Set-based Particle Swarm Optimization for Data Clustering

Lienke Brown
Department of Industrial Engineering
Stellenbosch, South Africa
21580790@sun.ac.za

Andries Engelbrecht
Department of Industrial Engineering, and Computer Science Division
Stellenbosch, South Africa
engel@sun.ac.za

ABSTRACT

Computational intelligence approaches to data clustering have been successful in producing compact and well-separated clusters. In particular, particle swarm optimization (PSO) is deemed an effective approach to data clustering. This paper develops and evaluates a discrete-valued variation of PSO, namely the set-based PSO (SBPSO) algorithm, to cluster data. The SBPSO algorithm is evaluated on six standard data sets and nine artificially generated data sets. The clustering results of the SBPSO algorithm is compared to the performance of established clustering algorithms and a PSO clustering algorithm. It is concluded that the results of the SBPSO algorithm varies with the data set characteristics. Nonetheless, the SBPSO is deemed a successful approach to clustering data.

CCS CONCEPTS

• **Theory of computation** → **Evolutionary algorithms**; • **Information systems** → **Clustering**.

KEYWORDS

Data Clustering; Particle Swarm Optimization; Set-based Particle Swarm Optimization; Mediod-based Clustering

ACM Reference Format:
Lienke Brown and Andries Engelbrecht. 2022. Set-based Particle Swarm Optimization for Data Clustering. In *ISMSI '22: 2022 6th International Conference on Intelligent Systems, Metaheuristics & Swarm Intelligence (ISMSI '22), April 9–10, 2022, Seoul, Republic of Korea.* ACM, New York, NY, USA, 7 pages. https://doi.org/10.1145/3533050.3533057

1 INTRODUCTION

Data clustering is the unsupervised grouping of data instances into clusters. Clusters are formed on the basic principle that data instances within the same cluster are similar to one another, and data points from different clusters are dissimilar. Clustering has various applications such as exploratory data analysis, statistical analysis, computer vision and pattern recognition, information retrieval, and object recognition [1] [9] [10].

It is possible to apply clustering to a wide variety of data domains, for example, text, multimedia, social networks, and biological data [7]. The context of the problem to which data clustering is applied varies with each application. Therefore, the clustering technique to be applied to a specific problem depends on a combination of the data domain and context of the problem at hand as the technique needs to be scalable and appropriate for the data domain [7].

The three main objectives of data clustering are [5]:

(1) To produce compact clusters: the distances between data points and the respective cluster centroid should be minimized;
(2) to produce well-separated clusters: the distances between cluster centroids should be maximized; and
(3) to produce an optimal number of clusters.

Therefore, clustering can be approached as a multi-objective optimization problem which yields an optimal set of centroids.

Particle swarm optimization (PSO) is a continuous optimization algorithm inspired by bird flocking behaviour [3]. While PSO has originally been developed to solve continuous-valued optimization problems, several adaptations of PSO have been proposed to solve discrete-valued problems, such as the binary PSO [6]. A discrete-valued PSO algorithm developed by Langeveld and Engelbrecht [8] is the set-based adaptation of the PSO, known as set-based particle swarm optimization (SBPSO). The candidate solutions are represented as sets, not as vectors as in standard PSO. Therefore, SBPSO can be applied to any discrete-valued optimization problem where solutions can be defined as sets.

The SBPSO algorithm can therefore also be applied to data clustering problems, provided that the clustering problem can be defined as a set-based optimization problem. SBPSO yields optimal sets as combinations of elements from the defined universe. Solutions to the clustering optimization problem can be represented as a set of centroids and the universe is a finite set of elements. Thus, the centroids are actual data points contained in the dataset. This implies that the set universe is the actual dataset. Therefore, clustering can be defined as a set-based optimization problem. The SBPSO clustering algorithm follows a medoid-based clustering approach. SBPSO is then applied to find the smallest number of medoids that yields compact and well-separated clusters.

Particle representation as a set has the implication that the size of the particle position can change during algorithm execution [8]. Consequently, SBPSO allows for clustering solutions of differing numbers of clusters. Therefore, the SBPSO algorithm can naturally find an optimal number of clusters within the data set.

This paper develops a new SBPSO algorithm for mediod-based data clustering. The algorithm is evaluated on fifteen different clustering problems, and its performance is compared to that of established clustering approaches such as the K-means and K-medoids algorithms. The results show that the SBPSO algorithm is a successful approach to data clustering.

The rest of the paper is organized as follows. Section 2 provides a short summary of literature related to the SBPSO algorithm. Section

ISMSI '22, April 9–10, 2022, Seoul, Republic of Korea

ACM ISBN 978-1-4503-9628-8/22/04.
https://doi.org/10.1145/3533050.3533057

3 describes the SBPSO algorithm and Section 4 presents the SBPSO clustering algorithm. Section 5 summarizes the empirical process, and Section 6 presents and discusses the results.

