sciendo RIGA TECHNICAL UNIVERSITY

Applied Computer Systems

ISSN 2255-8691 (online)
ISSN 2255-8683 (print)
May 2019, vol. 24, no. 1, pp. 49–60
https://doi.org/10.2478/acss-2019-0007
https://content.sciendo.com

Shared Subscribe Hyper Simulation Optimization (SUBHSO) Algorithm for Clustering Big Data – Using Big Databases of Iran Electricity Market

Mesbaholdin Salami[1], Farzad Movahedi Sobhani[2*], Mohammad Sadegh Ghazizadeh[3]
[1] *Department of Industrial Engineering, Central Tehran Branch, Islamic Azad University, Tehran, Iran*
[2] *Department of Industrial Engineering, Science and Research Branch, Islamic Azad University, Tehran, Iran*
[3] *Department of Electrical Engineering, Abbaspour School of Engineering, Shahid Beheshti University, Tehran, Iran*

***Abstract* – Many real world problems have big data, including recorded fields and/or attributes. In such cases, data mining requires dimension reduction techniques because there are serious challenges facing conventional clustering methods in dealing with big data. The subspace selection method is one of the most important dimension reduction techniques. In such methods, a selected set of subspaces is substituted for the general dataset of the problem and clustering is done using this set. This article introduces the Shared Subscribe Hyper Simulation Optimization (SUBHSO) algorithm to introduce the optimized cluster centres to a set of subspaces. SUBHSO uses an optimization loop for modifying and optimizing the coordinates of the cluster centres with the particle swarm optimization (PSO) and the fitness function calculation using the Monte Carlo simulation. The case study on the big data of Iran electricity market (IEM) has shown the improvement of the defined fitness function, which represents the cluster cohesion and separation relative to other dimension reduction algorithms.**

***Keywords* – Clustering, electricity demand forecasting, Monte Carlo simulation, particle swarm optimization, subspace search method.**

I. Introduction

In general, conventional data mining methods are not suitable for big data analysis since they pose serious challenges to a variety of distance measurements in a reasonable time. Accordingly, many techniques have been developed to establish how cluster subspaces can be found from a full data space. Although in most cases, a full data space can be very turbulent, the aim of clustering full-scale data requires to find similar data and put them in the same groups. The other goal is to find a set of attributes that clearly reflect the similarity of data in a dataset. To this end, many subspace clustering methods have developed to solve the problem of data analysis in a full-data space. These methods are limited to three general classes: correlation-based clustering method, biclustering method, and subspace search method [1]. The correlation-based clustering methods are used for a set of non-correlated dimensions, in which clusters are created in a new space or its subspaces. Different researchers have used this group of spatial transforms. For example, PCA has been used by Penkova et al. [2], Hough Transform by Vera et al. [3], and Fractal Dimension by Yang et al. [4]. Some relevant studies are mentioned in [5] to [8]. The biclustering method is a set of techniques to cluster data and attributes at the same time. Each data or attribute can be placed simultaneously in several clustering groups or neither of them. Some of the well-known algorithms are constant bicluster [9] and linear pattern biclustering [10]. Some other examples of relevant studies are mentioned in [11] to [14]. The third group includes the subspace search methods. These methods search different subspaces for clusters. Here, clusters are a set of similar data in the given space. The degree of similarity is measured using the common measurement methods, such as the distance or density. Several studies have been conducted and multiple algorithms developed in relation to this technique. Among the most important algorithms include PROCLUS [15], P3C [16], FSC [17], ESSC [18], FG-k-means [19], C-k-means [20], LAC [21], and CSSUB [22]. Other methods of this category are mentioned in [23] to [28]. In all of these problems, the aim is to select a set of subspaces as an appropriate substitute for the whole dataset. Although this category has a higher accuracy than other methods, but with a large amount of data, their precision is reduced because the subspaces cannot be good representation for the entire data. Therefore, they need to optimize cluster centres in subspaces with some smart mechanisms. This article proposes a hybrid clustering algorithm, namely, Shared Subscribe Hyper Simulation Optimization (SUBHSO). This algorithm is proposed to solve the problem of limiting the subspace method in the analysis of large data. In this algorithm, the initial solutions were generated by the CSSUB. This will be done by selecting the subspaces and clustering the data of the selected spaces. These solutions were introduced as the inputs of the PSO algorithm. The PSO algorithm optimizes the cluster centres, and generates new cluster centres (new offspring) in each repetition based on its mechanisms. The values of the objective function, defined in the PSO algorithm, were obtained from the Monte Carlo simulation. In that, a set of random pilot data was generated, and new cluster centres generated by PSO at each stage were compared to this random set to obtain the fitness function. This optimization or coordinate change of cluster centres continued until the conditions for the termination of PSO algorithm were fulfilled. To increase the chance of the selection of a set with better attributes, which led to the improvement of the fitness

* Corresponding author's e-mail: farzadmovahedi@gmail.com

function, these stages were used from beginning to end in a hyper procedure. Results from this algorithm were compared to other dimension reduction algorithms based on the subspace search; in addition, the superiority of this algorithm in cohesion and separation was revealed by comparing the calculated fitness function. This study used the IEM database. This market was established in 2002 as a competitive environment for trading electric power by micro and macro sellers and purchasers. Due to daily interactions in this market, a big amount of data is generated and stored. Twenty-four different parameters are continuously stored and updated in short time intervals in the Centre of Electricity Market Management. The electricity market manager (EMM) can predict the demand for electric power through clustering big data. In the next stages, the EMM can use the predicted electric power demand for fairly pricing of electricity and required controls on the electricity power supply chain. The rest of this paper includes model component (Section 2), Iran electricity market and big data (Section 3), data clustering with the proposed algorithm (Section 4), comparison of the proposed algorithm with predecessors in terms of execution (Section 5) and validation of the proposed algorithm (Section 6).

