Pure Appl. Geophys.

https://doi.org/10.1007/s00024-020-02618-6

Pure and Applied Geophysics

Robust Earthquake Cluster Analysis Based on K-Nearest Neighbor Search

Hamid Reza Samadi,[1] Roohollah Kimiaefar,[1,2] and Alireza Hajian[2]

Abstract—Grouping of earthquakes into distinct clusters is applied to improve mechanism identification and pattern recognition for active seismicity in a region. One of the important issues concerning earthquake data clustering is determining the optimum number of clusters (ONC) at the early stages of algorithms. In this paper a robust method based on K-nearest neighbor search (KNNS) is presented to achieve three goals: improving output accuracy, improving output stability, and adding the ability to weight the features used in ONC determination. By introducing a new formula, the proposed method utilizes the error calculated for clustered data based on the similarity between the members in each cluster. An outlier attenuation algorithm is also used to improve the performance of the method. Both the Krzanowski–Lai Index (KLI) and the silhouette coefficient (SC), as two conventional methods, were used to compare the results and evaluate the performance. Experiments on synthetic data sets verified the effectiveness of the method, with considerable differences found. The clustering of a real earthquake catalogue related to the seismogenic province of Zagros in Persia using our proposed methodology suggests using 13-cluster analysis for clustering based on the spatiotemporal features with the same weights, and seven-cluster analysis for a case where priority is given only to the spatial parameters of the epicenters. Under the same circumstances, the KLI and SC methods suggest three and 18 clusters, respectively. The results of the experiments on synthetic data sets indicate that the proposed method is quantitatively more stable and more accurate than the other two methods.

Keywords: KNN search, earthquake data clustering, number of clusters, outlier data, Zagros.

1. Introduction

Data mining generally refers to the extraction of knowledge from available information, in which the purpose is to discover the hidden patterns in a large database. Given recent advances in seismology data analysis, and thus the production of large data sets, the availability of powerful methods able to analyze a large amount of data is essential (Frawley et al. 1991). Clustering, as a method for data mining, is a technique that involves the grouping of observations into a certain number of clusters (Berkhin 2002). In seismology, although clustering is addressed extensively in seismicity analysis and aftershock identification (Zaliapin et al. 2008), there are other diverse uses including event matching of earthquake catalogues with geological evidence (Hall et al. 2018; Ansari et al. 2009), earthquake risk analysis (Mignan et al. 2016; Nazmfar 2019) and earthquake relocation studies (Trugman and Shearer 2017). A brief overview of previous studies is provided in Table 1.

Clustering methods can be categorized into two main classes: partitioning and hierarchical clustering (Dubes and Jain 1980). Partitioning methods divide an n-element set of data into k groups of elements so that $k \leq n$, and each group represents a cluster. Hierarchical methods divide the data set into clusters and sub-clusters in a tree form. In this group of methods, according to their interaction, clustering is progressively improved and therefore the results of clustering are qualified as more acceptable, while the execution time is shorter than with partitioning methods (Dubes and Jain 1980).

Two widely used algorithms amongst the partitioning methods are *K*-means, as a hard-clustering method (Hartigan and Wong 1979), and C-means clustering (Ren et al. 2016), which is a fuzzy clustering method. Both methods are based on minimizing the distance objective function given as:

$$J = \sum_{i=1}^{n} \sum_{j=1}^{n} \left\| x_j^i - c_i \right\|^2, \tag{1}$$

[1] Department of Physics, Central Tehran Branch, Islamic Azad University, Tehran, Iran. E-mail: r-kimiaefar@iaun.ac.ir

[2] Department of Physics, Najafabad Branch, Islamic Azad University, Najafabad, Iran.

Published online: 13 November 2020

Table 1

Overview of selected studies

Author(s)	Overview of the research
Zaliapin et al. (2008)	Using spatiotemporal parameters plus magnitude of the events, aftershocks were clustered separately from mainshocks in synthetic and real earthquake catalogues
Hall et al. (2018)	In order to perform computational analysis of the earthquake patterns, K-means clustering was used based on KLI cluster analysis. Correlation of the cluster boundaries with structural segmentation was used as an evaluation method
Ansari et al. (2009)	Seismotectonic models of Iran were compared with the output of Iran's earthquake clustered data. It was concluded that clustered data based on epicentral parameters were in good agreement with the seismotectonic models
Mignan et al. (2016)	From the methodological point of view, the combined effects of earthquake (regime) clustering and damage-dependent fragility on seismic risk were investigated
Nazmfar (2019)	Vulnerability of urban buildings against different earthquake intensities was evaluated by cluster analysis
Trugman and Shearer (2017)	Using cross-correlation data, station and event information, and velocity model, and based on hierarchical clustering, a new (relative) earthquake relocation method was proposed

In fuzzy clustering, unlike classical clustering (also called crisp or hard clustering), a data set is partitioned so that each data point belongs to one or more clusters with a membership degree which is not necessarily zero or 1, and can be a number between these values (Bezdek 1974; Hajian and Styles 2018). One of the most challenging issues in clustering is the problem of determining the optimum number of clusters (ONC). This is usually solved using statistical approaches (Sugar and James 2003), and many methods have been published in this context. Charrad et al. prepared an R package providing 30 indices for cluster analysis, called NbClust (Charrad et al. 2014). Unfortunately, the currently proposed methods are neither specialized for seismological data purposes (especially the case of severely overlapped clusters) nor adequately accurate and stable. This issue is strongly evaluated in this work by performing analysis on different synthetic data sets. As the main objective, in order to overcome the abovementioned drawbacks, a robust method of grouping analysis is presented here based on the basic philosophy of data clustering by proposing a new formula for calculating clustering error. In addition, in order to offer the option to prioritize the features used for error calculation, a built-in option is also provided for weighting the parameters during the grouping analysis procedure.

