Hindawi
Applied Computational Intelligence and Soft Computing
Volume 2022, Article ID 1036293, 10 pages
https://doi.org/10.1155/2022/1036293

Research Article

Enhanced Connectivity Validity Measure Based on Outlier Detection for Multi-Objective Metaheuristic Data Clustering Algorithms

Hossam M. J. Mustafa [1] **and Masri Ayob** [2]

[1]*Department of Computer System and Computer Science, Faculty of Computer Science and Informatics, Amman Arab University, Amman, Jordan*
[2]*Data Mining and Optimization Research Group, Center for Artificial Intelligence Technology, Faculty of Information Science and Technology, University Kebangsaan Malaysia, 43600 Bangi, Malaysia*

Correspondence should be addressed to Hossam M. J. Mustafa; h.mustafa@aau.edu.jo

Received 7 November 2021; Revised 15 February 2022; Accepted 5 March 2022; Published 28 March 2022

Academic Editor: Upaka Rathnayake

Data clustering algorithms experience challenges in identifying data points that are either noise or outlier. Hence, this paper proposes an enhanced connectivity measure based on the outlier detection approach for multi-objective data clustering problems. The proposed algorithm aims to improve the quality of the solution by utilising the local outlier factor method (LOF) with the connectivity validity measure. This modification is applied to select the neighbour data point's mechanism that can be modified to eliminate such outliers. The performance of the proposed approach is assessed by applying the multi-objective algorithms to eight real-life and seven synthetic two-dimensional datasets. The external validity is evaluated using the F-measure, while the performance assessment matrices are employed to assess the quality of Pareto-optimal sets like the coverage and overall non-dominant vector generation. Our experimental results proved that the proposed outlier detection method has enhanced the performance of the multi-objective data clustering algorithms.

1. Introduction

Data clustering intends to arrange collections of data points using similarity functions that can be employed next to understand the data. A diversity of applications utilised the data clustering algorithms to recognise the embedded structures within the data, and to analyse a precise collection of clusters to be additionally investigated and to recognise each cluster feature [1, 2]. Consequently, the quality of the clusters can be handled by utilising the internal validity/similarity measures, such as connectedness, compactness, and isolation. The data clustering validity measures serve as an important part in the development of the clustering algorithms, which are built based on distance measures such as the k-means partitioning algorithm. In general, the partitioning algorithms aim to identify spherically shaped clusters, but it is inefficient to recognise arbitrarily shaped clusters like non-convex or interlaced clusters that are studied in several applications. Moreover, the partitioning algorithms experience challenges in recognising data points that are either outlier or noise [3]. Unlike other validity measures, cluster connectivity works indifferently with the shape of clusters [4], which decides the degree to which neighbours of a data point have been located in the corresponding cluster. However, the robustness of the connectivity measure depends on the associated L-nearest neighbour [5, 6]. These neighbours concerned in quantifying the connectivity measure can contain outliers, which can extremely influence the accuracy of the connectedness based on non-reliable data points that can be a form of outliers [7]. Therefore, choosing a proper neighbour data point's mechanism can be adjusted to eliminate such outliers, to enhance the performance of the connectivity measure. Data clustering and outlier detection share a corresponding

relationship, in which a data point is recognised as a cluster member or an outlier. Data clustering algorithms commonly incorporate a mechanism for managing the outliers that eliminate these data points from the clusters. The applicability across the different problem fields is one significant problem for the outlier analysis [7–10]. Also, the effectiveness of an outlier analysis algorithm is quantified with the performance of the resolution of different thresholds for the outlier score.

The local distance methods have been applied in several outlier detection methods [7, 11]. The primary assumption of these methods is that the normal data points reside within dense neighbourhoods. In contrast to normal data, the outliers reside remotely out from the nearest neighbours. One of the most common local distance outliers detection algorithms is the local outlier factor (LOF) algorithm, which is used in several applications [7]. LOF is recognised as one of the widely applied local outliers detection algorithms and was introduced by [12], in which the local density of a point is associated with the surrounding neighbourhood points [7, 13]. Although the LOF geometric anticipation is employed in low-dimensional data, the LOF algorithm can be implemented in different dissimilarity functions [14]. The LOF algorithm has shown outperformance against different competitor algorithms in several disciplines such as fault detection [15] or network intrusion detection [16]. The LOF variants can be generalised and implemented in various applications, such as detecting outliers in big data [17], machine learning [18], and data streams [19]. Additionally, the LOF algorithm can be employed for different cluster shapes with different dissimilarity functions, while other local distance methods such as connectivity-based outlier factor (COF) deals with outliers differing from spherical density-based shapes such as lines, while the influenced outlierness (INFLO) method handles the clusters that reside near to each other, and the local outlier probability (LoOP) method utilises the measurement of data points in the corresponding dataset with other datasets. To solve the concerns explained above, this paper intended to address the multi-objective data clustering problems using an outlier detection approach. The contribution significance of the paper is twofold.

(1) We introduced a modified connectivity validity measure based on the outlier detection approach (coded as Conn_LOF) for multi-objective data clustering problems.

(2) We developed an algorithm that intends to enhance the quality of the solution generated by the multi-objective metaheuristic approach by utilising the LOF with the connectivity validity measure.

This paper is organised as follows: The related works of multi-objective metaheuristic clustering are briefly reviewed in Section 2. Section 3 discusses the theoretical background and concepts such as the data clustering problem, outlier detection methods, and the LOF method. In section 4, the description of the modified Conn_LOF approach is presented. Section 5 presents the experimental design of the modified Conn_LOF approach algorithm, and in Section 6 the experimental results of the introduced method are explained. Finally, Section 7 presents the paper's conclusions and future works.

