Research and Application of Intersection Clustering Algorithm Based on PCA Feature Extraction and K-Means

Wei Lv\textsuperscript{1}\textsuperscript{*}, Weiyu Tang\textsuperscript{2}, Hongliang Huang\textsuperscript{3}, Tao Chen\textsuperscript{1}

\textsuperscript{1} Alibaba Cloud Big Data Application College, Zhuhai College of Jilin University, Zhuhai 519041, China

\textsuperscript{2} School of Computing, Zhuhai College of Jilin University, Zhuhai 519041, China

\textsuperscript{3} School of Public Basic and Applied Statistics, Zhuhai College of Jilin University, Zhuhai 519041, China

\textsuperscript{*}Corresponding author’s e-mail: luwei@jluzh.edu.cn

Abstract. Traffic jam and traffic route design are important issues of people's livelihood. In this paper, machine learning algorithm is used to cluster intersection features. Based on the collected traffic intersection data, PCA dimension reduction technology is used for feature extraction, and then K-Means clustering algorithm is used to sum up the number of entrances, and each intersection is clustered. Thus, a new intersection clustering algorithm is established, studied and applied. The algorithm can match similar intersections, which provides theoretical support for the transportation department and related technical personnel.

1. Introduction

With the improvement of people's living standards, the continuous development of science and technology, more and more people have motor vehicles, at the same time, the phenomenon of traffic congestion is more and more serious. As the traffic hub of urban road, the problem of traffic jam is more serious. This not only increases the commuting time, causes the economic losses of drivers, affects the development of the national economy, but also wastes the fuel of cars and causes air pollution. Therefore, it is urgent to solve the problem of traffic jam at intersections.

In order to solve the problem of traffic jam, the principle is to reduce the number of vehicles per unit area of the road. In the detection of intersections, the commonly used data processing method of machine learning is used to extract features, reduce the dimension of feature space and eliminate the correlation between the original features. In addition, clustering combined with unsupervised machine learning algorithm can effectively summarize the situation of intersections.

Feature extraction starts from the initial set of measurement data, and constructs non redundant derived values (features) to provide information. It is a process of transforming the original data that cannot be recognized by machine learning algorithm into features that can be recognized by the algorithm. PCA is a very common data dimension reduction method. It uses orthogonal transformation to linearly transform the observed values of a series of possibly related variables, so as to project them into a series of linearly uncorrelated variables, which are called Principal Components. The basic idea of PCA is to calculate a group of new features from a group of features in order of importance. They are the linear combination of the original features, and there is no correlation between the new features. The mapping value of the original features on the new features is the new dimension reduced sample. In 1901, Karl Pearson invented principal component analysis and used it to analyze data and establish...
mathematical model[1], which is similar to the principal axis theorem in principle. According to different application fields, it is also called discrete Karhunen–Loève transform (KLT) in signal processing.

Cluster analysis[2][3], widely used in machine learning, data mining, pattern recognition, image analysis, biological information and other fields, is a statistical data analysis technology. Clustering is to divide similar objects into different groups or subsets by static classification, so that the member objects in the same subset have similar attributes, such as short space distance in the coordinate system. Common clustering methods include K-Means, Mean Shift, DBSCAN, GMM, Graph Community Detection, etc. K-Means clustering originated from a vector quantization method in signal processing, and now it is more popular in the field of data mining as a clustering analysis method. The purpose of K-Means clustering is to divide n points (which can be an observation or an instance of a sample) into K clusters, so that each point belongs to the cluster corresponding to the nearest mean value (cluster center), and take it as the clustering standard. In 1967, James MacQueen[4] first used the term "K-Means". He described a process of dividing N-dimensional population into K sets on the basis of samples. In 1957, Stuart Lloyd proposed a standard algorithm, which was published by Bell Laboratories in 1982[5]. In 1965, E.W.Forgy published essentially the same method, so K-Means algorithm is sometimes called Lloyd-Forgy method[6]. In 1975 and 1979, the more efficient version was proposed by Hartigan and Wong[7][8].

