Title:
Application of Cluster Analysis with Unsupervised Learning to Dockless Shared Bicycle Flow Control and Dispatching

Authors:
Shang-yuan Chen, shangyuanc@gmail.com, Feng chia University of Taichung
Tzu-tien Chen, catcherchen@gmail.com, The University of Hong Kong

Keywords:
Shared bicycles, R language, K-means clustering, kernel density spatial clustering, density-based spatial clustering of applications with noise

DOI: 10.14733/cadconfP.2019.122-126

Introduction:
The promotion of shared bicycles seeks to resolve the problem of short-distance urban transportation, and shared bicycles are an important means of contemporary green transportation [2]. However, after dockless shared bicycles are introduced on a large scale in the city, supply and demand problems involving different areas often emerge. This study applied cluster analysis with unsupervised learning to shared bicycle flow control and dispatching, and employed artificial intelligence to extract real, full-scale transportation rules from open data. First, this study proposes a model of shared bicycle control system and an incentive mechanism for reverse flow of bicycles based on threshold values. Then, we use the kernel density spatial clustering method to perform partitioning, grading, and incentives of check-out and check-in points' density in the area, furthermore, adopt the DBSCAN clustering method to establish dispersal and dispatching strategies. This study use Shanghai Open Data in modeling, verification, and used Rstudio software to produce visualized interactive graphics for demonstration.

Literature Review:
The scope of this study includes dockless shared bicycles and cluster analysis; the following is a review of the literature:

- *Dockless Shared Bicycles and Clustering Principles*

Unlike the case of shared bicycles with fixed docks, the clustering of dockless shared bicycles changes with time, and it is difficult to calculate check-out and check-in probabilities [6]. One feasible approach that can be employed to cluster dockless shared bicycles is to perform clustering of the coordinates of the bicycles' check-out and check-in locations. Moreover, the clustering of dockless shared bicycles must also reflect the precondition that users within each cluster must be able to easily find usable shared bicycles within a comfortable walking distance. In the case of Shanghai community residents, taking a walking distance of 787m as a dividing point, it will very seldom choose to walk when the distance is greater than 787m. This study therefore adopted the round figure of 700m as the scope of the most suitable walking distance; 700m serves as the baseline scope of clusters in the remainder of this paper, and since 700m is roughly equivalent to the distance that can be walked in 10 minutes, it is representative of local everyday walking.

- *Cluster Analysis*

Cluster analysis is defined as the division of heterogeneous objects into homogeneous subgroups [7]. Data clustering is ordinarily classified as unsupervised learning. Gan et al created the parsing tree based on the characteristics of clustering algorithms. From the top, this tree lists hard clustering and fuzzy clustering, where hard clustering includes partitional and hierarchical approaches, and hierarchical approaches may consist of either divisive or agglomerative methods [4].

Theory and Method:

As mentioned above, this study proposes a "shared bicycle control system," and applies cluster analysis with unsupervised learning to bicycle flow control and dispatching. This study also compares and analyzes two types of algorithms commonly used in hard clustering and partitioning: (1) kernel density spatial clustering, and (2) the density-based spatial clustering of applications with noise (DBSCAN).

- *Shared Bicycle Control System*

It is extremely important that shared bicycle companies possess real-time or ahead of time dispatching and management ability. This study proposes a shared bicycle control system model and an incentive mechanism for reverse flow of bicycles based on threshold values. As shown in Fig. 1, this system considers the number of check-outs within a unit time and a unit area to be the bicycle demand of that area, and the number of check-ins within a unit time and a unit area to be the bicycle supply of that area. In order to ensure that each area has a suitable amount of bicycle stock, incentives or compulsory rules must be used to control the rate of check-outs and check-ins. When stock is excessively high—higher than the upper threshold—the system will trigger the check-out incentive mechanism. When the stock is too low—below the lower threshold—the system will activate the check-in incentive mechanism. Apart from this, when the system detects that the bicycle density has reached the level of oversupply or shortage, the shared bicycle company will have to dispatch trucks to perform transport and force the system back within the normal range.