2. Related Works

Several multi-objective metaheuristics approaches have been introduced to solve data clustering problems [20–26]. The multi-objective data clustering approach was initially offered by [27], where they proposed a multi-objective data clustering algorithm that was based on one or more cluster quality measures. Their algorithm used the Pareto envelope-based selection algorithm (PESA-II), a multi-objective algorithm, to optimise the deviation and connectivity cluster quality measures. Their research was extended in [28], where they investigated the performance of four different pairs of criteria (cluster quality measures) in multi-objective clustering. Reference [29] introduced a new dynamic multi-objective evolutionary algorithm (MOEA) for data clustering, which applies a chromosome with variable length scheme to search for optimal cluster number and cluster centre. Reference [30] proposed a multi-objective optimisation algorithm for solving the categorical data clustering problem (MOGA). Reference [31] offered a multi-objective evolutionary ensemble algorithm for addressing texture image segmentation (MECEA). Reference [32] introduced an enhanced multi-objective evolutionary approach for data clustering (EMCOC), which aims to determine the overlapping complex shape dataset problem. Reference [33] offered a multi-objective genetic fuzzy clustering (MOVGA) for the segmentation of multispectral magnetic resonance imaging (MRI). Reference [34] proposed a multi-objective clustering algorithm (MOCA) for data clustering.

Recently, [35] proposed a multi-objective algorithm based on the artificial bee colony optimisation algorithm and the non-dominated sorting (NSABC) to solve the data clustering problems. Reference [21] offered a particle swarm optimisation using the multi-objective approach (MOPSO) to increase the diversity of the solutions. Later, [36] presented an improved binary gravitational search algorithm using the multi-objective approach for feature selection (IMBGSAFS). The Pareto-based approach is used in the algorithm to obtain better solutions diversity, by optimising the silhouette index and feature cardinality validity measures. Reference [37] introduced the multi-objective clustering algorithm based on a reduced-length representation. Reference [23] proposed a kernel-based, attribute-weighted multi-objective optimisation data clustering algorithm, in which they used the compactness and the separation cluster quality measures to find an optimal clustering solution.

Table 1 demonstrates that most of the offered multi-objective clustering approaches were based on the NSGA-II multi-objective algorithm, which was widely used to achieve high-quality solutions. Several multi-objective clustering algorithms employ more than one validity measure to be optimised simultaneously, which minimises two validity

Table 1: Summary of the popular multi-objective metaheuristic algorithms for data clustering with their related details.

Algorithm	Objective functions	Algorithm	Reference
VIENNA	Dev(C) & Conn(C)	PESA-II	[27]
MOEA	Dev(C) & Conn(C)	NSGA-II	[29]
MOCK	Dev(C) & Conn(C)	PESA-II	[5]
VRJGGA	Entropy & separation	NSGA-II	[32]
MOGA-medoid	Dev(C) & silhouette	NSGA-II	[30]
MECEA	Dev(C) & Conn(C)	PESA-II	[31]
EMCOC	Entropy & separation	NSGA-II	[32]
MOVGA	J_m & separation	NSGA-II	[33]
MOCA	Avg. Dev(C) & Conn(C)	NSGA-II	[34]
TSMPSO	SSE & Conn(C)	NSGA-II	[21]
NSABC	SSE & Conn(C)	NSGA-II	[35]
IMBGSAFS	Silhouette & cardinality	NSGA-II	[36]
MOKCW	Compactness & separation	NSGA-II	[23]

measures such as cluster connectivity (Conn) and overall cluster deviation (Dev).

According to the related studies of data clustering algorithms, which are based on the multi-objective metaheuristic algorithms, further enhancements are required to tackle the rapid growth of data complexity with the consideration of preserving the accuracy of the clustering algorithm [7]. Although the majority of the clustering algorithms attempt to detect outliers during the clustering analysis stage [7], few algorithms offer validity measures that can tackle the detection of these outliers [38]. The connectivity measure of the cluster, which is commonly used in most multi-objective clustering algorithms, can measure the level of the connectedness of the neighbour data objects that are located in the same cluster [6, 35] and may measure the amount of connectedness based on non-reliable data objects that can be a form of outliers [7]. Therefore, the selection of a suitable neighbour data objects mechanism can be modified to exclude such outliers, and consequently improve the performance of the connectivity measure.

3. Background

This section introduces the concepts of the data clustering problems, the outlier detection methods, and the LOF method.

3.1. Data Clustering Problems. Data clustering is an essential task of data mining that intends to group N data objects $X = \{x_1, x_2, \ldots, x_N\}$ into a set of clusters $C = \{C_1, C_2, \ldots, C_K\}$, where all data objects in the same clusters are similarly based on a specified similarity measure. The clustering methods must ensure the following hard constraints [39]:

(i) Each cluster should not be empty and hold at least one data object:

$$C_j \neq \phi, \quad \forall j \in \{1, 2, \ldots, K\}. \tag{1}$$

(ii) Various clusters should not share data objects:

$$C_j \cap C_i = \phi, \quad \forall j \neq i \text{ and } j, i \in \{1, 2, \ldots, K\}. \tag{2}$$

(iii) Every data object should be included in a cluster:

$$\bigcup_{j=1}^{k} C_j = X. \tag{3}$$

The mathematical representation of a multi-objective data clustering problem with M-objectives is given in equation (4) [40]:

$$\textbf{Optimize } f(X, C) = \left(f_1(X, C), f_2(X, C), \ldots f_M(X, C)\right),$$
$$\textbf{subject to } \begin{cases} g_i(X, C) \leq 0, & i = 1, 2, \ldots, p, \\ h_j(X, C) = 0, & j = 1, 2, \ldots, q. \end{cases} \tag{4}$$

The $f(X, C)$ is the objective function that measures the partitions' quality produced by the clustering algorithm, where the objective function can be minimised or maximised depending on the similarity/dissimilarity measure employed. $g_i(X, C)$ denotes the p inequality constraints, and $h_j(X, C)$ denotes the q equality constraints.