2 RELATED WORK

This paper is the first attempt at developing a SBPSO algorithm for data clustering. The following pieces of work are related to SBPSO for data clustering:

- **PSO for data clustering:** The first PSO approach for data clustering was developed and analyzed in [13].
- **SBPSO applied to the knapsack problem:** A generic SBPSO algorithm is developed and tested on the multidimensional knapsack problem. SBPSO is empirically shown to outperform other discrete-valued PSO algorithms [8].
- **SBPSO applied to feature selection:** It is empirically shown that a SBPSO k-nearest neighbour wrapper-based algorithm can effectively solve the feature selection problem (FSP). In particular, SBPSO statistically significantly outperforms three PSO algorithms on the FSP [4].
- **SBPSO applied to polynomial approximation:** SBPSO works well to find polynomial approximations in low dimensions with promise for improved performance in higher dimensions [14].

3 SET-BASED PARTICLE SWARM OPTIMIZATION

PSO was originally created to solve problems with continuous-valued decision variables. The PSO algorithm has been adapted for discrete-valued optimization problems [11]. Set-based particle swarm optimization is an example of an adaption of PSO for discrete-valued optimization problems where solutions can be represented as sets. This section discusses the SBPSO algorithm proposed by Langeveld and Engelbrecht [8].

3.1 Set-based Particle Swarm Optimization Concepts

The position and velocity of particles in SBPSO is defined as mathematical sets. The position is therefore a set of elements from the universal set, as opposed to a vector of fixed dimensions as for the classic PSO algorithm. The velocity is made up of operation pairs, each of which involves the addition or deletion of a single element. The solution of the SBPSO yields a set of elements corresponding to the best position found by the swarm.

For the application of SBPSO to a maximization problem, let

- $U = \{e_n\}_{n \in |D|}$ be the universal set which contains all the elements, e_n, of the data set of a finite number of elements, $|D|$,
- $X_i(t)$ be the position of particle i at iteration t,
- $V_i(t)$ be the velocity of particle i at iteration t,
- f be the objective function to be maximized,
- $Y_i(t)$ be the personal best of particle i, and
- $\widehat{Y}(t)$ be the global best position of the swarm at iteration t.

The paradigm of PSO is based on the idea of movement through the entire search space by using velocity. The attraction towards the personal best position of a particle influences the velocity of the particle. In SBPSO, the particle position $X_i(t)$ is also attracted to the personal best position $Y_i(t)$. However, because mathematical sets are used in SBPSO, the movement of a particle towards its personal best is achieved by removing elements from $X_i(t)$ that are not in $Y_i(t)$ and adding elements to $X_i(t)$ that are contained in $Y_i(t)$. Thereby, the sets are made more similar.

The particle position, $X_i(t)$, is updated by adding the velocity $V_i(t)$. The velocity in SBPSO is a set of operation pairs denoted as $(\pm, e)$, where $(+, e)$ refers to an addition of an element $e \in U$ and $(-, e)$ refers to the removal of element e. The velocity of a particle i is defined as $V_i(t)$ which is then written as $\{v_{i,1}, \ldots, v_{i,z}\} = \{(\pm, e_{n_{i,1}}), \ldots, (\pm, e_{n_{i,z}})\}$ where z represents the number of operation pairs and $e_{n_{i,1}}$ is an element contained in U.

Denote $\mathscr{P}(U)$ as the power set, i.e. the set of all subsets, of U. The position of a particle i, $X_i(t)$, is an element of $\mathscr{P}(U)$. The objective function f is a mapping which assigns a position a quality score in $\mathbb{R}$ which is written as $f : \mathscr{P}(U) \rightarrow \mathbb{R}$. The velocity $V_i(t)$ is a function which maps the current particle position $X_i(t)$ to a new position $X_i(t+1)$, written as $V_i(t) : \mathscr{P}(U) \rightarrow \mathscr{P}$.

To enhance exploration in SBPSO, two special operators are applied. The special operators ensures that elements that are not contained in the personal best position, $Y_i(t)$, of the particle nor in the global best position, $\widehat{Y}_i(t)$, are added to the position. The special operators also remove elements from the current position that are contained in $X_i(t)$ as well as in $Y_i(t)$ and $\widehat{Y}_i(t)$ to enhance exploration of the search space.

3.2 Set-based Particle Swarm Optimization Operators

The operators required for SBPSO are listed in this section. Arithmetic operators used in PSO are transformed to set-based operators. The following set-based operators are defined (for detailed definitions, please refer to [8]):

- *The addition of two velocities*, denoted as $V_1 \oplus V_2$.
- *The difference between two positions*, denoted as $X_1 \ominus X_2$.
- *The multiplication of a velocity by a scalar*, denoted as $\eta \otimes V$.
- *The addition of a velocity and a position*, denoted as $X \boxplus V$.
- *The removal of elements* which are in $X(t) \cap Y(t) \cap \widehat{Y}(t)$ from position $X(t)$, denoted as $\odot^-$.
- *The addition of elements* which are outside of $X(t) \cup Y(t) \cup \widehat{Y}(t)$, denoted as $\odot^+$.