II. Model Components

A. CSSUB Method

Zhu et al. [22] developed the clustering by shared subscribe (CSSUB) method as a subspace search algorithm. This algorithm is more accurate than other subspace algorithms [22]. This algorithm selects subspaces from general spaces of the problem and then clusters them using a conventional method. Then, it shares the selected subspaces with each other and re-clusters the cluster centres based on the closeness criterion. Next, this cluster is introduced as the final clustered solution for the whole dataset. Figure 1 presents these stages.

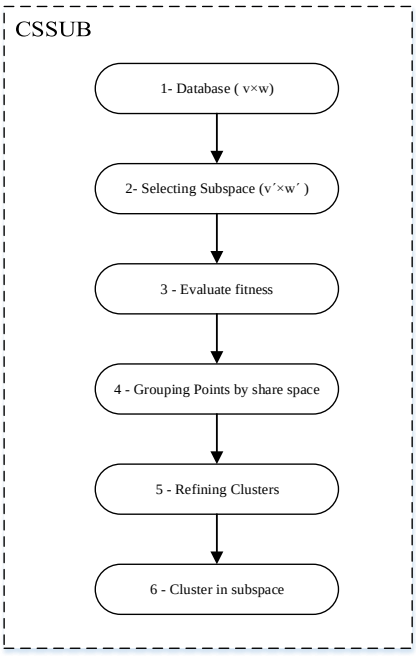

Fig. 1. Stages of CSSUB method.

This algorithm is comprised of the following stages:

Stage 1. Assuming that D is a set with v points and w parameters, the dataset dimension is $v \times w$.

Stage 2. Selecting u subspaces randomly from this dataset, $v_i^{'}$ points and $w_i^{'}$ parameters are selected in each stage. v' is the overall set of $v_i^{'}$ ($v' = \{v_1^{'}, v_2^{'}, \ldots, v_n^{'}\}$) and w´ is the overall set of $w_i^{'}$ ($w' = \{w_1^{'}, w_2^{'}, \ldots, w_n^{'}\}$).

Stage 3. Determining the cluster centres set for each selected subspace through k-means algorithm. To calculate the objective function, Davies–Bouldin index proposed by Davies et al. [28] is used. Equation (1) shows the calculation mechanism of this index, which includes cohesion and separation.

Given n dimensional points, let C_i be a cluster of data points. Let X_j be an n-dimensional feature vector assigned to cluster C_i:

$$S_i = \left(\frac{1}{T_i}\sum_{j=1}^{T_i}\left|X_j - A_i\right|^p\right)^{1/p} \quad (1)$$

Here A_i is the centroid of C_i and T_i is the size of the cluster i. S_i is a measure of scatter within the cluster. Usually the value of p is 2, which makes this a Euclidean distance function between the centroid of the cluster, and the individual feature vectors. Many other distance metrics can be used, in the case of manifolds and higher dimensional data, where the Euclidean distance may not be the best measure for determining the clusters. It is important to note that this distance metric has to match with the metric used in the clustering scheme itself for meaningful results. Here k indexes the features of the data and this is essentially the Euclidean distance between the centres of clusters i and j when p equals 2.

$$M_{i,j} = \left\|A_i - A_j\right\|_p = \left(\sum_{k=1}^{n}\left|a_{k,i} - a_{k,j}\right|^p\right)^{1/p} \quad (2)$$

$M_{i,j}$ is a measure of separation between C_i cluster and C_j cluster, $a_{k,i}$ is the k-th element of A_i, and there are n such elements in A for it is an n dimensional centroid.

Let $R_{i,j}$ be a measure of how good the clustering scheme is. This measure, by definition has to account for $M_{i,j}$, the separation between the i-th and the j-th cluster, which ideally has to be as large as possible, and S_i, within cluster scatter for cluster i, which has to be as low as possible. Hence, the Davies–Bouldin index is defined as the ratio of S_i and $M_{i,j}$ such that these properties are conserved:

$$R_{i,j} \geq 0 \quad (3)$$

$$R_{i,j} = R_{j,i} \quad (4)$$

$$\text{when } S_j \geq S_k \text{ and } M_{i,j} = M_{i,k} \text{ then } R_{i,j} > R_{i,k} \quad (5)$$

$$\text{when } S_j = S_k \text{ and } M_{i,j} \leq M_{i,k} \text{ then } R_{i,j} > R_{i,k} \quad (6)$$

With this formulation, the lower the value, the better the separation of the clusters and the "tightness" inside the clusters. A solution that satisfies these properties is:

$$R_{i,j} = \frac{S_i + S_j}{M_{i,j}} \quad (7)$$

This is used to define D_i:

$$D_i \equiv \max_{i \neq j} R_{i,j} \tag{8}$$

If N is the number of clusters, final DB index is:

$$DB = \frac{1}{N} \sum_{i=1}^{N} D_i \tag{9}$$

Stage 4. Sharing the subspace set randomly selected in previous stages. At this stage, the cluster centres obtained in Stage 3 are preserved and the points, which are not among the cluster centres, are not considered as the noise during the next stages. The similarity between the cluster centres in each subspace is measured using (10).

$$sim\left(x_i, x_j\right) = \frac{A(x_i) \cap A(x_j)}{A(x_i) \cup A(x_j)} \tag{10}$$

Stage 5. Re-clustering the point set maintained from the previous stage, cluster centres (x_i, x_j), based on the degree of similarity between them, using the k-means algorithm. The cluster centres obtained at this stage are changed during the next stage to improve the fitness criterion.