3.2. Connectivity of the Cluster. Connectivity of the cluster [27, 35] is an objective function used to measure the amount of neighbour data points that are placed in each cluster that should be minimised. The mathematical formulation of the cluster connectivity is shown in equations (5) and (6):

$$\text{connectivity}(C) = \sum_{i=1}^{N} \sum_{j=1}^{M} nn_i(j), \tag{5}$$

$$nn_i(j) = \begin{cases} \dfrac{1}{j}, & \text{if object } i \text{ is not in the same cluster of object } j, \\ 0, & \text{otherwise,} \end{cases} \tag{6}$$

where N is the number of data points, and parameter M represents the number of neighbour data points, which will be considered to measure the connectivity.

3.3. Outlier Detection Methods. The outlier detection methods are applied to overcome the influence of the outlier in creating descriptive or predictive models, and also to be adopted in the pre-processing stage in several applications of data mining. The common outlier detection techniques are classified into distance-based, density-based, distribution-based, clustering-based, and probabilistic-based methods. Besides, the outlier detection approaches are divided into local or global methods, in which global methods give each data point an anomaly score depending on the entire dataset points. On the contrary, the local distance methods assign an anomaly score to each data point depending on the surrounding neighbourhoods. Many variants of the local distance methods are introduced to produce simple anomaly score presentation and identify hidden outliers by the global methods. The variants of the local distance methods include the following methods:

(1) *Local Outlier Factor (LOF)* [12]**:** It is recognised as the most broadly adopted local methods that associates the local density of data objects with the average distance of the k-nearest-neighbour objects. The anomaly score of the LOF algorithm is defined as the ratio of the data points' local density to the neighbourhood points' average local density.

(2) *Connectivity-based Outlier Factor (COF)* [41]**:** It detects outliers of other density-based shapes like lines.

(3) *Influenced Outlierness (INFLO)* [42]**:** It was introduced to produce further reliable results involving the different clusters' densities that exist near each other.

(4) *Local Outlier Probability (LoOP)* [43]: It consists of statistical methods that define the anomaly score as a probability. These probabilities employ the analysis of data points in the dataset with other datasets.

The local distance methods have been utilised in several outlier detection methods [7, 44–46]. The primary assumption of these methods is that the points of normal data exist inside dense neighbourhoods. Unlike normal data, the outliers remain remotely out from the nearest neighbours. The nearest neighbour methods need a distance metric to identify the distance separating the two data points [7]. One of the popular local distance outliers detection algorithms is the LOF algorithm, which is applied in several applications [7].

3.4. Local Outlier Factor (LOF). LOF is one of the commonly used local outliers detection algorithms that was introduced by [12], in which the local density of a point is related to the surrounding neighbourhood points [7, 13]. The outlier factor is local which considers only each neighbourhood point. The local reachability distance of a point p is described as the inverse of the average reachability distance based on the minPts_nearest neighbours of p. Thus, minPts is a primary parameter needed by the LOF algorithm which indicates the number of nearest neighbours employed in discovering the local neighbourhood of each point. The local reachability distance (*lrd*) is defined by equation (7), and the reachability distance is defined by equation (8) [12]:

$$\mathrm{Ird}_{\mathrm{minpts}}(p) = \frac{1}{\left\{\left(\sum_{o \in N_{\mathrm{minpts}(p)}} \mathrm{reach\_dist}_{\mathrm{Minpts}}(p,o)\right)/N_{\mathrm{minpts}(p)}\right\}}, \tag{7}$$

$$\mathrm{reach\_dist}_{\mathrm{Minpts}}(p,o) = \max\{\mathrm{minPts\_distance}(o), \mathrm{dist}(p,o)\}, \tag{8}$$

where minPts denotes a positive integer, D denotes the dataset points, and $\{o, p\} \in D$. The $\mathrm{dist}_{\mathrm{Minpts}(p,\,o)}$ is defined as the distance between p and point o. Given the minPts_distance of p, the minPts_distance neighbourhood of p contains every point whose distance from p is not greater than the minPts_distance. The outlier factor of point p represents the level of point p to be considered an outlier, which is defined in equation (9) [12]**:**

$$\mathrm{LOF}_{\mathrm{minpts}}(p) = \frac{\sum_{o \in N_{\mathrm{minpts}(p)}} \left(\mathrm{Ird}_{\mathrm{minpts}}(o)/\mathrm{Ird}_{\mathrm{minpts}}(p)\right)}{\left|N_{\mathrm{minpts}(p)}\right|}. \tag{9}$$

The utilisation of distance ratios ensures that the local distance performance is properly assessed. Therefore, the $\mathrm{LOF}_{\mathrm{minpts}}$ for the points in density regions is close to 1 ($\mathrm{LOF} \simeq 1$). Otherwise, the $\mathrm{LOF}_{\mathrm{minpts}}$ of the outlier points will be much higher ($\mathrm{LOF} \gg 1$) because they are measured depending on the ratios to the average neighbour reachability distances. Essentially, the maximum value of $\mathrm{LOF}_{\mathrm{minpts}}$ over a variety of minpts amount is employed as the outlier score to identify the optimal neighbourhood size.

3.5. The Proposed Outlier Detection Approach. The proposed outlier detection approach of the connectivity measure (named Conn_LOF) is discussed in this section. The flowchart of the introduced outlier detection approach for the connectivity measure is shown in Figure 1, which includes the following stages:

(i) *Stage 1.* The pre-processing phase includes the gathering and cleaning of the needed datasets and then converting them into related nearest

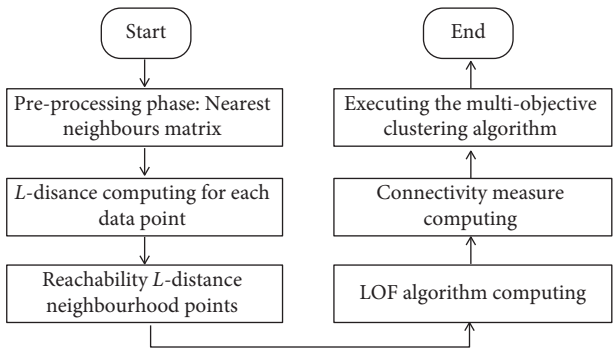

Figure 1: Flowchart of the proposed Conn_LOF algorithm.

neighbours matrix and other matrices, which are utilised throughout the generation of solutions.

