ANOMALY DETECTION IN LARGE-SCALE TRAJECTORIES USING HYBRID GRID-BASED HIERARCHICAL CLUSTERING

Feng Ding,* Jian Wang,* Jiaqi Ge,** and Wenfeng Li*

Abstract

The increasing availability of location-acquisition technologies (such as GPS and GSM networks) and mobile computing techniques has generated a lot of spatial-temporal trajectory data and indicates the mobility of diversified moving objects such as people, vehicles, and animals. This brings new opportunities to identify abnormal activities of moving objects. This paper describes our detection of anomalies in human trajectory data using a hybrid grid-based hierarchical clustering method based on Hausdorff distance, which is suitable for measuring the similarity between trajectories of different lengths. The trajectories were first transformed into grid-based trajectories using a grid structure. After that, the grid-based trajectories were clustered based on their pairwise Hausdorff distances by applying different versions of hierarchical clustering algorithms. We evaluated our research result using a real-life dataset (published by Microsoft Research Asia), ground truth reconstructed by us, and evaluation criteria widely used in data mining. The experimental results demonstrate that the proposed algorithm is more effective and much faster than the traditional hierarchical clustering algorithm according to the pairwise comparison results.

Key Words

Trajectory anomaly detection, grid-based trajectory, Hausdorff distance, hierarchical clustering

1. Introduction

In past decades, the analysis of trajectory data has been encouraged because of the widespread diffusion of new techniques and systems for monitoring, collecting, and storing of location-aware data generated by a wealth of technological infrastructures, such as GPS positioning and wireless networks [1]–[4]. Trajectory data is composed of a sequence of timestamped geospatial locations and carries real object movement information. These numerous data broaden trajectory research and consequent applications in many fields, such as trajectory pattern detection, with the purpose of movement behaviour mining, location prediction about users' mobile devices, location-based activity discovery for provisioning of location-based services, traffic analysis for transportation management, and community movement detection (*e.g.*, traffic) to enhance travel recommendations.

Recent years have witnessed increasing attention paid in trajectory anomaly detection [5]–[8], which aims to automatically detect suspicious moving patterns. Unusual moving object trajectory patterns can generally indicate abnormal road network traffic patterns: popular sporting events attract crowds, holidays create disruptions, and protests may result in road closures, *etc.* Therefore, the detection of trajectory anomalies can assist in sensing abnormal events and can improve traffic surveillance and management.

Hierarchical clustering has been widely applied in trajectory analysis [9]–[12]. The general idea is to create an initial cluster by putting all data points into a disjoint set of clusters. The proximity is calculated based on the distance between cluster centroids. After that, at each step, the nearest clusters are successively merged together, reducing the number of clusters. The iteration is stopped when no further merging is possible. It is not feasible to apply traditional hierarchical clustering in large-scale datasets, such as the trajectory data in this paper, but there exist improved hierarchical clustering methods, such as BIRCH [13] and CURE [14]. The idea of BIRCH is to pre-cluster all the data and create an initial in-memory CF tree for indexing. Only local information stored in the CF tree is used in clustering, so the time complexity is improved. CURE employs random sampling and a constant number of points to represent a cluster during partitioning. Both algorithms follow an approach designed to reduce

* School of Electronic Science and Engineering, Nanjing University, Nanjing, China; e-mail: dingfengatnju@gmail.com, wangjnju@nju.edu.cn, leewf_cn@hotmail.com

** Department of Computer and Information Science, Indiana University and Purdue University Indianapolis, Indianapolis, Indiana, USA; e-mail: gjqfishes@gmail.com

Corresponding author: Jian Wang

Recommended by Dr. Xiaonv Hu

(DOI: 10.2316/Journal.206.2018.5.206-0061)

the computation time by pre-grouping and indexing the data.

In this paper, we propose a novel algorithm to automatically detect trajectory anomalies for traffic surveillance and management. This work presents a hybrid clustering algorithm that combines a hierarchical method with a grid-based method. The performance of the proposed algorithm was then compared with a large-scale real-world dataset. The clustering results obtained from our algorithm were analysed for their effectiveness based on ground truth reconstructed by us and evaluation criteria used in data mining. We verified that the proposed algorithm significantly reduces the computational time associated with clustering, which is different from the traditional hierarchical clustering according to pairwise comparisons.