B. PSO-Monte Carlo (Simulation-Optimization)

Particle swarm optimization (PSO) is a global optimization algorithm for dealing with problems, in which the best solution can be represented as a point or surface in an n-dimensional space. This algorithm is one of the strongest population-based metaheuristic methods [30]. Hypotheses are formed in this space and seeded with an initial velocity, as well as a communication channel between the particles. Then, these particles are transferred to the response space and results are calculated based on a fitness function after each time step. The movement of these particles is accelerated towards particles in the same communication group with a higher fitness function. Although each method works well in the specific range of problems, this method showed very efficient in solving continuous optimization problems. Figure 2 shows the movement of these particles in this algorithm.

The particle position and their initial velocity are obtained using (11) and (12), respectively. The velocity, weight, and position of particles in each repetition were obtained using (13)–(15), respectively.

$$x_0^i = \text{The coordinates of the cluster centers derived from CSSUB} \tag{11}$$

$$v_0^i = \frac{x_{\min} + rand\,(x_{\max} - x_{\min})}{\Delta t} \tag{12}$$

$$v_{t+1}^i = w v_t^i + \phi_1 rand_1 \left(\frac{p^i - x_t^i}{\Delta t}\right) + \phi_2 rand_2 \left(\frac{p_k^g - x_t^i}{\Delta t}\right) \tag{13}$$

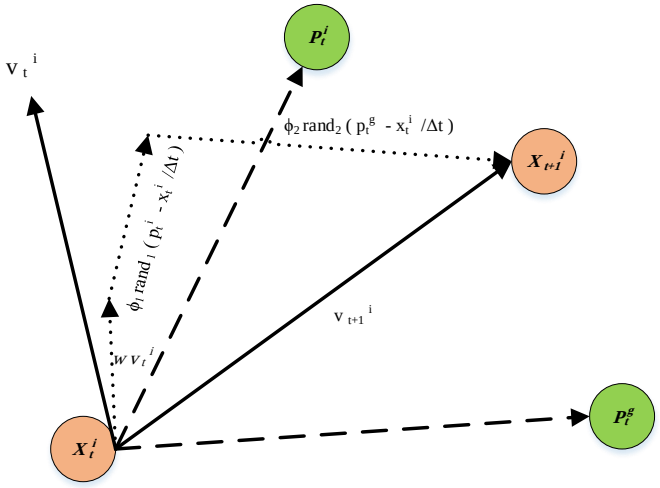

Fig. 2. PSO mechanism.

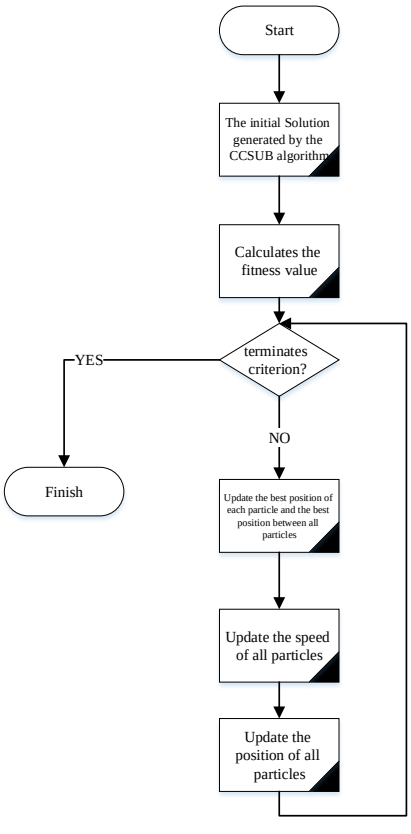

Fig. 3. PSO algorithm stages.

$$w = (w_1 - w_2)\frac{iter_{\max} - iter}{iter_{\max}} + w_2 \tag{14}$$

$$x_{t+1}^i = x_t^i + v_{t+1}^i \Delta t \tag{15}$$

Where x_0^i is the current position of the particle, $x_{\min}$ and $x_{\max}$ are, respectively, the minimum and maximum coordinates of the particle i, v_0^i is the initial velocity of particle i, v_{t+1}^i is the velocity of particle i at $t + 1$, ϕ_1 and ϕ_2 are the coefficients of motion tendency adjustment towards global optimal or the best

solution obtained by particle i, p^i is the best position experienced by particle i, p_t^g is the best position experienced by all particles until t, and x_t^i is the position of particle i at t.