(ii) *Stage 2.* The computation of L-distance: includes the computation of L-distances of each data point with the L-nearest neighbourhood data points based on the Euclidean distance [7].

(iii) *Stage 3.* The computation of reachability of L-distance neighbourhoods: consists of the computation of the local reachability distances along with the reachability of all L-distance neighbourhoods using equation (7).

(iv) *Stage 4.* The LOF algorithm computation: consists of labelling the outlier sequence of the entire L-distance neighbourhoods based on the outlier factor based on the chosen threshold value λ of LOF.

(v) *Stage 5.* The computation of the connectivity measure includes the computation of the connectivity validity measure using equation (5). The procedure of computing the connectivity measure excludes the outlier-labelled neighbourhoods' points.

(vi) *Stage 6.* The execution of the multi-objective clustering algorithm: executes the multi-objective clustering algorithm such as the non-dominated sorting genetic algorithm (NSGA-II) [47] and the strength Pareto evolutionary algorithm (SPEA-II) [48].

The algorithmic steps of the proposed method are shown in Algorithm 1, where λ denotes the threshold value used in the LOF algorithm that is set to 1, where the LOF value of each neighbourhoods point is approximated and then compared to the λ threshold value. The C_{label} matrix stores the labels of the neighbourhoods' points.

4. Experimental Design

The performance of the proposed Conn_LOF outlier detection method is examined using eight real-life datasets with a variety of complexity, obtained from the UCI repository of the machine learning databases [49], and seven synthetic two-dimensional datasets [5], as shown in Table 2.

Since most of the state-of-the-art multi-objective clustering algorithms are based on NSGA-II (as shown in Table 1), NSGA-II and SPEA-II algorithms are used to prove the contribution of this paper. Additionally, other multi-objective algorithms are not used since the proposed Conn_LOF method is performed before running the multi-objective clustering algorithm (as shown in Figure 1) and will not affect the algorithmic steps of any given algorithm.

To evaluate the performance and the effectiveness of the proposed Conn_LOF method, the NSGA-II algorithm [47] is modified by employing two conflicting objectives that include the intra-cluster distance [50] and the proposed Conn_LOF method (named as eNSGA-II) and compared with the NSGA-II algorithm with a pair of conflicting objectives that include the intra-cluster distance [50] and the standard connectivity of the cluster [27]. Similarly, SPEA-II [48] is modified by employing the intra-cluster distance and the Conn_LOF method (named as eSPEA-II) and then compared with the standard SPEA-II with a pair of conflicting objectives that include the intra-cluster distance [50] and the standard connectivity of the cluster [27].

The data clustering solutions are represented using a label-based representation that includes a one-dimensional array, where a solution is denoted as a set of N data objects. Figure 2 demonstrates a solution representation example of eight data objects and three clusters. The solutions are randomly generated. Each data object is randomly attached to a cluster.

The algorithm's external validity is evaluated using the F-measure [51]. The running time of the algorithms is not investigated since the Conn_LOF method runs before the execution of the multi-objective clustering algorithm (as shown in Figure 1), which will not affect the running time of these competing algorithms. The inference time is the same for a particular dataset depending on the number of attributes and instances.

Also, performance assessment indices (PI) are utilised to assess the Pareto-optimal sets' quality and to compare the performance between diverse multi-objective algorithms. Hence, to assess the multi-objective metaheuristic clustering algorithms, we followed the performance indices that have been used in recent data clustering researches [36, 52], including the Overall Non-dominated Vector Generation (ONVG) [53] and coverage [54]. The details of these indices are below:

1. Coverage of Two Sets (C) [54]: Coverage is employed to compare two solution sets based on domination. Assuming that S_1 and S_2 are two Pareto-fronts/sets, then $C(S_1, S_2)$ indicates the portion of set S_2 that is dominated by the solutions in set S_1. The mathematical formulation of the coverage is shown in equation (10).

$$C(S_1, S_2) = \frac{\left|\{b_2 \in S_2; \exists b_1 \in S_1 : b_1 \leq b_2\}\right|}{\left|S_2\right|}, \tag{10}$$

where higher values of C denote that the dominance is better, which must be within the range [0, 1].

2. Overall Non-dominant Vector Generation (ONVG) [53] represents the number of solutions in the Pareto-front set S; the mathematical formulation of the ONVG is shown in equation (11).

```
(i) //Inputs:
(ii) C//the nearest neighbours matrix that is generated from the stage (1)
(iii) L//number of nearest neighbours minPts in LOF algorithm
(iv) λ//The threshold used in the LOF algorithm
(v) C_label//the labels matrix generated by LOF
(vi) for each C_j in C do
(1)          //stage (2)
(vii)              Compute the L-distance neighbourhood points of C_j;
(2)          //stage (3)
(viii)             Compute the reachability distance for neighbourhood
(3)                points of C_j;
(xi)         //stage (3)
(x)                Compute the LOF of neighbourhood points of C_j;
(4)          //stage (4)
(ix)   for each neighbourhood point, P_i of C_j do
(5)          If LOF of P_i ≥ λ then
(6)                Label P_i as outlier and store it C_label;
(7)          Endif
(8) End for
(9) //stage (5)
(10) Compute connectivity of C by excluding outliers in C_label;
(11) //stage (6)
(12) Execute the multi-objective clustering algorithm;
```

Algorithm 1: Pseudo-code of the proposed LOF-based algorithm.

Table 2: The real-life and synthetic datasets used in the experiments of the proposed algorithm.

Real-life datasets	Synthetic datasets
CMC	2d-20c-no0
Ecoli	Elly-2d10c13s
Ionosphere	Engytime
Iris	Flame
Seeds	Sizes5
Sonar	Spherical_5_2
Soybean-small	Square1
Thyroid	

O_1	O_2	O_3	O_4	O_5	O_6	O_7	O_8
C1	C3	C2	C2	C1	C3	C3	C1

Figure 2: A candidate solution representation example.

$$\mathrm{ONVG}(S) = |S|. \tag{11}$$

To evaluate the performance of the multi-objective methods using the PI indices, a Pareto-front pool is generated utilising the whole Pareto-fronts of the competing multi-objective algorithms. The non-dominated solutions in N runs of every algorithm are joined. Some PIs require a Pareto-front pool such as the coverage measure.