2. Related Work

Numerous outlier-detection approaches are presented in the literature, including distance-based, density-based, direction-based, and classification and historical similarity-based approaches. Lee *et al.* [15] proposed a partition-and-detect framework to detect outlying sub-trajectories. In the first phase, each trajectory is partitioned into a set of trajectory partitions called t-partitions. In the second phase, outlying t-partitions are identified by distance-based and density-based measures. Ge *et al.* [16] took into account two types of outlying trajectories: outliers in terms of direction and outliers in terms of density. The continuous space is discretized into small grids, whereas a probabilistic model is employed to convert the direction information of trajectories in a grid into a vector with eight values that indicate the probabilities of moving in eight directions within this grid. The outlying score of a new trajectory can then be measured based on the density of the trajectories in the grids that actually passed by this trajectory. Li *et al.* [17] proposed a motion classifier for trajectory outlier detection that consists of the following three steps: (1) object movement features, called motifs, are extracted from the object paths. (2) To discover anomalies in object movements, motif-based generalization is performed to cluster similar object movement fragments and generalized the movements based on the associated motifs. (3) With motif-based generalization, objects are classified so that anomalous trajectories can be discriminated from normal ones.

Different from the above approaches of trajectory outlier detection, our work aims to detect all anomalous trajectories and pays more attention to the efficiency of the proposed algorithm. Our paper is organized as follows: Section 3 introduces the hybrid grid-based hierarchical clustering method for trajectory anomaly detection based on Hausdorff distance. The experimental results are displayed in Section 4 and our conclusion appears in Section 5.

3. Hybrid Grid-based Hierarchical Clustering

Hierarchical-based clustering is usually applied in clustering similar trajectories for anomaly detection. The traditional algorithm clearly follows an exhaustive pairwise comparison of the clusters, which is not feasible when employed in a large-scale dataset. Therefore, we improved the baseline algorithm by performing grid-based trajectory construction, which reduces the time complexity of hierarchical-based clustering. We will describe the details of the improved hybrid clustering algorithm in the following subsections.

3.1 Grid-based Trajectory Construction

In research into spatial data mining [18]–[23], grid indexing has been applied to realize faster processing of spatial accessing methods. Grid indexing partitions the data space into a certain number of cells, after which, additional operations are performed on the cells. Therefore, the algorithm performs efficiently when there are fewer cells (n) than in the total original data (N). In this work, GPS trajectories were transformed into *grid-based trajectories* before the clustering operation.

To process GPS trajectory data, the *haversine* formula [24] was applied to calculate the distances between the individual points of sets P and Q, which are denoted by $d(P, Q)$ in (1). By the haversine formula, the approximate distance between two points, $d(p, q)$, on the surface of the Earth (it is assumed the Earth is a perfect sphere) can be calculated as follows:

$$d(p, q) = 2R \arcsin(h) \tag{1}$$

where

$$h = \sqrt{\sin^2\left(\frac{\phi_q - \phi_p}{2}\right) + \cos(\phi_p)\cos(\phi_q)\sin\left(\frac{\lambda_q - \lambda_p}{2}\right)}$$

Here, ϕ_p and ϕ_q are the latitudes, and λ_p and λ_q are the longitudes of GPS points p and q, respectively; and $R = 6{,}371$ km is the approximate radius of the Earth modelled as a sphere.

Given a predefined origin $O(lng_0, lat_0)$, each GPS position $O(lng_i, lat_i)$ in a trajectory dataset can be converted to a longitude distance (x_i) and a latitude distance (y_i) using (2), where lng_i is the longitude value and lat_i is the latitude value.

$$\begin{aligned} x_i &= d\,(O, p_i) \times \cos\left(\arctan \tfrac{lng_i - lng_0}{lat_i - lat_0}\right) \\ y_i &= d\,(O, p_i) \times \sin\left(\arctan \tfrac{lng_i - lng_0}{lat_i - lat_0}\right) \end{aligned} \tag{2}$$

For each point in trajectory dataset $D(x, y)$, the row and column number of the grid to which the point belongs are calculated as follows:

$$\begin{aligned} s &= \left\lceil \frac{x_{\max} - x}{x_{\max} - x_{\min}} \cdot n_x \right\rceil \\ t &= \left\lceil \frac{y_{\max} - y}{y_{\max} - y_{\min}} \cdot n_y \right\rceil \end{aligned} \tag{3}$$

where $x_{\max}$ and $y_{\max}$ are the maximum values among x and y coordinates, whereas, $x_{\min}$ and $y_{\min}$ are the minimum values 0. n_x and n_y are the numbers of grids in the grid

