Clustering framework based on multi-scale analysis of traffic flow time series

Ruifang Li, Jin Yu*
School of Chongqing University, Chongqing, China

*Corresponding author: 20180702034t@cqu.edu.cn

Abstract. Clustering analysis of traffic flow time series can improve the accuracy of traffic flow prediction, which is the basis of traffic planning. The correlations between traffic flow series imply that there are some potential patterns of traffic flow movements. In this work, a clustering framework based on multi-scale analysis of traffic flow time series is proposed to seek these potential patterns. The framework includes the selection of the optimum algorithm, the construction of quantitative indicators to evaluate the clustering effect, and the visual inspection. The experimental results of 24-hour long-term traffic flow time series clustering on large-scale road network in Shenzhen indicate that the model can clearly distinguish different classes of traffic flow time series.

Keywords: Clustering, Traffic flow time series, Fluctuation pattern.

1. Introduction
Long-term traffic flow prediction on large road network is the basis of traffic control strategy. In fact, there are different fluctuation patterns among traffic flow time series. The traffic flow prediction model can predict each series separately based on clustering, which can improve the accuracy of traffic flow prediction.

Traffic flow clustering researches on data mining and artificial intelligence technology have become popular due to the increasing availability of urban traffic data. Such as: traffic flow theory [1, 2] and macroscopic fundamental diagram (MFD) [3], use the relationship between traffic parameters to distinguish traffic states. Clustering the raw traffic flow parameters in the original data space [4]. Through feature extraction, the primary data space is abstracted into feature space to reduce the dimension [5].

However, most of the literatures are based on the short-term traffic flow sequence clustering on small-scale road network, and do not ensure the stability and effectiveness of the final clustering algorithm. Therefore, this paper proposes a multi-scale clustering framework based on traffic flow time series analysis, including algorithm selection, algorithm stability and clustering effect evaluation indicators, and adds a visual inspection process to ensure that the selected optimal clustering algorithm can effectively identify different fluctuation patterns between traffic flow time series.

The rest of the paper is organized as follows: In Section 2, the framework of traffic flow time series clustering is proposed. The third part introduces the detailed steps, results and related analysis of the experiment. The fourth part summarizes the paper.
2. Clustering frame

2.1. Clustering algorithms

Common time series clustering algorithms can be divided into three categories: Hierarchical, Partition, Model-based. Combined with the characteristics of large road network and long-time traffic flow time series samples, three representative clustering algorithms are selected: GMM, K Means, BIRCH.

GMM assumes that the data points are linear combination of multiple Gaussian distribution functions. The introduction of standard deviation makes the shape of clustering no longer limited to circle. It is available to complex data structure clustering.

K Means calculates a new centroid after each iteration according to the distance from each data point to the centroid until the position of the centroid does not change much. The time complexity of the algorithm is linear and applicable to long-time samples.

BIRCH dynamically constructs a clustering feature tree by continuously inserting new data points. It minimizes I/O cost and is suitable for large samples.

2.2. Evaluation indicators of clustering effect

Because of the randomness of the clustering results. Referring to Luxburg U V [6], the instability of clustering can be defined as:

\[ \text{Instab} = \frac{2}{n(n-1)} \sum_{1 \leq i < j \leq n} d(C_i, C_j) \]  

Where \( C_i \) and \( C_j \) represent the clustering results at the \( t \) th and \( t' \) th time respectively, \( d(C_i, C_j) \) is the distance between them.

The Silhouette coefficient can measure the similarity intra cluster and separation degree between clusters which can be defined as:

\[ s = \frac{b-a}{\max(a,b)} \]

Where \( a \) is the mean distance between a sample and all other points in the same class, \( b \) is the mean distance between a sample and all other points in the next nearest cluster.

3. Experiment

3.1. Data source and preprocessing

The data comes from an open-source data shared by the Didi GAIA Initiative, which records the geographical coordinates in WKT form, ID and name of main roads in Shenzhen, traffic flow data at intervals of 10 minutes in 2018.