2. Materials and Method

To date, various methods have been introduced for cluster analysis and for determining the reasonable number of clusters (i.e. the number for which the output of clustering best fits with the physics of the system). However, none of these methods is specifically dedicated to analysis of earthquake catalogue data. Earthquake data usually have a very wide spatiotemporal distribution, and the ability to properly determine the ONC is precluded by two properties inherent in these data sets. The first obstacle is the existence of outliers as, in fact, isolated earthquakes in the catalogue, and the second is the high probability of clusters overlapping in feature space.

Among the different methods for ONC determination, here the KLI and the SC were selected to compare their proficiency with the proposed method. In the next two sections, a brief explanation of each method is provided.

2.1. Krzanowski–Lai Index

Introduced by Krzanowski and Lai (1988), the KLI is one of the best performers among the cluster analysis methods (Mufti et al. 2005). Recently, Hall et al. (2018) used the method for earthquake epicenter clustering related to the Afro-Arabian Rift System.

The KLI is defined as Eq. 2 (Petrosyan and Proutiere 2016):

$$KL_k = \left| \frac{Diff_k}{Diff_{k+1}} \right|, \tag{2}$$

where

$$Diff_k = (k-1)^{\frac{2}{p}} W_{k-1} - k^{\frac{2}{p}} W_k. \quad (3)$$

Here, k is the number of clusters which maximizes Eq. 2 and is considered the ONC. W is the result of summation, applied on squares of the distance of each object (within the cluster) from the centroid of the cluster, and finally, p denotes the number of features in the data set. The k which provides a higher KL_k value is considered as the ONC. Practically, the KLI calculation procedure is performed by averaging for a set of iterations for each k.

2.2. Silhouette Coefficient

The silhouette method is based on measuring the quality of clustering results by calculating the similarity of an object to its own cluster objects. The SC is calculated as (Savaş et al. 2019):

$$S_k = \frac{1}{n} \sum\nolimits_{j=1}^{n} \frac{b_j - a_j}{\max\{a_j, b_j\}}, \quad (4)$$

where a_j is the averaged distance (i.e. sum of distances, divided by the number of objects in the cluster minus 1) between the current object and all other within-cluster objects, and b_j is the minimum average distance from the jth point to points in the nearest cluster. Hence, in the best case that belongs to the ONC, the difference $bj - aj$ is closer to max $\{a_j, b_j\}$, and the k which maximizes Eq. 4 represents the ONC.

2.3. Cluster Analysis Based on K-Nearest Neighbor Search

The method presented in this work is based simply on the main philosophy of clustering; that is, in well-clustered data, the vectors or objects with maximum similarities are labeled with the same cluster; otherwise the situation can be considered as an error (Edson 1932). Based on the above, to calculate the ONC in a set of vectors, the following steps are worth considering in the presented method:

Step 1: From an initial to a final number of clusters (NC), determined by user, the clustering algorithm (any) is performed on the set of data.

Step 2: For any round of the above and for every object within the data, using KNNS introduced by Altman (1992), the most K similar objects are determined and grouped.

Step 3: Sum of error (SE) for each NC is calculated through Eq. 5:

$$\mathrm{SE_{NC}} = \sum_{i=1}^{N} \sum_{j=1}^{K} \frac{1}{\mathrm{NC}} Cri_{ij}, \quad (5)$$

where N is the total number of objects in the data, and the term Cri_{ij}, the error criterion, is zero everywhere except where it could be 1 by satisfying the following two conditions: the two compared objects (one from looping over all data and the other from the group of KNNS) cannot belong to the same cluster, and the distance between the two points must not be greater than the distance between the point in the KNNS group and the centroid of the labeled cluster for this point.

Step 4: NC with minimum SE is indicative of the ONC.

To reduce the effect of random centroid selection by clustering algorithms, the aforementioned steps are iterated (empirically, a number between 5 and 20 is sufficient to stabilize the output). Therefore, the process of error calculation can be continued by averaging over all $\mathrm{SE_{NC}}$ in all iterations:

$$\mathrm{Mean}(\mathrm{SE_{NC}}) = \sum_{i=1}^{\mathrm{itr}} \frac{(\mathrm{SE_{NC}})_i}{\mathrm{itr}}, \quad (6)$$

where itr is the number of all iterations. K (in Step 2) consists of a fixed value (e.g. constant percentage of the total data size that is empirically determined) and an incremental variable part, increasing based on the NC increment. This is mainly for solution convergence in the case where the data are noisy or the clusters overlap.

One of the most challenging topics in clustering is the problem of outlier presence (Gan and Ng 2017). Earthquake data clustering definitely suffers from this problem (Shi and Pun Cheng 2019). Hence, in the very first step of the algorithm, by a method based on Hampel filtering (Liu et al. 2004; Yao et al. 2019) in the subdivided zones resulting from regular gridding of the data, the amount of the outlier data is reduced. However, identification of the outliers is just to determine the correct number of clusters and does not necessarily mean that the outliers have to be deleted

from the data set. Nevertheless, this procedure is crucial and must be supervised by an expert to prevent any major changes in the structure of the data. A Hampel identifier, which is regarded as one of the most efficient methods for outlier detection (Pearson 2002), is originally based on the rolling median (RM) and median absolute deviation (MAD) of the data, which are both studied locally in a symmetric window around each element of the data set. The method marks a data point as an outlier if the following criteria are satisfied for element x.

$$x - \mathrm{RM}(X_N) \geq \mathrm{tf}.\mathrm{MAD}(X_N), \tag{7}$$

where

$$\mathrm{MAD}(X_N) = \mathrm{gd}.\mathrm{median}\{|x_1 - \mathrm{RM}(X_N)|, \ldots, |x_N - \mathrm{RM}(X_N)|\}. \tag{8}$$

Generally, the thresholding factor (tf) is set equal to 3, and the unbiased estimation of the Gaussian distribution (GD) is equal to 1.4826. Here, N stands for the window length and for preserving symmetric criteria, and is an odd number (Yao et al. 2019).

