Research on a Distribution-Outlier Detection Algorithm Based on Logistics Distribution Data

Sheng Lu¹,², Tiantian He¹, Qinkai Zhou³, Jiefeng Wen⁴, Yang Liu¹ and Meixia Zhang⁵,*

¹ School of Advanced Manufacturing Engineering, Chongqing University of Posts and Telecommunications, Chongqing 400065, China
² State Key Laboratory for Strength and Vibration of Mechanical Structures, Xi’an Jiaotong University, Xi’an 710049, China
³ School of Information and Electronic, Beijing Institute of Technology Beijing 100081, China
⁴ Shenzhen Pohoo Credit Financial Services Ltd, Chengdu 610041, China.
⁵ College of Landscape Architecture and Life Science, Chongqing University of Arts and Sciences, Yongchuan Chongqing 402160, China.

*Corresponding author email: 17197764@qq.com

Abstract. In this paper, we propose an outlier detection algorithm, which improves the existing algorithm from three aspects: the time outlier, the outlier detection of logistics distribution path, the time complexity. Aiming at the time outlier, this paper proposes an abnormal objective function based on the average speed of the logistics distribution process and realizes the detection of the outlier time of logistics distribution. In the aspect of the outlier detection of logistics distribution path, this paper combines the distance- and the angle-based outlier detection algorithms to realize the detection of the local abnormal situation of the logistics distribution trajectory. At the same time, the high time complexity of the algorithm is reduced by only retrieving the path sub-sections in the neighborhood of the proper radius. Experimental results indicate that the efficiency and speed of the proposed algorithm are better than those of the trajectory outlier detection (TRAOD) algorithm, and the proposed algorithm is suitable for outlier detection of logistics distribution.

1. Introduction

As Neghabadia et al. (2018) described, the impact of logistics on people's life is becoming more and more important, so more and more scholars pay attention to it. For example, Loos et al. (2016) construct a mapping of existing academic publications regarding ergonomics within the field of logistics. At present, the outlier detection technology of logistics distribution is still in its infancy, but the research on outlier detection algorithms locally and abroad has played a great role in the outlier detection of the logistics distribution trajectory. Hawkins (1980) first described the concept of outliers: data sets that are significantly different or completely inconsistent with other data. Li et al. (2007) improved the basic comparison unit of classification and proposed a Rule- and Motif-based Anomaly Detection in Moving Objects (ROAM) detection framework based on the Motion-Classifier method; Maciá-Pérez et al. (2015) proposed an extended mathematical framework by analyzing the time complexity of the rough set theory and proposed an outlier detection method based on rough sets by using that framework. In China, on the basis of the unit space division, Liu L et al. (2009) proposed an
improved R-tree-based trajectory outlier detection (R-TRAOD) algorithm because of the poor execution efficiency of the TRAOD algorithm; Dong et al. (2013) proposed a trajectory outlier detection algorithm based on an extended probability suffix tree (EPST); Zhu et al. (2017) combined the distance- and density-based clustering methods to achieve the similarity computation of trajectory segments.

In this paper, the outlier target function is established by the average speed of the logistics distribution process to achieve the outlier detection of the distribution time. At the same time, the detection of local abnormal conditions of the distribution trajectory is realized by combining the distance- and the angle-based outlier detection algorithms. In addition, to reduce the time complexity, the algorithm only retrieves the sub-segments in the neighborhood within an appropriate radius.

2. Overall Process of the Algorithm

With the assumption that there is a distribution sub-segment $s$, the neighborhood sub-segment set of this distribution sub-segment is $N(s) = \{seg_1, seg_2, seg_3, ..., seg_n\}$, so the equations used in delay detection of the delivery time are as follows:

\[
\bar{V} = \frac{1}{n} \sum_{i=1}^{n} V_{seg_i} \quad (1)
\]

\[
\sigma_v = \sqrt{\frac{1}{n} \sum_{i=1}^{n} (V_{seg_i} - \bar{V})^2} \quad (2)
\]

\[
\bar{V}_t \leq \bar{V} - 2\sigma_v \quad (3)
\]

As illustrated in Figure 1, there are two sub-segments, $L_i = s_i e_i$ and $L_j = s_j e_j$, of which $L_i$ is the longer one; $p_s$ and $p_e$ are the projections of $s_j$ and $e_j$ on $L_i$, respectively, and $s_i, s_j$ represent the start of the distribution of goods $i, j$ respectively, $e_i$ and $e_j$ represent the end of the distribution of goods $i, j$ respectively.