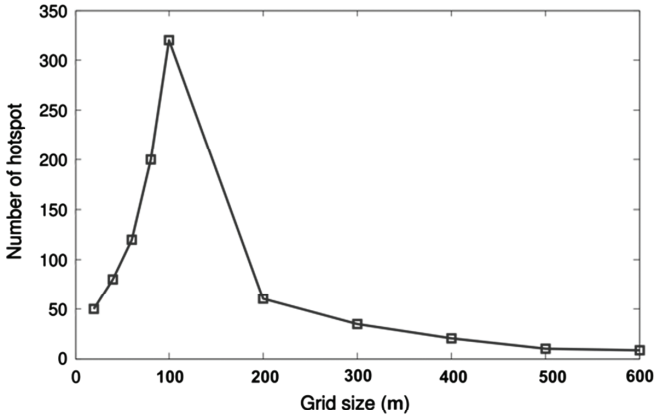

Figure 1. Effect of grid size on the number of hotspots.

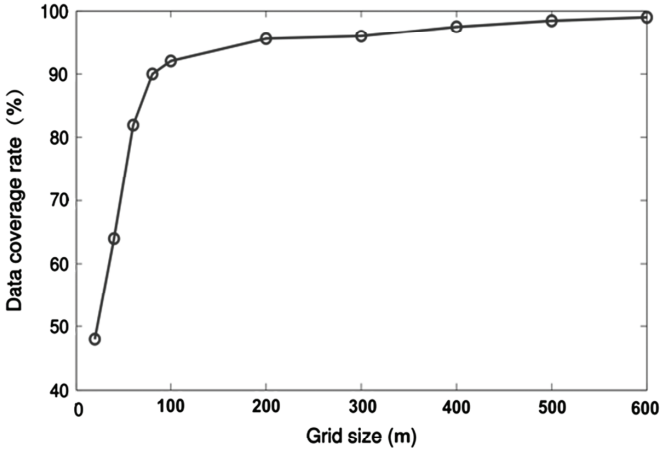

Figure 2. Effect of grid size on data coverage.

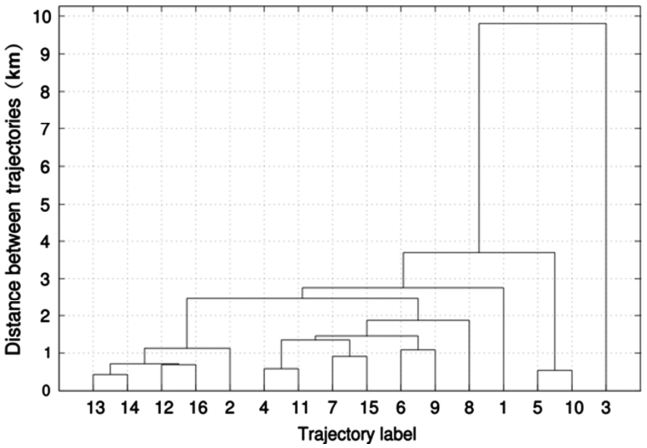

Figure 3. Maximum linkag.

structure. The grid is represented by (s,t). Then the grid structure is constructed. We define the capacity of grid (s,t) as $c_{s,t}$, which represents the number of points that are located in the grid. For each point, $c_{s,t}$ is increased by one after its assignment to the grid. A *hotspot* occurs when a grid contains at least $MinPts$ number of trajectory points, where $MinPts$ is a pre-defined threshold. The definition of grid-based trajectory is given in the following.

Definition 1 (Grid-based Trajectory). *Given a GPS trajectory* $T = \langle (p_1,t_1),\ldots,(p_i,t_i),\ldots,(p_n,t_n)\rangle$ *and a grid structure, a grid-based trajectory is defined as* $G = \langle (g_1,t'_1),\ldots,(g_j,t_j),\ldots,(g_m,t'_m)\rangle$ *where each grid* g_j *in* G *is a hotspot. Each grid* g_j *corresponding to a subsequence* $\langle (p_a,t_a),\ldots,(p_e,t_e)\rangle$ *where* $t'_j = t_a, (p_a,t_a)$ *is the arriving point and* (p_e,t_e) *is the exiting point.*