Applying an outlier removal procedure may cause some original cluster objects to be deleted; therefore, this step requires expert supervision for checking that clusters critically do not vanish.

In addition to increasing the accuracy of the algorithm, one of the reasons for proposing a new algorithm in the present work was to have the ability to apply different weights for all features in the process of determining ONC. In some circumstances, including earthquake risk analysis (Mignan et al. 2016) or earthquake migration studies (Chen et al. 2012), it is preferred to use the set of features with different weights in the algorithm in order to give priority to certain features or dimensions (in the special case of clustering earthquake data, this set of features is usually provided in the earthquake catalogues). In doing so, in the method proposed here, the sum of the error is calculated separately for specific features by finding the group of similar points by KNNS, using only the mentioned features. Finally, the error is calculated by weighting and averaging in each direction and adding them. The flowchart of the K-nearest neighbors search cluster analysis (KNNCA) is depicted in Fig. 1.

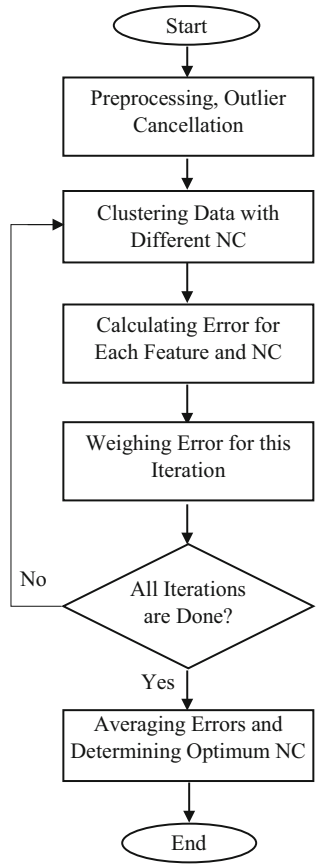

Figure 1
Flowchart of KNNCA for determination of the ONC in a multidimensional data set

3. Experiments

The performance of the proposed method was assessed by performing the algorithm in three sets of synthetic data, one of which was contaminated with 5% of additive outlier events. In the first data set, five randomly generated clusters were created without any additive outliers. This four-dimensional data were created based on five centroids and some randomly generated points with a variance equal to 1. The minimum and maximum of each feature (dimension) was bounded between 0 and 30, and as shown in Fig. 2a–c, the clusters are well separated from each other. With no preprocessing steps (e.g. data normalizing), and in accordance with the details provided in Table 2, the methods were applied over the range of NC = *2, 3, …, 14*. The result of 100 iterations of the methods are illustrated in Fig. 2.

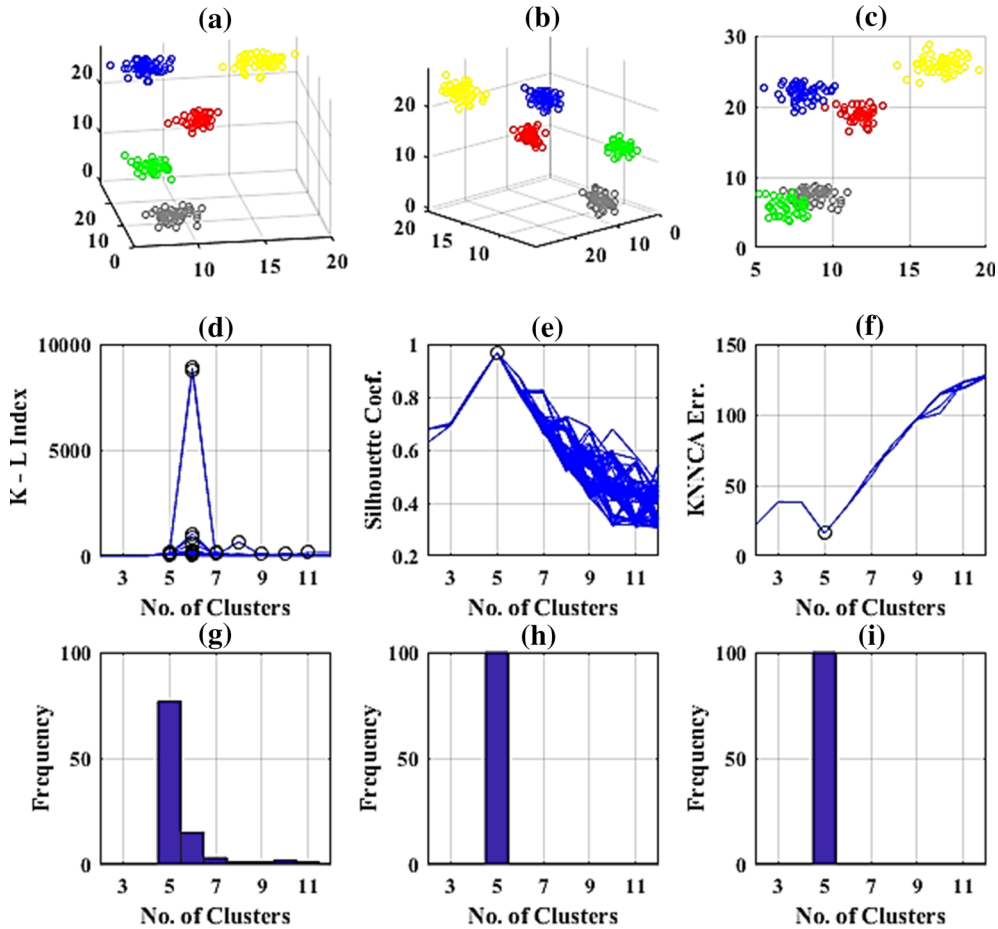

Figure 2
a–c Plot of a five-cluster synthetic data set from different view angles; 100 iterations of ONC calculations by **d** KLI, **e** SC, and **f** KNNCA. **g–i** Histograms of the outputs for the mentioned methods, respectively. Here, in the case of no additive outliers, the performance of silhouette and KNNCA are the same, whereas the KLI is slightly lagging

Table 2

Specifications of the methods for the first experiment

Method	Specifications
KLI	No. of internal iterations = 100, distance = "Euclidian", clustering algorithm = "K-means"
SC	Distance = "Euclidian", clustering algorithm = "K-means"
KNNCA	Outlier cancellation = 10%, distance = "Euclidian", clustering algorithm = "K-means"

Histograms of the outputs, i.e. the repetition frequency of the ONC resulting from each iteration, are also shown for stability comparison. The exact ONC for the KNNCA and the SC versus 77% correctness for the KLI output is indicative of the superiority of the first two methods in this experiment.

