A COMPARISON OF EFFICIENCY OF THE OPTIMIZATION APPROACH FOR CLUSTERING OF TRAJECTORIES

A. Moayedi [1, *], R. Ali Abbaspour [1], A. Chehreghan [2]

[1] School of Surveying and Geospatial Engineering, College of Engineering, University of Tehran, Tehran, Iran – (alimoayedi2013, abaspour)@ut.ac.ir
[2] Mining Engineering Faculty, Sahand University of Technology, Tabriz, Iran - Chehreghan@sut.ac.ir

Commission VI, WG VI/4

KEY WORDS: Trajectories, Clustering, K-means, Optimization, Meta-heuristic Algorithms, DTW

ABSTRACT:

Clustering is an unsupervised learning method that used to discover hidden patterns in large sets of data. Huge data volume and the multidimensionality of trajectories have made their clustering a more challenging task. K-means is a widely used clustering algorithm applied in the trajectory computation field. However, the critical issue with this algorithm is its dependency on the initial values and getting stuck in the local minimum. Meta-heuristic algorithms with the goal of minimizing the cost function of the K-means algorithm can be utilized to address this problem. In this paper, after suggesting a cost function, we compare clustering performance of seven known metaheuristic population-based algorithms including, Grey Wolf Optimizer (GWO), Particle Swarm Optimization (PSO), Sine Cosine Algorithm (SCA), and Whale Optimization Algorithm (WOA). The results obtained from the clustering of several data sets with class labels were assessed by internal and external clustering validation indices along with computation time factor. According to the results, PSO, and SCA algorithms show the best results in the clustering regarding the Purity, and computation time metrics, respectively.

1. INTRODUCTION

Today, trajectories of moving entities such as people equipped with GPS devices, taxis, vessels, aircraft, and even animals can be recorded, stored, and processed (Frattasi and Della Rosa, 2017). In recent years, resulting from high accessibility to a massive volume of data and their complexities, process and knowledge extraction has become a case of research attention.

Cluster analysis is one of the mining tasks associated with extracting relevant outcome invisible in data. Previous research in trajectory clustering can be divided into four general categories of partitioning, hierarchical, density-based, and optimization-based. K-means and spectral methods are two important algorithms in the partitioning methods that have been mentioned more than other methods in the previous research. Tork was one of the first researchers who developed clustering of spatio-temporal data using K-means (Tork, 2012). Atev et al., using a spectral clustering and similarity function based on Hasdorf distance presented a framework for clustering trajectories (Atev et al., 2010). Fu et al. proposed a framework to detect the anomaly and to categorize vehicle trajectories in traffic applications using hierarchical and spectral clustering (Fu et al., 2005). Palma et al. addressed discovering points of interest using density-based clustering (Palma et al., 2008). In their research, Nanni et al., Lee et al., Akasapu et al., also used density-based clustering for groping the trajectories (Akasapu et al., 2011; Lee et al., 2007; Nanni and Pedreschi, 2006). Ahmadyfard and Modares by mixing K-means and PSO, and Lu et al. by integrating K-means and genetic algorithm developed new methods for clustering, and their proposed methods proved superiority to K-means (Ahmadyfard and Modares, 2008; Lu et al., 2004). Also, Izakian et al., presented an automatic approach for trajectory clustering using PSO (Izakian et al., 2016).

K-means algorithm is a simple clustering method with low computation cost that classifies data into different groups to form clusters with respect to a specified metric. According to the NP-hard nature of the clustering problem and its sensitivity to the initial cluster centroids, K-means might be trapped in the local minimum (Likas et al., 2003), and this leads to low reliability of the achieved results particularly in complex, multidimensional data like trajectories (Ossama et al., 2011).

Optimization relates to the process of discovering optimal values from all potential value for the parameters of a specified scheme to maximize or minimize its output. Population-based optimization algorithms which are a kind of metaheuristic algorithms, try to reach an appropriate trade-off between exploration (exploring different regions of search space) and exploitation (local search) to search for a solution close to the optimum value in challenging problems like clustering. Optimization of the clustering problem has been used to solve many issues including image clustering (Omran et al., 2005), document clustering (Mahdavi et al., 2008), traffic management (Bacquet et al., 2011), and smart city (Logesh et al., 2018).