The Monte Carlo simulation is used to calculate the fitness function. This algorithm uses random sampling to calculate the results. The Monte Carlo methods are generally used for the simulation of physical, mathematical, and economic systems and one of the most used simulation algorithms [31]. On the other hand, the Monte Carlo is a class of computational algorithms that rely on iterative random sampling to calculate the results. Thus, a set of random points $G = \{g_1, g_2, \ldots., g_o\}$ is selected from the main dataset. Then, these points are allocated to the closest cluster centres. After the allocation of the points to the closest cluster centres, created by the PSO algorithm, the value of DB_l was calculated for this new configuration using (16). This execution process was repeated for as the number of time the Monte Carlo simulation was repeated (μ). Then, the mean value of the index in each iteration was calculated and considered as the objective function (17). Figure 4 presents the schematic of the stages of objective function calculation.

$$DB_l = \frac{1}{N}\sum_{i=1}^{N} D_{il} \quad \forall l \in \{1, \cdots, \mu\} \tag{16}$$

$$\overline{DB} = \frac{\sum_{l=1}^{\mu} DB_l}{\mu} \tag{17}$$

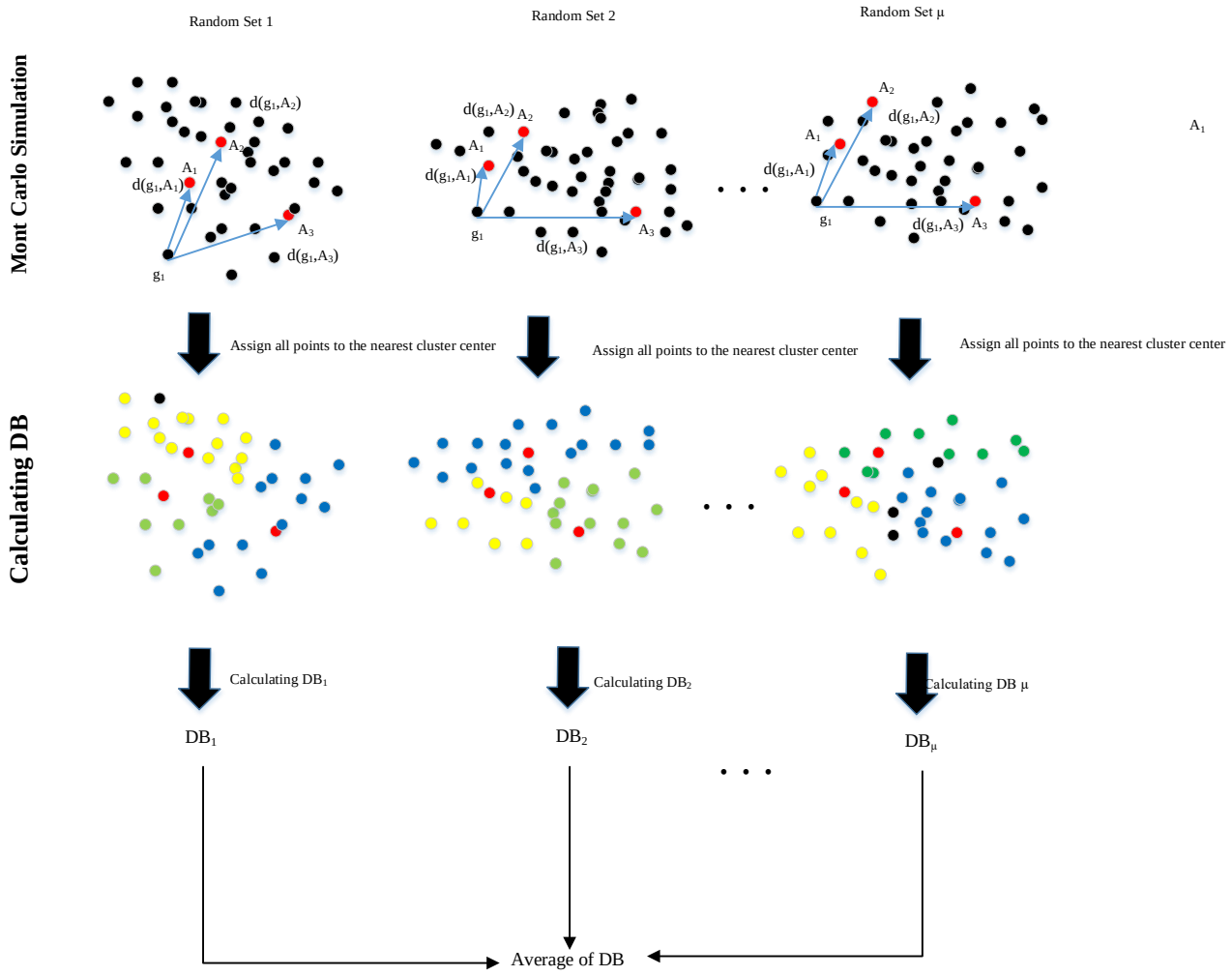

Fig. 4. Calculation mechanism of fitness function with Monte Carlo simulation.

The PSO stages are as follows (Fig. 3):

1. Seeding particles with an initial value (cluster centres from CSSUB outputs);
2. Calculating a fitness function based on Monte Carlo method (17);
3. Investigating PSO termination criterion $iter \geq iter_{\max}$ or $Fitness\ Value\ (\overline{DB}) \leq DB_m$ algorithm terminates if the criterion is fulfilled; otherwise, it returns to Step 4;
4. Recording the best position for each particle (p^i) and the best position between all particles (p_t^g);
5. Updating velocity vector of all particles using (13);
6. Moving particles to new position using (15) and returning to Step 3.

