the data clustering. Experiments show that the length of half a day is suitable for the size of the search step. 1248 Because it is a 40-min phased step over eight steps, which is usually considered a medium-term forecast rather than a short-term one. Half-day traffic data in a five-minute group selects 144 data as search steps in the time series, which is expressed as

$$x_i = \{t_i, t_{i+1}, \ldots, t_{i+n}\} \quad n \text{ is the search step size} \tag{9}$$

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

$$y_i = \{t_{i+n+1}, t_{i+n+2}, t_{i+n+3}, t_{i+n+4}\} \quad n \text{ is the search step size} \tag{10}$$

The sequence of datasets for the production experiment is shown in Fig. 4.

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. 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 of Euclidean Distance. The Kmeas algorithm used in this section is data clustering. Under the condition of unsupervised learning, the algorithm pays more attention to the similarity of sequences, and obtains the distribution of time series roughly. In the Kmeas algorithm, most people use Euclidean distance and Manhattan distance to represent the similarity measure of time series.

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

Fig. 5. Cluster number distribution of distance

Classification. As can be seen, the following figure shows the distribution of the predicted data according to the distance based clustering method is shown in the figure below

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 actual scene, showing whether there is a good generalization effect and then better application of improvement and optimization in the real situation.

The result of clustering analysis shows that the clustering data has changed obviously. Taking into account only the morphological characteristics of traffic flow time series and the original purpose of clustering them for seven days, the values of each class should tend to be averaged.

Similarity. After clustering all traffic flow time series, we can learn about grouping and find the average of the characteristics and tags of each cluster to represent the total value in the current cluster. The similarity between the predicted sequence and the seven clusters can be calculated. A distance matrix can be generated from the formula 1. The following is the distance of the sequence

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

Fig. 6. Distance between sequence and each cluster

Knowing the similarity, we can use the relationship between the similarity to calculate the weight W, which is used to predict the data value. That is to say, there is a mapping relationship between similarity and weight.In order to better show the mapping relationship between similarity and distance, the method of XGB will be used in this paper.

3.4 XGB's Method to Calculate the Weight

In fact, since we do not know the exact relationship between similarity and weight, the relationship between similarity and weight of each category may not be the positive correlation that we think the greater the similarity is, the higher the weight is.

Therefore, the method of XGB model was chosen in order to better fit the relationship between the two. As long as the distance between each category is required, the weight of each category can be obtained, and the predicted value can be obtained directly.

Construction of New Data Set. Calculate the distance of 7 categories of data, and sort them from small to large. These 7 clusters are taken as features of the training data. Label is the weight of each of the seven clusters.

Table 1. New data set

dist1	dist2	...	dist7	w_1	w_2	...	w_7
	...	...	...	...	...	...	

In the above experiment, after the similarity of a sample data is obtained, the distance between the predicted sequence and 7 clusters is sorted, and the set parameters are determined by Grid Search, and the sum of the 7 weight parameters is 1. The parameter with the least difference between the predicted value and the actual value, that is, the one with the best effect, is selected as the label of the current data. The distance of seven classes and the parameter of weight W are combined as a sample data.

After all the original test set data processing is completed, the data set of Table 1 is constructed. It is used to calculate the weight directly. 80% of the data are divided into training model XGB, and the rest are tested to see whether the model is good or bad.

Prediction of XGB. The training set of the data set is put into the training model, and the test set is used to evaluate the model and check the effect of the model.

Because the predicted weight has no constraint that the sum is 1, it is necessary to normalize the predicted samples, set one well, and express the relationship with each category.

Analysis of the Model. Since the size of the XGB data set will be much smaller than the previous data set, it is possible that the prediction of the data will be less effective.

However, the XGB method has its own advantages, which can more accurately get the appropriate weight, and get the proportion relationship between each category and the predicted value. Forecasts are much more accurate. The XGB model preprocessing and data storage after each step will not consume too much time and space, which is about the same as the Euclidian distance. But if the use of DTW, enough time, there will be a better effect. Due to time constraints, only XGB is used for comparative analysis of Euclidean distance here.