A set of 4D points, as much as 5% of the total number of objects in the previous data set, with variance equal to 30 were produced and added to the data as outliers (Fig. 3a–c). Using the same setup as in the prior experiment, the methods were applied on the data and the results are plotted in Fig. 3. In contrast to the case with no additive noise, where the SC was unsuccessful in providing the correct answer and the KLI lagged (about 34% in terms of performance), the proposed method preserves the performance due entirely to the outlier removal option.

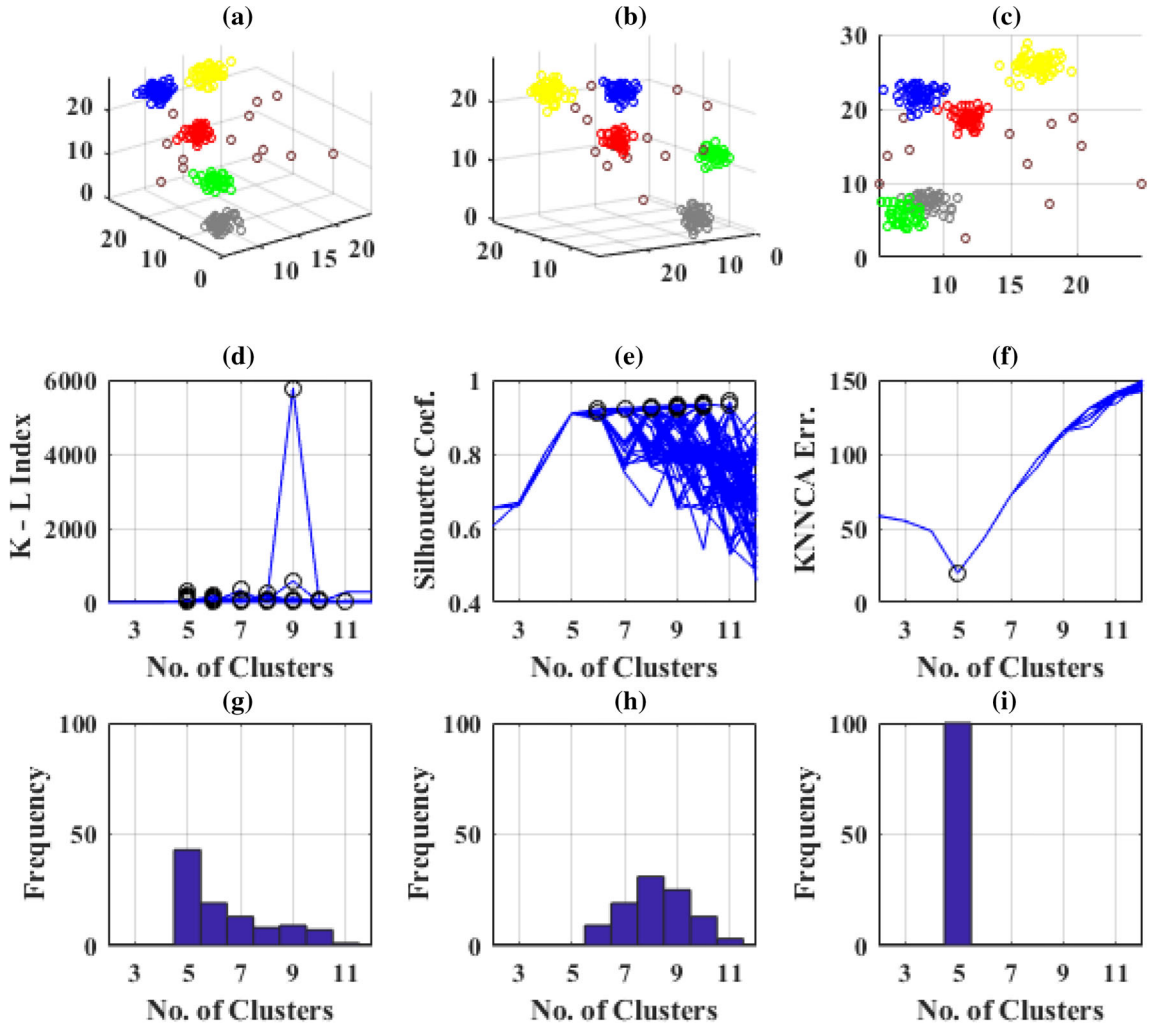

Figure 3
a–c The data in Fig. 2a, with 5% additive outliers plotted from different view angles. One hundred iterations of ONC calculations by **d** KLI, **e** SC, **f** KNNCA, and **g–i** histograms calculated for the methods' output, respectively. Obviously the proposed method has preserved the performance (with reference to previous experiment) completely while KL and SC has declined their performance 34 and 100% respectively. Comparing the histograms is also indicative of a notable difference between the stability of the methods

The experiments are continued with a synthetic 10-cluster data set that is also bounded to the same range as the previous data set (Fig. 4a–c). This is mostly because of decreasing cluster distances and assessing the ability to distinguish overlapped clusters. Also, the variance in the generated data points is doubled for the same reason. The output of the KLI is comparatively useless, as the pick of the histogram does not match the true NC.