**Figure 1.** A schematic view of a line segment’s Hausdorff distance.
Figure 2. Overall process of the algorithm

Definition 1: The parallel distance of track sub-segments \( L_i \) and \( L_j \) is defined as:

\[
d_{jj}(L_i, L_j) = \min(l_{jj1}, l_{jj2})
\]

Where \( l_{jj1} = \min(\text{dist}(p_s, s_j), \text{dist}(p_e, e_j)) \) and \( l_{jj2} = \min(\text{dist}(p_s, s_j), \text{dist}(p_e, e_j)) \).

Definition 2: Set \( l_{11} \) and \( l_{12} \) as Euclidean distances between \( s_j \) and \( p_s \) between \( e_j \) and \( p_e \), respectively, i.e. \( l_{11} = \text{dist}(s_j, p_s) \), and \( l_{12} = \text{dist}(e_j, p_e) \). The vertical distance between the track sub-segments \( L_i \) and \( L_j \) is

\[
d_{1}(L_i, L_j) = \frac{l_{11}^2 + l_{12}^2}{l_{11} + l_{12}}
\]

Definition 3: The angle distance between the track sub-segments \( L_i \) and \( L_j \) is as follows:

\[
d_{\theta}(L_i, L_j) = \begin{cases} \|L_j\| \times \sin \theta & 0^\circ \leq \theta < 90^\circ \\ \|L_j\| & 90^\circ \leq \theta < 180^\circ \end{cases}
\]

Where \( \theta \) is the angle between \( L_i \) and \( L_j \) and \( \|L_j\| = \min(\|L_i\|, \|L_j\|) \).

Definition 4: The distance function \( \text{dist}(L_i, L_j) \) between track sub-segments \( L_i \) and \( L_j \) consists of three component groups:

\[
\text{dist}(L_i, L_j) = \omega_{jj} \cdot d_{jj}(L_i, L_j) + \omega_{1} \cdot d_{1}(L_i, L_j) + \omega_{\theta} \cdot d_{\theta}(L_i, L_j)
\]

where \( \omega_{jj} \), \( \omega_{1} \) and \( \omega_{\theta} \) are determined by the application background.

The method based on angular distribution. According to the angle-outlier detection algorithm proposed by Karypis et al. (2008), the set consisted by the terminal point of the trajectory sub-segment and its neighbourhood sub-segment is assumed to be \( S(\{S\} = n) \), and the terminal point of the tested trajectory segment is assumed to be \( p \). By randomly selecting a pair of sample points \( a, b \in S\{p\} \), if \( \theta_{apb} \) represents the angle between vector \( \bar{pa} \) and \( \bar{pb} \), then the angle-based outlier factor \( VOA(p) \) is the variance of all \( \theta_{apb} \). The computation formula is as follows:

\[
VOA(p) = \frac{1}{n} \sum_{a,b \in S\{p\}}(\theta_{apb} - \bar{\theta}_p)^2
\]

Where
According to (8) and (9), the calculation of the angle-based outlier factor $VOA(p)$ does not require any parameters.

The algorithm’s steps are as follows:

Step 1: Getting the distribution sub-segment $s$ and indexing the neighbourhood sub-segment set $N(s)$;

Step 2: Using formulas (1)-(3) to detect whether segment $s$ has a time outlier. If formula (3) is true, segment $s$ has a time-delayed distribution, then the set $Time\_O$ is updated, and step 1 is repeated;

Step 3: Using formulas (4)-(7) to calculate the segment’s Hausdorff distance between sub-segment $s$ and every sub-segment in the neighbourhood sub-segment set $N(s)$. The DB $(p, d)$ criterion of outlier detection algorithm based on distance is used to analyse whether sub-segment $s$ is the process’s abnormal segment. If sub-segment $s$ is the process’s abnormal segment, the set $Pro\_O$ is updating, and step 1 is repeated;

Step 4: Using formulas (8) and (9) to calculate the value of $VOA(p)$ of the endpoint of sub-segment $s$. If $VOA(p)$ is less than the threshold angle VOA, the sub-segment $s$ had a process outlier, then set $Pro\_O$ is updated, and step 1 is repeated;

Step 5: Determining if the goods have been distributed to the actual terminal city, according to the attributes of the city (number, level and the location coordination) used to analyse whether the goods have been distributed to the right destination. If the goods have not been distributed to the right destination, set $Ter\_O$ is updated.

Step 6: Returning to step 1 until all the sub-segments have been detected.