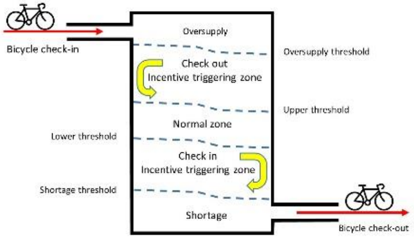

Fig. 1: Shared bicycle control system model.

- *The Two Types of Cluster Analysis Method*

[1] Kernel density spatial clustering

Density clustering methods are density-based, and this method involves clustering on the basis of density levels, where areas with a high density of objects are surrounded by relatively low-density areas [4]. This section uses the density clustering method known as the "kernel density estimation" method; this method involves use of the known locations of discrete events to estimate the overall density of a two-dimensional research domain. This estimated density is used to cover the top of the research domain with a grid cell, and estimation of density is performed based on each grid cell's center point. Weighting of each distance between a discrete event and grid cell center point is then performed employing a kernel function and bandwidth [9] (Fig. 2).

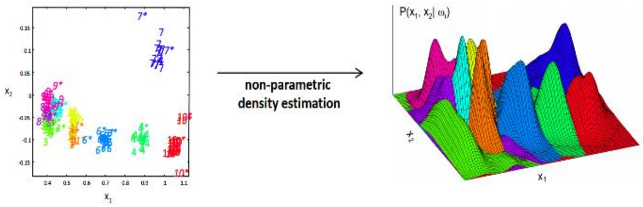

Fig. 2: Weighting based on a specific kernel function and bandwidth.

[2] Density-based spatial clustering of applications with noise (DBSCAN)

The DBSCAN method is also based on density: When a set of points is placed in a given space, this algorithm can assign nearby points to individual groups, and mark points located in low density areas as outside points, which constitute noise points, and are not connected with any high-density areas.

In Fig. 3, point n is a noise point. The goal of DBSCAN is to find the largest set of density-connected objects; in Fig. 4, the areas around data points p, q, and t are mutually connected [3]. When control of the scanning radius and minimum number of points minimize noise in the ultimate clustering results, this is generally felt to be an optimal DBSCAN result.

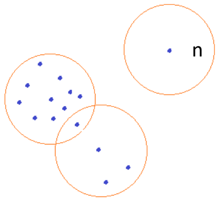

Fig. 3: Point n is a noise point.

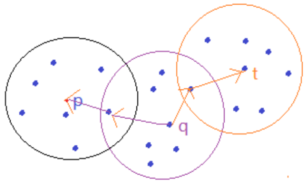

Fig. 4: Finding the largest set of density-connected objects.

Comparing the kernel density spatial clustering and DBSCAN methods, the former offers the advantage of easy density estimation, and can be used to perform partitioning and grading. However, the latter can only be used to mark those clusters that meet a preset density requirement. However, due to the need to set a MinPt parameter, DBSCAN can effectively avoid the single-link effect, which occurs when different clusters are connected by a single point or relatively few lines, which causes them to be seen as a single cluster. Because of this, the control radius ε and MinPts quantity can be used to create a control threshold. This study will use these two types of density-based clustering methods in subsequent applied research. This study used the R programming language as a data processing and analysis tool [5].

Realization and Verification:

This section realizes and tests a shared bicycle control system in accordance with the foregoing theory and method. We first perform scenario modeling of the incentive mechanisms controlling reverse movement based on threshold values in order to gain an understanding of problems that might be faced in actual operation. Afterwards, we employ open data items issued in August 2016 for the Shanghai Innovative Application Competition to perform modeling and verification.

- *Scenario Modeling Incentive Mechanisms Controlling Reverse Movement Based on Threshold Values*

Within any one cluster, the distance users needed to go to find shared bicycles did not exceed the most suitable walking distance, which implied that the minimum length of the diagonal lines across grid squares must be 700m, and the sides of the squares was 550m. This section expresses the unit density of bicycles within clusters in terms of numeric values (Fig. 5). We assumed a state of balance and normal stock at the initial time t_0 (00:00 on 8/1/2016). The bicycle density within a unit area was 6 units in all cases (left side, Fig. 6). After one hour, (01:00, designated t_1) we calculated changes in unit density of the bicycles stock (right side, Fig. 6). As a consequence, we were able to calculate the bicycle check-out rate (left side, Fig. 7) and check-in rate (right side, Fig. 7) of each grid square at a time t_0. (The areas with underlined numbers in these figures have a check-out rate ≥ check-in rate, which implies that $X_{t_{n-1}} \geq Y_{t_{n-1}}$ when n=1)

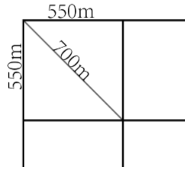

Fig. 5: Schematic diagram of scenario modeling of clustering.

