GPS trajectory clustering method for decision making on intelligent transportation systems

Gary Reyes Zambrano [a,*], Laura Lanzarini [b,**], Waldo Hasperué [b,***], and Aurelio F. Bariviera [c,d,****]

[a] *Universidad de Guayaquil, Facultad de Ciencias Físicas y Matemáticas, ECG 352,11000 Guayaquil, Ecuador*

[b] *Universidad Nacional de La Plata, Facultad de Informática, Instituto de Investigación en Informática LIDI (Centro CICPBA) 1900 La Plata, Buenos Aires, Argentina*

[c] *Universitat Rovira i Virgili, Department of Business, Av. Universitat 1 43204 Reus, Spain*

[d] *Universidad del Pacífico, Lima, Perú*

AbstractTechnological progress facilitates recording and collecting information on vehicles' GPS trajectories on public roads. The intelligent analysis of this data leads to the identification of extremely useful patterns when making decisions in situations related to urbanism, traffic and road congestion, among others. This article presents a GPS trajectory clustering method that uses angular information to segment the trajectories and a similarity function guided by a pivot. In order to initialize the process, it is proposed to segment the region to be analyzed in a uniform way forming a grid. The obtained results after applying the proposed method on a real trajectories database are satisfactory and show significant improvement in comparison with the methods published in the bibliography

Keywords: segmentation, clustering, GPS trajectories, intelligent transportation systems.

[*] gary.reyesz@ug.edu.ec
[**] laural@lidi.info.unlp.edu.ar
[***] whasperue@lidi.info.unlp.edu.ar
[****] aurelio.fernandez@urv.cat
[*] gary.reyesz@ug.edu.ec
[**] laural@lidi.info.unlp.edu.ar
[***] whasperue@lidi.info.unlp.edu.ar
[****] aurelio.fernandez@urv.cat

1. Introduction

The growing use of GPS devices and the evolution in the transportation field, demand increasingly efficient techniques for data analysis and decision-making. Intelligent transportation systems process large amounts of GPS trajectory data generated from vehicles on the roads in real time [1,2,3]. The collected data must be analyzed to convert them into knowledge in order to use them as support data in decision-making. The detection of traffic congestion, anomalous patterns in traffic that help to predict accidents and the evaluation of the performance of main roads and avenues are some of the main application scenarios.

As part of data processing, intelligent transportation systems use different algorithms to group GPS trajectories based on different criteria [4,5,6]. The bibliography discusses several methods that perform clustering based on data segmentation and the similarity calculation of these segments. There are different approaches to evaluate the similarity between segments of trajectories according to the type of object and context considered. In the case of GPS trajectories, the function must take into account the underlying graph of the road network and the graphs connectivity or compliance with the sequence order [7].

Among the most used similarity functions in the literature [8] are network-limited distance and distances based on shape and warping. For the purposes of this paper, shape-based distance measurements are of interest as they seek to identify the geometric characteristics of trajectories by emphasizing their shape. Among the shape-based distance measurements, this article uses the Hausdorff distance defined in [9] as the distance between sets of vectors and in [10] for the sequence of vectors.

This paper is organized as follows: section 2 analyzes some previous work that has sought to solve this problem, section 3 describes the proposed method, section 4 details the tests carried out and the obtained results and finally section 5 contains the conclusions and future lines of work.

2. Previous works

Clustering techniques aim to bring together elements with common characteristics using a descriptor or centroid associated with each group. Conventional solutions operate on numerical vectors and use a distance measurement to quantify the similarity between pairs of elements.

Literature proposes solutions to perform trajectory clustering that adapt conventional clustering algorithms emphasizing on the information representation, as well as the distance measurement to be used. Such is the case of the Tra-DBscan algorithm [11], which indicates a way to apply the classic DBscan clustering algorithm [12] to group GPS trajectories. This algorithm indicates how to segment the trajectories into sections and then, use the Hausdorff distance for sets of line segments defined in [13] as a metric to establish their similarity. This metric takes into account the perpendicular, angular and parallel distance of the different locations that make up the segments of the trajectories.

On the other hand, TRACLUS is a GPS trajectory-clustering algorithm defined in [14] that uses the same similarity metric for trajectory sections as TRA-DBscan, but it indicates a clearer strategy for partitioning trajectories taking into account angle variations between different locations. With the partitioning strategy used, not all connected sets can become clusters.