3.3 Set-based Particle Swarm Optimization Update Equations

The operators as defined in Section 3.2 is combined to form the velocity update equation for SBPSO as [8]

$$V_i(t+1) = a_1 r_1 \otimes (Y_i(t) \ominus X_i(t)) \oplus a_2 r_2 \otimes (\widehat{Y}_i(t) \ominus X_i(t)) \oplus (a_3 r_3 \odot_k^+ A_i(t)) \oplus (a_4 r_4 \odot^- S_i(t)) \quad (1)$$

where $A_i(t)$ represents the elements not contained in $X(t) \cup Y(t) \cup \widehat{Y}(t)$, $S_i(t)$ represents the elements which are contained in $X(t) \cap Y(t) \cap \widehat{Y}(t)$, $a_1, a_2 \in [0, 1]$, $a_3, a_4 \in [0, |U|]$ [8], and random numbers r_i are independently drawn from the uniform distribution on $(0, 1)$.

The position update equation is defined as [8]:

$$X_i(t+1) = X_i(t) \boxplus V_i(t+1) \tag{2}$$

4 SBPSO CLUSTERING

SBPSO represents candidate solutions as sets. The solution to the clustering optimization problem is consequently a set of centroids and the universe is a finite set of data points. Accordingly, the set of centroids found by the SBPSO algorithm is actual data points from the provided data set. The set of data points yielded by the algorithm is the cluster mediods. Consequently, the SBPSO clustering algorithm is based on the principles of K-medoids clustering.

In the context of clustering, a single particle represents the cluster medoids of the clusters. Each particle, X_i, is constructed as

$$X_i = \{\mathbf{m}_{i,1}, \ldots, \mathbf{m}_{i,k}, \ldots, \mathbf{m}_{i,n_i}\} \tag{3}$$

where $n_i = |X_i|$; $\mathbf{m}_k$ represents the mediod of cluster k. The swarm of the SBPSO algorithm is therefore composed of N sets of mediods yielding a variety of cluster partitioning arrangements. The number of particles is thus denoted by N. The particles are initialized with varying lengths ranging from a minimum of two mediods to a user specified maximum of number of mediods. It is important to note that the size of the particles can vary due to solutions being represented as sets. Thereby, SBPSO has the ability to optimize the number of clusters, K.

The SBPSO clustering algorithm uses the star topology, where information about best positions are exchanged among all particles; thus, all particles are attracted to the best position as found by the swarm. The same swarm topology is implemented for both the SBPSO and PSO clustering algorithms.

In order to evaluate the clustering capabilities of the SBPSO clustering algorithm in comparison to that of the PSO clustering algorithm, the same objective function is implemented for both algorithms. The objective function for set-based particles is

$$f(X_i) = s(X_i) + Dunn(X_i) \tag{4}$$

where $s(X_i)$ denotes the Silhouette index value and $Dunn(X_i)$ denotes the Dunn index value of the clustering structure yielded by position X_i. The fitness of the SBPSO particles is then measured using Equation (4). Consequently, by simultaneously maximizing the indices, an optimal number of clusters is produced.

The SBPSO algorithm for data clustering is presented in Algorithm 1. The SBPSO clustering algorithm first initialises the particles' position as random subsets of elements in the data set. Thereafter, the personal best position and global best position objective function values are set to be minus infinity, because the objective function is to be maximized.

After particles have been initialised, the algorithm runs for t_{max} iterations; t_{max} is the maximum number of iterations. For each particle, the Euclidean distance to each medoid is calculated. The remaining data points, which are not contained in the position of the particle, are assigned to the corresponding closest cluster medoid. The fitness of each particle is then determined with Equation (4) and the personal best and global best positions are updated accordingly.

The velocity of each particle is updated according to Equation (1). The updated velocity is then used to update the position of each particle using Equation (2). The global best position found at the end of t_{max} iterations is then the optimal cluster medoids as found by the SBPSO algorithm.

Input : Dataset, D; N; t_{max}; a_1; a_2; a_3; a_4
Output: K cluster medoids

```
for i = 1, ..., N do
    Set U = D ;
    Initialize iteration = 0 ;
    Initialize X_i as a random subset of U ;
    Initialize V_i = ∅ ;
    Initialize f(Y_i) = −∞ ;
    Initialize f(Ŷ) = −∞ ;
    Calculate the distance matrix with Euclidean distance ;
end
while iterations < t_max do
    for i = 1, ..., N do
        Assign p_j to cluster C_k such that
          d(p_j, m_{i,k}) = min_{∀k=1,...,K} {d(p_j, m_{i,k})};
        Calculate the fitness of the particle X_i using
          Equation (4) ;
        Set the personal best position ;
        if f(X_i) > f(Y_i) then
            Y_i = X_i ;
        end
        Set the global best position ;
        if f(Y_i) > f(Ŷ) then
            Ŷ = Y_i ;
        end
    end
    for i = 1, ..., N do
        Update V_i according to Equation (1) ;
        Update X_i according to Equation (2) ;
        Calculate the fitness of the updated particle X_i using
          Equation (4);
    end
    iterations = iterations + 1 ;
end
```

Algorithm 1: SBPSO Algorithm for Data Clustering

5 EMPIRICAL PROCESS

This section describes the procedure used to compare the clustering results of SBPSO against other established clustering algorithms. The section opens with a presentation of the clustering algorithms employed. The data sets used are then listed. Thereafter, the control parameter tuning process is discussed. Lastly, the criteria utilized for evaluating the quality of the clustering results are discussed.