Without loss of generality, and to improve the completeness of the movement behaviour, detection of as many hotspots as possible is expected. More discovered hotspots can improve the grid-based trajectory to represent the original GPS trajectory. However, the grid size affects the number of hotspots and the accuracy of its representation of the actual trajectories. Figures 1 and 2 present examples of the effect of grid size on the number of hotspots and data coverage rate. Notably, the data coverage rate is the ratio between the trajectory data covered in detected hotspots and the original trajectory data. Given a minimum number of trajectory points $MinPts = 20$, the number of hotspots and the data coverage rate is detected for various grid sizes. As shown in Fig. 2, the hotspot detected by the bigger grid size results in few hotspots to describe the actual trajectory movement although the most original trajectory data can be included in detected hotspots as shown in Fig. 3. A bigger grid size may easily detect the hotspots but may lose the granularity of movement behaviours. Conversely, a smaller grid size can increase the number of hotspots and obtain the granularity but may lead to low data coverage. Furthermore, a very small grid size can hardly detect hotspots at all, and the corresponding data coverage rate would be very low because the capacity of most grids is inferior to $MinPts$. According to the analysis above, the optimal grid size should create a trade-off between the number of hotspots and the data coverage, rate which are determined by the distribution of trajectories.

3.2 Hausdorff Distance

The Hausdorff distance [25], [26] is used to measure the dissimilarity of two sets of points in a metric space. It is defined as the maximum distance of the first set to the nearest point of the second set. For instance, given two sets of points $P = \{p_1,p_2,\ldots,p_m\}$ and $Q = \{q_1,q_2,\ldots,q_n\}$, the directed Hausdorff distance from P to Q is defined by

$$d_H(P,Q) = \max_{p\in P}\left\{\min_{q\in Q}\{d(p,q)\}\right\}, \tag{4}$$

where $d(p,q)$ is the distance between points p and q under any chosen distance metric. However, the distance function $d_H(p,q)$ is not symmetric and $d_H(p,q)$ is generally not equal to $d_H(q,p)$. Multiple ways have been introduced for combining the directed Hausdorff distances to obtain an undirected distance metric. Generally, the undirected Hausdorff distance is obtained by taking the maximum of the two directed distances.

$$d_H(p,q) = d_H(q,p) = \max\{d_H(p,q), d_H(q,p)\} \tag{5}$$

3.3 Hierarchical Clustering

A hierarchical clustering approach is a method of cluster analysis that seeks to build a hierarchy of clusters. Strategies for hierarchical clustering generally fall into two types: agglomerative and divisive. In our method, we focus on agglomerative clustering methods [27], [28]. In agglomerative clustering, each data point starts out as its own cluster. Then, the closest pair of clusters is merged based on a distance measure given by the user. This is repeated until all the points are merged into one root cluster. The criterion used to determine the distance between two clusters as a function of the pairwise distances between data points in these clusters is called the linkage criterion [29], [30]. In this work, we have experimented with the three criteria including single linkage, maximum linkage, and average linkage clustering.

4. Experiments

This paper describes a series of experiments that were carried out to evaluate the performance of the proposed trajectory anomaly detection approach. All the experiments were implemented in Java JDK 1.7 on an Intel Core CPU i5-4570 3.20 GHz machine with 8 GB of memory running Microsoft Windows 8.

4.1 Evaluation Dataset

The GeoLife dataset [31]–[33] consists of 17,621 trajectories from 178 users in a period of more than 4 years (from April 2007 to October 2011). The trajectories cover a total length of 1,251,654 km and a total duration of 48,203 h. The GPS positions were collected with a high frequency. Over 90% of the positions were recorded less than every 5 s with a distance of less than 10 m from their previous positions. This dataset captures the complexity of human movement patterns and includes a wide range of user activities, such as daily routines (commuting from home to work and *vice versa*), entertainment (shopping and dining), and sporting activities (hiking and cycling). Almost all trajectories are located in Beijing (China), although the GPS positions are distributed in more than 30 cities.