3. Experimental Results and Analysis

The operating principles are defined as follows: Orders are distributed from a starting point, and during the distribution process, they are distributed in accordance with established path in a high probability but not in accordance with an established path or stay in place in a low probability, until a complete distribution path is formed. The interface of the distribution outlier detection module is illustrated in Figure 3.

![Figure 3](image-url)

Figure 3. The interface of the distribution-outlier detection module.

In Figure 3, “☆”, “○” and “+” represent the different track points, the number represents the number of the track points, and the cyan dotted line represents the relationship between the logistics centers. The detection results are presented in the table in the upper-right corner. The related operation buttons are presented in the table in the lower-right corner.

In Figure 4, (a), (b) and (c) are the results of experiments on 100 distribution paths with different start points and destinations, and the paths include 500 sub-segments. The green lines are the original paths, the blue lines (1#) are the distribution’s time-delayed sub-segments, the red lines (2#) are the abnormal distribution path sub-segments, and the black lines (3#) are the wrong destination sub-segments during the distribution.
Figure 4. The detection results of the experiments on different starting points and destinations
In Figure 4 (a), this paper’s algorithm finds six time-delayed distribution sub-segments, three error-path distribution sub-segments and four destination-error distribution sub-segments. In Figure 4 (b), this paper’s algorithm finds six time-delayed distribution sub-segments, four error-path distribution sub-segments and one destination-error distribution sub-segment. Figure 4 (c) indicates that this paper’s algorithm finds one time-delayed distribution sub-segment, three error-path distribution sub-segments and eight destination-error distribution sub-segments. In other words, Figure 4 indicates that the algorithm we propose in this paper is effective in the detection of the express logistics distribution.

4. Conclusion
Aiming at the problem that existing outlier detection methods cannot detect the time delay in a logistics distribution, we established a time-delayed objective function based on the average speed of the distribution process to achieve the time-delayed detection. In the distribution-path outlier detection, aiming at the existing outlier detection methods cannot detect the local outlier in distribution trajectories, the idea of combining distance- and angle-based outlier detection algorithm is proposed,
and the anomaly detection of local distribution trajectories is achieved. In addition, the time complexity of the algorithm is reduced by only indexing the similarity of each path sub-section in the neighbourhood of the proper radius. Experimental results indicate that the algorithm in this paper can better detect the abnormal situation of the express distribution process.

Acknowledgments
This work is supported by the Science and Technology Research Program of Chongqing Municipal Education Commission (Grant No. KJQN201900632, KJZD-K201900604), and the Open Projects of State Key Laboratory for Strength and Vibration of Mechanical Structures (Grant No. SV2018-KF-29).

References
[1] Neghabadia, P. D., K. E. Samuelb, and M. L. Espinousea. 2018. “Systematic Literature Review on City Logistics: Overview, Classification and Analysis.” International Journal of Production Research.
[2] Loos, Mauricio Johnny, E. Merino, and C. M. T. Rodriguez. "Mapping the state of the art of ergonomics within logistics." Scientometrics 109.1(2016):85-101.
[3] Enderlein, G. (2010). “Hawkins, D. M.: Identification of Outliers. Chapman and Hall, London – New York 1980, 188 s. £ 14, 50.” Biometrical Journal 29 (2): 198-198.
[4] Li, X., J. Han, S. Kim, and Gonzalez, H. (2007). “ROAM: Rule- and Motif-Based Anomaly Detection in Massive Moving Object Data Sets.” Siam International Conference on Data Mining, April 26-28, 2007, Minneapolis, Minnesota, USA DBLP.
[5] Maciá-Pérez, F. M., J. V. Berna-Martinez, A. F. Oliva, and M. A. A. Ortega. 2015. “Algorithm for the detection of outliers based on the theory of rough sets.” Decision Support Systems 75 (C): 63-75.
[6] Liu, L. X., S. J. Qiao, B. Liu, J. J. Le, and C. J. Tang. (2009). “Efficient Trajectory Outlier Detection Algorithm Based on R-Tree.” Journal of Software 20 (9): 2426-2435.
[7] Dong, G. B., A. R. Xue, and B. T. Zhao. (2013). “Mining Abnormal Path from RFID Path Data Sets.” Application Research of Computers. 2452-2454
[8] Zhu, Y., H. W. Li, C. Fan, D. H. Xu, and F. Shi. (2017). “Clustering-Based Taxi Trajectory Outlier Detection.” Computer Engineering. 3-19
[9] Karypis, G, E. H. Han, and V. Kumar. (2008). “CHAMELEON A Hierarchical Clustering Algorithm Using Dynamic Modeling.” Computer 32 (8): 68-75.