Received: 8 June 2017 | Revised: 8 April 2018 | Accepted: 25 May 2018

DOI: 10.1111/exsy.12318

ORIGINAL ARTICLE

WILEY Expert Systems

A geodesic distance-based approach for shape-independent data clustering using coalitional game

Behrouz Beik Khorasani | Mohammad Hossein Moattar | Yahya Forghani

Department of Software Engineering, Mashhad Branch, Islamic Azad University, Mashhad, Iran

Correspondence
Mohammad Hossein Moattar, Department of Software Engineering, Mashhad Branch, Islamic Azad University, Mashhad, Iran.
Email: moattar@mshdiau.ac.ir

Abstract

Different clustering approaches consider different aspects of quality including accuracy of cluster formation, speed of execution, and resource requirements. A major issue in data clustering is to consider the density of clusters and their structure. A challenge is to relax any assumption on the clusters shape such as sphericity, which is common in most partitioning approaches such as K-means. In this article, with the help of the coalitional game theory and the concept of geodesic distance calculation, a density-based shape-independent clustering approach is proposed. In addition to the emphasis on the application of game theory, we also pay attention to the relative neighbourhood of data, which is depicted using geodesic distance. Geodesic distance is a well-known measure for finding the manifold of data. The new idea supposes to discover any embedded structure of data and avoid finding only spherical clusters. The proposed approach is evaluated on a number of standard University of California, Irvine (UCI) datasets, and the results show the effectiveness of the proposed approach in comparison with some other approaches.

KEYWORDS

coalitional game, data clustering, data manifold, geodesic distance

1 | INTRODUCTION

Clustering is one of the basic human mental activities and helps to extract information from the unlabelled data. Clustering has different names based on the context in which it is used, such as unsupervised learning in pattern recognition, classification in biology and social sciences, and partition typology in graph theory. One of the problems in data clustering is the impact of data features. These features can result in an overlapping between the groups of data. In these circumstances, evolutionary methods are one of the favourite options for cluster formation. Game theory is a multifaceted approach for learning, which is able to give an explanation for the behavioural characteristics of a wide range of existing learning algorithms (Li, Chen, He, & Jiang, 2010). Ideal clustering needs to guarantee an internal condition which is minimizing the within cluster distances and an external condition which is maximum between cluster distances (Zheng, Liu, & Qi, 2012). It has been claimed that game theory is a good means for achieving the mentioned conditions and forming good quality clusters (Li et al., 2010; Heileman, Caudell, Hush, Krause, & Verzi, 2015). Unlike classic games, where players show quite logical behaviour and always choose the optimal strategy in complex environments, coalitional game theory combines the traditional ones with the idea of evolutionary game theory and assumes rationalities such as limited information and cognitive environment.

Also, the distance or affinity metric between data points plays a pivotal role in clustering accuracy and speed. The aim of this paper is to propose an approach which not only constructs good quality clusters but also avoids merely spherical clusters. For this purpose, in this paper, complex data structure discovery is achieved by finding the data manifold using geodesic distance. As the main contribution of the proposed approach, geodesic distance between data points is used as the clustering score. The remainder of this paper is as follows. Section 2 summarizes the related

works. Section 3 introduces materials and methods. The proposed approach is detailed in Section 4. Section 5 discusses the experiments and results, and finally, the conclusion and future works are mentioned in Section 6.

2 | RELATED WORKS

Clustering has different application from unsupervised learning (Pelillo & Bulo, 2014) to outlier removal and instance reduction (Hamidzadeh, Monsefi, & Sadoghi Yazdi, 2014; Hamidzadeh, Monsefi, & Sadoghi Yazdi, 2015). There are many different approaches from which partitioning methods are the most popular. Also, modelling approaches such as support vector data description (Tax & Duin, 2004) have been also proposed for clustering. Some of the recent works have proposed game theory concepts as a clustering approach. Hsiao and Chang (2013) proposed a game theory clustering approach in which each data point is a player. After applying the cooperative rules of betrayal and isolation, a partition graph is created. Garg and Narahari (2013) proposed a new game theory clustering method using Shapely index. In Garg and Narahari (2013), the similarity between two members of a cluster is determined by the distance from the centroid. Guruacharya, Niyato, Bennis, and Kim (2013) used a new coalitional game model for data clustering. Kumar and Reddy (2016) pointed out the well-known density-based spatial clustering of applications with noise (DBSCAN) method for density-based and shape-independent clustering. Rota Bulo and Pelillo addressed a hypergraph clustering problem that extracts groups with maximum dependency (Rota Bulo & Pelillo, 2013). This approach proposes a novel clustering approach based on the game theory. Unlike standard clustering methods, the approach proposed in Rota Bulo and Pelillo (2013) does not need to know the number of clusters beforehand. Koloniari and Pitoura (2012) proposed a game theoretic approach to cluster the social network's users. Zheng et al. (2012) proposed the Bayesian games for effective routing and energy consumption in wireless sensor networks. Another work which proposed game theory for clustering is Li et al. (2010) in which automatic formation of clusters is considered. Imen, Radjef, and Kechadi (2014) introduced a clustering method with the help of multiobjective games. The number of clusters is calculated automatically. This method deals with data clustering in three stages.

In contrast with classic games where players have shown quite logical behaviour and always choose the optimal strategy in complex environments, the coalitional games, which have limited information and limited environment, do not guarantee quite logical decisions. In these games, perfect rationality might create so-called backward induction and cause stucking in the paradox of infinite repeated games. Thus, different dynamic rules can be used to define logical boundaries and describe the behaviour of the players in the coalitional game theory.

There are some works concerning shape independent data clustering which try to avoid spherical cluster formation. Galluccio, Michel, Comon, Kliger, and Hero (2013) proposed a new distance calculation using Prime algorithm for finding nonconvex clusters. Stuetzlea and Nugenta (2012) proposed a nonparametric method to detect clusters through data density. This approach finds an embedded graph approach to estimate the density of data. Fukui, Ono, Megano, and Numao (2013) proposed a semisupervised method of evolutionary distance metric learning. They also developed their method and provided an evolutionary approach for multiobjective clustering (Megano, Fukui, Numao, & Ono, 2015). There are also modification on support vector data description such as weighting or using evolutionary algorithms (Hamidzadeha, Sadeghi, & Namaei, 2017; Sadeghi & Hamidzadeh, 2018) to enhance the approach for clustering and classification problems.