In this paper, in order to study and apply the intersection clustering algorithm, according to the collected traffic intersection data, the entrance direction, intersection features and entrance features, feature extraction function and transformation matrix set are discussed. Through machine learning algorithm clustering, PCA dimension reduction technology feature extraction, combined with K-Means clustering algorithm to cluster the number of intersections, a new intersection feature clustering algorithm is established. Hope to bring inspiration for the transportation department and related technical personnel, and can be used in the future traffic congestion problem.

2. Basic Concept and Definition of Intersection Clustering Algorithm

According to the research topic, we propose five definitions for the intersection similarity matching model proposed in this paper.

Definition 1: Import Direction

In this paper, the import direction \(\alpha\) is defined as the import direction, and its range is \(\alpha = 1,2,3,\ldots,8\). They correspond to eight directions: East, South, West, North, Northeast, Southeast, Northwest and Southwest.

Definition 2: Intersection Characteristics and Entrance Characteristics

Let \(n_\alpha = N(T_\alpha)\) be the characteristic number of an import \(\alpha\).

The characteristic of the import \(\alpha\) is an \(n_\alpha\)-dimensional vector \(T_\alpha = (t_1^{(\alpha)}, t_2^{(\alpha)}, \ldots, t_{n_\alpha}^{(\alpha)})\).

The feature of an intersection is defined as a high-dimensional vector composed of all the entrance features, \(T = (T_1, T_2, \ldots, T_8)\).

Because all imports are homogeneous and the characteristic number of each import is equal, denoted as \(N(T_1) = N(T_2) = \ldots = N(T_8) = n_0\), \(T\) is an \(8 \times n_0\) vector \(N(T) = 8 \times N(T)\) (1)

Definition 3: Feature Extraction Function

The process of extracting intersection features from intersection channelization data is as follows:

\[ f(\mathcal{X}) \rightarrow \mathcal{T} \]

Where \(f\) is a mapping from intersection channelized data \(\mathcal{X}\) to \(N(\mathcal{T})\) dimensional space. Generally, \(\mathcal{X}\) is a higher dimensional space than \(\mathcal{T}\). If \(f\) is closer to bijection, the reduction degree of \(f\) is higher. But the higher the reduction degree is, the more likely it is to cause dimension explosion. In practice, there will be a balance.

Definition 4: Transformation Matrix Set
Consider the following case: after all channelization elements of one intersection rotate a certain angle, they will coincide with the channelization characteristics of another intersection. When comparing the similarity of two intersections, the rotation angle of intersection must be considered. The expression of rotation on eigenvector is eigenvector multiplied by a transposition matrix.

There are eight cases of circular arrangement from 1 to 8:

\((1,2,\ldots,8), (8,1,\ldots,7)\ldots(2,3,\ldots,1)\)

Consider a permutation cyclic group in an \(8n_0 \times 8n_0\) matrix:

\[
D = \begin{pmatrix}
0 & I_a & 0 & \cdots & 0 \\
0 & 0 & I_a & \cdots & 0 \\
0 & 0 & 0 & \cdots & I_a \\
\vdots & \vdots & \vdots & \ddots & \vdots \\
I_a & 0 & 0 & \cdots & 0
\end{pmatrix}
\]

\(3\)

Where \(n = 1,2,\ldots,8; I_a\) is an identity matrix of \(n_0 \times n_0\).

Note that when \(n = 1\), the value of the above matrix is \(D_1\), and there are only eight elements in permutation Cyclic Group \(D\):

\[
D = \{D_1, D_2, \ldots, D_8\}
\]

\(4\)

Definition 5: Intersection Similarity

An intersection is regarded as a vector in \(N(T)\) dimensional Euclidean space, and the cosine of the angle between the two vectors is used as the similarity index. However, due to the need to consider the problem of intersection rotation, the similarity of two intersections \(i, j\) is as follows:

\[
S(i,j) = \max_{g \in D} \frac{\langle T_i^g, T_j^g \rangle}{\|T_i^g\|\|T_j^g\|} \times 100\%
\]

\(5\)

It is easy to know that the similarity relation is symmetric, that is \(S(i,j) = S(j,i)\).