4.2 Ground Truth and Evaluation Metric

Without prior knowledge of a person's movement patterns, it is impossible to clearly identify specific trajectories as anomalous [34]–[36]. The GeoLife dataset comprises data collected from regular people during their daily life but does not provide detailed information about the activities or circumstances that created the behaviours and resulted in the recorded trajectory. To quantify our anomaly detection performance [37], we reconstructed ground-truth labels with the assistance of 10 volunteers. These volunteers intuitively labelled the ordinary and anomalous behaviours based on their life experience. After reconstructing the ground truth, we evaluated the effectiveness of our approach based on some common metrics. A true positive (TP) rate and a true negative (TN) rate are sufficient to describe the performance of trajectory outlier detection. Alternatively, a false positive (FP) rate and a false negative (FN) rate could also be defined. It should be noted that TP rate + FN rate = 1 and TN rate + FP rate = 1. In addition, accuracy, precision, recall, and f-measure are considered. Here, TP indicates the number of trajectories labelled as "normal" that are detected as "normal," TN indicates the number of trajectories labelled as "abnormal" that are detected as "abnormal," FP indicates the number of trajectories labelled as "normal" that are detected as "abnormal" and FN indicates the number of trajectories labelled as "abnormal" that are detected as "normal."

4.3 Grid-based Trajectory Construction

Based on the distribution of the GeoLife dataset, we segmented grids within the metropolitan area of Beijing based on longitude [116.09, 116.70] and latitude [39.70, 40.18]. Incomplete records and those beyond this area were discarded. We divided the area into 520×530 grid cells, wherein the area of each cell corresponded to $100 \times 100\,\mathrm{m}^2$.

4.4 Anomaly Detection Results and Discussion

First, we validated the effectiveness of the proposed trajectory anomaly detection approach using the GeoLife dataset. Table 1 shows the evaluation result of trajectory anomaly detection performance on trajectory data from user 000 consisting of 80 trajectories. According to the reconstructed ground truth, the 80 trajectories include 65 "normal" trajectories and 15 "abnormal" trajectories. Note that λ_T is a predefined, distance threshold used to identify normal or abnormal trajectory, while P and N represent "normal" and "abnormal," respectively. As shown in Table 1, a smaller λ_T or a bigger λ_T fails to separate normal trajectories from abnormal ones. A smaller λ_T results in more false negatives and a bigger one results in more false positives. In this work, the optimal value of λ_T is set at 5 km.

For convenience of demonstration, Figs. 3–5 illustrate the hybrid clustering results of 16 trajectories in the form of a dendrogram. Taking Fig. 3 as an example, for a specified λ_T, trajectories can be identified as "normal" or "abnormal." If we suppose that $\lambda_T = 4$, all trajectories except trajectory 3 are located in the same cluster, which can be labelled as "normal." Obviously, trajectory 3 is identified as an anomalous trajectory. Moreover, if $\lambda_T = 3$, trajectories 3, 5, and 10 are located in the same cluster and other trajectories are in the other cluster. Trajectories in the small cluster are notable because they represent some digression from the normal cluster, which contains the majority of the trajectories. Therefore, trajectories 3, 5, and 10 are identified as anomalous trajectories. Figures 4 and 5 can be analysed in the same way.

In addition, running time performances of the hybrid grid-based hierarchical clustering algorithm and traditional hierarchical clustering algorithm were compared by applying them on trajectory datasets of different scale. Due to the exhaustive pairwise comparison, the time complexity

Table 1
Trajectory Anomaly Detection Performance

λ_T (km)	P	N	TP	TN	FP	FN	TPR	TNR	FPR	FNR	Accuracy	Precision	Recall	F
1	40	40	40	15	0	25	0.62	1.00	0.00	0.38	0.69	1.00	0.62	0.77
2	44	36	44	15	0	21	0.68	1.00	0.00	0.32	0.74	1.00	0.68	0.81
3	53	24	53	15	0	12	0.82	1.00	0.00	0.18	0.85	1.00	0.82	0.90
4	56	24	55	14	1	10	0.85	0.93	0.07	0.15	0.86	0.98	0.85	0.91
5	63	17	62	14	1	3	0.95	0.93	0.07	0.05	0.95	0.98	0.95	0.96
6	66	13	64	12	2	3	0.98	0.80	0.20	0.02	0.95	0.97	0.98	0.97
7	71	9	64	8	7	1	0.98	0.53	0.47	0.02	0.90	0.90	0.98	0.94
8	75	5	64	4	11	1	0.98	0.25	0.73	0.02	0.85	0.85	0.98	0.91
9	76	4	64	3	12	1	0.98	0.20	0.80	0.02	0.84	0.84	0.98	0.90
10	78	2	65	2	13	0	1.00	0.13	0.87	0.00	0.84	0.83	1.00	0.90