About the geodesic distance popularity among new articles, we should refer to Murari et al. (2013). Király, Vathy-Fogarassy, and Abonyi (2016) assumed a non-linear manifold using the geodesic distance between data points and used C-Medoid as the clustering algorithm. Chen et al. (2015) proposed K-means clustering with geodesic distance. Ester, Kriegel, Sander, and Xu (1996) proposed DBSCAN approach to discover clusters with different shapes. Also, a recent approach using geodesic distance as the metric is proposed in Heidari and Moattar (2017), which has applied this measure for improve Gaussian process latent variable model for clustering and classification.

3 | MATERIALS AND METHODS

3.1 | Game theory

Game theory was introduced in Neumann and Morgenstern (1944) and then different algorithms were developed to solve problems with this theory. Methods such as asymmetric games (Veller & Hayward, 2016) and simultaneous games (Hornborg, 2003; Zhang, Ramirez-Marquez, & Wang, 2015) are of this kind. By 1950, John Nash published a famous article and suggested the non-cooperative game theory, in which the Nash equilibrium concept was proposed and the equilibrium was introduced. Given that the non-cooperative games were based on mathematics, players in the game must act completely rational or even logical. Otherwise, they may not find the Nash equilibrium. Coalitional games are defined with the concept of evolutionarily strategy and are often used to describe the evolution of the social behaviours in animals (Li et al., 2010). The main concept of game theory is that the players are always competing with each other to for more gain.

3.1.1 | Coalitional games for clustering

A clustering method based on the coalitional game theory is proposed in Li et al. (2010). The best players usually try to help each other to earn points. A higher profit or gain makes a player out of a coalition to join another coalition or form a new coalition. In coalitional games, the players try to help each other make their profit. So the strategy of each player affects the other teams' rating and the rating of each player influences the strategy of the coalition. The approach which is introduced in Li et al. (2010) applies inverse Euclidean distance to rate the players in a coalition.

However, we suppose that using Euclidean metric may lead to spherical clusters, Hence, other shapes and cluster formations may be difficult to be recovered using Euclidean metric. Hence, in the proposed approach, coalitions are formulated and scored using a geodesic distance, to consider the unknown distribution of data points.

3.2 | Locally linear embedding

Locally linear embedding (LLE; Rovis & Saul, 2000) is a method to reduce the size of the feature space in two steps. In the first stage, using Equation 1 weight matrix (W) is calculated on data points X.

$$E(W) = \min\left(\sum_i \left|\vec{x}_i - \sum_j w_{ij}\vec{x}_j\right|^2\right). \tag{1}$$

In which $\vec{x}_j$ are the neighbouring points of $\vec{x}_i$.Then, weights, w_{ij}, calculated from Equation 1 are fixed, so that the relational distance between the lower dimension data (Y) and original data (X) is preserved as in Equation 2.

$$\phi(W) = \min\left(\sum_i \left|\vec{y}_i - \sum_j w_{ij}\vec{y}_j\right|^2\right). \tag{2}$$

3.3 | Geodesic distance

Clustering measures will have direct impact on the quality of clustering. Various articles use different criteria for this purpose such as the Euclidean distance, which is denoted in Equation 3.

$$D_e(x_i, x_j) = \sqrt{\sum_{k=1}^{n}(x_{ik} - x_{jk})^2}. \tag{3}$$

In this regard, n is the dimensionality of data points. Another possible criterion is Geodesic distance, which seeks to find the shortest path between two data points through a graph. The geodesic distance between the two points will be equal to the sum distance in the shortest path between the points. To calculate the geodesic distance, the shortest path between the points is obtained using scanning algorithms. Different algorithms have been proposed in this regard. In this study, we will use the Floyd–Warshall algorithm.

For example, as denoted in Figure 2, to obtain the geodesic distance between two vertices V1 and V3, we traverse the shortest path through V2 (red lines). The shortest path between V1 and V5 (blue line) passes through V2. Considering that the red path passes through one point (i.e., V2) whereas the blue path has two intermediate vertices (i.e., V2 and V4), the density through the blue path is higher, which expresses that vertices V1, V2, V4, and V5 are more densely distributed as compared with V1, V2, and V3. Having these observations, we can conclude that V1 and V5 are geometrically closer than V1 and V3. However, considering the Euclidean distance (denoted with green lines in Figure 1), V1 and V3 seem to be closer.

3.3.1 | Floyd–Warshall algorithm

Floyd–Warshall algorithm seeks to find the shortest path between all pairs of vertices in a single run. Repeating this process ultimately results in matrix of shortest paths between vertices in the graph. Equation 4 explains how to calculate the path in a recursive manner.

$$\text{Shortest\_Path}(i, j, k) = \min(\text{Shortest\_Path}(i, j, k-1), \text{Shortest\_Path}(i, k, k-1) + \text{Shortest\_Path}(k, j, k-1)) \tag{4}$$

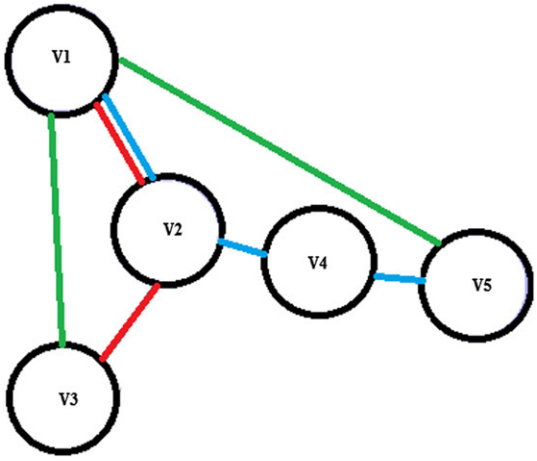

FIGURE 1 An example of geodesic distance calculation

In this study, Algorithm (1) is applied in order to calculate the geodesic distance between the data points. The inputs of the algorithm are the Euclidean distances between all pairs of data points.