To evaluate the effect of outlier data on ONC determination, an incremental schema was chosen for the number of additive outliers in each iteration. In doing so, a synthetic four-cluster data set was generated so that the clusters were well separated from each other in order to reduce the effect of overlapping clusters in this experiment. The outlier data, from 1 to 30% of the size of the original data set, were added and ONC determination methods were performed in each iteration. The results of this experiment are illustrated in Fig. 5, indicating the robustness of the KNNCA based on only two wrong predictions (out of 30). Also, it is concluded that the SC is not feasible

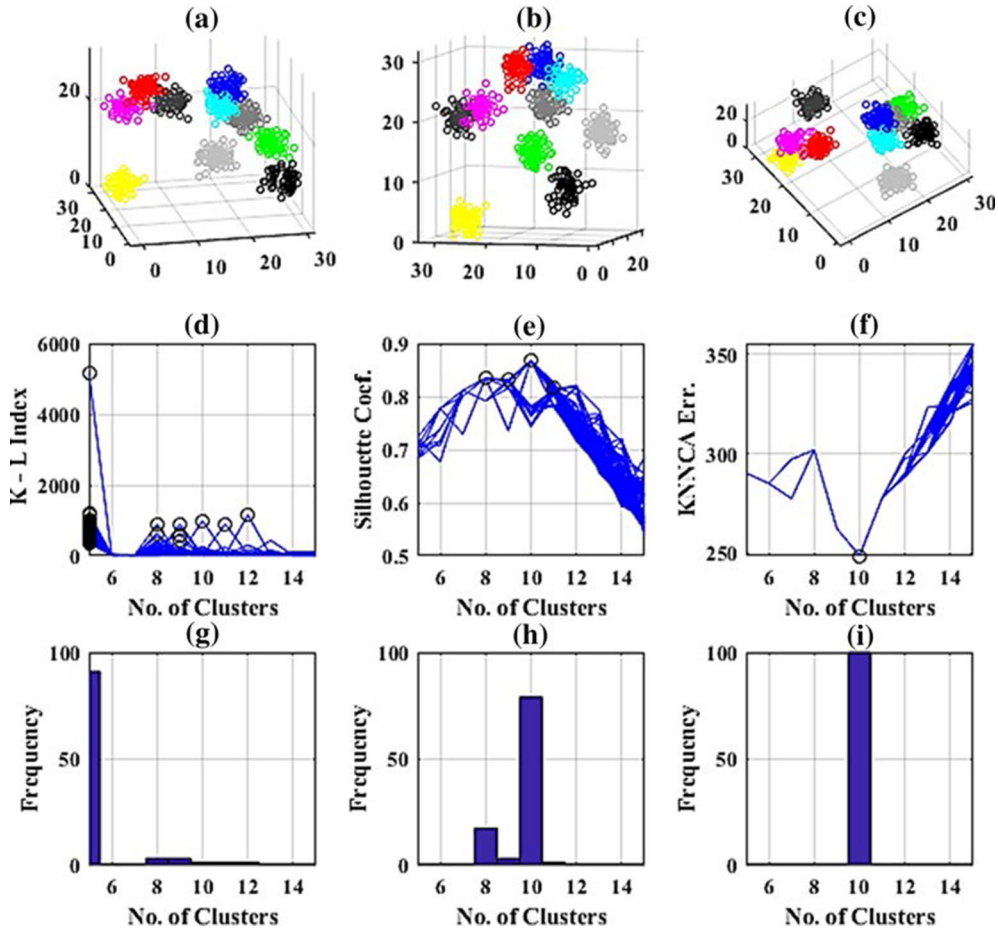

Figure 4
a–c A synthetic 10-cluster data set with some overlapped clusters from different view angles. One hundred iterations of ONC calculations by **d** KLI, **e** SC and **f** KNNCA, and **g–i** histograms calculated for the output of the methods, respectively. The KLI has 2%, SC has 78% and KNNCA has 100% correct output predictions. Unlike previous tests, in the situation with clean data (no additive outlier), the SC recovers the performance even with overlapped clusters

for the cases where the input data are contaminated with outliers.

4. *Clustering Earthquake Catalogue of Zagros*

After investigating the performance of the methods using synthetic data sets, we used information from a real earthquake catalogue related to the events recorded in the seismogenic province of Zagros in Iran. The Zagros thrust folded belt is located between the Arabian plate and the central Iranian plate. This belt is over 1600 km in length and 200–300 km in width. This region is known as an active seismic zone, and more than half of Iran's instrumented earthquakes happen along this belt (Tatar et al. 2002). Some of the most important faults located in this region are the Zagros main reverse fault, high Zagros fault, Zagros fore-deep fault and Kazeroon fault. The study of the area shows that a series of steep reverse faults are the source of the seismic events in this seismogenic zone (Talebian and Jackson 2004). This earthquake catalogue contains 554 events with $M_L \geq 4.0$ reported by the International Institute of Earthquake Engineering and Seismology between 2006 and 2019. The plot of the catalogue using epicentral spatial parameters is shown in Fig. 6. The first experiment on this data set was an all-even-weights

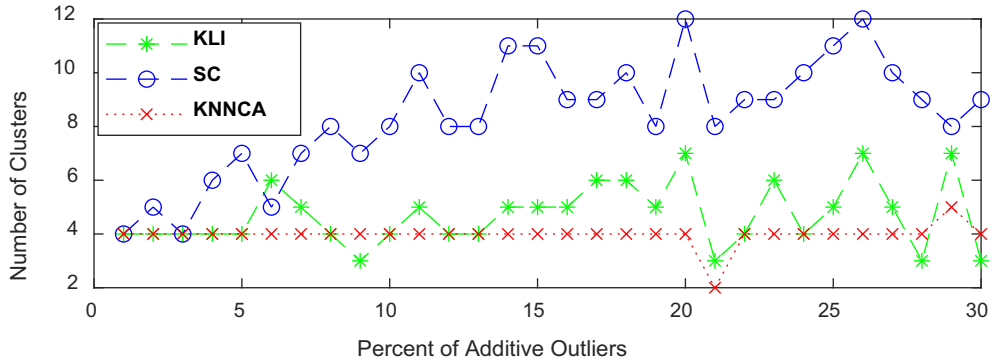

Figure 5
The ONC calculation in a synthetic four-cluster data set with varying outlier amount, for evaluating the effect of outlier presence on the accuracy of the methods