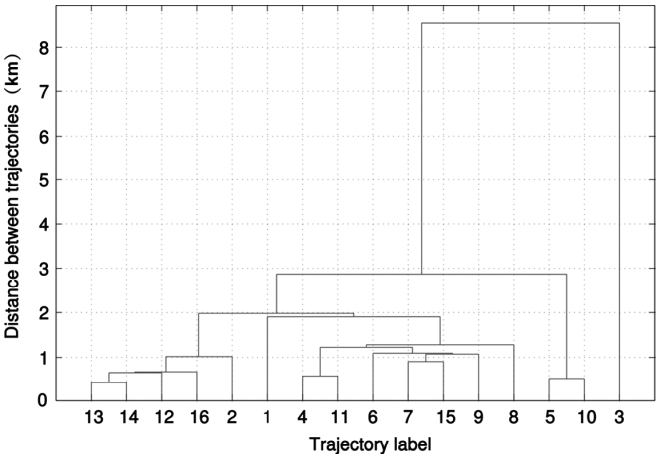

Figure 4. Single linkage.

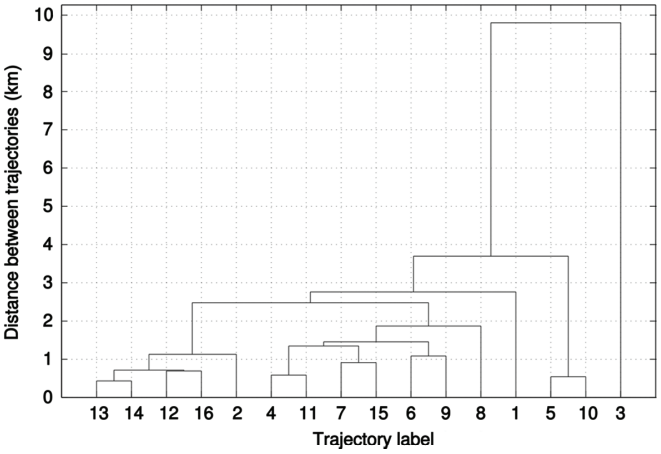

Figure 5. Average linkage.

in each iteration of the basic agglomerative clustering algorithm is $O(N^2)$, where N refers to the size of the dataset. The time complexity of our hybrid grid-based hierarchical clustering algorithm is correlated with the number of

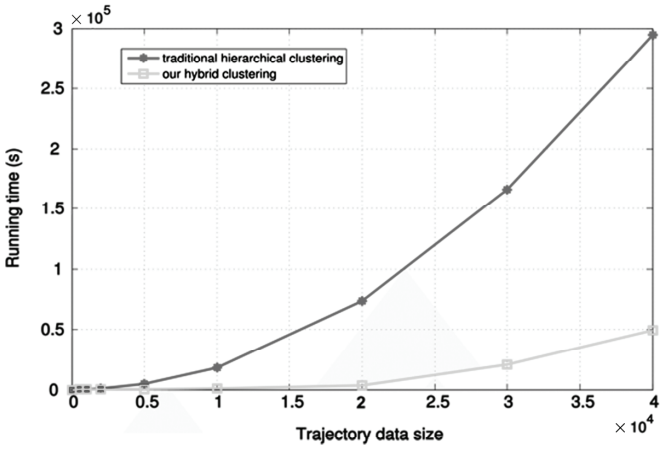

Figure 6. Running time comparison.

grid indices. Without loss of generality, and assuming a uniform distribution of data points, the number of data points in a grid cell is N/M where M is the total number of grid indices. Consequently, the time complexity of our algorithm is $O(N^2/M^2)$. As shown in Fig. 6, we compared the execution time of both algorithms on eight different trajectory data sizes with the grid size set to 100×100 for all trajectories. The results demonstrate that the hybrid grid-based hierarchical clustering algorithm is much faster than the traditional hierarchical clustering algorithm.

5. Conclusion

In this paper, a hybrid grid-based hierarchical clustering algorithm was proposed to discover trajectory anomalies in large-scale GPS trajectory datasets. First, GPS trajectories were transformed into grid-based trajectories based on a grid indexing procedure. Then, during the hierarchical clustering, only the distances between different grids were

compared, replacing an exhaustive pairwise comparison of all the data points. Grid indexing is advantageous in significantly decreasing the size of data space needed to run the distance computation. Given the appropriate grid size, it is necessary to significantly reduce the amount of distance computation needed for each step without sacrificing accuracy. We reconstructed the ground truth and applied some common metrics in data mining to evaluate the performance of our proposed method. Experimental results show that the algorithm can significantly reduce computation time while guaranteeing the effectiveness of anomaly detection when it is applied in large-scale real-world GPS datasets.