Algorithm (1) Geodesic distance calculation with Floyd–Warshall method

```
Nearest_Path = Floyd_warshall (Euclidean_Distance)
{
   Number_Of_Nodes = size (Euclidean_Distance,1);
   Nearest_Path = Euclidean_Distance;
   for k = 1:Number_Of_Nodes.
     for i = 1:Number_Of_Nodes.
       for j = 1:Number_Of_Nodes.
         Nearest_Path (i,j) = min (Nearest_Path (i,j),Nearest_Path (i,k) + Nearest_Path (k,j));
}
```

4 | THE PROPOSED APPROACH

As discussed before, in the proposed coalitional game, any given point of data X is defined as a player and the coalition represents a cluster. We assume that each coalition has a set X with N players in an n-dimensional space. In this space, it is desirable to define a measure of proximity between the data points (i.e., are close to each other) so that to satisfy a specific condition. In this study, we have applied the geodesic distance between the data points instead of the conventional methods of calculating the Euclidean distance. Geodesic distance aims to capture the structural neighbourhood of data. For example, it can be seen in Figure 2 that O1 is farther from O5 than X1, but O1 and O5 are both located in the same neighbourhood considering the manifold and may fall in the same cluster. Linking all data points together, there is a complete graph in which the weight of edges is the direct (Euclidean) distance between the two ends of the edge.

Therefore, having found the shortest path between two data points, the normalized geodesic distance between the points is calculated as follows:

$$D_g(x_i, x_j) = \frac{Nearest\_Path(x_i, x_j)}{n+1}. \tag{5}$$

In which, n is the number intermediate data points along the shortest path between x_i and x_j and aim to discount the impact of straight path between the two points. As already mentioned, rating the players is the most important part of the coalitional games. One of the important issues of this method is how to calculate the score of each player. The score of player i (the ith data point) is denoted by u (i) in Equation 6.

$$u(i) = \sum_{j \in \Gamma(i)} R(i,j) \tag{6}$$

$$R(i,j) = p(x_i, x_j) \times Deg(j) / D_g(x_i, x_j). \tag{7}$$

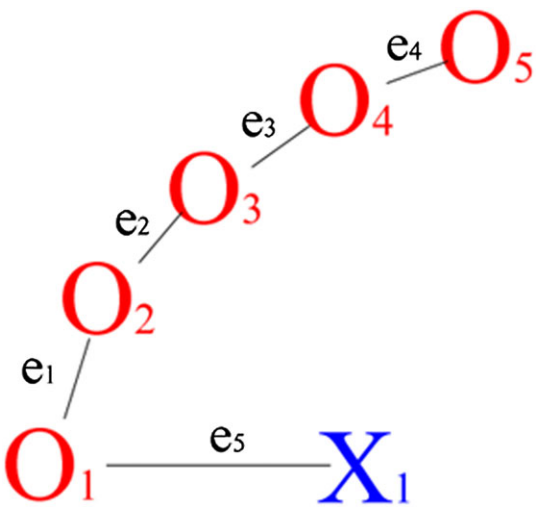

FIGURE 2 Placement of the data on a manifold

In which $p(x_i, x_j)$ denotes the inverse number of members in the cluster containing both x_i and x_j, Deg(j) is the number of players in coalition with jth data point and Γ(i) is the cluster which include ith data point. Therefore, the strength of a coalition is proportional to the inverse number of data points and geodesic distance between data laying in the same cluster. In the first phase, initial clusters are formed using k-Nearest Neighbors (KNN) which mean that

$$p(x_i, x_j) = \frac{1}{k}. \tag{8}$$

Concluding on the scoring function of each player, the proposed geodesic coalition-based game theory clustering (GCGC) is summarized as follows.

```
Clustering with GCGC (X: dataset, k: number of clusters)
{
   Euc_dist = Calculate Euclidean Distance (X)
   GeoDifference = Calculate Geodesic Distance with Floyd_warshall method (Euc_dist)
   Y = LLE Dimensionality reduction (GeoDifference) % If necessary
   Euc_dist = Calculate Euclidean Distance (Y)
   GeoDifference = Calculate Geodesic Distance with Floyd_warshall method (Euc_dist)
   [Clusters0, Edges0] = Compute KNN with GeoDifference
   While noChange (Clusters0, Edges0)
      Degik = Compute Degree for Yik
      PayOffsik = Compute Payoffs for Yik
      [Clusters,Edges] = Expand (Clusters, PayOffs, Edges,GeoDifference)
      Edges = CutCluster (Clusters, Edges);
   }
}
```

The algorithm ends if the cluster membership of data points is not changed in two successive iterations. In the above algorithm, LLE method is used to find the embedding manifold and also reduce the data dimensionality to decrease the computational cost of distance calculation.

As a simple illustrated example, Figure 3 shows two clusters of points, that is, red points and blue points. These clusters are not spherical, and the data points are laid on two curves. In this example, the Euclidean distance between O1 and O5 is greater as compared with the distance between O5 and X1. Also, considering the centroid of red cluster, X1 is more likely to be included in red cluster than O1 or O5 points, because the Euclidean distance between X1 and the centroid of red cluster is less than the Euclidean distance between O1 or O2 and the centroid of red cluster. We have

$$D_e(O5, O1) = \sqrt{4^2 + 0^2} = 4$$

$$D_e(O5, X1) = \sqrt{2^2 + 1^2} = 2.236.$$

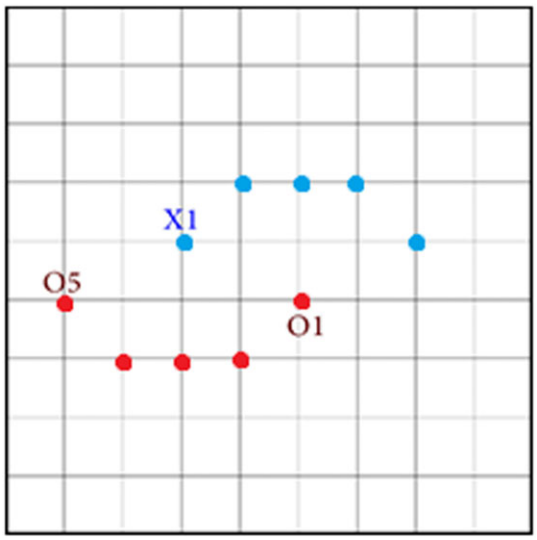

FIGURE 3 A toy example to show the contribution of the proposed geodesic-based coalition game clustering approach

However, considering the proposed Geodesic distance, the distances between these points are as follows:

$$D_g(O5, X1) = \frac{1.416 + 1 + 2}{3} = 1.472$$

$$D_g(O5, O1) = \frac{1.416 + 1 + 1 + 1.416}{4} = 1.208.$$

Therefore, if Geodesic distance is used, O1 is closer to O2 than X. Thus, O5 and O1 are grouped in the same clusters whereas blue points will be grouped together, when using our proposed geodesic-based coalition game clustering approach.