3. Intersection Clustering Algorithm

3.1. Algorithm flow chart

In this paper, based on PCA dimension reduction technology and K-Means clustering algorithm, we get intersection clustering algorithm, and carry out experiments and research. The flow chart is shown in Figure 1.

![Figure 1. Research flow chart.](image)
3.2. Data preprocessing

After preprocessing the collected traffic intersection data, some data tables are listed as follows: Mold type table (see Table 1), Feature table (see Table 2), and Mold feature weight corresponding table (see Table 3 and Table 4).

| Type | Type Descr | fac |
|------|------------|-----|
| 1    | Pedestrian crossing | lpw |
| 2    | Full screen light    | 1   |
| ...  | ...         |     |
| 77   | Straight ahead non motorized lamp | z |

Table 2. Feature table

| fIndex | feature | fDescr |
|--------|---------|--------|
| 1      | p_count | Number of people crossing the street at one time |
| 2      | l1p_count | Number of left crossing |
| ...    | ...     | ...    |
| 78     | z_count | Number of molds for non motor vehicles |

Table 3. Mold feature weight corresponding table

\[
\begin{array}{|c|c|c|}
\hline
\text{Type} & \text{fIndex} & \text{weight} \\
\hline
2 & 7 & 0.3 \\
3 & 8 & 0.3 \\
... & ... & ... \\
75 & 78 & 0.1 \\
\hline
\end{array}
\]

Table 4. Mold feature weight corresponding table

\[
\begin{array}{|c|c|c|c|}
\hline
\text{Type} & \text{guide} & \text{fIndex} & \text{weight} \\
\hline
1 & 16 & 1 & 1 \\
1 & 32 & 1-2 & 0.5-0.5 \\
1 & 64 & 3 & 1-3 \\
1 & 128 & 4 & 1 \\
\hline
\end{array}
\]

In addition, the data used in this paper also are aggregation index description and the field description of moldinfo in the ledger system. For Type, see Table 3 and Table 4.

Based on the above data, this paper verifies the intersection clustering algorithm. We preprocess the dataset according to the following rules.

1. The angle of some dies is reversed during initialization, so the exit needs to be adjusted;
2. No straight left, no left, no straight right, no right, no left, no right;
3. The mold angle is scaled: \( \text{Angle} = (\text{Angle} + 180) \mod 360 \);
4. For all molds: \( \text{Angle} = \text{Angle} \times 0.005 \).

3.3. Data cleaning

In Table 1, all molds are divided into several types, see fac field for details.

1. \( lpw \): Lane pedestrian crossing waiting area (excluding right turn channelized crossing, it needs to be judged by combining the guide field).
2. \( rp \): turn right and cross the street channelized (it needs to be judged according to the guide field).
3. \( e \): export.
4. \( f \): prohibited sign.
5. \( l \): signal light.
6. \( h \): facilities divided into areas, such as double spring line, guardrail, etc.
7. \( i \): channelized Islands.
8. \( z \): non motor vehicle mold.
9. \( d \): coil.
Generally speaking, except \( lpw \) and part of \( l \) that have binding relationship with \( lpw \) have definite approach value, others are not or are not trusted. This step is to reassign values to the \( Approach \) fields of all molds to facilitate feature extraction by import.

3.4. K-Means clustering algorithm induces the number of imports

K-Means algorithm is one of the most commonly used algorithms in clustering analysis, and it is also an important distance clustering algorithm in machine learning. In this paper, in order to get the best number of imports, we use K-Means clustering algorithm for induction.

We assume that all \( lpw \) molds belonging to the same import will be clustered in \textit{visio}, and vice versa. Assume that the number of different entries in \( lpw \) and partially bound \( l \) is K-Means clustering operation with \( 2 \sim n \) categories for \( X \), \( Y \) and \textit{Angle} fields of \( lpw \). Assuming that the number of categories in \( 2 \sim n \) makes the contour coefficient of clustering result maximum, it is the final number of entrances of the intersection. The voting method is used from K-Means output to new import:

\[
C \rightarrow \text{NewApproach} = \arg \max_{\alpha} \sum_{i=1}^{\text{lpw}} I(i.\text{Approach} = \alpha)
\]  

(6)

Where \( C \) is the class tag value.