For the traffic time series data, denoising can distinguish the volatility pattern between traffic flow series more accurately, such as: highlight the morning and evening peak characteristics or stable trend of traffic flow sequence. There are many methods to transform raw data, including discrete Fourier transform, discrete wavelet transforms, and clipping. In this paper, we choose a semi soft threshold function which takes into account the advantages of soft threshold and hard threshold. The threshold formula is as follows:

\[ T = \left( \frac{\text{median}(|CD1|)}{0.6745} \right) \sqrt{2 \ln N} \]

Where \( CD1 \) is the detail coefficient of the first level decomposition, \( N \) is the data length.
3.2. The optimal number of clusters of the algorithm

The programming package used by all programs is Scikit-learn. For GMM, BIC coefficient is served to measure the similarity within the cluster. The lower the score, the higher the similarity within the cluster; For BIRCH, Calinski-Harabasz coefficient is employed to measure the inter group dispersion and intra group similarity. The larger the value is, the better the clustering effect is; For K_Means, Contour analysis is adopted to measure the degree of separation between clusters. The values of 1 and 0 respectively indicate the best and worst clustering results.

![Figure 1. Selection of optimal cluster number for GMM and BIRCH.](image1)

![Figure 2. Selection of optimal cluster number for K_Means.](image2)

Optimal clustering number of GMM is 5, which indicates that GMM is more capable of screening fine-grained traffic fluctuation characteristics. Considering that the optimal cluster number of K_Means is 2, BIRCH chooses the cluster number 3 with the second highest score.

Table 1 records the scores of algorithms with the best number of clusters calculated using the evaluation indicators in Section 2.3. Silhouette coefficient indicates that K_Means has the best clustering effect, and the Instab coefficient show that GMM has the optimum stability.
Table 1. Results of two evaluation indicators on three clustering algorithms.

| coefficient | GMM (5)  | K_Means (2) | BIRCH (3) |
|-------------|----------|-------------|-----------|
| Silhouette  | 0.077    | 0.120       | 0.041     |
| Instab      | 0.043    | 0.048       | 0.036     |

3.3. Clustering Visualization of traffic flow time series

The second cluster has the most violent swing range, with the most peaks and troughs. Cluster 1, 4, 5 have 1-2 large and several small wave crest. The peak of the first cluster is very steep, and the fifth is relatively flat. The results show that GMM can effectively identify the difference of wave patterns between sequences.

Figure 3. Typical traffic flow time series in GMM clustering results.

Figure 4. Typical traffic flow time series in K_Means clustering results.
There is no significant difference in the fluctuation pattern between clusters, which indicates that K_Means can’t identify the traffic flow movement information of traffic flow. After the visualization of the clustering results in section 3.2, it is found that the traffic flow time series data cluster is a long strip, which does not conform to the high-dimensional spherical distribution.

In the first cluster, there are the most peaks and valleys. The wave of the second cluster is smoother than that of the third cluster. The results show that BIRCH can also identify the volatility difference between sequences, but the effect is slightly worse than GMM.

After the analysis of the whole framework, we get the following conclusions: GMM is the best, BIRCH is the second, both of them can effectively capture the traffic flow movement, K_Means is the worst.

4. Conclusion
In conclusion, we propose a clustering framework based on multi-scale analysis of traffic flow time series to find potential traffic flow patterns. The clustering framework includes the selection of clustering algorithm, the quantitative index used to evaluate the clustering effect, and the visual inspection. Besides, we use real data to verify our clustering framework, which is the 24-hour long-time traffic flow time series are clustered on a large road network in Shenzhen, China, and find that our proposed framework can clearly distinguish different trends of traffic flow time series. On this basis, neural network can be used to establish a prediction model for traffic flow time series.

References
[1] Kerner, B.S., Klenov, S.L.: A microscopic model for phase transitions in traffic flow. J. Phys. A: Math. Gen. 35(3), L31–L43 (2002).
[2] Kerner, B.S., et al.: Cellular automata approach to three-phase traffic theory. J. Phys. A: Math. Gen. 35(47), 9971–10013 (2002)
[3] F, Xu., et al.: Traffic state evaluation based on macroscopic fundamental diagram of urban road network. Procedia Social Behav. Sci. 96 (1), 480 – 489 (2013).
[4] Di, B.: A multi-index fusion clustering strategy for traffic flow state identification. IEEE Access 7, 166404–166409 (2019).
[5] Constantinos, A., et al.: Dynamic data-driven local traffic state estimation and prediction. Transp. Res. Pt. C-Emerg. Technol. 34, 89 – 107 (2013).
[6] Luxburg U V. Clustering Stability: An Overview [J]. Foundations and Trends® in Machine Learning, 2010, 2 (3): 2010.