5 | EXPERIMENTS

5.1 | Datasets

Eight UCI repository datasets are applied for the experiments and assessments, which are detailed in Table 1. Figure 4 depicts datasets distribution and formations in this study.

The experiments are performed on a 64 bit core i5 2.5 GHz processor with 6 GB of RAM and Windows 7 Ultimate operating system.

5.2 | Evaluation criteria

The superiority of the clustering methods is usually evaluated in terms of speed, hardware and software limitations, and also simplicity and the complexity of the calculations. In this study, precision of the method is considered and therefore, the proposed GCGC method is compared with traditional methods such as game-based clustering (Li et al., 2010), Kmeans, PCA_Kmeans and DBSCAN (Kumar & Reddy, 2016) in terms of the purity index as denoted in Equation 8.

$$\text{Purity}(\Omega, C) = \frac{1}{N} \times \sum_{n=1}^{k} \max_j \left| w_k \cap c_j \right|. \tag{8}$$

In which, N is denoted as the number of data points, Ω is the ground truth clustering, and C is the resulting clustering. As mentioned before, LLE method is used to reduce the data size.

5.3 | Experimental results

In the proposed approach, K nearest neighbours are grouped in the first stage. For a small value of K when the network grows, a large number of small clusters appear. On the other hand, a big value of K provides more neighbours, which causes larger clusters. In these experiments, K is selected as the number of data samples divided by the preset number of cluster. The experimental results are indicated in Table 2. As denoted in Table 2, the proposed approach is almost superior to other methods in terms of accuracy. The proposed approach is even superior as compared with DBSCAN algorithm (Kumar & Reddy, 2016), which shows its appropriate performance. A main drawback of DBSCAN algorithm is the tuning of its parameters which is not an issue in the proposed approach. The only dataset for which the accuracy of the proposed method is not satisfactory is Iris dataset, which may be due to its cluster simplicity.

Figure 5 visualizes the output of two clustering techniques. As can be seen, K-means methods tend to form circular clusters whereas the proposed GCGC approach tries to fit the data to the manifold.

TABLE 1 The experimental datasets in detail

Dataset	Number of features	Number of clusters	Number of records
Breast	10	2	699
Ionosphere	34	2	351
Soybean	35	4	47
Iris	3	3	150
Wine	13	3	178
Path-based	2	3	300
Spiral	2	3	312
Flame	2	2	240
Anuran Calls	22	4	7,195

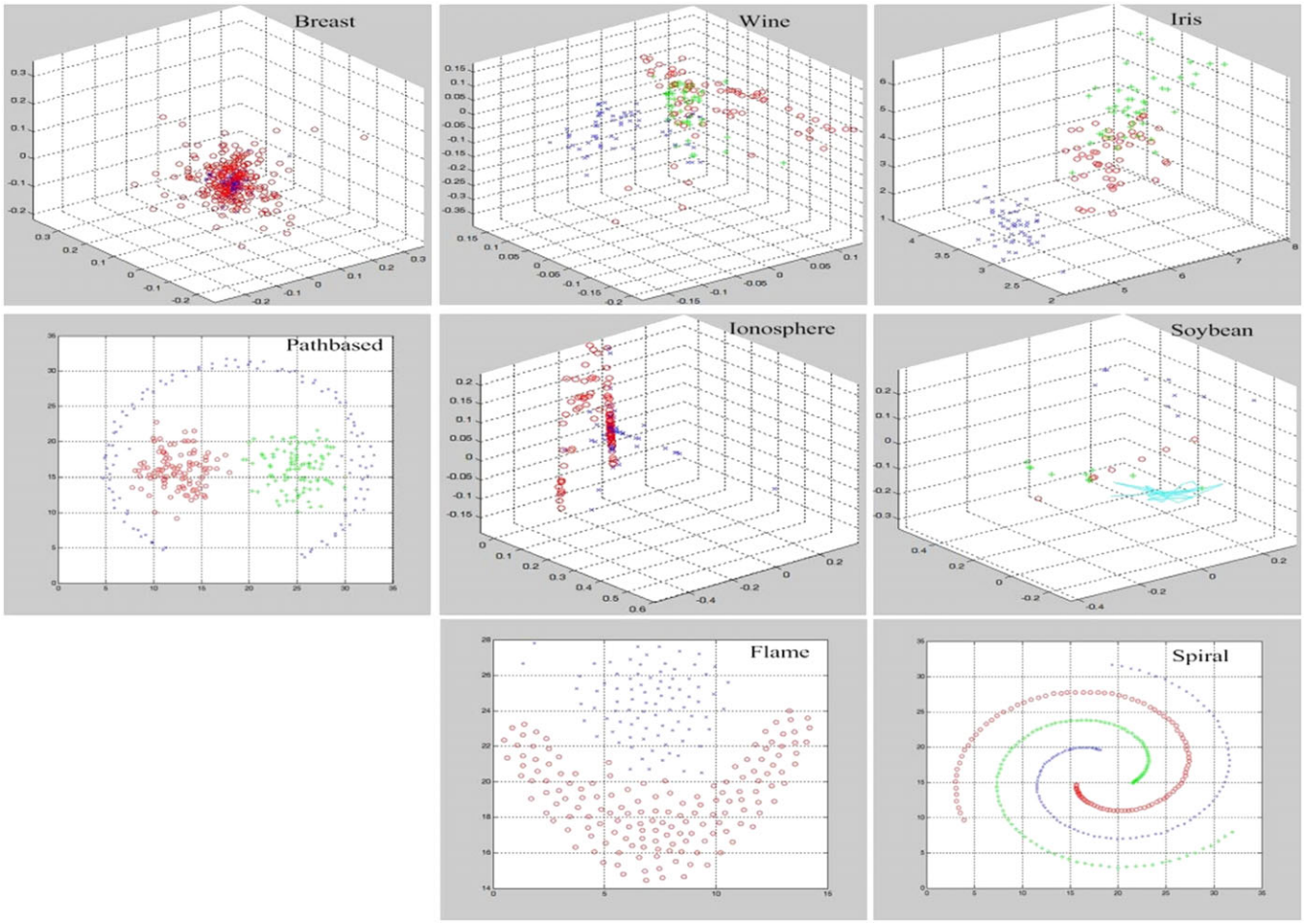

FIGURE 4 Visualization of the reduced dimension datasets

TABLE 2 Experimental purity results of the proposed approach compared with other approaches