References

[1] X.Y. Wang, *et al.*, Trajectory tracking control of a hydraulic parallel robot manipulator with lumped disturbance observer, *International Journal of Robotics and Automation*, *28*(2), 2013, 103–111.

[2] J. Kim and W. Chung, Efficient placement of beacons for localization of mobile robots considering the positional uncertainty distributions, *International Journal of Robotics and Automation*, *30*(2), 2015, 119–127.

[3] P. He and S. Dai, Real-time stealth corridor path planning for fleets of unmanned aerial vehicles in low-altitude penetration, *International Journal of Robotics and Automation*, *30*(1), 2015, 60–69.

[4] C.-C. Chen, *et al.*, A framework of barcode localization for mobile robots, *International Journal of Robotics and Automation*, *28*(4), 2013, 317–330.

[5] Z. Fu, W. Hu, and T. Tan, Similarity based vehicle trajectory clustering and anomaly detection, *Proc. IEEE Conf. on Image Processing*, Genova, Italy, 2005, II-602-5.

[6] C. Piciarelli, C. Micheloni, and G.L. Foresti, Trajectory-based anomalous event detection, *IEEE Transactions on Circuits and Systems for Video Technology*, *18*(11), 2008, 1544–1554.

[7] P. Claudio and G.L. Foresti, On-line trajectory clustering for anomalous events detection, *Pattern Recognition Letters*, *27*(15), 2006, 1835–1842.

[8] L. Rikard and G. Falkman, Online learning and sequential anomaly detection in trajectories, *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *36*(6), 2014, 1158–1173.

[9] Y. Zheng, *et al.*, Mining interesting locations and travel sequences from GPS trajectories, *Proc. 18th international Conf. on World Wide Web*, ACM, Madrid, Spain, 2009, 791–800.

[10] Y. Zheng, Trajectory data mining: An overview, *ACM Transactions on Intelligent Systems and Technology*, *6*(3), 2015, 29.

[11] P.-R. Lei, A framework for anomaly detection in maritime trajectory behaviour, *Knowledge and Information Systems*, *47*(1), 2015, 1–26.

[12] L. Ssebazza and Y.-J. Pan, DGPS-based localization and path following approach for outdoor wheeled mobile robots, *International Journal of Robotics and Automation*, *30*(1), 2015, 13–25.

[13] T. Zhang, R. Ramakrishnan, and M. Livny, BIRCH: An effective data clustering method for very large database, *Proc. ACM SIGMOD International Conf. on Management of Data*, Montreal, Canada, 1996.

[14] S. Guha, R. Rastogi, and K. Shim, CURE: An efficient clustering algorithm for large databases, *Information Systems*, *26*(1), 1998, 35–58.

[15] J.G. Lee, J. Han, and K. Whang, Trajectory clustering: A partition-and-group framework, *Proc. ACM SIGMOD International Conf. on Management of Data*, ACM, Beijing, China, 2007, 593–604.

[16] Y. Ge, *et al.*, Top-eye: Top-k evolving trajectory outlier detection, *Proc. 19th ACM international Conf. on Information and knowledge management*, Toronto, Canada, 2010, 1733–1736.

[17] X. Li, J. Han, and S. Kim, Motion-alert: Automatic anomaly detection in massive moving objects, *Intelligence and Security Informatics*, Springer, Berlin, Heidelberg, 2006, 166–177.

[18] N.H. Park and W.S. Lee, Statistical grid-based clustering over data streams, *ACM SIGMOD Record*, *33*(1), 2004, 32–37.

[19] A. Amini, *et al.*, A study of density-grid based clustering algorithms on data streams, *Proc. 8th IEEE Conf. on Fuzzy Systems and Knowledge Discovery*, Shanghai, China, 2011, 1652–1656.

[20] M.R. Ilango and V. Mohan, A survey of grid based clustering algorithms, *International Journal of Engineering Science and Technology*, *2*(8), 2010, 3441–3446.

[21] J. Zhu, *et al.*, Merging grid maps via point set registration, *International Journal of Robotics and Automation*, *28*(2), 2013, 180–191.

[22] S. Saeedi, *et al.*, Occupancy grid map merging for multiple robot simultaneous localization and mapping, *International Journal of Robotics and Automation*, *30*(2), 2015, 149–157.

[23] B.P. DeJong, Two-and three-dimensional auditory occupancy grids with a mobile robot, *International Journal of Robotics and Automation*, *29*(1), 2014, 14–22.