C. SUBHSO Algorithm

In the proposed hybrid method, the initial solutions (cluster centres) generated by CSSUS are calculated using the optimized PSO algorithm whose objective function is calculated from the simulation of random data selected by Monte Carlo method. To increase the chance of selecting an attribute set with the highest fitness value without concentrating on a specific attribute combination, the hyper form of this structure was used. Figure 5 shows the SUBHSO steps.

Steps of the new algorithm are defined as follows:

1. Some subspaces are selected from the initial databases;
2. For each selected subspace set, CSSUB is executed;
3. PSO-Monte Carlo algorithm (simulation-optimization) is used to improve cluster centres until the termination conditions are fulfilled.

Among all solutions from each hyper calculation of the algorithm, the best one is selected and considered as the best cluster of the whole dataset (At this stage, the EMM predicts the power demand based on the maximum similarity of daily data with cluster centres).

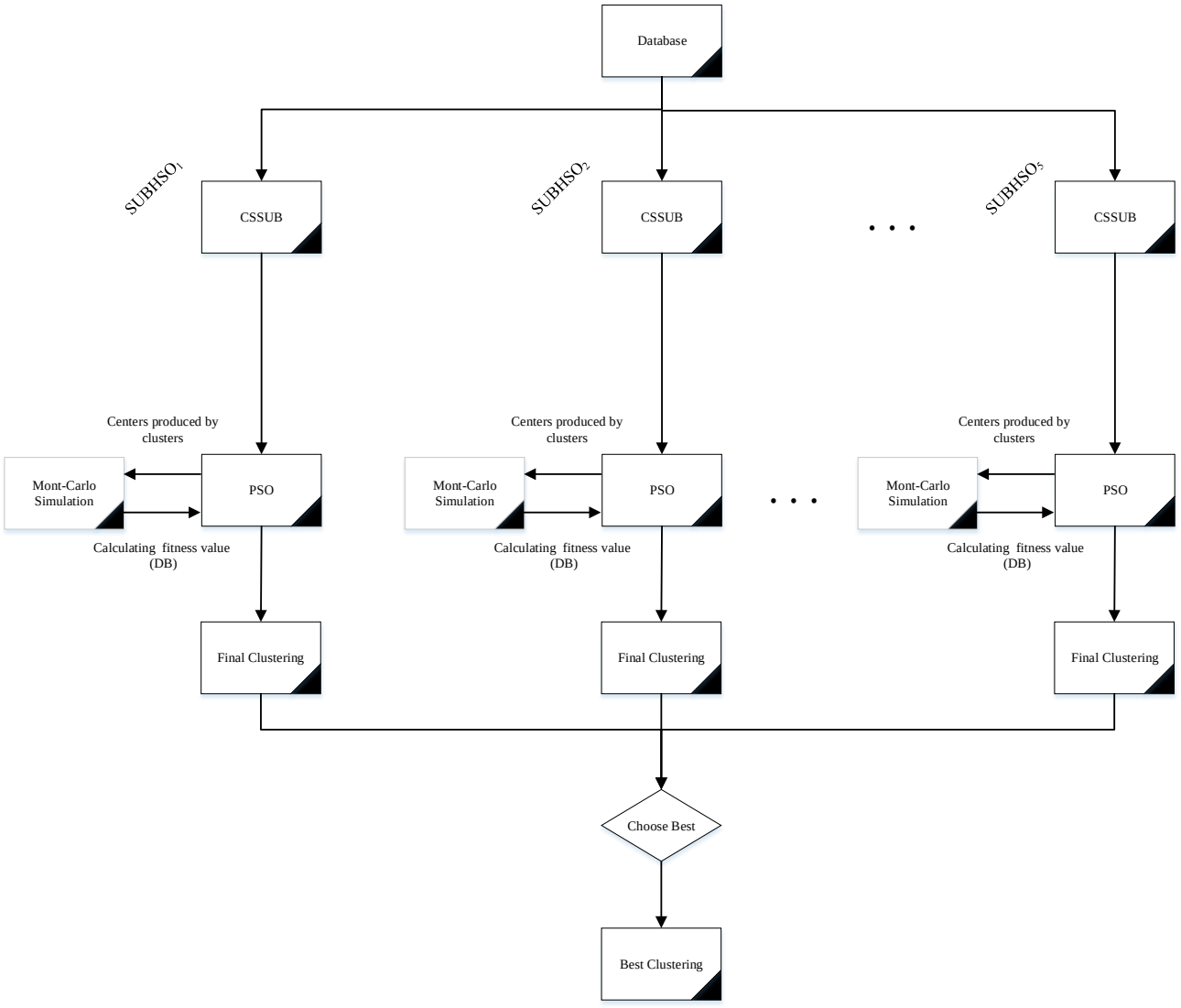

Fig. 5. SUBHSO algorithm.

III. Iran Electricity Market and Big Data

The electricity market is a place for trading electric power, like any other product, in a supply-and-demand system between the micro and macro sellers and purchasers. In contrast to the conventional electricity market structures, which unified the production, distribution and transfer sectors, these sectors act independently in the new structure. Meanwhile, the EMM is responsible for monitoring these interactions and pricing mechanism for the future periods. The EMM also predicts the electricity power price based on the predicted demand. As a result, electricity power demand prediction is the main variable for accurate prediction of its price. Since 2004, 24 parameters have been generated and recorded every minute or hour in the electricity market database. The data of IEM were extracted

from website [32], a database that had been developed by Iran Grid Management Company to control the electricity market in Iran. According to Fig. 6, these parameters are generated and recorded at a high rate. Considering the big data, including high-volume of records (more than 8-million records occupying 4 500 GB) and high-volume of attributes (more than 10 parameters), the dataset dimension is generally greater than 8000000 × 24. As a result, the dimension reduction methods should be used for data clustering.

