A novel centroids initialisation for K-means clustering in the presence of benign outliers

Amin Karami*

Department of Architecture, Computing and Engineering (ACE),
University of East London (UEL),
Docklands Campus, UK
Email: a.karami@uel.ac.uk
*Corresponding author

Shafiq Urréhman

China Euro Vehicle Technology AB (CEVT),
Theres Svenssons Gata 7, SE-41755 Göteborg, Sweden
Email: shafiq.urrehman@cevt.se

Mustansar Ali Ghazanfar

Department of Architecture, Computing and Engineering (ACE),
University of East London (UEL),
Docklands Campus, UK
Email: m.ghazanfar@uel.ac.uk

Abstract: K-means is one of the most important and widely applied clustering algorithms in learning systems. However, it suffers from centroids initialisation that makes K-means algorithm unstable. The performance and the stability of the K-means algorithm may be degraded if benign outliers (i.e., long-term independence data points) appear in data. In this paper, we developed a novel algorithm to optimise K-means performance in the presence of benign outliers. We firstly identified the benign outliers and executed K-means across them, then K-means runs over all data points to re-locate clusters' centroids, providing high accuracy. The experimental results over several benchmarking and synthetic datasets confirm that the proposed method significantly outperformed some existing approaches with better accuracy based on applied performance metrics.

Keywords: clustering; K-means; centroid initialisation; benign outlier.

Reference to this paper should be made as follows: Karami, A., Urréhman, S. and Ghazanfar, M.A. (xxxx) 'A novel centroids initialisation for K-means clustering in the presence of benign outliers', *Int. J. Data Analysis Techniques and Strategies*, Vol. x, No. x, pp.xxx–xxx.

Biographical notes: Amin Karami is currently a Senior Lecturer at the University of East London (UEL) in UK. He has been extensively working on big data technologies, computational intelligence, optimisation and network analysis. He received his MSc in Informatics from the University of Skövde,

Sweden in 2011. He also completed his PhD from the Computer Architecture Department at the Universitat Politécnica de Catalunya Barcelona Tech (UPC), Spain in February 2015. He has been carried out several international research collaborations, several funds and grants, and several invited talks and presentations.

Shafiq Urréhman is Tech Lead AI/ML and Senior Technical Expert at China Euro Vehicle Technology AB (CEVT), Sweden. He has 15 years of experience in developing AI/ML/technical projects for various industries spreading from mining, automotive, forestry, communication, assistive and medicine. Before joining CEVT, he has been working as an Associate Professor and the Director of i2lab at the Umea University, Sweden, and an Associate Professor at the Linköping University, Sweden and the Founder/Team Lead of Intelligent Systems Group at UEL, UK.

Mustansar Ali Ghazanfar is currently a Lecturer at the University of East London (UEL), UK. He has been extensively working on machine learning, deep learning, artificial intelligence, and business analytics. He receive his PhD in Machine Learning from the University of Southampton UK in 2012. He has more than 10 years of industrial and academic experience. He has more than 50 international publications and according to Google Scholar, his research attracts more than 0.5 citation per day in 2019.

1 Introduction

One of the most primitive human actions is grouping and classifying similar and heterogeneous objects into distinct categories (Karami and Guerrero-Zapata, 2014). In exploratory data analysis and data mining domain, this task is known as clustering. Clustering is the unsupervised classification technique that divides a set of given data into different clusters, in which the similar data are grouped into a same cluster (Karami and Guerrero-Zapata, 2015a; Li, 2011). Clustering techniques have been employed in many applications, such as wireless sensor networks, medicine, biology, psychology, statistics, computer networking, program comprehension, software visualisation and engineering (Celebi et al., 2013; Khanmohammadi et al., 2017; Alaei et al., 2018).