In this paper, a proper cost function for trajectory clustering has been proposed and seven metaheuristic optimization methods, i.e., GWO (Mirjalili et al., 2014), PSO (Kennedy and Eberhart, 1995), SCA (Mirjalili, 2016), WOA (Mirjalili and Lewis, 2016), were used to enhance the results of trajectory clustering. The results were evaluated by Purity indices along with the computation time factor.

* Corresponding author

2. METHODOLOGY

2.1 Problem Definition

The clustering problem assigns the members of a set of trajectories $TD=\{T_1,T_2,\dots,T_N\}$ to k clusters $C = \{C_1, C_2,\dots,C_k\}$. Each trajectory T_i has multi-dimensional information and can simply be shown as $T = [(x_1,y_1,t_1),\dots,(x_m,y_m,t_m)]$. (x,y) refers to the location of the recorded points and t is the recording time. Clusters C must be defined in a way that trajectories within a cluster have the most similarity with each other and the most difference with the trajectories in other clusters.

Clustering can be solved as an optimization problem. Considering S and f(s) as the set of feasible solutions and the objective function to optimize respectively, optimization problem seeking a global optimum $s^* \in S$ where:

$$\forall s \in S, f(s^*) \leq f(s) \tag{1}$$

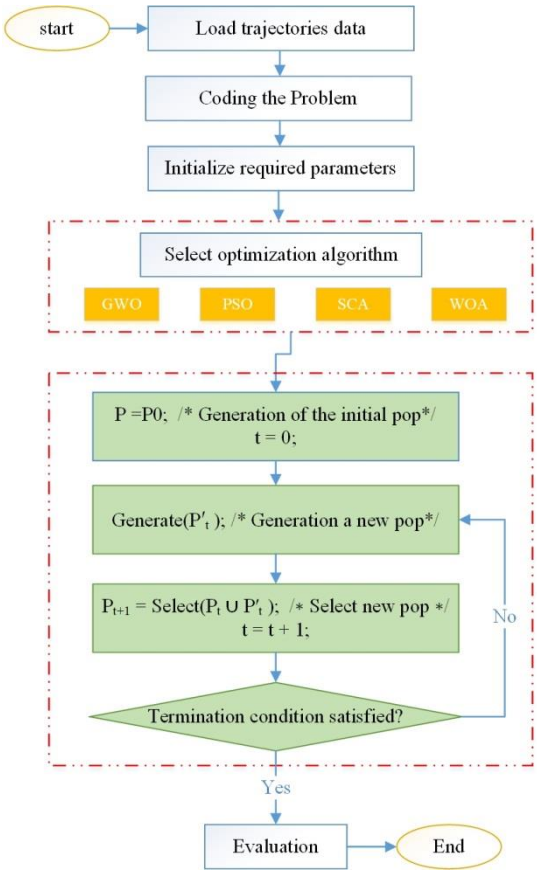


Figure 1. Flowchart of the proposed methodology

2.2 Encoding

Before optimization starts, the problem ought to be simplified and modeled. Population-based metaheuristics could be considered as an iterative improvement in a population of solutions. Starting from an initial population of solutions, they continue to generate a new population. Subsequently, some selection procedures based on their fitness are used to integrate this new population into the current one. This process iterates until the criteria for a stop are specified. The general flowchart of the methodology steps is represented in Figure 1.

Therefore in our approach, a population set $P=\{p_1,p_2,\dots p_n\}$ is considered. Each member of this set is a vector of length k having IDs referring to the cluster centers of each of the k clusters and structures the final clustering solution. Figure 2 demonstrates the encoding of the problem of trajectory clustering.