6	6	6
6	6	6
6	6	6

6	***5***	6
5	10	***5***
6	***5***	6

Fig. 6: Changes in unit bicycle density in clusters: Initial state (left) and state at time t_1 (right).

1	*2*	2
3	2	*2*
2	*2*	4

Check-outs

1	*1*	2
2	6	*1*
2	*1*	4

Check-ins

Fig. 7: Check-out rate (left) and check-in rate (right) of bicycles in each cluster at period of t_0.

In accordance with the foregoing observations, the most important aspect is maintaining a balance between supply and demand within each cluster in this dynamic system, and incentives must seek to reverse imbalances in supply and demand during the previous period of time. This study consequently designed check-out incentive formula (1) and check-incentive formula (2) (in which X is the check-out rate, Y is the check-in rate, *t* is the time, and n is an integer number):

[1] If $X_{t_{n-1}} < Y_{t_{n-1}}$ at t_{n-1}, then

$$\frac{|Y_{t_{n-1}} - X_{t_{n-1}}|}{2^i} \text{ is the check-out incentive amount} \qquad (1)$$

Where *i* is an integer number, and is the check-out sequence at time t_n.

In accordance with formula (1), at time t_n, the incentive for check-out of the first bicycle is largest, and the incentive for the check-out of each subsequent bicycle will decrease progressively with the increasing root of $\frac{1}{2}$ with the increase in the number of bicycles checked out: $\frac{1}{2^1}, \frac{1}{2^2}, \frac{1}{2^3}$,,,, and so on.

[2] If $X_{t_{n-1}} > Y_{t_{n-1}}$ at time t_{n-1}, then

$$\frac{|X_{t_{n-1}} - Y_{t_{n-1}}|}{2^j} \text{ is the check-in incentive amount} \qquad (2)$$

Where *j* is an integer number, and is the check-in sequence at time t_n.

In accordance with formula (2), at time t_n, the incentive for check-in of the first bicycle is largest, and the incentive for the check-in of each subsequent bicycle will decrease progressively with the increasing root of $\frac{1}{2}$: $\frac{1}{2^1}, \frac{1}{2^2}, \frac{1}{2^3}$,,, and so on.

As described above, if a bicycle is checked out from area p at time t_{out} and checked in area q at time t_{in}, the incentive amount will be the sum of the check-in incentive in area p during at time t_{out} and the check-out incentive in area q during at time t_{in}. This can be represented as:

$$Sum = \frac{|Y_{t_{out(n-1)}} - X_{t_{out(n-1)}}|}{2^i} + \frac{|X_{t_{in(n-1)}} - Y_{t_{in(n-1)}}|}{2^j}, \text{ which is the total incentive amount} \qquad (3)$$

- *Use of Shanghai Open Data in Modeling and Verification*

This study used 101,843 data items issued in August 2016 for the Shanghai Open Data Innovative Application Competition [8] to perform modeling. We used this data to find optimal clustering methods for shared bicycles in Shanghai by examining relatively small areas (the area around Shanghai's Jingan Temple) and relatively large areas (the part of Shanghai within the fourth ring roads—121.75˚E-121.125˚E, 30.95˚N-31.45˚N), and also established models corresponding to the state of clustering. We employed the following four steps in this process:

[1] Compilation and processing of the needed data

As of August 2016, it was projected that 223,000 Mobike bicycles were in use in Shanghai (Soda Open data, 2018). Since each Mobike bicycle in China was ridden by 5.4 persons per day [1].

[2] Data visualization analysis by using the RStudio software.

