Computers, Environment and Urban Systems 55 (2016) 55–65

Contents lists available at ScienceDirect

Computers, Environment and Urban Systems

journal homepage: www.elsevier.com/locate/ceus

# Automated clustering of trajectory data using a particle swarm optimization

Zahedeh Izakian [a,*], Mohammad Saadi Mesgari [a], Ajith Abraham [b]

[a] Faculty of Geodesy & Geomatics, and Geoinformation Technology Center of Excellence, K. N. Toosi University of Technology, Tehran, Iran
[b] Machine Intelligence Research Labs (MIR Labs), Scientific Network for Innovation and Research Excellence, Auburn, USA

ARTICLE INFO

*Article history:*
Received 6 April 2014
Received in revised form 29 October 2015
Accepted 29 October 2015
Available online 6 November 2015

*Keywords:*
Fuzzy C-means clustering
Particle swarm optimization
Trajectory data
Discrete cosine transform
Dynamic time warping distance

ABSTRACT

Clustering trajectory data discovers and visualizes available structure in movement patterns of mobile objects and has numerous potential applications in traffic control, urban planning, astronomy, and animal science. In this paper, an automated technique for clustering trajectory data using a Particle Swarm Optimization (PSO) approach has been proposed, and Dynamic Time Warping (DTW) distance as one of the most commonly-used distance measures for trajectory data is considered. The proposed technique is able to find (near) optimal number of clusters as well as (near) optimal cluster centers during the clustering process. To reduce the dimensionality of the search space and improve the performance of the proposed method (in terms of a certain performance index), a Discrete Cosine Transform (DCT) representation of cluster centers is considered. The proposed method is able to admit various cluster validity indexes as objective function for optimization. Experimental results over both synthetic and real-world datasets indicate the superiority of the proposed technique to fuzzy C-means, fuzzy K-medoids, and two evolutionary-based clustering techniques proposed in the literature.


## 1. Introduction

Trajectories are spatio-temporal data, comprising the geographical positions of moving objects in various time steps. Moving objects can be vehicles in roads, airplanes in airline traffic, animals in forest areas, and so on. Capturing and analyzing this type of data may have numerous potential applications. Clustering as a strong instrument for exploratory data mining and knowledge discovery provides useful information about existing patterns in datasets. Scientists in various applications might use clustering techniques to reveal the available structure within data. For example, experts in traffic control and urban planning are interested in available structure within moving patterns of vehicles (e.g. cars, busses, airplanes) in different time intervals. They can use this information for road construction, or designing monitoring systems, etc.

With the recent development of data collection technologies like sensor networks and GIS, collecting trajectory data has never been easier. Consequently, there is a huge amount of this type of data collected in different applications (Mennis and Guo, 2009). To extract meaningful information from these datasets, designing and developing advanced data mining techniques is necessary. In this paper, an automated fuzzy clustering technique is proposed to discover and visualize the available structure within trajectory data.

Clustering is one of the most popular and powerful techniques in revealing and visualizing the available structure within data (Henriques, Bacao and Lobo, 2012, Eckley and Curtin, 2013). Fuzzy C-means (FCM) (Bezdek, 1981) is a commonly-used objective function-based clustering technique where, instead of a binary assignment, each object might belong to several clusters with a membership degree expressed in unit interval. Although this technique has been employed successfully in many applications, there are some challenges in applying this technique for clustering trajectory data as follows:

- FCM is sensitive to initialization and may produce results located in local optima. Since trajectories are high-dimensional data, falling into local optima in applying this technique is more likely.
- The most commonly used similarity measure in FCM is Euclidean distance, while in trajectory data using some more advanced similarity/dissimilarity measures (e.g. Dynamic Time Warping distance (DTW) (Berndt and Clifford, 1994)) is more appropriate. One has to be aware of challenges in employing DTW distance in FCM technique.
- FCM needs to know the number of available clusters in data, and in many real-world applications this value is not explicit.