Among many clustering algorithms developed in the past 60 years, K-means is one of the oldest and commonly used algorithms, first employed by James MacQueen in 1967. K-means is simple, easy to implement, suitable for large datasets, and very efficient with linear time complexity. However, it suffers from several drawbacks, one such drawback is that it is sensitive to clusters' centroids initialisation, particularly in the presence of low-frequent patterns called outliers (Karami and Guerrero-Zapata, 2015a; Gan and Ng, 2017; Min and Kai-fei, 2015; Karami, 2018).

The low-frequent data patterns are mostly *malignant* outliers, which adversely affect the clustering quality. Detecting and removing such malignant outliers improves the clustering accuracy. Much research on clustering attempts to remove these using 'outlier removal techniques' (Gan and Ng, 2017; Hautamäki et al., 2005) or combine meta-heuristic optimisation algorithms with machine learning techniques (Santhanam and Padmavathi, 2015; Marghny and Taloba, 2011; Shahreza et al., 2011) in order to improve K-means in the presence of such outliers.

In contrast, *benign* outliers have been getting less attention, most approaches simply ignore them during training (Karami and Guerrero-Zapata, 2015a; Shin et al., 2017). Benign outliers are the long-term independent data points that are mostly not included in a 90%–95% confidence interval of normally distributed data (Jach and Kokoszka, 2008). Since the benign outliers are inherently a part of original data (see Figure 1 in Section 2) and they must be available for training purposes (i.e., they are not severe or malignant outliers), removing them might result in inappropriate training, unstable and diverge modelling. It means that, the benign outliers exist in all the datasets and users must decide to take away a small portion of data (i.e., out of 95% of normal distribution) as benign or a large portion (i.e., out of 90% of normal distribution) as benign.

In this research work, we focus on the benign outliers to improve the stability and the robustness of K-means algorithm through novel centroid initialisation technique. In our approach we initially identify the benign outliers and initialise K-means centroids across them. Then, K-means runs over all data points to re-locate clusters' centroids, providing better classification and accuracy rate.

The rest of this paper is organised as follows. Section 2 discusses the importance of the benign outliers. Section 3 describes K-means clustering algorithm. The proposed method is presented and discussed in Section 4. Experimental results are presented in Section 5, and, conclusion is given in Section 6.

2 Benign outliers

Outliers are data points that are distant from other data and may indicate experimental error, often resulting in exclusion from the dataset. Outliers may occur due to several reasons, such as, measurement error, incidental systematic error, or by chance. It is often not trivial to ascertain the cause of an outlier, resulting in, not a straightforward way to express rules for their removal. For instance, a person with an IQ of 130 is not outlier. Outliers may or may not be a problem depending on several factors (Marr, 2015):

- some statistical tests are robust and can accommodate outliers, others may be severely influenced by outliers
- some data types will naturally contain extreme values which are entirely inherent
- the presence of outliers may, in fact, be of interest.

Figure 1 depicts a sample of data distribution with a set of data samples that are far from the 90%–95% of the normal distribution; however, they are not malignant or severe outliers. Hence, we cannot remove them because they are inherently a main part of the original data. In this research work, we call them *benign outlier*, by removing them might result in inappropriate training, unstable and diverge data modelling. To be able to deal accurately with the benign outliers, we would initially need to identify them. To do so, we employ Hotelling's T-squared distribution technique (Yi et al., 2016).

2.1 Hotelling's T-squared distribution

The Hotelling's T-squared distribution is a multivariate generalisation of the Student's t-test. The form of the Hotelling's T-squared is as follows:

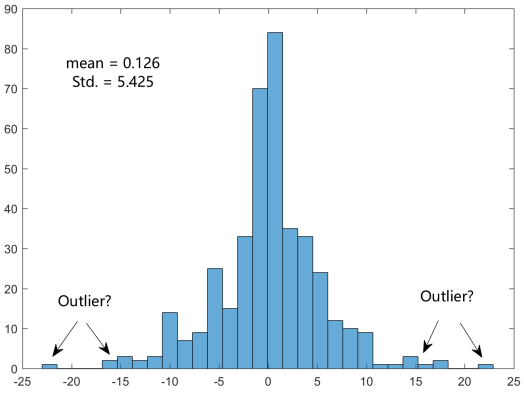
$$T^2 = (X - \bar{X})W^{-1}(X - \bar{X}) \tag{1}$$