The setting of the parameters for the competing algorithms was independently performed 31 times on each of the 15 datasets; then the average value and the standard deviation of the F-measure are computed. The population size is set to 20 and the maximum number of iterations is set to 1000. The nearest L data points are set to 21. Lastly, Java 1.8 is used to implement the algorithms and were run on a personal computer with a CPU of Intel Core i7 (2.6 GHz) that was equipped with 4 GB memory.

5. Experimental Results and Discussion

Table 3 shows the results of the coverage (C), where A, B, C, and D symbolise eNSGA-II, NSGA-II, eSPEA-II, and SPEA-II, respectively. The $C(A, B)$ values compared with $C(B, A)$ values obtained better coverage for the datasets 2d-20c-no0, CMC, Ecoli, engytime, Flame, Seeds, Sizes5, Sonar, Soybean-small, and Thyroid, which means that the entire solutions in the pool of NSGA-II at least have been dominated by a single solution of the eNSGA-II solutions pool. On the other hand, the $C(A, B)$ values compared to $C(B, A)$ mostly obtained better coverage for

Table 3: The coverage metric of the obtained Pareto-fronts by the competing algorithms from the combined pool of sets.

Dataset	C (A, B)	C (B, A)	C (C, D)	C (D, C)
2d-20c-no0	**0.85**	0	**1**	0
CMC	**0.71**	0.25	**1**	0
Ecoli	**0.50**	0	**1**	0
Elly-2d10c13s	0	**0.29**	**1**	0
Engytime	**0.67**	0.05	**1**	0
Flame	**0.38**	0.20	**0.89**	0
Ionosphere	0	**1**	0	**1**
Iris	0.13	**0.73**	0	**1**
Seeds	**0.64**	0.43	**0.41**	0.30
Sizes5	**0.40**	0	**0.43**	0
Sonar	**0.75**	0.73	0.54	**0.66**
Soybean-small	**0.55**	0.18	0.22	**0.41**
Spherical_5_2	0	**1**	**1**	0
Square1	0	**1**	**1**	0
Thyroid	**0.80**	0	**0.60**	0

Table 4: The average and standard deviation[a] of the obtained F-measure obtained by the competing algorithms.

Dataset	SPEA-II	eSPEA-II	NSGA-II	eNSGA-II
2d-20c-no0	0.208 (0.01)	**0.547** (0.04)	**0.57** (0.03)	0.551 (0.04)
CMC	**0.598** (0.04)	0.59 (0.037)	**0.583** (0.02)	0.583 (0.02)
Ecoli	0.783 (0.03)	**0.824** (0.02)	0.853 (0.03)	**0.858** (0.04)
Elly-2d10c13s	0.310 (0.01)	**0.556** (0.04)	0.580 (0.05)	**0.581** (0.04)
Engytime	0.645 (0.02)	**0.958** (0.10)	0.957 (0.10)	**0.957** (0.08)
Flame	0.864 (0.07)	**0.864** (0.05)	0.877 (0.05)	**0.877** (0.06)
Iris	**0.887** (0.02)	0.826 (0.01)	0.857 (0.01)	**0.863** (0.01)
Seeds	0.867 (0.02)	**0.880** (0.04)	0.876 (0.02)	**0.876** (0.02)
Sizes5	0.869 (0.01)	**0.87** (0.026)	**0.901** (0.03)	0.865 (0.02)
Soybean-small	**0.979** (0.06)	0.94 (0.063)	**0.957** (0.06)	0.875 (0.06)
Spherical_5_2	0.848 (0.07)	**0.885** (0.07)	0.888 (0.08)	**0.888** (0.07)
Square1	0.605 (0.03)	**0.978** (0.08)	0.940 (0.06)	**0.972** (0.06)
Thyroid	0.868 (0.01)	**0.868** (0.01)	0.861 (0.01)	**0.881** (0.01)

[a]The result shows the average F-measure and the results' standard deviation in brackets.

the datasets Elly-2d10c13s, Ionosphere, Iris, Spherical_5_2, and Square1. The C(C, D) values compared with C(D, C) values obtained better coverage for the datasets 2d-20c-no0, CMC, Ecoli, Elly-2d10c13s, engytime, Flame, Seeds, Sizes5, and Spherical_5_2, which means that the entire solutions in the pool of SPEA-II at least have been dominated by a single solution of the eSPEA-II solutions pool. In contrast, the C(D, C) values compared to C(C, D) mostly obtained better coverage for the datasets Ionosphere, Iris, Sonar, and Soybean-small.

Generally, this shows that the solutions in the modified algorithms with the Conn_LOF method's pool dominated the standard algorithms' solutions in a considerably high ratio. In conclusion, the modified algorithms with Conn_LOF method attained better performance amongst other standard algorithms based on the coverage PI.

Table 4 reveals the results of the obtained F-measure on the Pareto-fronts produced by the competing algorithms. The eNSGA-II algorithm achieves higher F-measure results than the NSGA-II algorithm for most of the datasets except 2d-20c-no0, CMC, Sizes5, and Soybean-small datasets. The eSPEA-II provides higher F-measure results than SPEA-II for most of the datasets excluding CMC, Iris, and Soybean-small datasets. The results verify that the average F-measure of the eNSGA-II and eSPEA-II is enhanced by adopting the Conn_LOF method compared to the corresponding standards NSGA-II, and SPEA-II.