In [15] the ATCGD algorithm was proposed, which is composed of 3 phases called partitioning, mapping and clustering. In the partitioning phase, the MDL partitioning method (AD-MDL) is applied based on the average angular difference of each segment. In the second phase, the trajectory segments are mapped into the corresponding cells and in the clustering phase, a DBscan-based algorithm is used to group the segments. Its application on different trajectory databases showed a significantly better performance than the one obtained using TRACLUS algorithm. In view of the good results achieved, it is considered extremely useful to use perpendicular, parallel and angular distance measurements in the representation of the different locations that make up each trajectory.

In [16] an algorithm was formulated for the detection of passengers boarding and disembarking points in taxis. For the clustering stage the GADBSCAN algorithm was used, which is an adaptation of the DBscan algorithm specifically designed to work with this type of data. For the trajectory selection, a weighted tree was incorporated which takes into account factors such as distance, driving time and vehicle speed.

Regarding the trajectories partitive clustering, an algorithm to group GPS trajectories using a variant of the Fuzzy C-Means (FCM) algorithm was proposed in [17]. An important aspect of this work is the proposal to partition GPS trajectories using a method based on the line segments angle that, according to the authors, reduces the loss of local information. Regarding clustering itself, K-means++ based on Hausdorff to produce initial centroids, and a Lagrange-based method to improve clustering were used. The results of this proposal were measured on GPS trajectories in real-world taxis.

Based on the aforementioned researches, it can be stated that the analysis of GPS trajectory segments has shown better results for trajectory clustering; therefore, the present research proposes a trajectory method based on the classical K-means basic clustering method, which uses the analysis of trajectory segments to construct a segmentation method. The Hausdorff distance is used as similarity metric and as contribution of the research, the clustering process is described with the particularity of being guided by a pivot. The method proposed in this article includes the detailed description of the trajectories segmentation method and the form of initializing the centers.

3. 3. GPS trajectories clustering method

A GPS trajectory is defined by a set of geographic locations, each of which is represented by its latitude and longitude. In other words, the ith trajectory is of the form $TR_i = (P_{i1}, P_{i2}, \cdots, P_{is})$ being each location P_{ij} a vector of the form $P_{ij} = (latitud_{ij}, longitud_{ij})$.

The GPS trajectory clustering method proposed in this research uses the K-means basic method for clustering, the Hausdorff distance as similarity metric and the pivot concept for centroid recalculation. The method has a trajectory segmentation component and a sub-trajectory clustering component. The first one divides GPS trajectories into sub-trajectories characterized by no significant changes in direction, which favors its analysis. The second proposed component is the sub-trajectory clustering component, which analyzes the sub-trajectories obtained in the previous component and groups them using the K-means basic algorithm from a given criterion.

An important aspect of any clustering technique is the proper initialization of the initial centers. In this case, it is proposed to make a grid of the plane that covers all the sub-trajectories. The Hausdorff distance is used as similarity measure between sub-trajectories. As a result, the proposed method returns the list of clusters formed.

3.1. GPS trajectory segmentation

GPS trajectory segmentation is the first component of the proposed method and it aims to create segments integrated by GPS points that share a common feature. The objective here is to identify segments that, in their composition, do not contain sudden, abrupt or significant direction changes, thus maintaining a stable orientation. To do this, the angles that each line segment forms with the reference plane will be calculated and compared in order. Angle values exceeding 180 degrees will be expressed as the difference between 360 degrees and the obtained value. In this way, all the angles of the trajectory will be expressed by values belonging to the interval $[-\pi, \pi]$.

Given the ith trajectory $TR_i = (P_{i1}, P_{i2}, \cdots, P_{is})$, the segmentation process begins by calculating the angle α formed by the line segment joining P_{i1} and P_{i2}, registering them as the first two points of the first sub-trajectory. The angle α is taken as a reference of the direction of the sub-trajectory to be formed. Then the angle determined by the line segment joining P_{i2} and P_{i3} is calculated and compared with α. If the difference between the two angles is below a threshold value previously defined, also called *angular tolerance*, then point P_{i3} is added to the sub-trajectory and the process continues with the next point. If not, the sub-trajectory is ended, added to the sub-trajectories list and a new sub-trajectory is started by calculating a new value of the reference angle. This process is repeated until all angles of the trajectory have been analyzed. As a result, a list formed by the different segments of the original trajectory will be obtained, which becomes the input of the GPS sub-trajectories clustering method.