IV. Data Clustering With the Proposed Algorithm

In this study, five iterations were considered for SUBHSO execution (r = 5). Table I represents the CSSUB parameters based on the study by Zhu et al. [22]. Table II presents the PSO parameters based on the studies by Amjady et al. [29]. Table III shows the Monte Carlo parameters based on the study by Chen et al. [26]. Results from the proposed algorithm based clustering of 1500 historical data of the electricity market are presented in Figs. 7–10.

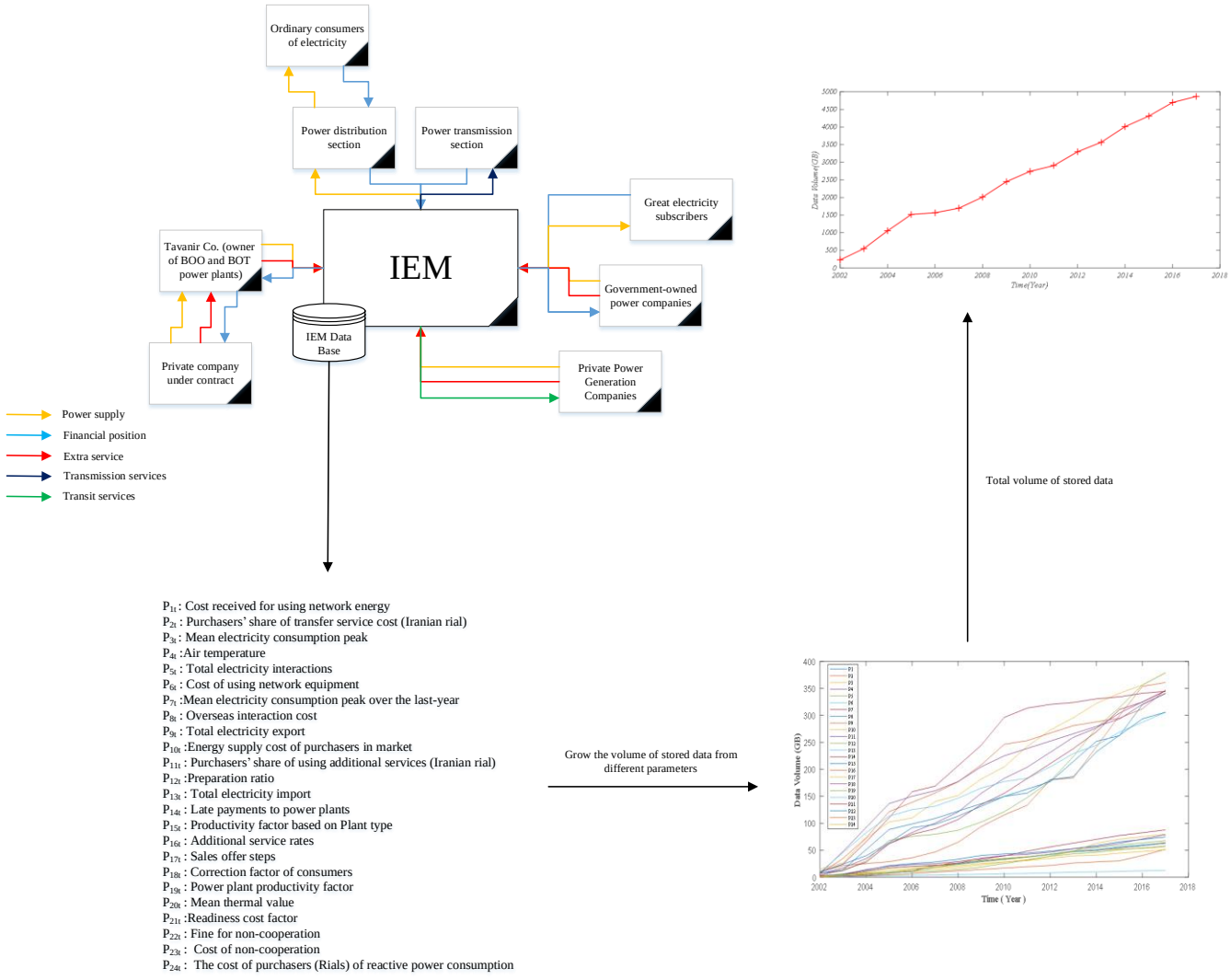

Fig. 6. Electricity market structure and its dataset dimension.