[24] C.C. Robusto, The cosine-haversine formula, *The American Mathematical Monthly*, *64*(1), 1957, 38–40.

[25] M.P. Dubuisson and A.K. Jain, A modified Hausdorff distance for object matching, *Proc. 12th IAPR International Conf. on Pattern Recognition*, IEEE, Jerusalem, Israel, 1994, 566–568.

[26] Z. Zhang, K. Huang, and T. Tan, Comparison of similarity measures for trajectory clustering in outdoor surveillance scenes, *Proc. 18th Conf. on Pattern Recognition*, Vol. 3, IEEE, Hong Kong, China, 2006.

[27] W.H. Day and H. Edelsbrunner, Efficient algorithms for agglomerative hierarchical clustering methods, *Journal of Classification*, *1*(1), 1984, 7–24.

[28] I. Davidson and S.S. Ravi, Agglomerative hierarchical clustering with constraints: Theoretical and empirical results, *Proc. ACM Conf. on Knowledge Discovery in Databases*, Springer, Berlin, Heidelberg, 2005, 59–70.

[29] Y. Zhao and K. George, Evaluation of hierarchical clustering algorithms for document datasets, *Proc. 11th Conf. on Information and Knowledge Management*, ACM, Virginia, USA, 2002, 515–524.

[30] F. Murtagh and C. Pedro, Algorithms for hierarchical clustering: An overview, *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, *2*(1), 2012, 86–97.

[31] Y. Zheng, H. Fu, X. Xie, W.-Y. Ma and Li, Q. Geolife GPS trajectory dataset – User Guide, *Geolife GPS trajectories 1.1*, 2011.

[32] Y. Zheng, X. Xing and W.-Y. Ma, GeoLife: A collaborative social networking service among user, location and trajectory, *IEEE Data Engineering Bulletin*, *33*(2), 2010, 32–39.

[33] B. Li, B. Hou, W. Yu, *et al.*, Applications of artificial intelligence in intelligent manufacturing: A review, *Frontiers of Information Technology & Electronic Engineering*, *18*(1), 2017, 86–96.

[34] J.Y. Zhuang, *et al.*, Personalized topic modeling for recommending user-generated content, *Frontiers of Information Technology & Electronic Engineering*, *18*(5), 2017, 708–718.

[35] J.A. Rincon, J. Bajo, A. Fernandez, *et al.*, Using emotions for the development of human-agent societies, *Frontiers of Information Technology & Electronic Engineering*, *17*(4), 2016, 325–337.

[36] B. Ju, Y. Qian, and M. Ye, Preference transfer model in collaborative filtering for implicit data, *Frontiers of Information Technology & Electronic Engineering*, *17*(6), 2016, 489–500.

[37] G. Song, X. Jin, G. Chen, *et al.*, Two-level hierarchical feature learning for image classification, *Frontiers of Information Technology & Electronic Engineering*, *17*(9), 2016, 897–906.

Biographies

Feng Ding was born in Jiangsu, China, in 1992. He received his B.S. degree in 2014 and his M.S. degree in 2017 from Nanjing University, Nanjing, China. His research interests are social networks, trajectory data mining, machine learning, computer vision, camera-based positioning, vehicular positioning, and pattern recognition.

Jian Wang was born in Jiangsu, China in 1978. He received his M.S. degree from Nanjing University of Science and Technology, Nanjing, China, in 2003, and a Ph.D. degree from Nanjing University, Nanjing, China, in 2006. He is currently an associate professor at the school of electronics and sciences, Nanjing University, Nanjing, China. His research interests include social networks, video coding, video transmission, and parallel computing.

Jiaqi Ge is currently a data scientist of Expedia inc. He received his Ph.D. in Computer Science from Purdue University West Lafayette, and a M.S. in Computer Engineering and a B.S. in Electrical Engineering from Nanjing University, Nanjing, China. His main research area is on data mining and machine learning, focusing on mining and modeling uncertain data and big data including sensor data, online shopping data and time series data. He also works on distributed data mining applications.

Wenfeng Li was born in Jiangsu, China in 1975. He received a Ph.D. degree in communication and information systems from Southeast University, Nanjing, China, in 2012. He is currently a Postdoctoral Fellow with the School of Electronic Science and Engineering, Nanjing University, Nanjing. His research interests include wireless and mobile communications, wireless sensor networks, and satellite communications and networks.