Additionally, the impact of adopting the Conn_LOF is perceived in the ONVG metric, as shown in Table 5, in which the eNSGA-II algorithm achieves higher ONVG results than the NSGA-II algorithm for most of the datasets except 2d-20c-no0, Ecoli, and Seeds. The eSPEA-II provides higher ONVG results than SPEA-II for most of the datasets except 2d-20c-no0, Sizes5, and Soybean-small. The table also shows a weak performance of other competing algorithms concerning the ONVG metric. Hence, the modified eNSGA-II and eSPEA-II achieve better ONVG performance.

Results shown in Table 4 are additionally analysed using Friedman's test ranking using the F-measure. As presented in Table 6, Friedman's test shows that eNSGA-II achieved the best F-measure rank. The NSGA-II achieved the second rank, and the eSPEA-II algorithm achieved the third rank. Finally, SPEA-II obtained the worst rank.

In general, eNSGA-II, and eSPEA-II are proven to be a reliable choice for data clustering in the multi-objective approach by adopting the Conn_LOF outlier detection method for providing Pareto-front solutions with efficient clustering measures for datasets with varying characteristics and complexity.

Table 5: The ONVG metric of the obtained Pareto-fronts by the competing algorithms from the combined pool of sets.

Dataset	SPEA-II	eSPEA-II	NSGA-II	eNSGA-II
2d-20c-no0	**19**	5	**16**	13
Cmc	4	**11**	7	**8**
Ecoli	6	**7**	**6**	3
Elly-2d10c13s	6	**9**	7	**10**
Engytime	8	**23**	6	**9**
Flame	6	**9**	8	**10**
Ionosphere	1	**7**	7	**13**
Iris	11	**15**	11	**15**
Seeds	10	**12**	**14**	11
Sizes5	**8**	7	5	**5**
Sonar	6	**13**	8	**11**
Soybean-small	**12**	9	11	**11**
spherical_5_2	1	**8**	2	**10**
Square1	6	**6**	2	**8**
Thyroid	5	**12**	5	**10**

Table 6: Friedman test ranking for eNSGA-II, eSPEA-II, NSGA-II, and SPEA-II algorithms using the F-measure.

Algorithm	Rank
eNSGA-II	2.19
NSGA-II	2.27
eSPEA-II	2.46
SPEA-II	3.08

6. Conclusions and Future Work

In this paper, an enhanced connectivity measure based on the LOF outlier detection method (Conn_LOF) is offered to enhance the performance of the connectivity measure by eliminating the outliers. To examine the efficiency of the proposed Conn_LOF method, it is employed within the competing algorithms and tested on eight real-life datasets with a variety of complexity obtained from the UCI repository of the machine learning database. Thus, the efficiency of the competing algorithms is tested on seven synthetic two-dimensional synthetic datasets with different cluster shapes and characteristics. The experimental results show that the performance of the modified eNSGA-II and eSPEA-II enhanced by adopting the Conn_LOF method concerning the average, and the standard deviation results of the *F*-measure. Thus, the multi-objective performance assessment matrices are used to evaluate the quality of the Pareto-optimal sets that include coverage and overall non-dominant vector generation. Furthermore, the Conn_LOF outlier detection method is proven to be effective when combined with the clustering algorithms to provide better Pareto-front solutions with efficient clustering measures for datasets with varying characteristics and complexity.

Data Availability

The real-life datasets used to support the findings of this study have been deposited in the UCI Data repository (URLs: https://archive.ics.uci.edu/ml/datasets/Contraceptive+Method+Choice; https://archive.ics.uci.edu/ml/datasets/ecoli; https://archive.ics.uci.edu/ml/datasets/ionosphere; https://archive.ics.uci.edu/ml/datasets/Iris; https://archive.ics.uci.edu/ml/datasets/seeds; https://archive.ics.uci.edu/ml/datasets/connectionist+bench+(sonar,+mines+vs.+rocks); https://archive.ics.uci.edu/ml/datasets/soybean+(small); https://archive.ics.uci.edu/ml/datasets/thyroid+disease). Additional synthetic datasets (such as 2d-20c-no0, Elly-2d10c13s, Engytime, Flame, Sizes5, Spherical_5_2, and Square1) were used to support this study and are available at [doi: 10.1109/TEVC.2006.877146]. These prior datasets are cited at relevant places within the text as references [5].

Conflicts of Interest

The authors declare no conflicts of interest regarding this paper.

Acknowledgments

This research was funded by a research grant from Universiti Kebangsaan Malaysia (Ref. No: DIP-2019-013).

References

[1] L. M. Abualigah, A. T. Khader, and E. S. Hanandeh, "A new feature selection method to improve the document clustering using particle swarm optimization algorithm," *Journal of Computational Science*, vol. 25, pp. 456–466, 2018.

[2] V. Boeva, "Clustering approaches for dealing with multiple DNA microarray datasets," *Journal of Computational Science*, vol. 5, no. 3, pp. 368–376, 2014.

[3] D. Mustafi, G. Sahoo, and A. Mustafi, "An improved heuristic K-means clustering method using genetic algorithm based initialization," *Advances in Intelligent Systems and Computing*, vol. 509, pp. 123–132, 2017.

[4] M. Garza-Fabre, J. Handl, and J. Knowles, "A new reduced-length genetic representation for evolutionary multiobjective clustering," *Lecture Notes in Computer Science*, Springer, Berlin, Germany, pp. 236–251, 2017.

[5] J. Handl and J. Knowles, "An evolutionary approach to multiobjective clustering," *IEEE Transactions on Evolutionary Computation*, vol. 11, no. 1, pp. 56–76, 2007.

[6] A. Mukhopadhyay, U. Maulik, and S. Bandyopadhyay, "A survey of multiobjective evolutionary clustering," *ACM Computing Surveys*, vol. 47, no. 4, pp. 1–46, 2015.

[7] C. C. Aggarwal, *Data Mining*, Springer, Switzerland, 2015.

[8] N. A. Jamil, S. L. Wang, and T. F. Ng, "Self-adaptive differential evolution based on best and mean schemes," in *Proceedings of the 2015 IEEE International Conference on Control System, Computing and Engineering (ICCSCE)*, pp. 287–292, Penang, Malaysia, November 2015.