TABLE I
CSSUB Parameters

Parameter	Value
u	4
v	100
w	2
k	15
n	4

TABLE II
PSO Parameters

Parameter	Value
Swarm size	60
Initial weight w_1	0.9
Final weight w_2	0.4
ϕ_1, ϕ_2	2, 2
$iter_{\max}$	1000
DB_m	800

TABLE III
MONTE CARLO SIMULATION PARAMETERS

Parameter	Description	Value	Parameter	Description	Value
μ	Number of simulation iterations	1000	D_{P12}	Preparation ratio	[100, 150]
o	Simulation size	100	D_{P13}	Total electricity import, MW	[10, 2000]
D_{P1}	Cost received for using network energy, Iranian Rial	[2500, 5600]	D_{P14}	Late payments to power plants, days	[1500, 2500]
D_{P2}	Purchasers' share of transfer service cost, Iranian Rial	[20000, 40000]	D_{P15}	Productivity factor based on Plant type	[0.3, 1]
D_{P3}	Mean electricity consumption peak, MW	[25000, 40000]	D_{P16}	Additional service rates	[2000, 6000]
D_{P4}	Air temperature, °C	[−15, 42]	D_{P17}	Sales offer steps, Iranian Rial	[7000, 22000]
D_{P5}	Total electricity interactions, MW	[100, 3000]	D_{P18}	Correction factor of consumers	[0.3, 0.9]
D_{P6}	Cost of using network equipment, Iranian Rial	[2500, 4000]	D_{P19}	Power plant productivity factor	[0.3, 0.98]
D_{P7}	Mean electricity consumption peak over the last-year, MW	[25000, 45000]	D_{P20}	Mean thermal value, Iranian Rial	[20000, 40000]
D_{P8}	Overseas interaction cost, Iranian Rial	[3000, 40000]	D_{P21}	Readiness cost factor, Iranian Rial	[0.3, 1]
D_{P9}	Total electricity export, MW	[100, 2000]	D_{P22}	Fine for non-cooperation, Iranian Rial	[100, 500]
D_{P10}	Energy supply cost of purchasers in market, Iranian Rial	[150000, 200000]	D_{P23}	Cost of non-cooperation, Iranian Rial	[2000, 4500]
D_{P11}	Purchasers' share of using additional services, Iranian Rial	[20000, 40000]	D_{P24}	The cost of purchasers of reactive power consumption, Iranian Rial	[1500, 4000]

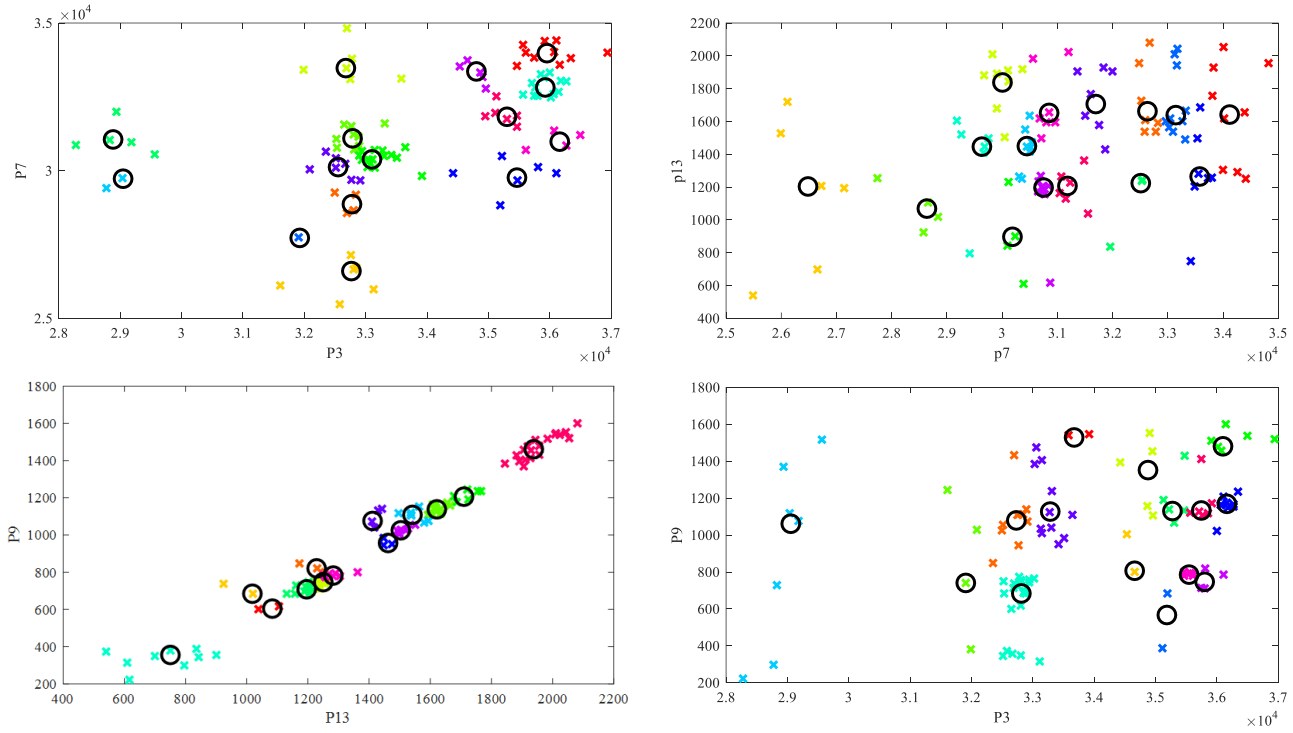

Fig. 7. Primary clustering by k-means algorithm.