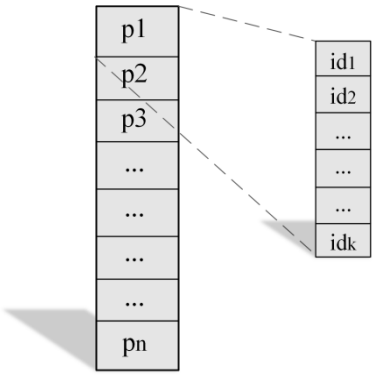


Figure 2. Encoding of the problem

2.3 Cost Function

After the population is created, each of the members has to be assessed. In this paper, a cost function for the clustering problem of trajectories is defined as Equation 2. The minimization of this cost function, which is one of the main goals of clustering, leads to the higher similarity of trajectories in each cluster and the difference with other clusters.

$$CostFunction = \sum_{j=1}^{k} \sum_{T_i \in C_j} \mathrm{D}\left(Z_{C_j}, T_i\right) \tag{2}$$

where Z_{Cj} is the representative trajectory for the cluster C_j and D is the DTW distance between each pair of trajectories and is obtained through Equation 3.

$$D\left(Z_{C_j}, T_i\right) = \begin{cases} 0 & , m = n = 0 \\ \infty & , m = 0 \,||\, n = 0 \\ euclidean\left(a_h^p, b_g^p\right) + & \\ min\left\{ \begin{matrix} D\left(R\left(Z_{C_j}\right), R(T_i)\right), \\ D\left(R\left(Z_{C_j}\right), T_i\right), \\ D\left(Z_{C_j}, R(T_i)\right) \end{matrix} \right\} & , Others \end{cases} \tag{3}$$

where $R(Z_{Cj})$ and $R(T_i)$ are the resulting trajectories after the removal of the first points from Z_{Cj} and T_i, respectively. m and n refer to trajectory lengths of Z_C and T_i. Also, a_h^p is the p^{th} dimension of the h^{th} point from the Z_{C_j} trajectory and b_g^p is the p^{th} dimension of the g^{th} point from the T_i trajectory.

2.4 Clustering Efficiency

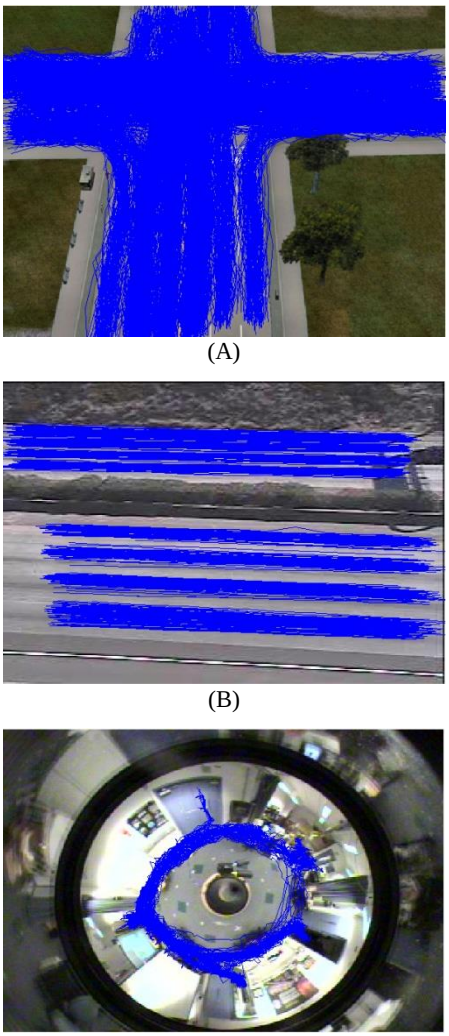
After reaching the loop's stopping condition and achieving the final results, the evaluation step needs to be taken. Two indicators, namely Silhouette and purity, along with the computation time factor are considered to evaluate the performance of different algorithms. Purity shows the percentage of properly clustered data, and it is computed according to Equation 4 (Manning et al., 2010).

$$Purity = \frac{1}{N}\sum_{k}\max_{j}\left|C_i \cap L_j\right| \quad (4)$$

where N represents the total number of data and L_j shows the class labels.

3. Results and Evaluations

In this section, seven well-known optimization algorithms were discussed in order to analytically evaluate the use of different metaheuristic algorithms for clustering trajectories. The results of the implementation were obtained from three datasets, including labomni, i5, and cross provided by the Computer Vision and Robotics Department of the University of California, San Diego (Figuer 3). (Morris and Trivedi, 2009).