This operation is helpful to correct the error caused by the deviation of drawing angle in some molds.

The objective function of K-Means algorithm is as follows:

\[
\min \sum_{i=1}^{K} \sum_{x \in X} \|x - c_i\|^2
\]  

(7)

For K-Means algorithm, the determination of \( K \) value is the most important part for the clustering algorithm. According to the industry experience, the number of clusters determined is too many, and it is not necessarily the real number of clusters we get. Therefore, we hope to determine the real number of clusters from the data itself, that is, the best number of clusters for the data. According to the objective function, we calculate the cost function of data generation when the number of clusters is \( 1 \sim 9 \), as shown in Figure 2. It can be seen that when the number of clusters reaches 3, the value of cost function tends to be stable, so the traffic intersection data collected in this paper is clustered into 3 clusters.

![Figure 2. Parameter selection.](image)

PCA algorithm is used to reduce the multidimensional traffic data to two dimensions. At the same time, the two-dimensional data are clustered into three categories by K-Means algorithm. In order to verify the rationality of the algorithm used in this paper, the clustering results are compared with those of DBSCAN density clustering algorithm, as shown in Figure 3. It can be found that K-Means has a better effect on traffic data processing, and can extract its features more accurately.
3.5. Feature extraction and weighting

There are 77 kinds of molds, and 78 features are extracted from each import (see Table 2 for details). The corresponding relationship and corresponding weight of molds and features in the mold table are shown in Table 3 and Table 4 for details.

4. Conclusion

In this paper, based on the collected intersection data of urban traffic network, combined with machine learning clustering model, a new intersection clustering algorithm is established. PCA was used to reduce the dimension of multi-dimensional data into two dimensions and extract data features. Then according to the cost function calculation, super parameter optimization and other methods to select parameters, get the $K$ value. The K-Means algorithm was used to cluster the intersections with high urban similarity, and the intersections with similar data features were labeled with the same label, and the number of entrances was summarized. Finally, the method of feature extraction and weighting was combined. Thus, the intersection clustering algorithm based on PCA feature extraction and K-Means was established. Through the clustering algorithm established in this paper, the features of different intersections can be extracted accurately. It provides theoretical guidance for solving the problems of traffic congestion control and traffic design.

Acknowledgments

This paper was supported by the project “Excellent School with Innovation” scientific research project of Guangdong Provincial Department of Education: Key Technology Research and Development of Traffic Operation Big Data Analysis, 2018KTSCX309, and the 2018 Innovation Cultivation Project of Zhuhai College of Jilin University: 2018XJCQ011.

References

[1] Pearson K. LIII. On lines and planes of closest fit to systems of points in space[J]. The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science, 1901, 2(11): 559-572.
[2] Abdi H. Additive tree representations (with an application to face processing)[J]. Lecture Notes in Biomathematics.-1995.-84.-PP, 1995: 43-59.
[3] Clatworthy J, Buick D, Hankins M, et al. The use and reporting of cluster analysis in health psychology: A review[J]. British journal of health psychology, 2005, 10(3): 329-358.
[4] MacQueen J. Some methods for classification and analysis of multivariate observations[C]//Proceedings of the fifth Berkeley symposium on mathematical statistics and probability. 1967, 1(14): 281-297.
[5] Lloyd S. Least squares quantization in PCM[J]. IEEE transactions on information theory, 1982, 28(2): 129-137.
[6] Forgy E W. Cluster analysis of multivariate data: efficiency versus interpretability of classifications[J]. biometrics, 1965, 21: 768-769.
[7] Hartigan J A. Clustering algorithms[M]. John Wiley & Sons, Inc., 1975.
[8] Hartigan J A, Wong M A. Algorithm AS 136: A k-means clustering algorithm[J]. Journal of the royal statistical society. series c (applied statistics), 1979, 28(1): 100-108.