5.1 Clustering Algorithms Employed For Evaluation

The clustering results of the SBPSO algorithm is compared to a selection of clustering algorithms representative of the main clustering categories, namely partitional and hierarchical clustering. The algorithms selected, along with the respective R functions, are

presented in Table 1. In addition, the SBPSO clustering results are also evaluated against the results of a standard PSO clustering algorithm [13]. The purpose of assessing the clustering results against that of the PSO is to critically evaluate whether the SBPSO variation of the PSO algorithm improves the clustering abilities of the swarm-based algorithm.

Table 1: Clustering Methods Considered for Analysis

Algorithm	Category	Function in R
K-means	Partitional	*kmeans*
K-medoids	Partitional	*pam*
Agnes	Hierarchical	*hclust*
Density-based	Density-based	*dbscan*

5.2 Benchmark Problems

The data sets used in the experiment consist of both standard clustering data sets from the UCI Repository of Machine Learning Databases [2] and artificially generated data sets. Six standard data sets are used for the purpose of this study, namely:

- **Iris plants:** There are 150 data points, 3 classifications, and 4 variables in this data set. The classes are equally distributed.
- **Wine:** There are 178 data points and 4 variables in this data set. There are no classifications present in the data set, therefore data points are unlabelled.
- **Breast cancer:** There are 569 data points, 2 classifications, and 31 variables in this data set. However, only 9 variables are deemed relevant to the classification of the data point. The classes are unequally distributed.
- **Cervical Cancer Behavior:** There are 72 data points, 2 classifications, and 19 variables in this data set.
- **Ceramics:** There are 88 data points, and 19 variables in this data set. There are no classifications present in the data set, therefore data points are unlabelled.
- **QCM Sensor:** There are 125 instances, 5 classifications and 8 variables in this data set.

Mixtures of Gaussians are used to generate artificial data sets. The artificial data sets serve the purpose of evaluating the impact of specific data set characteristics on the clustering results obtained. The impact of the following data set characteristics are inspected:

- noise,
- cluster density, and
- the number of clusters present.

Three levels of severity are produced for each characteristic in order to critically evaluate the impact thereof on the clustering results. The default parameters for the mixture of Gaussians are set to produce a two dimensional data set containing three equally sized, well-separated and compact clusters without noise. Each data set contains 150 data points.

5.3 Parameter Tuning

A grid search is implemented for the parameter tuning process. The domain of the parameters of an algorithm is divided into a discrete grid. Performance metrics based on the combinations of the values contained in the grid are evaluated to find the best combination in the domain. The performance metric utilized is the Silhouette index. The parameters and corresponding parameter ranges considered in the grid search for each clustering algorithm is presented in Table 2.

Table 2: Parameters Considered for Parameter Tuning

Algorithm	Parameters	Parameter Range
SBPSO	a_1	[0; 1]
	a_2	[0; 1]
	a_3	[0; $\|U\|$]
	a_4	[0; $\|U\|$]
PSO	a_1	[0; 5]
	a_2	[0; 5]
	ω	[0.5; 0.99]
	k	[2; 10]
K-means	k	[2; 10]
K-mediods	k	[2; 10]
Agnes	k	[2; 10]
DB	eps	[0.2; 10]
	min-pts	[2; 10]

For the purposes of this paper, the same number of particles is set to 10 for both the SBPSO and PSO clustering algorithms. The size of the k-tournament selection is not included in the grid search due to limited computational resources. The size of the k-tournament selection implemented in the adapted SBPSO algorithm is set to three and the maximum number of medoids initialized is set to ten.

The results of the parameter tuning process for the SBPSO algorithm is presented in Table 3.

Table 3: Tuned Parameters for Standard Data Sets

Data Set	a_1	a_2	a_3	a_4
Iris	0.63	0.65	0.73	0.58
Breast Cancer	69	0.75	0.64	0.52
Wine	0.76	0.52	0.78	0.51
Cervix	0.42	0.41	0.30	0.31
Ceramics	0.35	0.51	0.34	0.32
QCM Sensor	0.42	0.31	0.33	0.31
Noise Level 1	0.49	0.51	0.59	0.72
Noise Level 2	0.76	0.72	0.91	0.54
Noise Level 3	0.53	0.51	0.89	0.51
Cluster Density Level 1	0.51	0.41	0.30	0.31
Cluster Density Level 2	0.71	0.52	0.52	0.62
Cluster Density Level 3	0.62	0.51	0.71	0.81
Number of Clusters Level 1	0.51	0.53	0.49	0.50
Number of Clusters Level 2	0.71	0.52	0.52	0.51
Number of Clusters Level 3	0.62	0.51	0.71	0.81

5.4 Evaluation Criteria

The results of the clustering algorithms are measured as the averages and corresponding standard deviations over 30 independent runs. The quality of the respective clusterings are evaluated against to the following criteria:

- the Silhouette index, where it is the objective to maximize the Silhouette index;
- the average intra-cluster distance, where it is the objective to minimize the intra-cluster distance;
- the average inter cluster distance, where it is the objective to maximize the inter-cluster distance; and
- the number of clusters produced.