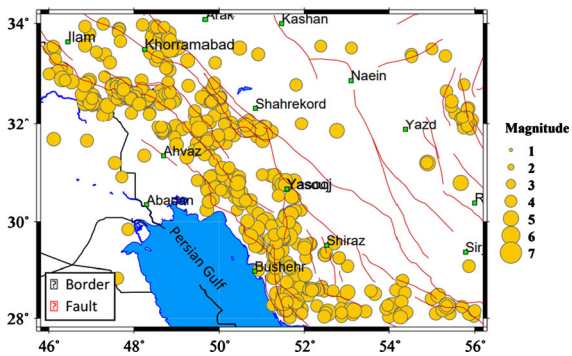

Figure 6
Distribution of the Zagros earthquakes reported by the International Institute of Earthquake Engineering and Seismology from 2006 to 2019

ONC determination. Here, 15% of the events were eliminated as outliers. This was done based on visual assessment, i.e. by checking that the data structure did not have major changes after outlier elimination. For the KNNS grouping and error calculation, 2% of the data were used, and the range of NC was considered to be between 3 and 20 clusters. The results of this test are illustrated in Fig. 7, and based on the histogram interpretation, although it is concluded that almost all methods performed steadily in ONC determination, there is a considerable difference between the KNNCA outputs and those of the other two methods. While the KLI and the SC indicate an ONC of three, the proposed method suggests 13 clusters. This experiment shows that outliers severely affect the KLI and the SC responses.

As the last experiment, giving priority to spatial parameters of the epicenters and setting the related weights equal to 0.425 (for longitude and latitude), 0.15 for depth, and zero for the time of the events, the proposed algorithm was applied on the data with no other changes in setup. For the KLI and the SC calculations, only latitude and longitude information was used, as there is no built-in option for weighting the input features. The results of the tests over a range of 5–30 clusters are depicted in Fig. 8, suggesting an ONC of 18 clusters for the KLI and SC, and seven for the KNNCA. The remarkable point about the results of this experiment is that the ONC determined by the first two methods was higher than the case of clustering using four features (in the previous test), exactly contrary to the case of the proposed method.

5. *Conclusion*

In this paper, we proposed a method for determining the optimum number of clusters for multidimensional data sets such as earthquake data. The developed method is equipped with built-in tools for outlier elimination and parameter weighting which can be used in the clustering procedure. Improving the accuracy of the algorithm and the ability for weighting of the features used in clustering was the main purpose in developing this method. In the first three experiments, all of which were performed on synthetic data sets, the results were

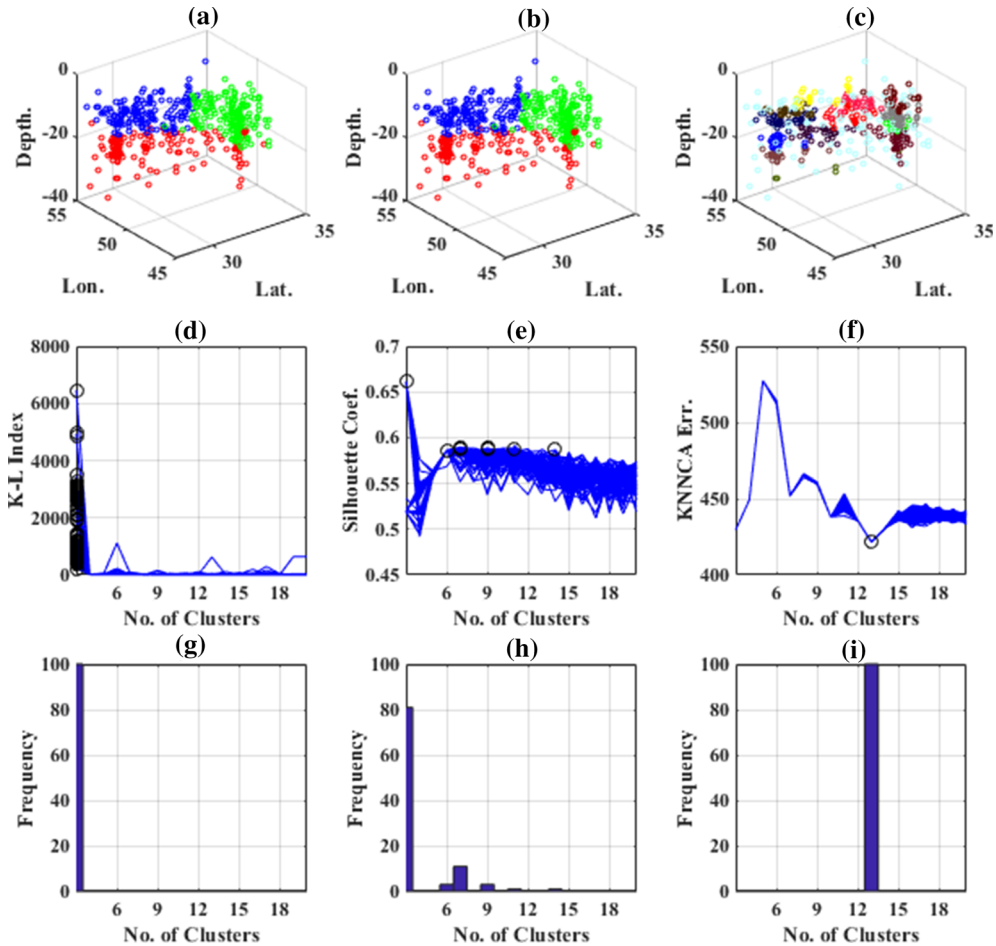

Figure 7

a–c K-means clustering of the Zagros earthquakes based on the spatiotemporal parameters with ONC determined by KLI, SC and KNNCA, respectively. One hundred iterations of ONC calculations by **d** KLI, **e** SC and **f** KNNCA, and **g–i** histograms calculated for the output of the methods, respectively