where X is the original data matrix, $\bar{X}$ is the mean of the dataset, and W is the covariance matrix of X. The Hotelling's T-squared statistic is approximately F-distributed as follows:

$$F_{p,n,\alpha} \sim T^2 \frac{(n-a)}{a(n-1)} \tag{2}$$

Any sample that has an F-value exceeding the critical F-value can be considered as an outlier. There is no an obvious F-value and must be empirically chosen. We setup $F\text{-}value = 3$ experimentally to find the most probable data points as benign outlier.

Figure 1 A sample of normal distribution (see online version for colours)



3 K-means clustering algorithm

K-means clustering algorithm groups the set of data points into a predefined number of clusters based on a distance function, most commonly used distance function is Euclidean distance (Karami and Guerrero-Zapata, 2015a). The standard K-means algorithm is summarised as follows:

1 randomly initialise K centroids

2 calculate Euclidean distance between each data points and centroids and select the smallest distance as the closest cluster centroid to data point

3 recalculate the cluster centroids using the mean of data points in each cluster

4 repeat step 2 and 3 until the centroids do not change any more in the predefined number of iteration or a maximum number of iterations have been reached.

4 The proposed method

The proposed method for optimal placement of cluster centroids in the presence of benign outliers with K-means algorithm is described in Algorithm 1.

Algorithm 1 The pseudocode of the proposed method

Input: Training dataset
Output: Well-separated clusters using K-means
N = The number of training data
T = The maximum number of iteration
K = The number of clusters
$F - Value$ = The threshold value for finding benign outliers
Phase 1: Identify Benign Outliers
while N **do**
(1) Identify $Benign.Outliers$ using equations (1) and (2)
end while
Phase 2: Initialise K-means centroids
(1) $Centers$ = Place randomly K centroids across $Benign.Outliers$ data points
Phase 3: Run K-means
while $Iter < T$ or cluster centroids do not change any more **do**
(1) Find the closest $Centers$ to each data point using Euclidean distance: $\sqrt{\sum_{i=1}^{N}\sum_{j=1}^{K}(Data_i - Centers_j)^2}$
(2) Recalculate $Centers$ using the mean of data points within a same cluster
end while
return Well-separated clusters

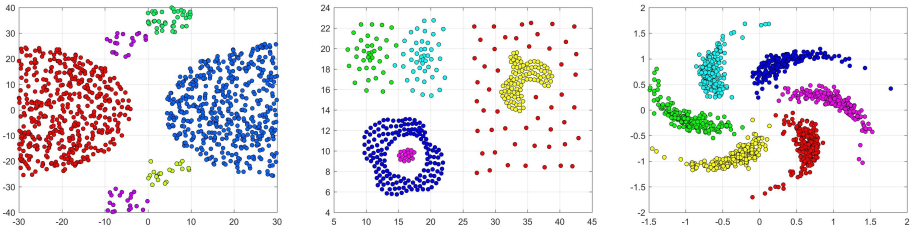
Our proposed algorithm modifies the original K-means clustering algorithm by introducing an initialisation stage where we identify the benign outliers and randomly place centroids across them. The time complexity of the proposed algorithm (Algorithm 1) is therefore calculated in two stages. The first stage (i.e., identifying benign outliers for centroids initialisation) has a complexity of $O(N)$, and the second stage (i.e., running K-means clustering) has a complexity of $O(T.N.K)$. The overall complexity of the proposed approach is $O(N + T.N.K)$, which reduces to $O(T.N.K)$. In practice, although the proposed technique increases the time complexity by $O(N)$ required for the initialisation stage, this additional cost results in improved clustering performance.