(A) (B) (C)

Figure 3. Trajectory datasets (A) cross, (B) i5, and (C) labomni

The calculated Purity value for various algorithms is presented in Table 1. According to this table, the best result in the labomni data set is related to the PSO algorithm with 0.9234 Purity value, and the next ranks were achieved by WOA and SCA algorithms. In this data, the GWO algorithm did not perform well, and its Purity value is even lower than that of K-means. PSO algorithms produced acceptable results in i5 as all trajectories were correctly clustered, and the Purity value of 1 is achieved. Also, in this data, GWO was unable to improve the results of the K-means and along with SCA achieved the most mediocre results in i5 data. Finally, trajectory clustering results in cross show better performance of WOA, and PSO compared to poor GWO and SCA results. As it can be seen from Table 1, PSO with mean Purity value of 0.9527 has achieved the best outcomes in all three datasets.

Method	Dataset		
	labomni	i5	cross
K-means	0.8421	0.8250	0.7752
GWO	0.8230	00.7500	0.6821
PSO	0.9234	1	0.9347
SCA	0.8612	0.7500	0.6968
WOA	0.8900	0.8750	0.9453

Table 1. Purity number in different methods

Figure 4 compares different algorithms concerning the running time (seconds), which is another essential factor in determining the appropriate algorithm. According to this figure, although the PSO algorithm has shown excellent performance in trajectory clustering in terms of the Purity index, it resulted in high computation time. From Figure 4, we can see that the SCA has proven to be the fastest algorithm among the discussed ones, and all the other algorithms demonstrated computation times close to one another.

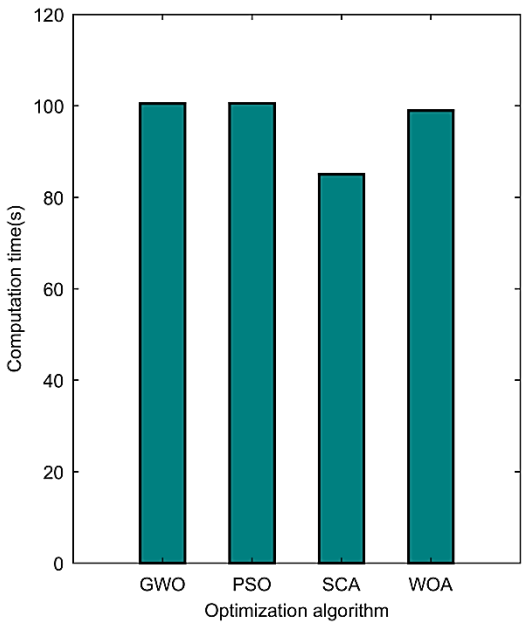


Figure 4. The computational time of the different algorithms

4. Conclusion

K-means algorithm is a simple clustering method with low computation time. However, it is sensitive to the initial cluster centroids. In this paper, after proposing an appropriate cost function, four well-known optimization algorithms in trajectory clustering were employed and their results in Purity index, as well as computation time, were compared and evaluated. Furthermore, to remove any data effects on results, three different data sets were used.

The results obtained in this paper have shown the superiority of PSO algorithm with regard to the purity index. However, the computation time of PSO was higher than other algorithms.

Furthermore, GWO and SCA have failed to show good performance in trajectory clustering.

REFERENCES

Ahmadyfard, A., Modares, H., 2008. Combining PSO and k-means to enhance data clustering, 2008 International Symposium on Telecommunications. IEEE, pp. 688-691.

Akasapu, A.K., Rao, P.S., Sharma, L., Satpathy, S., 2011. Density based k-nearest neighbors clustering algorithm for trajectory data. International Journal of Advanced Science and Technology 31.

Atev, S., Miller, G., Papanikolopoulos, N.P., 2010. Clustering of vehicle trajectories. IEEE transactions on intelligent transportation systems 11, 647-657.