quantitatively monitored and were found to support the ability of the method to perform more accurately even in the case of overlapped clusters or when outliers were added to the data sets. At least 21% and as much as 56% performance improvement was recorded for the KNNCA compared to the KLI and the SC, which confirms the superiority of the proposed method. Based on this performance, the method was also applied on a real earthquake catalogue related to the Zagros seismogenic zone containing 554 data points, all of which include hypocenter locations and time of the events. Based on the outputs of the methods, it is suggested that the data be clustered into 13 and seven groups for even-weighted features and epicentral clustering, respectively. A comparison the results of the histograms also confirms the greater stability of the KNNCA Thus, based on the above, it is concluded that the proposed method is more accurate and more stable than the KLI and the SC methods.

Data Availability

The earthquake catalogue is available online at: https://www.iiees.ac.ir/.

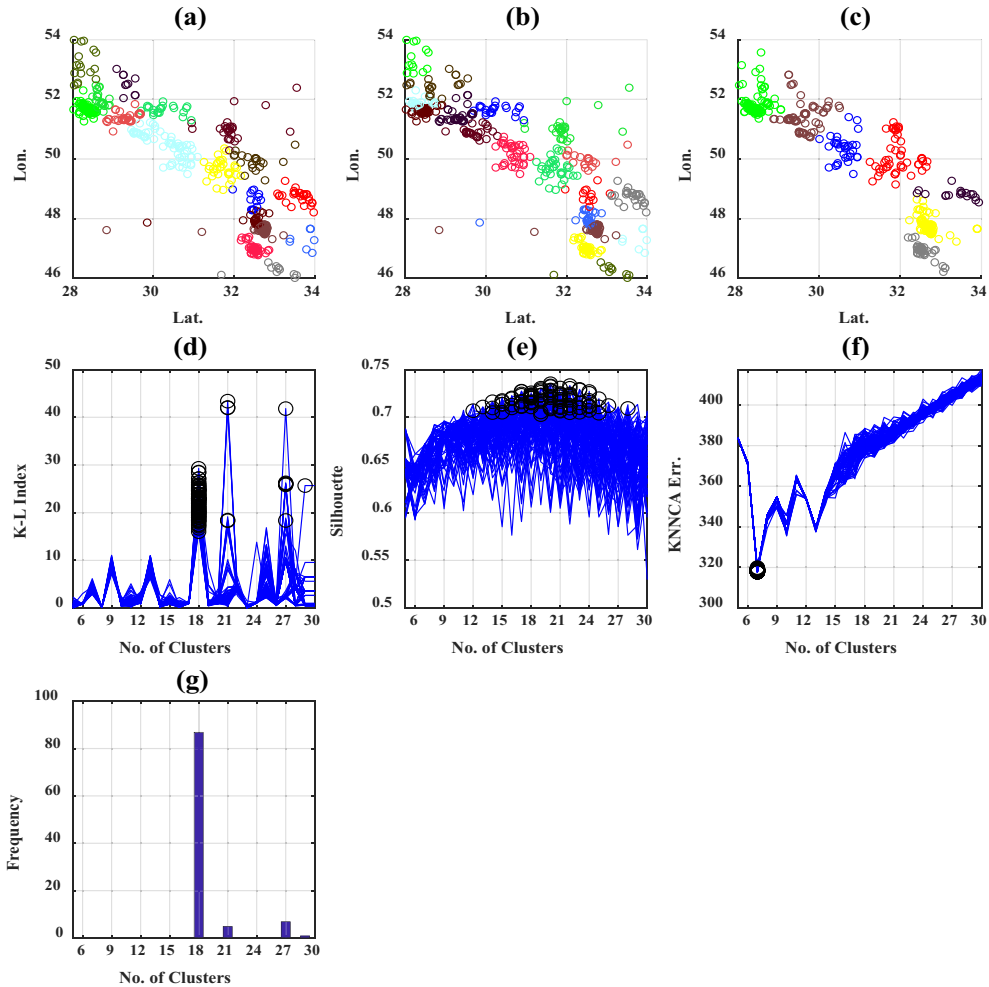

Figure 8
a–c K-means clustering of Zagros earthquakes giving priority to epicentral parameters with ONC determined by KLI, SC and KNNCA, respectively. One hundred iterations of ONC calculations by **d** KLI, **e** SC and **f** KNNCA, and **g–i** histograms calculated for the output of the methods, respectively

Code Availability

Requests for codes, after publication of this article, will be considered by the corresponding author.

Compliance with Ethical Standards

Conflict of Interest The authors declare that they have no conflict of interest.