5 Experimental results

To assess the performance and accuracy of the proposed method, we compare it against two existing widely-used algorithms, K-means++ (Arthur and Vassilvitskii, 2007) and density K-means (Yuan et al., 2015). We perform the comparison using several performance metrics, namely, mean square error (MSE), standard deviation (Std.), detection rate (DR), false positive rate (FPR), and purity. These are well-known metrics used for assessing and comparing the performance and the accuracy of clustering algorithms (Karami and Guerrero-Zapata, 2015a, 2015b).

We use eight datasets from two different sources for the comparison, ensuring that there are diversity in the datasets and a generality in the results. The first source of data has three 2D synthetic dataset and is currently being used to analyse learning algorithm (Karami and Johansson, 2014). Figure 2 shows these three synthetic datasets. Another source of data gives five classic benchmark problems from UCI machine learning repository, namely, Iris, Glass, Wine, Ionosphere, and Zoo. Table 1 shows the characteristics (features, classes and patterns) for these five benchmark datasets.

Figure 2 Synthetic datasets derived from Karami and Johansson (2014), (a) outlier (D = 1,020, K = 5) (b) compound (D = 399, K = 6) (c) pinwheel (D = 1,200, K = 6) (see online version for colours)



(a) (b) (c)

Notes: D – the number of data points; K – the number of classes.

Table 1 The five applied benchmark datasets

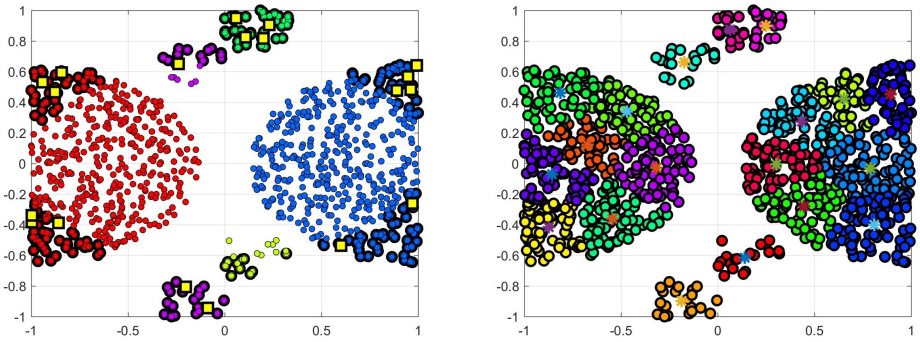
Dataset	*Features*	*Classes*	*Patterns*
Iris	4	3	150
Glass	9	6	214
Wine	13	3	178
Ionosphere	34	2	351
Zoo	17	7	101

5.1 Results of synthetic data

Figures 3(a), 5(a) and 7(a) show the first stage of the proposed algorithm. We selected a random K for each dataset to be able to visualise the functionality of the first stage of the proposed method. This stage discovered the benign outliers (e.g., the bold and highlighted data points) and initialised clusters' centroids across them. Initial centroids are drawn with yellow squares. Figures 3(b), 5(b) and 7(b) show the final placement of clusters' centroids to visualise the performance of the proposed method over applied 2D datasets.

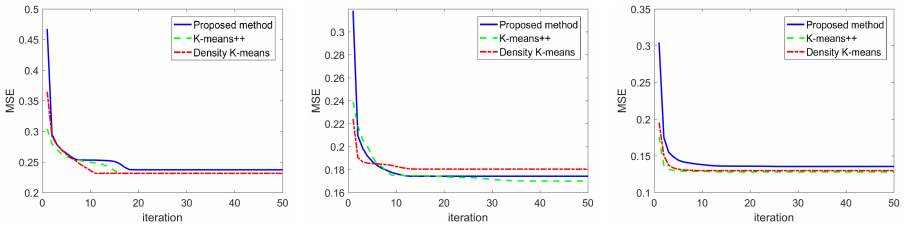
The MSE values of three methods over applied datasets are depicted in Figures 4, 6, and 8 with different random K values. As we expected, the initial MSE value from our method is high due to spreading initial centroids in far spots. However, after a while through K-means iterations, the proposed method could significantly provide less or reasonable MSE as compared to existing methods. This non-significant MSE result is a cost of considering benign outliers. In contrast, this provides significant results based on higher DR and lower FPR at the same time, that are the main attributes for considering clustering quality. To evaluate the accuracy and the robustness of the proposed method, we considered several numbers of centroids and presented the best results of 10 times individual runs in Tables 2–4. According to results, the proposed method performed well as compared to other methods in terms of high DR, low FPR and high purity at the same time.