Clusters that meet the above objectives are compact and well separated. The Silhouette index is utilized as the metric to compare the clustering performance of the clustering algorithms, because the index incorporates the inter- and intra-cluster distances.

The SBPSO and PSO clustering algorithms are compared to each other using a Mann-Whitney U test. The Mann-Whitey U test (with level of significance is set to $\alpha = 0.05$) is used to compare differences between two independent groups [12]. In this case, the independent groups are the Silhouette index values produced by the two different clustering algorithms. The hypothesis is defined as

- *Null hypothesis* H_0: There is no difference between the Silhouette index values.
- *Alternative hypothesis* H_1: There is a difference between the Silhouette index values.

The six clustering algorithms are ranked against each other based on the highest average Silhouette index produced over the 30 independent runs.

6 EMPIRICAL ANALYSIS

This section presents the results for the standard data sets, followed by the results of the artificial data sets.

6.1 Standard Data Sets Results

The clustering results obtained for the standard data sets are captured in Table 4. The results of the Mann-Whitney U test applied to the clustering results of the SBPSO and PSO algorithms are captured in Table 5. The results indicate that there is a significant difference between the clustering results of the SBPSO and PSO for all data sets, except the Wine data set. The SBPSO algorithm yields a higher average Silhouette index value than the PSO algorithm for the Iris and Breast Cancer data sets. For the Cervix, Ceramics and QCM Sensor data sets, the PSO yields a higher average Silhouette index value. However, in the case of the Wine data set, there is no significant difference between the Silhouette index values of the SBPSO and PSO algorithms.

All algorithms produce two clusters for the Iris data set. The number of clusters found does not correspond to the three classes present in the data set - which is the number of clusters expected to be produced. However, two clusters yields a higher Silhouette index value for the Iris data set and the objective of the parameter tuning process in this project is to maximize the Silhouette index. There are no class labels present in the Wine data set and thus no expected number of clusters to be produced. The PSO, K-means, and K-medoids algorithms yield three clusters. The Agnes and DB algorithms both yield two clusters. The SBPSO algorithm produced an average number of clusters between two and three.

The Breast Cancer data set contains two class labels and thus two clusters are expected to be detected by the clustering algorithms. However, only the PSO, K-means, and K-medoids algorithms produce two clusters. The SBPSO algorithm produces an average number of clusters of 2.2, which is close to the number of clusters expected to be found.

Table 4: Standard Data Sets Clustering Results

Problem	Algorithm	Intra-Cluster Distance	Inter-Cluster Distance	Silhouette Index	K
Iris	SBPSO	1.506±0.000	3.647±0.000	0.581±0.000	2
	PSO	1.511±0.122	3.642±0.009	0.579±0.003	2
	K-means	1.506±0.000	3.647±0.000	0.581±0.000	2
	K-mediods	1.506±0.000	3.647±0.000	0.581±0.000	2
	Agnes	1.506±0.000	3.647±0.000	0.581±0.000	2
	DB	1.537±0.000	3.461±0.000	0.563±0.000	2
Wine	SBPSO	3.953±0.256	5.603±0.100	0.262±0.008	2.5±0.759
	PSO	3.727±0.112	5.496±0.152	0.259±0.022	3
	K-means	3.626±0.000	5.517±0.000	0.284±0.000	3
	K-mediods	3.674±0.000	5.510±0.000	0.267±0.000	3
	Agnes	4.867±0.000	7.069±0.000	0.298±0.000	2
	DB	4.815±0.000	6.250±0.000	0.224±0.000	2
Breast Cancer	SBPSO	6.758±0.229	14.869±3.328	0.506±0.097	2.2±0.523
	PSO	5.875±0.262	9.567±0.2.665	0.356±0.071	2
	K-means	5.445±0.016	8.742±0.0253	0.313±0.006	3
	K-mediods	5.795±0.000	9.047±0.000	0.349±0.000	2
	Agnes	6.906±0.000	19.489±0.000	0.633±0.000	2
	DB	6.829±0.000	-	0.569±0.000	1
Cervix	SBPSO	5.386±0.347	6.917±0.555	0.199±0.035	2.36±1.129
	PSO	5.331±0.061	6.705±0.089	0.200±0.018	2
	K-means	6.686±0.112	5.336±0.069	0.198±0.022	2
	K-mediods	4.186±0.006	6.312±0.005	0.174±0.003	9
	Agnes	5.911±0.000	8.719±0.000	0.310±0.000	2
	DB	5.018±0.000	6.304±0.000	0.233±0.000	3
Ceramics	SBPSO	4.664±0.422	6.595±0.494	0.260±0.023	2.63±1.066
	PSO	4.678±0.122	6.509±0.030	0.277±0.001	2
	K-means	6.495±0.113	4.672±0.000	0.275±0.031	2
	K-mediods	6.468±0.000	4.685±0.000	0.268±0.000	2
	Agnes	5.526±0.000	8.735±0.000	0.342±0.000	2
	DB	5.277±0.000	5.836±0.000	0.565±0.000	2
QCM Sensor	SBPSO	1.923±0.323	5.288±0.260	0.557±0.034	2.50±0.629
	PSO	2.093±0.000	5.413±0.000	0.588±0.000	2
	K-means	2.093±0.000	5.413±0.000	0.588±0.000	2
	K-mediods	2.093±0.000	5.413±0.000	0.588±0.000	2
	Agnes	2.093±0.000	5.413±0.000	0.588±0.000	2
	DB	2.455±0.000	3.123±0.000	0.826±0.000	11