3.2. GPS sub-trajectories clustering

Once obtained the sub-trajectories list, a partitive clustering is carried out using a winner-take-all style algorithm where first, the examples are assigned to the nearest centroids and then the centers are updated.

Algorithm 1 Pseudocode of the proposed method

Input: Trajectory $T = (P_1, P_2, \cdots, P_n)$ formed by GPS locations and the THRESHOLD value representing the angular tolerance.
Output: The list of sub-trajectories L
Process
Calculate the alpha angle determined by P_1 and P_2.
Add P_1 and P_2 to the sub-trajectory.
for i= 3 to n **do**
  Calculate the beta angle determined by P_{i-1} and P_i.
  if (beta - alpha) < THRESHOLD **then**
    Add Pi to the sub-trajectory.
  else
    // sub-trajectory is finished
    Add sub-trajectory to L
    Calculate the alpha angle determined by P_i and P_{i+1}.
    Initiate a new sub-trajectory formed by P_i and P_{i+1}.
    $i = i + 1$
  end if
end for
return L

This research proposes the use of a grid to define initial centroids. For this purpose, the area covered by the trajectories is divided into uniformly distributed sectors. In each sector, 4 centroids are defined dividing it in 8 equal parts. The first centroid cuts the sector in half and the remaining 3 do the same, increasing the angle by 45 degrees each time. With these initial positions, the objective is to detect the different inclinations that can appear in the groups of sub-trajectories. The assignment of sub-trajectories to the centroids is performed using the Hausdorff distance defined for line segments [9]. Once this step has been completed, the centroids are recalculated.

Figure 1 represents the centroid recalculation process using the pivot where T_1, T_2 and T_3 represent the trajectories and P_1 represents the first point of the pivot to be analyzed. The calculation of the distance of all points of the trajectory T_1 to point P_1 is made using the Euclidean distance, and the point with the shortest distance, denoted as A_1, is selected. This point is taken as a reference to calculate the distance of all the points of the trajectory T_2 to point A_1. As a result of this operation, the point with the shortest distance is selected and denoted as B_1, this point is taken as a reference to calculate the distance of all the points of the trajectory T_3 to B_1 and the point with the shortest distance is selected, denoted as C_1. This process is carried out for all the trajectories and the analyzed information is stored in the form of $(A_1, B_1, C_1 \cdots, Z_1)$. This information allows the calculation of the mean value of the stored data for longitude and latitude, thus constructing the first point, called NP_1, of the new centroid. The remaining points of the centroid are calculated in the same way but starting at the remaining points of the pivot $(P_2, P_3, P_4, \cdots, P_n)$.

The process of calculation and recalculation of the centroid may contain serious errors if a randomly selected trajectory would be used as the chosen trajectory to start the analysis, and it was a trajectory with very few points, in other words, the least similar trajectory of the group. In order to avoid or reduce these errors, it is proposed to use the pivot (plotted according to the characteristics of the trajectories that make up a group) as initial trajectory or guide, to perform the calculations of the points of the new centroid, as shown in Figure 1.

The pivot plotting is constructed taking into account two important aspects of all the trajectories that make up a group: the displacement of the group of trajectories, determining whether the pivot will be plotted vertically or horizontally, and the mean value of the amount of points of all the trajectories in the group.

4. Results

The proposed method was used to identify travel patterns in taxi trajectories registered in the city of Beijing, China. The choice of the base is due to the fact that Beijing is the fourth most populated city in the world, with a population density of approximately 21.7 million people. This data set consists of 27385 line segments and 71375 GPS locations and it has been previously used in [17], research that was briefly described in 2.

As a result of applying the segmentation method described in section 3.1 to the original data (71375 locations) approximately 100 sub-trajectories were obtained. Each sub-trajectory was formed by approximately 700 locations of the form (latitude, longitude) and the threshold used for the difference between angles was 5 degrees. The difference between the number of line segments obtained by the authors of the paper [17] and those obtained in the present investigation lies in the segmentation method used. The objective of the research is to evaluate the clustering method in a profound way, so it is decided to continue with the execution of the method, having as input the obtained sub-trajectories.

To initialize the centroids, the area covered by the trajectories was divided using a regular grid. Although in this case, the selection of the place where the grid was located was made based on the trajectories, cartographic information referring to the distribution of the most important circulation routes of the place can also be used, thus making the method independent of the input data.