Figure 3 The results of proposed method for outlier dataset (K = 20), (a) centre initialisation (b) after training (see online version for colours)



(a) (b)

Figure 4 MSE results with different K for outlier dataset, (a) K = 7 (b) K = 12 (c) K = 20 (see online version for colours)

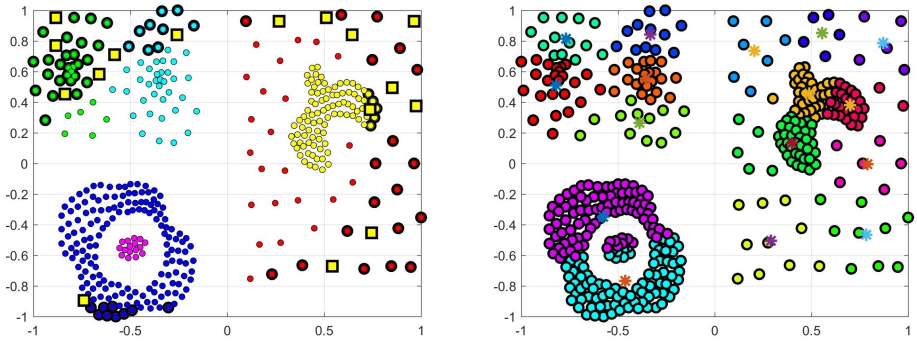


(a) (b) (c)

Table 2 The average results of clustering for outlier dataset

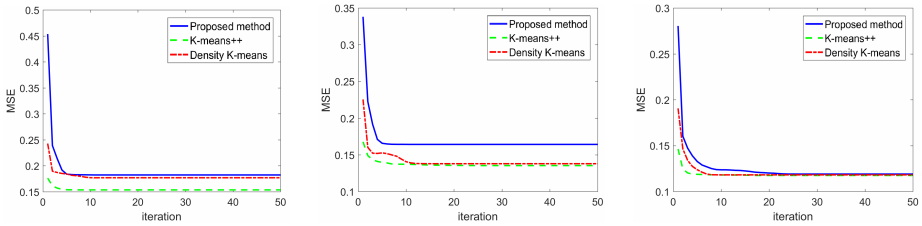
K (centre)	*Methods*	*MSE*	*Std.*	*DR (%)*	*FPR (%)*	*Purity*
7	Proposed method	0.237	0.1023	83.17	3.43	0.96
	K-means++	0.2312	0.1001	82.66	3.45	0.955
	Density K-means	0.2313	0.1002	82.53	3.88	0.954
12	Proposed method	0.174	0.073	92.86	1.40	0.98
	K-means++	0.169	0.0758	88.30	2.46	0.96
	Density K-means	0.18	0.0796	90.68	2.71	0.96
20	Proposed method	0.1354	0.0555	100	0	1
	K-means++	0.1279	0.0533	98.33	0.45	0.995
	Density K-means	0.1298	0.0541	94.73	0.84	0.98

Figure 5 The results of proposed method for compound dataset (K = 17), (a) centre initialisation (b) after training (see online version for colours)



(a) (b)

Figure 6 MSE results with different K for compound dataset, (a) K = 12 (b) K = 17 (c) K = 20 (see online version for colours)



(a) (b) (c)