The Cervix data set contains two class labels. The SBPSO algorithm yielded an of average 2.36 clusters. The PSO, K-means, and Agnes algorithms yielded two clusters. The K-medoids algorithm yielded a total of nine clusters and the DB algorithm a total of three.

The rankings of the clustering algorithms are provided in Table 6. The clustering result of the DB function applied to the Breast Cancer data set is excluded from the rankings as the algorithm is unable to produce at least two clusters. The Agnes function yields the best average Silhouette index values for four of the standard data sets. The SBPSO is ranked higher than the PSO for three of the standard data sets.

6.2 Artificial Data Sets Results

The clustering results for the three severity levels of the noise, cluster density, and number of clusters data sets are presented in Tables 7, 8, and 9, respectively. The results of the Mann-Whitney U tests are presented in Table 10.

For the three noise data sets, there is significant difference between the clustering results of the SBPSO and PSO for severity

Table 5: Statistical Significance: Standard Data Sets

Iris	**Wine**	**Breast Cancer**
p-value < 0.001 Significant Reject H_0	p-value = 0.5792 Not Significant Do Not Reject H_0	p-value < 0.001 Significant Reject H_0
Cervix	**Ceramics**	**QCM Sensor**
p-value = 0.03837 Significant Reject H_0	p-value = < 0.001 Significant Reject H_0	p-value < 0.001 Significant Reject H_0

Table 6: Ranks: Standard Data

Algorithm	**Iris**	**Wine**	**Breast Cancer**	**Cervix**	**Ceramics**	**QCM Sensor**
SBPSO	1	4	2	4	6	3
PSO	2	5	3	3	3	2
K-means	1	2	5	5	4	2
K-medoids	1	3	4	6	5	2
Agnes	1	1	1	1	2	2
DB	3	6	-	2	1	1

Table 7: Artificial Data Sets Clustering Results: Noise

Problem	Algorithm	Intra-Cluster Distance	Inter-Cluster Distance	Silhouette Index	K
Level 1	SBPSO	0.417±0.032	2.340±0.014	0.780±0.014	3.40±0.681
	PSO	0.388±0.028	2.327±0.012	0.789±0.071	3.75±0.444
	K-means	0.376±0.022	2.295±0.048	0.753±0.071	4
	K-mediods	0.365±0.000	2.324±0.000	0.797±0.000	4
	Agnes	0.364±0.000	2.319±0.000	0.794±0.000	4
	DB	0.725±0.000	2.304±0.000	0.901±0.000	3
Level 2	SBPSO	0.392±0.017	2.374±0.005	0.802±0.005	3.10±0.447
	PSO	0.396±0.014	2.372±0.003	0.801±0.002	3.05±0.224
	K-means	0.338±0.013	2.355±0.010	0.805±0.029	4
	K-mediods	0.332±0.000	2.348±0.000	0.811±0.000	4
	Agnes	0.333±0.000	2.357±0.000	0.815±0.000	4
	DB	0.538±0.000	2.334±0.000	0.876±0.000	2
Level 3	SBPSO	0.330±0.024	2.356±0.008	0.827±0.006	3.55±0.686
	PSO	0.308±0.007	2.352±0.006	0.833±0.004	4
	K-means	0.375±0.021	2.292±0.054	0.754±0.068	4
	K-mediods	0.301±0.000	2.348±0.000	0.837±0.000	4
	Agnes	0.313±0.000	2.351±0.000	0.830±0.000	4
	DB	0.601±0.000	2.302±0.000	0.699±0.000	4

levels two and three. For a severity level of one, there is no significant difference between the algorithms. The algorithm ranks are given in Table 11. SBPSO outperforms the PSO for severity level two, whereas the PSO outperforms SBPSO for severity level three.

There are three clusters present for all three severity levels of the noise data sets. As indicated in Table 7, only the DB algorithm was able to successfully detect the correct number of clusters for severity level one. Both the SBPSO and PSO algorithms yield an average number of clusters between three and four for severity levels one and two. The SBPSO yields a lower average than PSO for severity levels one and three. Thus, only for some of the independent runs the SBPSO algorithm correctly detects three clusters.