[3] Application of kernel density spatial clustering, partitioning, grading, and incentives

Applying the kernel density spatial clustering method to perform partitioning and grading of check-out and check-in density in the area around Shanghai's Jingan Temple on 8/1/2016. (Fig. 8)

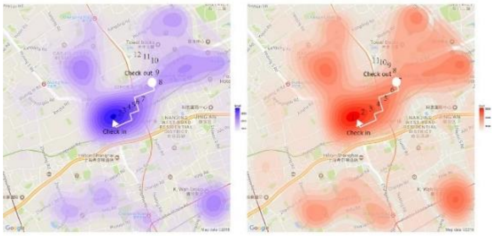

Fig. 8: Partitioning of check-out density (left), check-in density (right).

[4] Establishment of a dispersal and dispatching strategy

The DBSCAN clustering method can be used in conjunction with control radius ε and MinPts values to find ε - adjacent areas and the largest sets of density-connected objects, while avoiding noise. When the radius ε was entered as 700m and the MinPts value was set at 36,640 bicycles, we could find clusters with an oversupply of bicycles. In Fig. 9 By the same principle, if a shortfall of 800 bicycles in each 550m*550m grid square indicates a shortage, 4,071 bicycles in an area with a radius of 700m can be considered the shortage threshold. Employing the DBSCAN method, the radius ε was entered as 700m and the MinPts value was set as 4,071 bicycles. As shown in Fig. 10, after subtracting the brown cluster among light green noise points, those green noise points were seen to have a bicycle shortage.

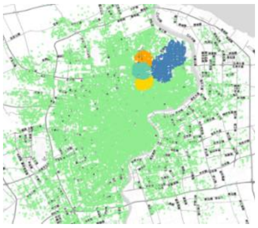

Fig. 9: Analysis of areas with a bicycle density oversupply.

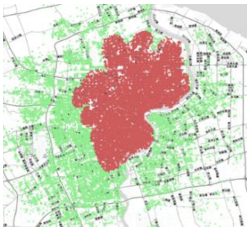

Fig. 10: Analysis of areas with bicycle shortages.

Conclusions:

This study established a prototype control system for dockless shared bicycles, with an incentive mechanism promoting the reverse flow of bicycles, encouraging bicycle users to autonomously maintain a balance between bicycle supply and demand. In addition, this study also applied the kernel density spatial cluster method to partitioning and grading, and developed an incentive system; and also applied the DBSCAN clustering method to finding clusters with an oversupply of bicycles and noise indicating bicycle shortages, which allowed the development of a concrete dispersal and dispatching strategy.

References:

[1] Beijing Planning Design Research Institute, 2017 Sharing Bike and Urban Development White Paper, 03-05, http://www.199it.com/archives/581592.html

[2] Chang, S. K.; Chang, H. W.; Chen, Y. W.: Green Transportation, Slow, Friendly, and Sustainable: People-Oriented Transportation Environment Makes The City Smoother and Life Better, Neonaturalism, 2013, 80-81.

[3] Ester, M.; Kriegel, H.-P.; Sander, J.; Xu, X.: A density-based algorithm for discovering clusters in large spatial databases with noise, International Conference on Knowledge Discovery and Data Mining, 1996, 226-231.

[4] Gan,G.; Ma, C.; Wu, J.: Data Clustering: Theory, Algorithms, and Applications, Chapman & Hall/CRC, 2007, 10. https://doi.org/10.1137/1.9780898718348.fm

[5] Lander, J.P.: R for Everyone: Advanced Analytics and Graphics, Addison-Wesley Professional; 1 edition, 2013.

[6] Li, Y.X.; Zheng, Y.; Zhang, H. C.; Chen, L.: Traffic prediction in a bike-sharing system, 23rd SIGSPATIAL International Conference on Advances in Geographic Information Systems Seattle, 33, 2015, 78-88.

[7] Negnevitsky, M.: Artificial Intelligence: A Guide to Intelligent Systems, Pearson Education Limited, 2011, 303.

[8] Soda Open data, 2018/05, http://shanghai.sodachallenges.com/data.html

[9] Timothy, H.; Zandbergen P.: Kernel density estimation and hotspot mapping: Examining the influence of interpolation method, grid cell size, and bandwidth on crime forecasting, Policing: An International Journal of Police Strategies & Management, 37(2), 2014, 305-323. https://doi.org/10.1108/PIJPSM-04-2013-0039