Figure 2 shows the location of the centroids, which indicate the usual circulation zones using $k = 20$. To measure the performance of the proposed method, the pbm-index was used as indicated in [17,18]. This index is calculated from the product of three factors: the proportion of groups formed, a factor that quantifies the cohesion of the groups, and the greatest distance between pairs of centers. The higher the value obtained in this index, the better the clustering. The obtained results are displayed in Table 1, alongside the results obtained by other classical methods of the literature such as: K-means, K-median, Fuzzy C-Means (FCM) and FCML defined in [17].

As can be seen in table 1, the pbm-index metric values obtained by the proposed method are better than the values proposed in the bibliography. The main reason for the improvement of the results lies in the initialization of the centroids and the way to update the centroids position from this initial information.

5. Conclusions and future work

A GPS trajectory clustering method that uses angular information to identify sections with stable orientation and a pivot-based clustering technique have been presented. The proposed method segments GPS trajectories into shorter trajectories, called sub-trajectories, so that do not contain sudden, abrupt or significant changes in direction. For this purpose, a previously defined threshold is used, which constitutes a parameter of the proposed method. The initial pivots are evenly distributed within a grid applied over the area of interest.

The results obtained by applying the proposed method to a database of real trajectories have been satisfactory. Current work aims to improve the performance of the distance measurement used to assign sub-trajectories to centers. In the future, it would be interesting to have a real-time visualization of the

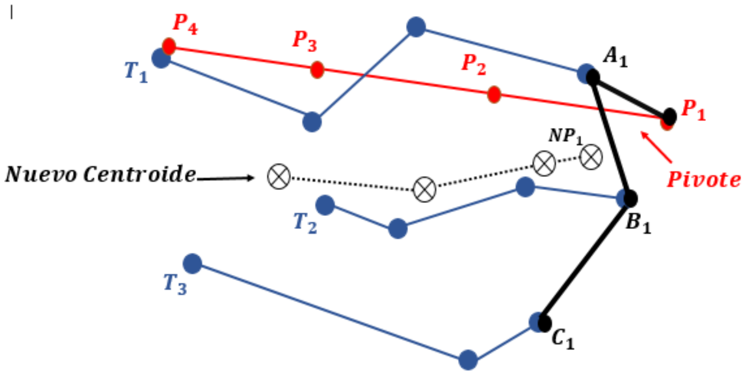

Figure 1. Representation of the centroid recalculation process using a pivot

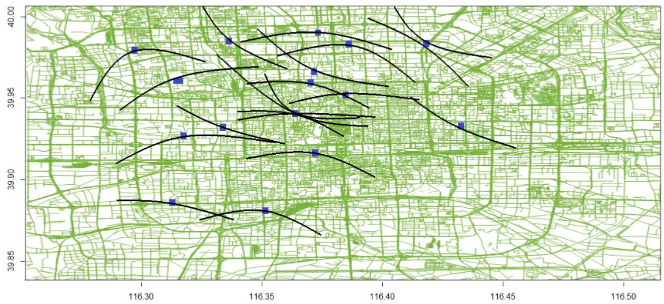

Figure 2. Representation of results for $k = 20$

Table 1

Comparison between experimental data from the bibliography and the current research

Values	K-means	K-median	FCM	FCML [17]	Proposed method
K = 20					
Maximum	0.07237	0.07639	0.080722	0.09033	0.466803
Mean	0.059763	0.0615	0.08001	0.086792	0.19846
Minimum	0.045748	0.046031	0.07909	0.086792	0.120109
K = 40					
Maximum	0.053466	0.054535	0.062047	0.066194	0.233402
Mean	0.047736	0.045803	0.060685	0.067674	0.075338
Minimum	0.040186	0.027937	0.060046	0.067124	0.044605

changes that occur in the centroids as new trajectories are added.

References

[1] H. Menouar, I. Guvenc, K. Akkaya, A. S. Uluagac, A. Kadri, and A. Tuncer, UAV-Enabled Intelligent Transportation Systems for the Smart City: Applications and Challenges, *IEEE Commun. Mag.*, **55**(3)(2017), 22–28.

[2] J. Guerrero-Ibáñez, S. Zeadally, and J. Contreras-Castillo, Sensor Technologies for Intelligent Transportation Systems, *Sensors*, **18** (2018), 1–24.

[3] N. Markovic, P. Sekuła, Z. Vander Laan, G. Andrienko, and N. Andrienko, Applications of Trajectory Data from the Perspective od a Road Transportation Agency: Literature Review and Maryland Case Study, *IEEE Trans. Intell. Transp. Syst.*, 2018.