The following figure shows the weights of the predicted results.

```
array([[0.09259243, 0.503311  , 0.09004612, ..., 0.1951583 , 0.02927206,
        0.1178335 ],
       [0.16224124, 0.2085919 , 0.15434733, ..., 0.10209591, 0.08707408,
        0.08930299],
       [0.16820015, 0.17057562, 0.14777775, ..., 0.08558471, 0.11823782,
        0.17418884],
       ...,
       [0.2544834 , 0.23196718, 0.14363924, ..., 0.08511011, 0.07576442,
        0.1274816 ],
       [0.11999956, 0.11634909, 0.14853086, ..., 0.5316235 , 0.101652  ,
        0.02974296],
       [0.08982415, 0.1501841 , 0.20426954, ..., 0.16044119, 0.03047957,
        0.0235962 ]], dtype=float32)
```

Fig. 7. Weight of prediction

Due to the advantages of XGB itself, the accuracy of the prediction data is much better. So, if you have a larger amount of data, it will work better for the model store, and XGB has a loss function to prevent overfitting.

```
9.97287936201832        -2.6165895439404885        0.9350044508417967
```

Fig. 8. MAE,R2,MAPE

The Q-Q diagram of XGB is as follows

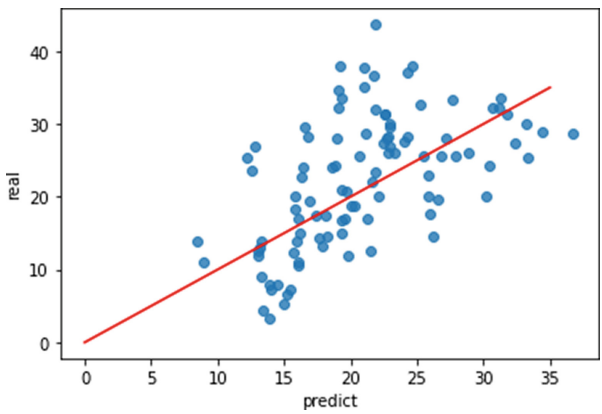

Fig. 9. XGB's Q-Q figure

Even if the data set is relatively small, the effect is still good, indicating the feasibility of this method is relatively strong.

4 Conclusion and Prospect

Euclidean distance will still be widely used in practice as a similarity, and XGB can be used if necessary, because the grid search will make the parameters more reasonable, and it will have better prediction and generalization effect.

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.

As a highly intensive integration and comprehensive application of a new generation of technology, the Internet of Things has the characteristics of high knowledge intensity, wide application range, great growth potential, strong driving force and good comprehensive benefits. China's 12th Five-Year Plan has accumulated nine fields of projects: smart industry, smart agriculture, smart logistics, smart transportation, smart power grid, smart environmental protection, smart security, smart medical care, and smart home. These areas of concentrated development almost cover all aspects of our social production and life.

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. Weiwei, W.A.N.G., Xinghua, S.H.A.N.: Study on regular pattern of railway passener flow in three-daw 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 Intell. 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. 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/ http://ieeexplore.ieee.org/abstract/document/8009664/
6. 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, UK, pp. 2283–2289. IEEE (2018)
7. 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)

8. 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: ACM International Conference Proceeding Series, pp. 68–74. ACM (2017)
9. Zhang, G.: Research and Application on Interval Time Series Clustering Based on DTW. Northwest Normal University (2020)
10. 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)
11. Chen, H., Liu, C., Sun, B.: Survey on similarity measurement of time series data mining. Control Decision **32**(001), 1–11 (2017)
12. Sun, B., Ma, L., Shen, T., et al.: A robust data-driven method for multi-seasonal and heteroscedastic IoT time series preprocessing. Wirel. Commun. Mobile Comput. 6692390 (2021)
13. Lai, Y.: Study on Real-Time Prediction of Arrival Time for Floating Transit Vehicle. Chongqing University (2011)
14. Sun, B., Cheng, W., Bai, G., Goswami, P.: Correcting and complementing freeway traffic accident data using mahalanobis distance based outlier detection. Tehn. Vjesn. Techn. Gazette **24**(5), 1597–1607 (2017)
15. Lyu, Z.: Price Forecast and Comparative Study of Stock Index Futures Based on Machine Learning Algorithms. Zhejiang University (2020)
16. Sun, B., Cheng, W., Goswami, P., Bai, G.: Short-term traffic forecasting using self-adjusting k-nearest neighbours. IET Intell. Transp. Syst. **12**(1), 41–48 (2018). https://doi.org/10.1049/iet-its.2016.0263
17. Jiang, D., Pei, J., Zhang, A.: DHC: a density-based hierarchical clustering method for time series gene expression data. BIBE 393–400 (2003)
18. Ashish, S., Dale, E.: Clustering for multivariate time series data. In: Proceeding of the American Control Conference Anehorage, May, 2002, pp. 586–591 (2002)
19. Zheng, C.Z.L.: Shape clustering on time series data. In: Proceedings of Information Technology and Environmental System Sciences (ITESS), vol. 3, pp. 1249–1253 (2008)