	Iris	Wine	Breast	Soybean	Ionosphere	Path-based	Spiral	Flame	Anuran Calls
GCGC	80.41	76.40	94.85	97.83	86.61	81.33	99.68	96.67	60.07
Game (Li et al., 2010)	80.41	61.80	94.95	93.62	74.64	80.33	98.08	91.67	59.50
K-means	89.30	70.20	93.85	68.10	71.02	43.00	30.13	84.17	57.98
PCA-KM	88.70	70.20	93.85	72.30	71.00	46.33	30.13	84.17	57.98
DBSCAN (Kumar & Reddy, 2016)	93.40	62.92	65.09	38.29	75.41	81.06	99.80	72.43	45.08

Note. GCGC: geodesic coalition-based game theory clustering.

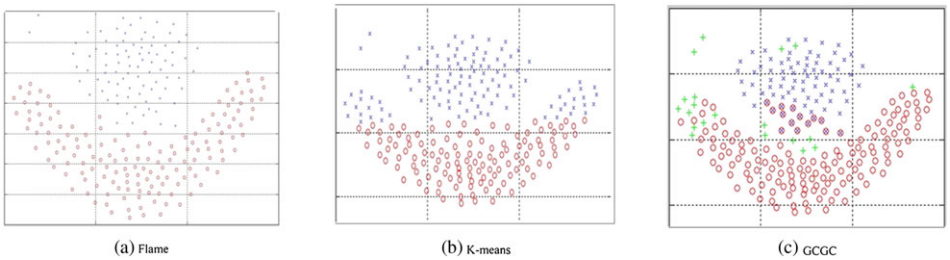

FIGURE 5 Clustering result of K-means and GCGC on flame dataset. GCGC: geodesic coalition-based game theory clustering

As it is clearly shown in Figure 6, the approach proposed in Li et al. (2010) fails to capture the spatial relation between clusters while in the proposed method clusters span the manifold of data. In this way, all the immediate neighbours of each player, regardless of the distribution, are placed in the same cluster.

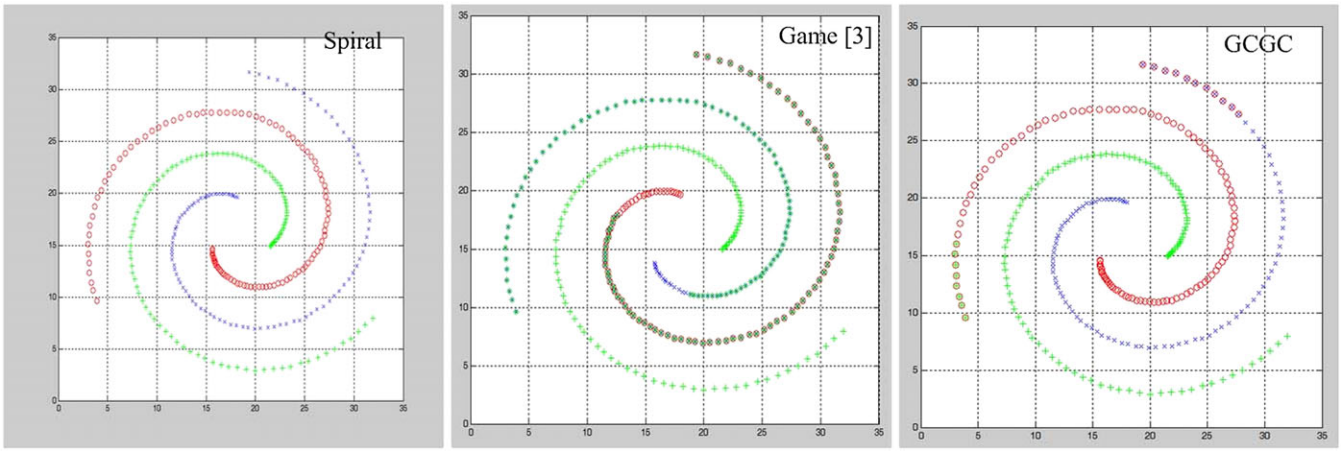

FIGURE 6 Clustering result of method in Li et al. (2010) and the proposed GCGC on flame dataset. GCGC: geodesic coalition-based game theory clustering

Our main aim for proposing geodesic distance is to cluster the data on the manifold. As it is seen in Figure 7, the data manifold is fully recognized and our main goal is satisfied. However, in Figure 7, because of the proximity of data to other clusters, a number of data is simultaneously identified in other clusters.

The proposed method causes a gradual increase in the value of K and avoids forming small clusters. Figure 8 shows the effect of changing K for flame dataset. The larger the value of K is, the more coherent the clusters will be formed. Therefore, selecting a proper value for K is of great importance. In this study, we tested the method with different values of K and reported the best results.