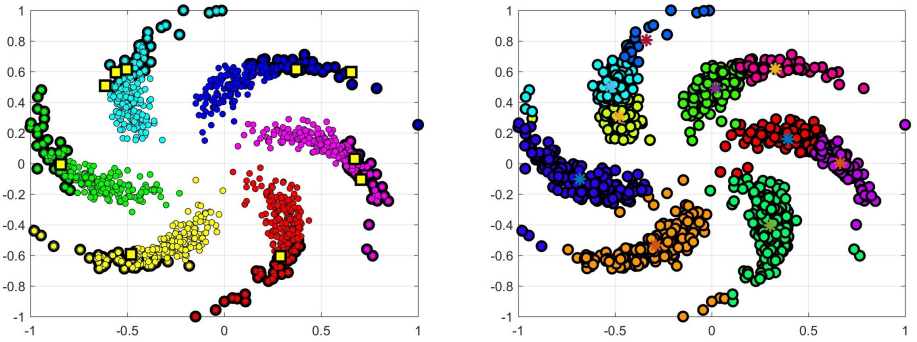
Table 3 The average results of clustering for compound dataset

K (centre)	*Methods*	*MSE*	*Std.*	*DR (%)*	*FPR (%)*	*Purity*
12	Proposed method	0.1824	0.088	94.22	5.28	0.925
	K-means++	0.1537	0.0925	91.09	5.33	0.904
	Density K-means	0.177	0.0868	89.77	3.86	0.897
17	Proposed method	0.1641	0.0807	96.10	3.04	0.9323
	K-means++	0.1353	0.0669	94.27	6.15	0.9223
	Density K-means	0.1379	0.0933	93.28	3.97	0.929
20	Proposed method	0.119	0.067	97.73	1.19	0.9674
	K-means++	0.1175	0.0634	95.76	2.22	0.9599
	Density K-means	0.1181	0.068	95.35	1.44	0.9574

5.2 *Results of benchmarking data*

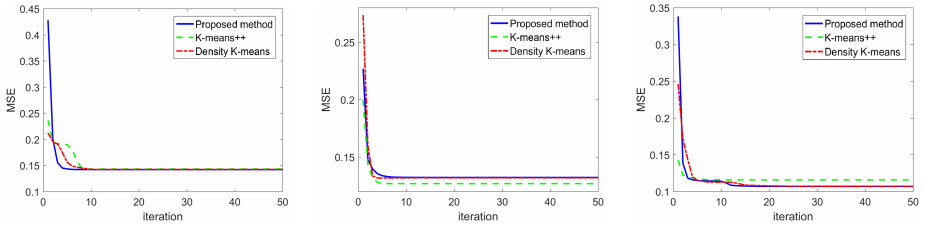
We added a white Gaussian noise to benchmarking datasets for creating low frequency noises as benign outliers. In this experiment, the signal-to-noise ratio is 10 dB. Figure 9 shows a sample of added white Gaussian noise to sawtooth signal.

Figure 7 The results of proposed method for pinwheel dataset (K = 10), (a) centre initialisation (b) after training (see online version for colours)



(a) (b)

Figure 8 MSE results with different K for Pinwheel dataset, (a) K = 8 (b) K = 10 (c) K = 14 (see online version for colours)



(a) (b) (c)

Table 4 The average results of clustering for Pinwheel dataset

K (centre)	*Methods*	*MSE*	*Std.*	*DR (%)*	*FPR (%)*	*Purity*
8	Proposed method	0.1421	0.0998	97.73	1.32	0.975
	K-means++	0.1432	0.102	97.02	1.61	0.9717
	Density K-means	0.1424	0.1	96.83	1.80	0.9708
10	Proposed method	0.1322	0.0942	98.26	1.22	0.979
	K-means++	0.1268	0.0897	97.71	1.46	0.976
	Density K-means	0.1319	0.0939	97.12	1.52	0.972
14	Proposed method	0.107	0.0725	99.03	0.59	0.989
	K-means++	0.1157	0..0834	98.19	0.9	0.983
	Density K-means	0.1069	0.0764	98.47	1.011	0.98