For the cluster density data sets, there is no significant difference between the clustering results of SBPSO and PSO for any of the three severity levels tested. As shown in Table 8, all six clustering algorithms are able to correctly detect the three clusters present in the cluster density data sets for severity levels one and two. However, SBPSO and PSO detect two clusters for some of the independent runs performed on severity level three. Hence, the average number of clusters found for severity level three is less than three. The DB algorithm also detects two clusters for severity level three.

Table 8: Artificial Data Sets Clustering Results: Cluster Density

Level 1	Algorithm	Intra-cluster Distance	Inter-cluster Distance	Silhouette Index	K
Level 1	SBPSO	0.245±0.000	2.399±0.000	0.882±0.000	3
	PSO	0.245±0.000	2.399±0.000	0.882±0.000	3
	K-means	0.508±0.187	2.369±0.048	0.749±0.091	3
	K-mediods	0.245±0.000	2.399±0.000	0.882±0.000	3
	Agnes	0.245±0.000	2.399±0.000	0.882±0.000	3
	DB	0.376±0.000	2.574±0.000	0.905±0.000	3
Level 2	SBPSO	0.251±0.000	2.569±0.000	0.892±0.000	3
	PSO	0.251±0.000	2.569±0.000	0.892±0.000	3
	K-means	0.282±0.137	2.564±0.021	0.877±0.068	3
	K-mediods	0.251±0.000	2.569±0.000	0.892±0.000	3
	Agnes	0.251±0.000	2.569±0.000	0.892±0.000	3
	DB	0.376±0.000	2.574±0.000	0.905±0.000	3
Level 3	SBPSO	0.586±0.239	3.298±0.046	0.820±0.059	2.70±0.470
	PSO	0.499±0.093	3.152±0.113	0.821±0.035	2.95±0.223
	K-means	0.517±0.176	3.134±0.422	0.787±0.163	3
	K-mediods	0.433±0.000	3.277±0.000	0.858±0.000	3
	Agnes	0.433±0.000	3.277±0.000	0.858±0.000	3
	DB	0.433±0.000	3.267±0.000	0.863±0.000	2

Table 9: Artificial Data Sets Clustering Results: Number of Clusters

Problem	Algorithm	Intra-Cluster Distance	Inter-Cluster Distance	Silhouette Index	K
Level 1	SBPSO	0.417±0.032	2.340±0.014	0.780±0.014	3.40±0.681
	PSO	0.388±0.028	2.327±0.012	0.789±0.071	3.75±0.444
	K-means	0.376±0.022	2.295±0.048	0.753±0.071	4
	K-mediods	0.365±0.000	2.324±0.000	0.797±0.000	4
	Agnes	0.364±0.000	2.319±0.000	0.794±0.000	4
	DB	0.725±0.000	2.304±0.000	0.901±0.000	3
Level 2	SBPSO	0.392±0.017	2.374±0.005	0.802±0.005	3.10±0.447
	PSO	0.396±0.014	2.372±0.003	0.801±0.002	3.05±0.224
	K-means	0.338±0.013	2.355±0.010	0.805±0.029	4
	K-mediods	0.332±0.000	2.348±0.000	0.811±0.000	4
	Agnes	0.333±0.000	2.357±0.000	0.815±0.000	4
	DB	0.538±0.000	2.334±0.000	0.876±0.000	2
Level 3	SBPSO	0.330±0.024	2.356±0.008	0.827±0.006	3.55±0.686
	PSO	0.308±0.007	2.352±0.006	0.833±0.004	4
	K-means	0.375±0.021	2.292±0.054	0.754±0.068	4
	K-mediods	0.301±0.000	2.348±0.000	0.837±0.000	4
	Agnes	0.313±0.000	2.351±0.000	0.830±0.000	4
	DB	0.601±0.000	2.302±0.000	0.699±0.000	4

There is a significant difference between the clustering results of SBPSO and PSO for the number of clusters data sets. For each of the severity levels, the SBPSO yields a higher average Silhouette index value than PSO. Table 9 indicates that none of the clustering algorithms were able to produce the correct number of clusters for the data sets for any of the severity levels, except for the DB algorithm at severity level three. At severity level one, the average number of clusters detected by SBPSO is lower than the actual

Table 10: Statistical Significance: Artificial Data

Noise Level 1	**Noise Level 2**	**Noise Level 3**
p-value = 0.05966 Not Significant Do Not Reject H_0	p-value = 0.01093 Significant Reject H_0	p-value < 0.001 Significant Reject H_0
Cluster Density Level 1	**Cluster Density Level 2**	**Cluster Density Level 3**
Yields the exact same results	Yields the exact same results	p-value = 0.3269 Not Significant Do Not Reject H_0
Number of Clusters Level 1	**Number of Clusters Level 2**	**Number of Clusters Level 3**
p-value < 0.001 Significant Reject H_0	p-value < 0.001 Significant Reject H_0	p-value < 0.001 Significant Reject H_0

Table 11: Ranks: Artificial Data Sets

Algorithm	**Noise Level 1**	**Noise Level 2**	**Noise Level 3**
SBPSO	5	5	4
PSO	4	6	2
K-means	6	4	5
K-medoids	2	3	1
Agnes	3	2	3
DB	1	1	6
Algorithm	**Cluster Density Level 1**	**Cluster Density Level 2**	**Cluster Density Level 3**
SBPSO	2	2	4
PSO	2	2	3
K-means	3	3	5
K-medoids	2	2	2
Agnes	2	2	2
DB	1	1	1
Algorithm	**Number of Clusters Level 1**	**Number of Clusters Level 2**	**Number of Clusters Level 3**
SBPSO	4	2	4
PSO	5	5	5
K-means	2	3	2
K-medoids	1	1	1
Agnes	1	1	3
DB	3	3	-

number of clusters present in the data set. Only K-mediods and Agnes correctly produce three clusters. For severity levels two and three, SBPSO also detects less clusters than there are present in the data set.