Table 3 shows the clustering times of the proposed approach and other clustering approaches on different datasets. As it can be seen, the proposed approach is computationally expensive, which is due to its game-based nature. The same observations can be seen for the approach

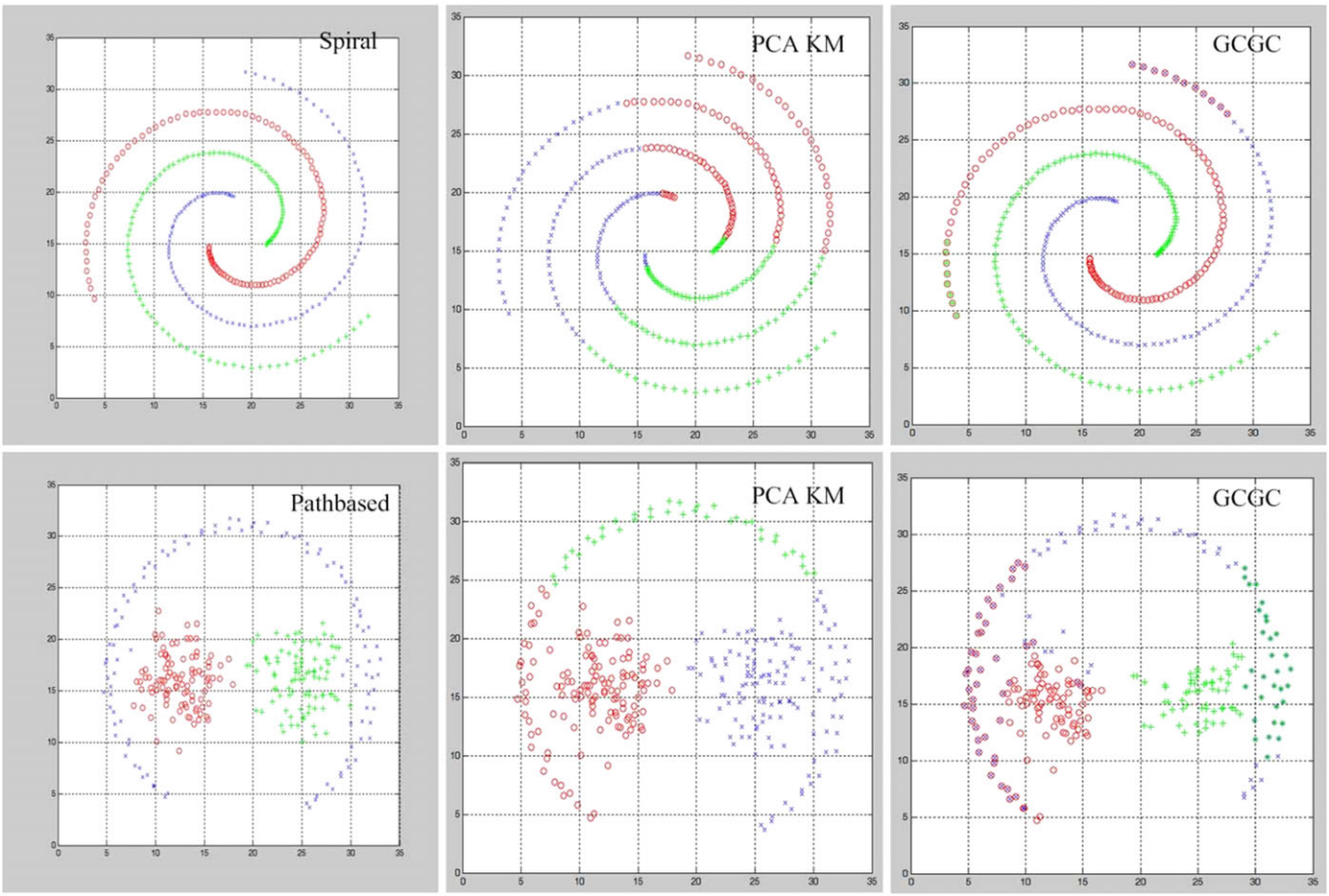

FIGURE 7 Clustering result of PCA-Kmeans and GCGC on spiral and path-based datasets. GCGC: geodesic coalition-based game theory clustering

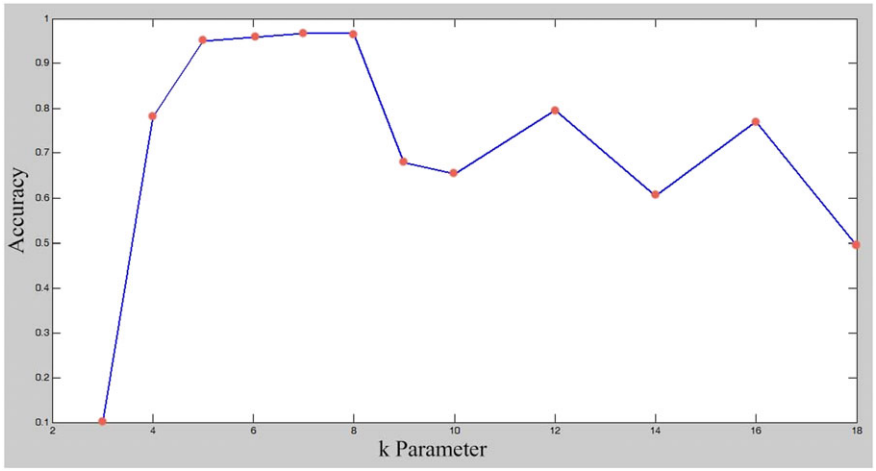

FIGURE 8 Effect of K on clustering accuracy of GCGC for flame dataset. GCGC: geodesic coalition-based game theory clustering

TABLE 3 Processing time of the proposed approach compared with other approaches on evaluation data (in seconds)

	Iris	Wine	Breast	Soybean	Ionosphere	Path-based	Spiral	Flame
GCGC	2.44	0.57	7.37	0.11	15.26	0.79	0.92	0.57
Game (Li et al., 2010)	2.16	0.55	6.89	0.11	15.12	0.69	0.89	0.53
K-means	0.01	0.01	0.11	0.01	0.02	0.04	0.02	0.03
PCA-KM	0.04	0.02	0.17	0.05	0.19	0.02	0.04	0.15

Note. GCGC: geodesic coalition-based game theory clustering.

proposed in Li et al. (2010). In comparison, the approaches based on K-means have considerably lower complexity, which is the main drawback of the proposed approach.