All experiments over benchmarking datasets were run 10 times with different K values, and the average classification error (Ave.) and the standard deviation (SD) were computed. In the conducted experiments, 70% of dataset is used for training and the rest is considered as test data in order to validate the quality of the proposed method. After training, the label (class) of each formed cluster comes from the largest number of a class within a same cluster. For instance, if a cluster contains two instances from class A, 12 instances from class B, and one instance from class C, the label of this cluster is considered as class B and the three instances from classes A and C are

considered as misclassification. The results have been summarised in Table 5. The results show that the proposed method tends to obtain more accurate classification rate (Ave.) and lower SD as compared to other methods.

Figure 9 Added white Gaussian noise to sawtooth signal (see online version for colours)

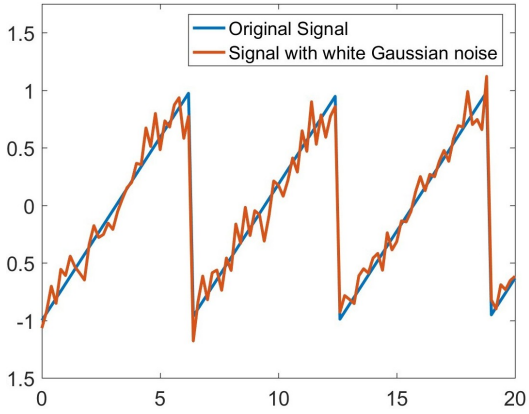


Table 5 Classification error (%) of applied methods over benchmarking datasets

Method	*Type*	*Criteria*	*Dataset*				
			Iris	*Glass*	*Wine*	*Ionosphere*	*Zoo*
Proposed method	Training	Ave.	3.804	12.684	13.854	8.375	7.51
		SD	0.83	1.958	1.747	1.892	2.035
	Test	Ave.	3.361	11.345	12.91	7.286	6.631
		SD	0.858	2.02	1.927	1.892	1.688
K-means++	Training	Ave.	5.211	16.689	16.011	9.533	9.727
		SD	2.129	2.836	2.86	2.593	2.648
	Test	Ave.	4.407	15.437	16.203	9.464	9.592
		SD	2.118	2.958	2.675	2.532	2.157
Density K-means	Training	Ave.	4.461	16.474	15.875	10.658	9.201
		SD	1.563	2.739	2.442	2.941	2.287
	Test	Ave.	4.496	15.279	15.523	9.078	7.899
		SD	1.572	2.765	2.23	2.578	2.311
K-means	Training	Ave.	4.74	15.043	17.112	10.103	10.01
		SD	2.165	3.401	3.545	4.001	3.877
	Test	Ave.	6.192	16.123	15.455	9.912	9.136
		SD	2.152	3.131	4.031	4.002	3.563

6 Conclusions

In this paper, a new centroid initialisation method for K-means clustering algorithm was introduced. The proposed method firstly considered and discovered benign outliers which exist inherently in almost all datasets. Benign outliers are usually out of 90%–95% of confidential interval of normal distribution of data. According to the experimental results, clusters' centroids initialisation through these tangible data points constructed clusters with higher accuracy as compared to some existing methods.

References

Alaei, A., Conte, D., Martineau, M. and Raveaux, R. (2018) 'Blind document image quality prediction based on modification of quality aware clustering method integrating a patch selection strategy', *Expert Systems with Applications*, Vol. 108, pp.183–192.

Arthur, D. and Vassilvitskii, S. (2007) 'K-means++: the advantages of careful seeding', *Proceedings of the Eighteenth Annual ACM-SIAM Symposium on Discrete Algorithms, SODA '07*, pp.1027–1035, Society for Industrial and Applied Mathematics.

Celebi, M.E., Kingravi, H.A. and Vela, P.A. (2013) 'A comparative study of efficient initialization methods for the k-means clustering algorithm', *Expert Systems with Applications*, Vol. 40, No. 1, pp.200–210.