To deal with the above-mentioned challenges, in this paper a particle swarm optimization approach (Kennedy and Eberhart, 1995, Kennedy and Eberhart, 1997) for fuzzy clustering of trajectory data has been proposed. In this method, PSO is employed to find near

* Corresponding author at: No. 1346, ValiAsr Street, Mirdamad cross, Geomatics Engineering Faculty, K.N. Toosi University of Technology, Tehran, Iran.
*E-mail addresses:* zizakian@mail.kntu.ac.ir (Z. Izakian), mesgari@kntu.ac.ir (M. Saadi Mesgari), ajith.abraham@ieee.org (A. Abraham).

http://dx.doi.org/10.1016/j.compenvurbsys.2015.10.009
0198-9715/

optimal number of clusters as well as near optimal cluster centers based on a certain performance index. For this purpose, a Discrete Cosine Transform (DCT) representation of trajectories is considered and PSO estimates DCT coefficients of cluster centers.

Hybridizing FCM with evolutionary algorithms (e.g. PSO) to prevent falling into local optima and improving the quality of clusters (in terms of a certain performance index) has been employed successfully in some literature (e.g., see (Izakian and Abraham, 2011)). Furthermore, applying evolutionary algorithms in clustering data allows us to use more advanced objective functions. This study is organized as follows:

In Section 2 we present a brief literature review. In Section 3 we formulate the problem and describe the performance index used in this paper for clustering. Section 4 briefly describes FCM and PSO clustering techniques. In Section 5 we focus on some concepts in dealing with trajectory data and explain the proposed method for clustering trajectory data. In Section 6 experimental studies are reported and Finally Section 7 concludes this study.

## 2. Literature review

Piciarelli et al. (2005) proposed an online trajectory clustering method for video surveillance. First, the moving objects were detected in video frames and then their tracks were compared to existing cluster centers. The distance between a trajectory and a cluster center was calculated using a normalized distance between every point of trajectory and the nearest point of the corresponding cluster center. If a match was found (in terms of a threshold), the corresponding cluster center was updated, otherwise a new cluster was created. Vlachos et al. (2002) introduced a hierarchical clustering technique using a longest common subsequence distance function. They showed that longest common subsequence is more efficient than Euclidean distance and dynamic time warping in their techniques when dealing with noisy trajectories. Yanagisawa and Satoh (2006) proposed two hierarchical techniques for clustering trajectory data. A combination of DTW distance and Euclidean distance function was employed in the process of clustering.

Li and Hu (2006) proposed a trajectory clustering framework comprising trajectory smoothing, feature extraction, trajectory coarse clustering and trajectory fine clustering. Wavelet decomposition for reducing the effect of noise in trajectory smoothing step was used. In the feature extraction step, a trajectory directional histogram was proposed to represent the statistical directional distribution of a trajectory. Coarse clustering and fine clustering steps were performed using a new graph theoretic clustering algorithm called dominant-set clustering. Anjum and Cavallaro (2007) presented a three-step unsupervised approach for fuzzy clustering of trajectory data. In the first step, a mean-shift technique to detect coarse clusters was employed. Next, adjacent clusters were combined by analyzing the clusters' attributes and finding similar behaviors. And finally, sparse clusters were considered as outliers and were eliminated.

Lee et al. (2007) proposed a trajectory clustering algorithm (called TRACLUS) based on a new partition-and-group framework with the use of discovering the common sub-trajectories from a trajectory database. In this method, first the trajectories were divided into a set of line segments at specific points and then similar line segments were detected using a suitable distance function and were grouped together into a cluster. Three distance functions namely the perpendicular distance, the parallel distance and the angle distance were used. Morris and Trivedi (2009) evaluated some different similarity measures including HU, PCA, DTW, LCSS, PF, MODH, and some clustering algorithms, such as Divisive, Direct, Agglomerative, Hybrid, Graph and Spectral over trajectory data and examined their strengths and weaknesses. They concluded that the performance of a method depends on the trajectories characteristics in a dataset. Zhang and Pi (2009) proposed a trajectory clustering algorithm based on a symmetric neighborhood technique (called BSNTC). In their proposed method, the neighbors and reverse neighbors of a trajectory were considered in computing its density distribution. Unlike the existing trajectory clustering algorithms which use global parameters to discover common trajectories, the proposed method requires only one input parameter. The experimental studies showed the efficiency of the proposed method in detecting clusters with different shapes. Chen et al. (2011) divided trajectories into a set of sub-trajectories, and computed the similarity between trajectories using a Hausdorff distance. A DBSCAN clustering algorithm was applied for clustering.