References

[1] H. Menouar, I. Guvenc, K. Akkaya, A. S. Uluagac, A. Kadri, and A. Tuncer, UAV-Enabled Intelligent Transportation Systems for the Smart City: Applications and Challenges, *IEEE Commun. Mag.*, **55**(3)(2017), 2228.

[2] J. Guerrero-Ibáñez, S. Zeadally, and J. Contreras-Castillo, Sensor Technologies for Intelligent Transportation Systems, *Sensors*, **18** (2018), 1–24.

[3] N. Markovic, P. Sekuła, Z. Vander Laan, G. Andrienko, and N. Andrienko, Applications of Trajectory Data from the Perspective od a Road Transportation Agency: Literature Review and Maryland Case Study, *IEEE Trans. Intell. Transp. Syst.*, 2018.

[4] Yulong Wang, K. Qin, Y. Chen, and P. Zhao, Detecting Anomalous Trajectories and Behavior Patterns Using Hierarchical Clustering from Taxi GPS Data, *Int. J. Geo-Information*, **7**(25) (2018), 1–20.

[5] X. Jiang, E. N. de Souza, A. Pesaranghade, B. Hu, D. L. Silver, and S. Matwin, *TrajectoryNet: An Embedded GPS Trajectory Representation for Point-based Classification Using Recurrent Neural Networks*, 2017 conference of the Center for Advanced Studies on Collaborative Research, 2017, 19.

[6] M. Fountoulakis, N. Bekiaris-liberis, C. Roncoli, I. Papamichail, and M. Papageorgiou, Highway traffic state estimation with mixed connected and conventional vehicles: Microscopic simulation-based testing q, *Transp. Res. Part C* **78** (2017), 13–33.

[7] E.D. Andrea, D. Di Lorenzo, B. Lazzerini, F. Marcelloni, F. Schoen, and V.S. Marta, *Path Clustering based on a Novel Dissimilarity Function for Ride-Sharing Recommenders*, 2016 IEEE Int. Conf. Smart Comput., 2016.

[8] P. Besse, B. Guillouet, J. Loubes, R. Franois, P. Besse, B. Guillouet, J. Loubes, and R. F. Review, *Perspective for Distance Based Trajectory Clustering*, HAL, 2015.

[9] F. Hausdorff, Bemerkung uber den Inhalt von Punktmengen, *Math. Ann.*, **75**(3) (1914), 428–433.

[10] F. Treves, Topological Vector Spaces, Distributions and Kernels, F. Treves, Topological Vector Spaces, Distributions and Kernels, *Pure and Applied Mathematics*, **25** (2016), Elsevier.

[11] L. Liu, J. Song, B. Guan, Z. Wu, and K. He, Tra-DBScan: a Algorithm of Clustering Trajectories, *Appl. Mech. Mater.*, (2012), 4875–4879.

[12] M. Ester, K. H-P, S. J, and X. X, *A density-based algorithm for discovering clusters in large spatial databases with noise*, 2nd Int'l Conf. on Knowledge Discovery and Data Mining, 1996, 226–231.

[13] Y. Gao and M. K. H. Leung, Line segment Hausdorff distance on face matching, *Pattern Recognit.*, **35**(2) (2002), 361–371.

[14] J. Lee, J. Han, and K.-Y. Whang, *Trajectory Clustering: A Partition-and-Group Framework*, SIGMOD/PODS '07 International Conference on Management of Data Beijing, China, 2007.

[15] Y. Mao, H. Zhong, H. Qi, P. Ping, and X. Li, An Adaptive Trajectory Clustering Method Based on Grid and Density in Mobile Pattern Analysis, *Sensors* (2017), 1–19.

[16] Y. Shen, L. Zhao, and J. Fan, Analysis and Visualization for Hot Spot Based Route Recommendation Using Short-Dated Taxi GPS Traces, *Inf. 2015*, **6** (2015), 134–151.

[17] X. Zhou, F. Miao, H. Ma, H. Zhang, and H. Gong, A Trajectory Regression Clustering Technique Combining a Novel Fuzzy C-Means Clustering Algorithm with the Least Squares Method,*Int. J. Geo-Information*, **7**(164) (2018), 9–16.

[18] S. Pakhira, M.K. Bandyopadhyay and U. Maulik, Validity index for crisp and fuzzy clusters, *Pattern Recognit.*, **37** (2004), 487–501.