To evaluate the significance of the resulting accuracies using the proposed approach compared with other methods, we applied Wilcoxon ranked sum test, which is calculated with the following formulas:

$$W^- = \text{sign}\left(\sum_{\text{sign}(X_{2,i}-X_{1,i})<0}\left[\text{sign}(X_{2,i}-X_{1,i})^*R_i\right]\right) \tag{9}$$

$$W^+ = \sum_{\text{sign}(X_{2,i}-X_{1,i})>0}\left[\text{sign}(X_{2,i}-X_{1,i})^*R_i\right] \tag{10}$$

$$W = \min(W^-, W^+). \tag{11}$$

In which $X_{2,\ i}$ is the ith output of the jth sample (i.e., evaluated approach) and R_i is the rank of the output in the tail. Also *sign(.)* is the sign function and $W^-$and $W^+$are the negative and positive Wilcoxon value, respectively. The outputs of the test are denoted in Table 4. As usual, H0 hypothesis is "There is no significant difference between approaches" whereas H1 hypothesis is "There is significant difference between approaches."

Having nine experiments for each approach, the Wilcoxon critical value for two-tails test with α = 0.05 level of significance is 6. In Table 4, in every places that *W* is lower than the critical value (denoted in bold), the H0 hypothesis is rejected, which means that there is a significant difference between the approaches. Hence, although other approaches are not much different pairwisely, the proposed approach significantly outperforms other approaches.

TABLE 4 Outputs of Wilcoxon ranked sum test (W) for the evaluated approaches

Algorithm	GCGC	Game (Li et al., 2010)	K-means	PCA-KM	DBSCAN (Kumar & Reddy, 2016)
GCGC	—	2	4	4	4
Game (Li et al., 2010)		—	11	11	8
K-means			—	17	20
PCA-KM				—	20
DBSCAN (Kumar & Reddy, 2016)					—

Note. GCGC: geodesic coalition-based game theory clustering.

5.4 | Discussions

Having the above observations, we can claim that the proposed approach performs well for datasets with geometrical forms and not necessarily spherical. Of course, considering the intrinsic behaviour of the proposed algorithm, the approach also performs satisfactory on datasets with spherical clusters. As denoted in the experimental results, the clustering results of the proposed approach is considerably lower on Iris dataset. We suppose that other than the distribution of data points in Iris dataset, the main cause of this observation is the balance of Iris dataset. With these observations, we suppose that the approach highly outperforms other approaches on imbalance datasets. However, we emphasize that the approach is still appropriate for many test benchmarks. We suppose that other than inconsiderable improvement on balance datasets or spherical shape clusters, the main drawback of the proposed approach is its high complexity, which is denoted in Table 3. Using instance selection approaches can be applied in the proposed framework as the future works to decrease the computational complexity of the approach.

In the proposed approach, a coalition game-based clustering approach is proposed, which clusters data points in a game optimization framework. In this framework, each player searches for better positions considering the strategy and position of other players to gain better score. Because in the proposed approach, data points are considered as the players; to begin the algorithm, all data points (players) should be ready. It means that the order of the instances does not have any impact on the result of the proposed approach. Also, the approach is not appropriate for online clustering and the data points should be given in batch. Adapting the approach for online clustering may be a direction for future works.

6 | CONCLUSION

Distance measure plays a key role in the accuracy and speed of clustering. In this study, the distance calculation between the data points is discussed. The method of calculating the Euclidean distance is substituted by finding geodesic distance to avoid circular and spherical clusters. The proposed approach needs to search for the data manifold to achieve this purpose. As denoted in the experiments, due to the proximity of the clusters, a number of data which do not lie on the same manifold have been clustered incorrectly.

The membership of each data point to any cluster is perceived as a function of the members of that cluster. In other words, the more the number of neighbours in a cluster, the greater the probability of falling in that cluster. In this regard, the membership degree of each data point and the increment in the consistent members has a direct impact on cluster consistency. So this term controls the growth of clusters and prevents joining too many nodes to one cluster.

In some experiments, data may be felt into more than one cluster. Future works may include finding some measure to prevent this problem. Considering the high processing time of the proposed approach as compared with other methods, reducing the time complexity of the proposed approach through using appropriate heuristics may be an important direction of future works. Also, as mentioned in the previous section, the approach is not designed for online clustering due to the batch nature of the proposed game algorithm. Another suggestion for the future works may be to adapt the game algorithm for online clustering in the proposed framework.

ORCID

Mohammad Hossein Moattar http://orcid.org/0000-0002-8968-6744
Yahya Forghani http://orcid.org/0000-0002-3218-5666

REFERENCES

Chen, X., Zhu, Y., Li, F., Zheng, Z., Chang, E., & Ma, J. (2015). Accurate segmentation of touching cells in multi-channel microscopy images with geodesic distance based clustering. *Neurocomputing*, *149*, 39–47. https://doi.org/10.1016/j.neucom.2014.01.061

Ester, M., Kriegel, H. P., Sander, J., & Xu, X. (1996). A density-based algorithm for discovering clusters. *KDD-96 Proceedings on AAI*, 226–231.

Fukui, K.I., Ono, S., Megano, T., & Numao, M. (2013). Evolutionary distance metric learning approach to semi-supervised clustering with neighbor relations. IEEE 25th International Conference on Tools with Artificial Intelligence, Herndon, VA.

Galluccio, L., Michel, O., Comon, P., Kliger, M., & Hero, A. O. (2013). Clustering with a new distance measure based on a dual-rooted tree. *Information Sciences*, *251*, 96–113. https://doi.org/10.1016/j.ins.2013.05.040

Garg, V. K., & Narahari, Y. (2013). Novel biobjective clustering (BiGC) based on cooperative game theory. *IEEE Transactions on Knowledge and Data Engineering*, *25*(5), 1070–1082. https://doi.org/10.1109/TKDE.2012.73

Guruacharya, S., Niyato, D., Bennis, M., & Kim, D. I. (2013). Dynamic coalition formation for network MIMO in small cell networks. *IEEE Transactions on Wireless Communications*, *12*(10), 5360–5372. https://doi.org/10.1109/TWC.2013.090513.130516

Hamidzadeh, J., Monsefi, R., & Sadoghi Yazdi, H. (2014). LMIRA: Large margin instance reduction algorithm. *Neurocomputing*, *145*, 477–487. https://doi.org/10.1016/j.neucom.2014.05.006

Hamidzadeh, J., Monsefi, R., & Sadoghi Yazdi, H. (2015). IRAHC: Instance reduction algorithm using hyper-rectangle clustering. *Pattern Recognition*, *48*, 1878–1889. https://doi.org/10.1016/j.patcog.2014.11.005

Hamidzadeha, J., Sadeghi, R., & Namaei, N. (2017). Weighted support vector data description based on chaotic bat algorithm. *Applied Soft Computing*, *60*, 540–551. https://doi.org/10.1016/j.asoc.2017.07.038

Heidari, M., & Moattar, M. H. (2017). Discriminative geodesic Gaussian process latent variable model for structure preserving dimension reduction in clustering and classification problems. *Neural Computing and Applications*, 1–14. https://doi.org/10.1007/s00521-017-3273-4

Heileman, G., Caudell, T., Hush, D., Krause, K., & Verzi, S. (2015). *A game-theoretic support vector machine classifier*. New Mexico: The University of New Mexico.