Bacquet, C., Zincir-Heywood, A.N., Heywood, M.I., 2011. Genetic optimization and hierarchical clustering applied to encrypted traffic identification, 2011 IEEE Symposium on Computational Intelligence in Cyber Security (CICS). IEEE, pp. 194-201.

Frattasi, S., Della Rosa, F., 2017. Mobile positioning and tracking: from conventional to cooperative techniques. John Wiley & Sons.

Fu, Z., Hu, W., Tan, T., 2005. Similarity based vehicle trajectory clustering and anomaly detection, IEEE International Conference on Image Processing 2005. IEEE, pp. II-602.

Izakian, Z., Mesgari, M.S., Abraham, A., 2016. Automated clustering of trajectory data using a particle swarm optimization. Computers, Environment and Urban Systems 55, 55-65.

Kennedy, J., Eberhart, R., 1995. Particle swarm optimization (PSO), Proc. IEEE International Conference on Neural Networks, Perth, Australia, pp. 1942-1948.

Lee, J.-G., Han, J., Whang, K.-Y., 2007. Trajectory clustering: a partition-and-group framework, Proceedings of the 2007 ACM SIGMOD international conference on Management of data. ACM, pp. 593-604.

Likas, A., Vlassis, N., Verbeek, J.J., 2003. The global k-means clustering algorithm. Pattern recognition 36, 451-461.

Logesh, R., Subramaniyaswamy, V., Vijayakumar, V., Gao, X.-Z., Indragandhi, V., 2018. A hybrid quantum-induced swarm intelligence clustering for the urban trip recommendation in smart city. Future Generation Computer Systems 83, 653-673.

Lu, Y., Lu, S., Fotouhi, F., Deng, Y., Brown, S.J., 2004. Incremental genetic K-means algorithm and its application in gene expression data analysis. BMC bioinformatics 5, 172.

Mahdavi, M., Chehreghani, M.H., Abolhassani, H., Forsati, R., 2008. Novel meta-heuristic algorithms for clustering web documents. Applied Mathematics and Computation 201, 441-451.

Manning, C., Raghavan, P., Schütze, H., 2010. Introduction to information retrieval. Natural Language Engineering 16, 100-103.

Mirjalili, S., 2015. Moth-flame optimization algorithm: A novel nature-inspired heuristic paradigm. Knowledge-Based Systems 89, 228-249.

Mirjalili, S., 2016. SCA: a sine cosine algorithm for solving optimization problems. Knowledge-Based Systems 96, 120-133.

Mirjalili, S., Lewis, A., 2016. The whale optimization algorithm. Advances in engineering software 95, 51-67.

Mirjalili, S., Mirjalili, S.M., Hatamlou, A., 2016. Multi-verse optimizer: a nature-inspired algorithm for global optimization. Neural Computing and Applications 27, 495-513.

Mirjalili, S., Mirjalili, S.M., Lewis, A., 2014. Grey wolf optimizer. Advances in engineering software 69, 46-61.

Morris, B., Trivedi, M., 2009. Learning trajectory patterns by clustering: Experimental studies and comparative evaluation, 2009 IEEE Conference on Computer Vision and Pattern Recognition. IEEE, pp. 312-319.

Nanni, M., Pedreschi, D., 2006. Time-focused clustering of trajectories of moving objects. Journal of Intelligent Information Systems 27, 267-289.

Omran, M., Engelbrecht, A.P., Salman, A., 2005. Particle swarm optimization method for image clustering. International Journal of Pattern Recognition and Artificial Intelligence 19, 297-321.

Ossama, O., Mokhtar, H.M., El-Sharkawi, M.E., 2011. An extended k-means technique for clustering moving objects. Egyptian Informatics Journal 12, 45-51.

Palma, A.T., Bogorny, V., Kuijpers, B., Alvares, L.O., 2008. A clustering-based approach for discovering interesting places in trajectories, Proceedings of the 2008 ACM symposium on Applied computing. ACM, pp. 863-868.

Tork, H.F., 2012. Spatio-temporal clustering methods classification, Doctoral Symposium on Informatics Engineering. Faculdade de Engenharia da Universidade do Porto Porto, Portugal, pp. 199-209.