Gan, G. and Ng, M.K-P. (2017) 'K-means clustering with outlier removal', *Pattern Recognition Letters*, Vol. 90, pp.8–14.

Hautamäki, V., Cherednichenko, S., Kärkkäinen, I., Kinnunen, T. and Fränti, P. (2005) 'Improving k-means by outlier removal', *Scandinavian Conference on Image Analysis*, pp.978–987.

Jach, A. and Kokoszka, P. (2008) 'Wavelet-based confidence intervals for the self-similarity parameter', *Journal of Statistical Computation and Simulation*, Vol. 78, No. 12, pp.1181–1200.

Karami, A. (2018) 'An anomaly-based intrusion detection system in presence of benign outliers with visualization capabilities', *Expert Systems with Applications*, Vol. 108, pp.36–60.

Karami, A. and Guerrero-Zapata, M. (2014) 'Mining and visualizing uncertain data objects and named data networking traffics by fuzzy self-organizing map', *Proceedings of the Second International Workshop on Artificial Intelligence and Cognition (AIC 2014)*, pp.156–163.

Karami, A. and Guerrero-Zapata, M. (2015a) 'A fuzzy anomaly detection system based on hybrid PSO-kmeans algorithm in content-centric networks', *Neurocomputing*, Vol. 149, Part C, pp.1253–1269.

Karami, A. and Guerrero-Zapata, M. (2015b) 'A hybrid multiobjective RBF-PSO method for mitigating DoS attacks in named data networking', *Neurocomputing*, Vol. 151, Part 3, pp.1262–1282.

Karami, A. and Johansson, R. (2014) 'Choosing dbscan parameters automatically using differential evolution', *International Journal of Computer Applications*, Vol. 91, No. 7, pp.1–14.

Khanmohammadi, S., Adibeig, N. and Shanehbandy, S. (2017) 'An improved overlapping k-means clustering method for medical applications', *Expert Systems with Applications*, Vol. 67, pp.12–18.

Li, C.S. (2011) 'Cluster center initialization method for k-means algorithm over data sets with two clusters', *Procedia Engineering*, Vol. 24, pp.324–328.

Marghny, M. and Taloba, A.I. (2011) 'Outlier detection using improved genetic k-means', *International Journal of Computer Applications*, No. 11, pp.33–36.

Marr, P. (2015) *SPSS Mini-Lab: Outliers* [online] http://webspace.ship.edu/pgmarr/Geo441/Lectures/OPT%201%20-%20Outlier%20Detection.pdf (accessed 1 July 2017).

Min, Z. and Kai-fei, D. (2015) 'Improved research to k-means initial cluster centers', *Ninth International Conference on Frontier of Computer Science and Technology (FCST)*, pp.349–353.

Santhanam, T. and Padmavathi, M. (2015) 'Application of k-means and genetic algorithms for dimension reduction by integrating SVM for diabetes diagnosis', *Procedia Computer Science*, Vol. 47, pp.76–83.

Shahreza, M.L., Moazzami, D., Moshiri, B. and Delavar, M. (2011) 'Anomaly detection using a self-organizing map and particle swarm optimization', *Scientia Iranica*, Vol. 18, No. 6, pp.1460–1468.

Shin, W., Cho, K.H. and Kim, J.J. (2017) 'How to estimate k value without domain knowledge in k-means', *3rd International Conference on Control, Automation and Robotics (ICCAR)*, pp.701–704.

Yi, J., Huang, D., Fu, S., He, H. and Li, T. (2016) 'Optimized relative transformation matrix using bacterial foraging algorithm for process fault detection', *IEEE Transactions on Industrial Electronics*, Vol. 63, No. 4, pp.2595–2605.

Yuan, Q., Shi, H. and Zhou, X. (2015) 'An optimized initialization center k-means clustering algorithm based on density', *IEEE International Conference on Cyber Technology in Automation, Control, and Intelligent Systems (CYBER)*, pp.790–794.