Hornborg, A. (2003). Cornucopia or zero-sum game? The epistemology of sustainability. *Journal of World Systems Research*, *9*(2), 205–216. https://doi.org/10.5195/jwsr.2003.245

Hsiao, P. C., & Chang, L. W. (2013). Image denoising with dominant sets by a coalitional game approach. *IEEE Transactions on Image Processing*, *22*(2), 724–738. https://doi.org/10.1109/TIP.2012.2222894

Imen, H., Radjef, M. S., & Kechadi, M. T. (2014). Clustering based on sequential multi-objective games. In *Data warehousing and knowledge discovery* (pp. 369–381). Switzerland: Springer. https://doi.org/10.1007/978-3-319-10160-6_33

Király, A., Vathy-Fogarassy, A., & Abonyi, J. (2016). Geodesic distance based fuzzy c-medoid clustering—Searching for central points in graphs and high dimensional data. *Fuzzy Sets and Systems*, *286*, 157–172. https://doi.org/10.1016/j.fss.2015.06.022

Koloniari, G., & Pitoura, E. (2012). A game-theoretic approach to the formation of clustered overlay networks. *IEEE Transactions on Parallel and Distributed Systems*, *23*(4), 589–597. https://doi.org/10.1109/TPDS.2011.155

Kumar, K. M., & Reddy, A. R. M. (2016). A fast DBSCAN clustering algorithm by accelerating neighbor searching using Groups method. *Pattern Recognition*, *58*, 39–48. https://doi.org/10.1016/j.patcog.2016.03.008

Li, Q., Chen, Z., He, Y., & Jiang, J. (2010). A novel clustering algorithm based upon game on evolving network. *Expert System with Applications*, *37*, 5621–5629. https://doi.org/10.1016/j.eswa.2010.02.050

Megano, T., Fukui, K.I., Numao, M., & Ono, S. (2015). Evolutionary multi-objective distance metric learning for multi-label clustering. 2015 IEEE Congress on Evolutionary Computation (CEC), Sendai.

Murari, A., Boutot, P., Vega, J., Gelfusa, M., Moreno, R., & Verdoolaege, G. (2013). Clustering based on the geodesic distance on Gaussian manifolds for the automatic classification of disruptions. *Journal of IOP science on Nuclear Fusion*, *53*(3), 033006. https://doi.org/10.1088/0029-5515/53/3/033006

Neumann, J., & Morgenstern, O. (1944). *Theory of games and economic behavior*. Princeton University Press: Princeton.

Pelillo, M., & Bulo, S. (2014). *Registration and recognition in image and videos*. Berlin Heidelberg: Springer-Vverlag.

Rota Bulo, S., & Pelillo, M. (2013). A game-theoretic approach to hypergraph clustering. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *35*(6), 1312–1327. https://doi.org/10.1109/TPAMI.2012.226

Rovis, S. T., & Saul, L. K. (2000). Nonlinear dimensionality reduction by locally linear embedding. *Science*, *290*(5500), 2323–2326. https://doi.org/10.1126/science.290.5500.2323

Sadeghi, R., & Hamidzadeh, J. (2018). Automatic support vector data description. *Soft Computing*, *22*(1), 147–158. https://doi.org/10.1007/s00500-016-2317-5

Stuetzlea, W., & Nugenta, R. (2012). A generalized single linkage method for estimating the cluster tree of a density. *Journal of Computational and Graphical Statistics*, *19*, 397–418.

Tax, D. M., & Duin, R. P. (2004). Support vector data description. *Machine Learning*, *54*(45–66), 2004–2066. https://doi.org/10.1023/B:MACH.0000008084.60811.49

Veller, C., & Hayward, L. K. (2016). Finite-population evolution with rare mutations in asymmetric games. *Journal of Economic Theory*, *162*, 93–113. https://doi.org/10.1016/j.jet.2015.12.005

Zhang, C., Ramirez-Marquez, J., & Wang, J. (2015). Critical infrastructure protection using secrecy—A discrete simultaneous game. *European Journal of Elsevier Operational Research*, *242*(1), 212–221. https://doi.org/10.1016/j.ejor.2014.10.001

Zheng, G., Liu, S., & Qi, X. (2012). Clustering routing algorithm of wireless sensor networks. *Journal of Systems Engineering and Electronics*, *23*(1), 154–159. https://doi.org/10.1109/ICICTA.2010.343

Behrouz Beik Khorasani has received his M.S. degree from Islamic Azad University, Mashhad Branch, Mashhad, Iran. His research interests include artificial intelligence and machine learning.

Mohammad Hossein Moattar has received his PhD in 2010 from Amirkabir University of Technology, Iran. Presently, he is working as assistant professor of the Department of Computer Engineering in Islamic Azad University, Mashhad branch, Iran. His research areas include artificial intelligence, machine learning, and pattern recognition.

Yahya Forghani has received his PhD in 2013 from Ferdowsi University of Mashhad, Iran. He has 12 years of teaching experience. Presently, he is working as assistant professor of the Department of Computer Engineering in Islamic Azad University, Mashhad branch, Iran. His research areas include artificial intelligence, machine learning, and image processing.

How to cite this article: Beik Khorasani B, Moattar MH, Forghani Y. A geodesic distance-based approach for shape-independent data clustering using coalitional game. *Expert Systems*. 2018;e12318. https://doi.org/10.1111/exsy.12318