## 3. Problem formulation and performance index

Assume that there is a set of $N$ trajectories $\mathbf{r}_1, \mathbf{r}_2, \ldots, \mathbf{r}_N$ where, each trajectory is an ordered sequence including a number of 2D or 3D points. If $i$th object, $\boldsymbol{r}_i$, is a 3D trajectory, it is in the form

$$\boldsymbol{r}_i = (x_{i1}, y_{i1}, z_{i1}), (x_{i2}, y_{i2}, z_{i2}), \ldots, (x_{in}, y_{in}, z_{in}) \tag{1}$$

where $n$ is the length of trajectory and $(x_{ik}, y_{ik}, z_{ik})$ represents the spatial coordinates of object $\boldsymbol{r}_i$ in time step $k$. The objective of this paper is to cluster this type of data to optimize a certain performance index. For this purpose, a particle swarm optimization technique has been proposed. This technique is able to find (near) optimal number of clusters as well as (near) optimal cluster centers automatically. There is numerous performance indexes proposed in the literature for clustering. These techniques are known as cluster validity index (for example see (Wang and Zhang, 2007)). Note that selecting an appropriate performance index is application-dependent and can be based on the application purpose and the nature of data. In this paper we consider the cluster validity index proposed by Xie and Beni (XB) (Xie and Beni, 1991). This technique is one of the most widely-used performance indexes in the literature and takes into account compactness and separation of clusters. This index is expressed as follows:

$$V_{XB} = \frac{J}{Sep} = \frac{\sum_{i=1}^{c}\sum_{k=1}^{N} u_{ik}^{m} d^2(\mathbf{v}_i, \mathbf{r}_k)}{N \times \min_{i \neq j} d^2(\mathbf{v}_i, \mathbf{v}_j)} \tag{2}$$

where, $d(.)$ stands for a distance function, $c$ is number of clusters, $N$ is number of trajectories, $\boldsymbol{v}_i$ is $i$th revealed cluster center, $u_{ik}$ is membership degree of $\boldsymbol{r}_k$ to cluster center $\boldsymbol{v}_i$ and $m$ ($m > 1$) is fuzzification coefficient. In (2), $J$ evaluates the compactness of clusters and $Sep.$ estimates separation between cluster centers. A lower value of $V_{XB}$ is desired.

## 4. Clustering trajectory data

In this section we review two clustering techniques: Fuzzy C-means clustering and PSO-based clustering. These techniques are the basis of the proposed method in this paper.

### 4.1. Fuzzy C-means clustering

Let us consider $N$ objects $\boldsymbol{r}_1, \boldsymbol{r}_2, \ldots, \boldsymbol{r}_N$. The FCM technique partitions these $N$ objects into $c$ clusters by minimizing the following objective function.

$$J = \sum_{i=1}^{c}\sum_{k=1}^{n} u_{ik}^{m} d^2(\mathbf{v}_i, \mathbf{r}_k). \tag{3}$$

FCM produces a set of $c$ cluster centers $\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_c$ and a partition matrix $U = [u_{ik}]$, $i = 1,2,\ldots,c$, $k = 1,2,\ldots,n$ describing the membership degrees of each object to different clusters, with the following properties:

$$u_{ik} \in [0, 1], \sum_{i=1}^{c} u_{ik} = 1 \forall k, \text{and } 0 < \sum_{k=1}^{n} u_{ik} < n \forall i. \tag{4}$$