[9] N. A. Harun, M. Makhtar, A. Abd Aziz, Z. A. Zakaria, F. S. Abdullah, and J. A. Jusoh, "The application of apriori algorithm in predicting flood areas," *International Journal of Advanced Science, Engineering and Information Technology*, vol. 7, no. 3, p. 763, 2017.

[10] M. Sammour and Z. Othman, "An agglomerative hierarchical clustering with various distance measurements for ground level ozone clustering in putrajaya, Malaysia," *International Journal of Advanced Science, Engineering and Information Technology*, vol. 6, no. 6, p. 1127, 2016.

[11] G. Gan and M. K.-P. Ng, "*k*-means clustering with outlier removal," *Pattern Recognition Letters*, vol. 90, pp. 8–14, 2017.

[12] M. M. Breunig, H.-P. Kriegel, R. T. Ng, and J. Sander, "LOF: identifying density-based local outliers," *ACM SIGMOD Record*, vol. 29, no. 2, pp. 93–104, 2000.

[13] N. Malini and M. Pushpa, "Analysis on credit card fraud identification techniques based on KNN and outlier detection," in *Proceedings of the 2017 Third International Conference on Advances in Electrical, Electronics, Information, Communication and Bio-Informatics (AEEICB)*, pp. 255–258, Chennai, India, Febuary 2017.

[14] G. O. Campos, A. Zimek, J. Sander et al., "On the evaluation of unsupervised outlier detection: measures, datasets, and an empirical study," *Data Mining and Knowledge Discovery*, vol. 30, no. 4, pp. 891–927, 2016.

[15] L. Wang and X. Deng, "Multimode process fault detection method based on variable local outlier factor," in *Proceedings of the 2017 9th International Conference on Modelling, Identification and Control (ICMIC)*, pp. 175–180, Kunming, China, July 2017.

[16] J. Auskalnis, N. Paulauskas, and A. Baskys, "Application of local outlier factor algorithm to detect anomalies in computer network," *Elektronika ir Elektrotechnika*, vol. 24, no. 3, pp. 96–99, 2018.

[17] Y. Yan, L. Cao, C. Kulhman, and E. Rundensteiner, "Distributed local outlier detection in big data," in *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 1225–1234, Halifax, NS, Canada, August 2017.

[18] L. Qi and L. Ting, "Active semi-supervised affinity propagation clustering algorithm based on local outlier factor," in *Proceedings of the 2018 37th Chinese Control Conference (CCC)*, pp. 9368–9373, Wuhan, China, July 2018.

[19] S. Seo, S. Park, I. Hwang, and J. Kim, "ADSTREAM: anomaly detection in large-scale data streams using local outlier factor based on micro-cluster," *Advanced Science Letters*, vol. 23, no. 10, pp. 10204–10209, 2017.

[20] S. Das, S. Chaudhuri, and A. K. Das, "Optimal set of overlapping clusters using multi-objective genetic algorithm," in *Proceedings of the 9th International Conference on Machine Learning and Computing 2017*, pp. 232–237, Singapore, Febuary 2017.

[21] J. Prakash and P. K. Singh, "An effective multiobjective approach for hard partitional clustering," *Memetic Computing*, vol. 7, no. 2, pp. 93–104, 2015.

[22] S.-T. Wang, "An analysis of the optimal customer clusters using dynamic multi-objective decision," *International Journal of Information Technology and Decision Making*, vol. 17, no. 02, pp. 547–582, 2018.

[23] Z. Zhou and S. Zhu, "Kernel-based multiobjective clustering algorithm with automatic attribute weighting," *Soft Computing*, vol. 22, no. 11, pp. 3685–3709, 2018.

[24] E. Gajda-Zagórska, R. Schaefer, M. Smołka, D. Pardo, and J. Álvarez-Aramberri, "A multi-objective memetic inverse solver reinforced by local optimization methods," *Journal of Computational Science*, vol. 18, pp. 85–94, 2017.

[25] D. E. Hernández, E. Clemente, G. Olague, and J. L. Briseño, "Evolutionary multi-objective visual cortex for object classification in natural images," *Journal of Computational Science*, vol. 17, pp. 216–233, 2016.

[26] K. K. Bharti and P. K. Singh, "A three-stage unsupervised dimension reduction method for text clustering," *Journal of Computational Science*, vol. 5, no. 2, pp. 156–169, 2014.

[27] J. Handl and J. Knowles, "Evolutionary multiobjective clustering," *Lecture Notes in Computer Science*, Springer, vol. 3242, pp. 1081–1091, , Berlin, Germany, 2004.

[28] J. Handl and J. Knowles, "Clustering criteria in multiobjective data clustering," *Lecture Notes in Computer Science*, Springer, Berlin, Germany, pp. 32–41, 2012.

[29] E. Chen and F. Wang, "Dynamic clustering using multiobjective evolutionary algorithm," in *Computational Intelligence and Security*, Y. Hao, J. Liu, Y. Wang et al., Eds., Springer, Berlin, Germany, pp. 73–80, 2005.

[30] A. Mukhopadhyay and U. Maulik, "Multiobjective approach to categorical data clustering," in *Proceedings of the 2007 IEEE Congress on Evolutionary Computation*, pp. 1296–1303, Singapore, September 2007.

[31] X. Xiaoxue Qian, X. Xiangrong Zhang, L. Licheng Jiao, and W. Wenping Ma, "Unsupervised texture image segmentation using multiobjective evolutionary clustering ensemble algorithm," in *Proceedings of the 2008 IEEE Congress on Evolutionary Computation (IEEE World Congress on Computational Intelligence)*, pp. 3561–3567, Hong Kong, China, June 2008.

[32] K. S. N. Ripon and M. N. H. Siddique, "Evolutionary multiobjective clustering for overlapping clusters detection," in *Proceedings of the 2009 IEEE Congress on Evolutionary Computation*, pp. 976–982, Trondheim, Norway, May 2009.