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

References

Altman, N. S. (1992). An introduction to kernel and nearest-neighbor nonparametric regression. *American Statistician, 46*(3), 175–185. https://doi.org/10.1080/00031305.1992.10475879

Ansari, A., Noorzad, A., & Zafarani, H. (2009). Clustering analysis of the seismic catalogue of Iran. *Computers & Geosciences, 35,* 475–486. https://doi.org/10.1016/j.cageo.2008.01.010

Berkhin, P. (2002). *Survey of clustering data mining techniques. Technical report, Accrue Software*. Berlin, Heidelberg: Springer. https://doi.org/10.1007/3-540-28349-8_2

Bezdek, J. C. (1974). Cluster validity with fuzzy sets. *J Cybern, 3,* 58–73. https://doi.org/10.1080/01969727308546047

Charrad, M., Ghazzali, N., Boiteau, V., & Niknafs, A. (2014). NbClust: an R package for determining the relevant number of

clusters in a data set. *Journal of Statistical Software, 61,* 1–36. https://doi.org/10.18637/jss.v061.i06

Chen, X., Shearer, P. M., & Abercrombie, R. E. (2012). Spatial migration of earthquakes within seismic clusters in Southern California: evidence for fluid diffusion. *Journal of Geophysical Research, 117,* B04301. https://doi.org/10.1029/2011JB008973

Dubes, R. C., & Jain, A. K. (1980). Clustering methodology in exploratory data analysis. *Advances in Computers, 19,* 113–228. https://doi.org/10.1016/S0065-2458(08)60034-0

Frawley, W. J., Piatetski, S. G., & Matheus, C. J. (1991). Knowledge discovery in databases. *AI Mag, 13,* 57–70. https://doi.org/10.1609/aimag.v13i3.1011

Gan, G., & Ng, M. K. P. (2017). K-means clustering with outlier removal. *Pattern Recognition Letters, 90,* 8–14. https://doi.org/10.1016/j.patrec.2017.03.008

Hajian, A., & Styles, P. (2018). *Application of soft computing and intelligent methods in geophysics*. Springer International Publishing. https://doi.org/10.1007/978-3-319-66532-0

Hall, T. R., Nixon, C. W., Burton, P. W., & Ayele, A. (2018). Earthquake clustering and energy release of the African–Arabian rift system. *Bulletin of the Seismological Society of America, 108,* 155–162. https://doi.org/10.1785/0120160343

Hartigan, J. A., & Wong, M. A. (1979). Algorithm AS 136: a K-means clustering algorithm. *Applied Statistics, 28,* 100–108. https://doi.org/10.2307/2f2346830

Krzanowski, W., & Lai, Y. (1988). A criterion for determining the number of groups in a data set using sum-of-squares clustering. *Biometrics, 44*(1), 23–34. https://doi.org/10.2307/2531893

Liu, H., Shah, S., & Jiang, W. (2004). On-line outlier detection and data cleaning. *Computers & Chemical Engineering, 28,* 1635–1647. https://doi.org/10.1016/j.compchemeng.2004.01.009

Mignan, A., Danciu, L., & Giardini, D. (2016). Considering large earthquake clustering in seismic risk analysis. *Natural Hazards, 91,* 149–172. https://doi.org/10.1007/s11069-016-2549-9

Mufti, G. B., Bertrand, P., & Moubarki, L. E. (2005). Determining the number of groups from measures of cluster stability. *Proc Int Symp Appl Stoch Models Data Anal, 1,* 404–413.

Nazmfar, H. (2019). An integrated approach of the analytic network process and fuzzy model mapping of evaluation of urban vulnerability against earthquake. *Geomat Nat Hazards Risk, 10,* 1512–1528. https://doi.org/10.1080/19475705.2019.1588791

Perarson, R. K. (2002). Outliers in process modeling and identification. *IEEE Transactions on Control Systems Technology, 10,* 55–63.

Petrosyan, V., & Proutiere, A. (2016). Viral clustering: a robust method to extract structures in heterogeneous datasets. *Proc Thirtieth AAAI Conf Artifi Intell, 1,* 1986–1992.

Ren, M., Liu, P., Wang, Z., & Yi, J. (2016). A self-adaptive fuzzy c-means algorithm for determining the optimal number of clusters. *Comput Intell Neurosci, 3,* 1–12. https://doi.org/10.1155/2016/2647389

Savaş, C., Yıldız, M. S., Eken, S., İkibaş, C., & Sayar, A. (2019). Clustering earthquake data: Identifying spatial patterns from non-spatial attributes. In Gyamfi, A., & Williams, I. (Eds.), *Big data and knowledge sharing in virtual organizations* (pp. 224–239). IGI Global. https://doi.org/10.4018/978-1-5225-7519-1.ch010

Shi, Z., & Pun Cheng, L. S. C. (2019). Spatiotemporal data clustering: a survey of methods. *ISPRS Int J Geo-Inf, 8*(3), 1–16. https://doi.org/10.3390/ijgi8030112

Sugar, C. A., & James, G. M. (2003). Finding the number of clusters in a data set: an information-theoretic approach. *Journal of American Statistical Association, 98,* 750–763. https://doi.org/10.1198/016214503000000666

Talebian, M., & Jackson, J. (2004). A reappraisal of earthquake focal mechanisms and active shortening in the Zagros mountains of Iran. *Geophysical Journal International, 156,* 506–526. https://doi.org/10.1111/j.1365-246X.2004.02092.x

Tatar, M., Hatzfeld, D., Martinod, J., Walpersdorf, A., Ghafory-Ashtiany, M., & Ch´ery, J. (2002). The present-day deformation of the central Zagros from GPS measurements. *Geophysical Research Letters*. https://doi.org/10.1029/2002GL015159

Trugman, D., & Shearer, P. (2017). GrowClust: a hierarchical clustering algorithm for relative earthquake relocation, with application to the Spanish springs and sheldon, nevada, earthquake sequences. *Seismological Research Letters, 88,* 379–391. https://doi.org/10.1785/0220160188

Yao, Z., Xie, J., Tian, Y., & Huang, Q. (2019). Using hampel identifier to eliminate profile-isolated outliers in laser vision measurement. *J Sens*. https://doi.org/10.1155/2019/3823691

Zaliapin, I., Gabrielov, A., Keilis-Borok, V., & Wong, H. (2008). Clustering analysis of seismicity and aftershock identification. *Physical Review Letters, 101*(1), 1–4. https://doi.org/10.1103/PhysRevLett.101.018501

(Received March 10, 2020, revised October 21, 2020, accepted October 23, 2020)