As shown in Table 11, the SBPSO is not ranked the best for any of the artificial data sets. The DB algorithm is ranked best for the noise data sets severity level one and two, and all three levels of severity of the cluster density data sets. For the noise data set at severity level three, K-medoids yields the highest average Silhouette index value. Agnes is ranked the highest for the number of clusters severity levels one and two. K-medoids also yields the highest average Silhouette index value for the number of clusters at severity level three.

Overall, none of the clustering algorithms dominate over all the data sets. The performance of each algorithm varies with the data set characteristics - this refers to the ability to produce an optimal number of clusters and to produce a high Silhouette index value.

7 CONCLUSIONS

This paper investigated the application of the set-based particle swarm optimization (SBPSO) algorithm to cluster data. Mediod-based clustering is formulated as a set-based optimization problem, and a SBPSO algorithm was developed find the optimal set of mediods to server as cluster centroids. The mediod-based SBPSO clustering algorithm was compared to four established clustering algorithms as well as a PSO clustering algorithm. While the SBPSO clustering algorithm did not provide the best performance for all of the datasets used, it provided good rankings with respect to the clustering performance metrics. The SBPSO succeeded in finding optimal numbers of clusters.

Future work will further explore the mediod-based SBPSO clustering algorithm to implement mechanisms to improve its performance, specifically approaches to maintain exploration for longer periods of time and approaches to ensure that good mediods are not removed from set-based particles. The SBPSO approach to clustering will also be adapted to clustering on non-stationary data.

REFERENCES

[1] HC Andrews. 1972. *Introduction to Techniques in Pattern Recognition.* John Wiley & Sons, New Yrok.

[2] Dheeru Dua and Casey Graff. 2017. UCI Machine Learning Repository. (2017). https://archive.ics.uci.edu/ml/index.php

[3] Russ Eberhart, Pat Simpson, and Roy Dobbins. 1996. *Computational Intelligence PC Tools.* Academic Press Professional, Inc., USA.

[4] Andries P. Engelbrecht, Jacomine Grobler, and Joost Langeveld. 2019. Set based particle swarm optimization for the feature selection problem. *Engineering Applications of Artificial Intelligence* 85 (2019), 324–336. https://doi.org/10.1016/j.engappai.2019.06.008

[5] A K Jain, M N Murty, and P J Flynn. 1999. Data Clustering: A Review. *Comput. Surveys* 31, 3 (1999), 265–323.

[6] James Kennedy and Russell C. Eberhart. 1997. Discrete Binary Version of the Particle Swarm Algorithm. *IEEE International Conference on Systems, Man, and Cybernetics* 5 (1997), 4104–4108. https://doi.org/10.1109/ICSMC.1997.637339

[7] Vipin Kumar. 2014. *Data Clustering: Algorithms and Applications.* Chapman and Hall/CRC.

[8] Joost Langeveld and Andries P. Engelbrecht. 2012. Set-Based Particle Swarm Optimization Applied to the Multidimensional Knapsack Problem. *Swarm Intelligence* 6, 4 (2012), 297–342. https://doi.org/10.1007/s11721-012-0073-4

[9] T Lillesand and R Keifer. 1994. *Remote Sensing and Image Interpretation.* John Wiley & Sons.

[10] M. R. Rao. 1971. Cluster Analysis and Mathematical Programming. *J. Amer. Statist. Assoc.* 66, 335 (1971), 622–626. http://www.jstor.org/stable/2283542

[11] Ahmad Rezaee Jordehi and Jasronita Jasni. 2015. Particle Swarm Optimisation for Discrete Optimisation Problems: A Review. *Artificial Intelligence Review* 43, 2 (2 2015), 243–258. https://doi.org/10.1007/s10462-012-9373-8

[12] B Rosner and D Grove. 1999. Use of the Mann-Whitney U-test for Clustered Data. *Statistics in Medicine* 18 (1999), 1387–1400.

[13] DW Van der Merwe and Andries Petrus Engelbrecht. 2003. Data clustering using particle swarm optimization. In *The 2003 Congress on Evolutionary Computation, 2003. CEC'03.*, Vol. 1. IEEE, 215–220.

[14] Jean Pierre van Zyl and Andries P. Engelbrecht. 2021. Polynomial Approximation Using Set-Based Particle Swarm Optimization. In *Advances in Swarm Intelligence*, Ying Tan and Yuhui Shi (Eds.). Springer International Publishing, 210–222.