[33] A. Mukhopadhyay and U. Maulik, "A multiobjective approach to MR brain image segmentation," *Applied Soft Computing*, vol. 11, no. 1, pp. 872–880, 2011.

[34] O. Kirkland, V. J. Rayward-Smith, and B. de la Iglesia, "A novel multi-objective genetic algorithm for clustering," in *Proceedings of the 12th International Conference on Intelligent Data Engineering and Automated Learning - IDEAL 2011*, H. Yin, W. Wang, and V. Rayward-Smith, Eds., pp. 317–326, Springer, Norwich, UK, Lecture Notes in Computer Science, September 2011.

[35] A. Kishor, P. K. Singh, and J. Prakash, "NSABC: non-dominated sorting based multi-objective artificial bee colony algorithm and its application in data clustering," *Neurocomputing*, vol. 216, pp. 514–533, 2016.

[36] J. Prakash and P. K. Singh, "Gravitational search algorithm and K-means for simultaneous feature selection and data clustering: a multi-objective approach," *Soft Computing*, vol. 23, no. 6, pp. 2083–2100, 2017.

[37] H. S. Jangwan and A. Negi, "A swarm optimization based power aware clustering strategy for WSNs," *International Journal of Advanced Science, Engineering and Information Technology*, vol. 7, no. 1, p. 250, 2017.

[38] F. De Morsier, D. Tuia, M. Borgeaud, V. Gass, and J.-P. Thiran, "Cluster validity measure and merging system for hierarchical clustering considering outliers," *Pattern Recognition*, vol. 48, no. 4, pp. 1478–1489, 2015.

[39] S. Das, A. Abraham, and A. Konar, "Automatic clustering using an improved differential evolution algorithm," *IEEE Transactions on Systems, Man, and Cybernetics-Part A: Systems and Humans*, vol. 38, no. 1, pp. 218–237, 2008.

[40] H. M. J. Mustafa, M. Ayob, M. Z. A. Nazri, and G. Kendall, "An improved adaptive memetic differential evolution optimization algorithms for data clustering problems," *PLoS One*, vol. 14, no. 5, Article ID e0216906, 2019.

[41] J. Tang, Z. Chen, A. W.-c. Fu, and D. W. Cheung, "Enhancing effectiveness of outlier detections for low density patterns," in *Advances in Knowledge Discovery and Data Mining*, M.-S. Chen, P. S. Yu, and B. Liu, Eds., Springer, Berlin, Germany, pp. 535–548, 2002.

[42] W. Jin, A. K. H. Tung, J. Han, and W. Wang, "Ranking outliers using symmetric neighborhood relationship," in *Advances in Knowledge Discovery and Data Mining*, W.-K. Ng,

M. Kitsuregawa, J. Li, and K. Chang, Eds., Springer, Berlin, Germany, pp. 577–593, 2006.

[43] H.-P. Kriegel, P. Kröger, E. Schubert, and A. Zimek, "LoOP: local outlier probabilities," in *Proceeding of the 18th ACM Conference on Information and knowledge management-CIKM '09*, p. 1649, Hong Kong, China, November 2009.

[44] G. Jaradat, M. Ayob, and I. Almarashdeh, "The effect of elite pool in hybrid population-based meta-heuristics for solving combinatorial optimization problems," *Applied Soft Computing*, vol. 44, pp. 45–56, 2016.

[45] E. T. Yassen, M. Ayob, M. Z. A. Nazri, and N. R. Sabar, "An adaptive hybrid algorithm for vehicle routing problems with time windows," *Computers & Industrial Engineering*, vol. 113, pp. 382–391, 2017.

[46] N. R. Sabar, M. Ayob, and G. Kendall, "A hybrid of differential evolution and simulated annealing algorithms for the capacitated arc routing problems," in *Proceedings of the 6th Multidisciplinary International Conference on Scheduling: Theory and Applications (MISTA 2013)*, pp. 549–554, Ghent, Belgium, August 2013.

[47] K. Deb, A. Pratap, S. Agarwal, and T. Meyarivan, "A fast and elitist multiobjective genetic algorithm: NSGA-II," *IEEE Transactions on Evolutionary Computation*, vol. 6, no. 2, pp. 182–197, 2002.

[48] E. Zitzler, M. Laumanns, and L. Thiele, "SPEA2: improving the strength Pareto evolutionary algorithm," in *Proceedings of the Evolutionary Methods for Design, Optimization and Control with Applications to Industrial Problems-EUROGEN'2001*, pp. 95–100, Athens. Greece, September 2001.

[49] C. L. Blake and C. J. Merz, *UCI Repository of Machine Learning Databases*, University of California, Los Angeles, CA, USA, 1998, https://archive.ics.uci.edu/ml/.

[50] S. Das, A. Abraham, and A. Konar, "Metaheuristic pattern clustering - an overview," *Metaheuristic Clustering*, Springer, Berlin, Germany, pp. 1–62, 2009.

[51] A. Topchy, A. K. Jain, and W. Punch, "A mixture model for clustering ensembles," in *Proceedings of the 2004 SIAM International Conference on Data Mining (SDM)*, pp. 379–390, Lake Buena Vista, FL, USA, April 2004.

[52] M.-G. Martínez-Peñaloza, E. Mezura-Montes, N. Cruz-Ramírez, H.-G. Acosta-Mesa, and H.-V. Ríos-Figueroa, "Improved multi-objective clustering with automatic determination of the number of clusters," *Neural Computing & Applications*, vol. 28, no. 8, pp. 2255–2275, 2017.

[53] T. Okabe, Y. Yaochu Jin, and B. Sendhoff, "A critical survey of performance indices for multi-objective optimisation," in *Proceedings of the 2003 Congress on Evolutionary Computation, 2003. CEC '03*, vol. 2, pp. 878–885, Canberra, ACT, Australia, December 2003.

[54] E. Zitzler and L. Thiele, "Multiobjective evolutionary algorithms: a comparative case study and the strength Pareto approach," *IEEE Transactions on Evolutionary Computation*, vol. 3, no. 4, pp. 257–271, 1999.