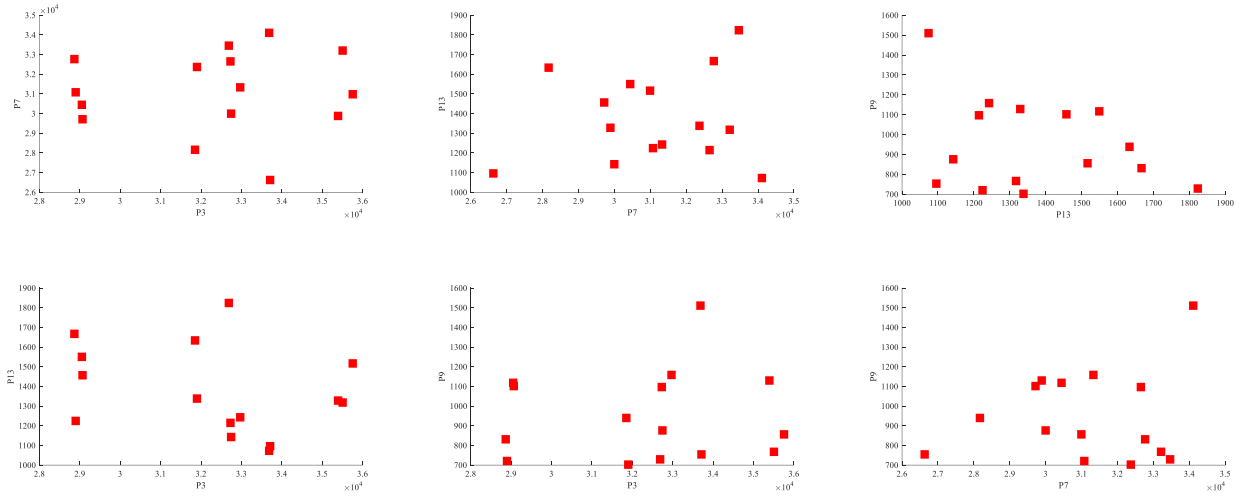

Fig. 8. Sharing sub-space and re-calculating cluster centres.

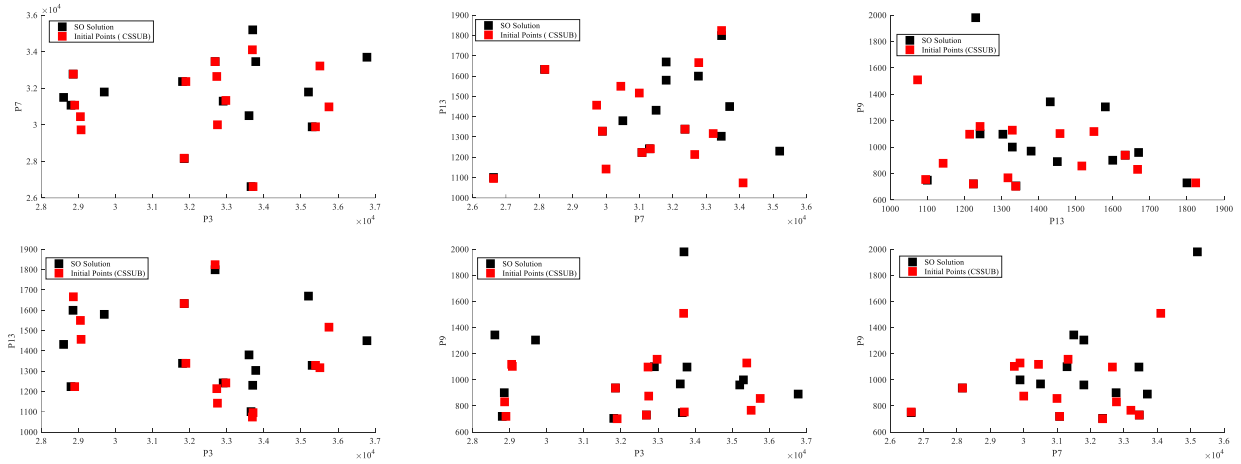

Fig. 9. Optimization of cluster centres by implementing SO algorithm.

V. Comparison of the Proposed Algorithm with Predecessors in Terms of Execution

In this section, the fitness function of the proposed algorithm was compared with the latest subspace search methods for different points of the overall dataset. The results are presented in Table IV.

TABLE IV
Comparison of Execution Accuracy ($\overline{DB}$)

Number of points	Model					
	LAC	PROCLUS	P3C	ESSC	CSSUB	SUBHSO
500	1350	1290	920	890	892	860
1500	1650	1580	1530	1459	1456	1235
2500	2104	1920	1896	1792	1789	1360
3500	1950	1900	1893	1870	1857	1667
4500	3100	2800	2450	2355	1980	1690
5500	2560	2456	2380	2349	2264	1705
Average	2119	1991	1844.83	1785.83	1706.33	1419.5

Results show that the fitness function of the proposed algorithm improved by 286.83 as compared to the best algorithm available. This improvement was due to the combination of different attributes with hyper use of the algorithm and optimization of cluster centres from the execution of simulation-optimization algorithm on the coordinates of the cluster centres. Figure 11 shows a greater improvement of the proposed algorithm relative to its predecessor with increasing the number of points. As a result, the distance between lines increases. This is because the likelihood of losing deleted data in the proposed algorithm reduces through the optimization of cluster centres and combination of more attributes.

To compare the ratio of normal changes of algorithm execution time to normalized value of the fitness function, variable S was defined as $S = Runtime' \cdot \overline{DB}'$ when $Runtime'$ and $\overline{DB}'$ are normalized. The lower value of this variable is more desirable. Figure 12 shows changes of S with increasing the number of points. Accordingly, the proposed algorithm performs better than its predecessor in dealing with datasets of higher dimensions.

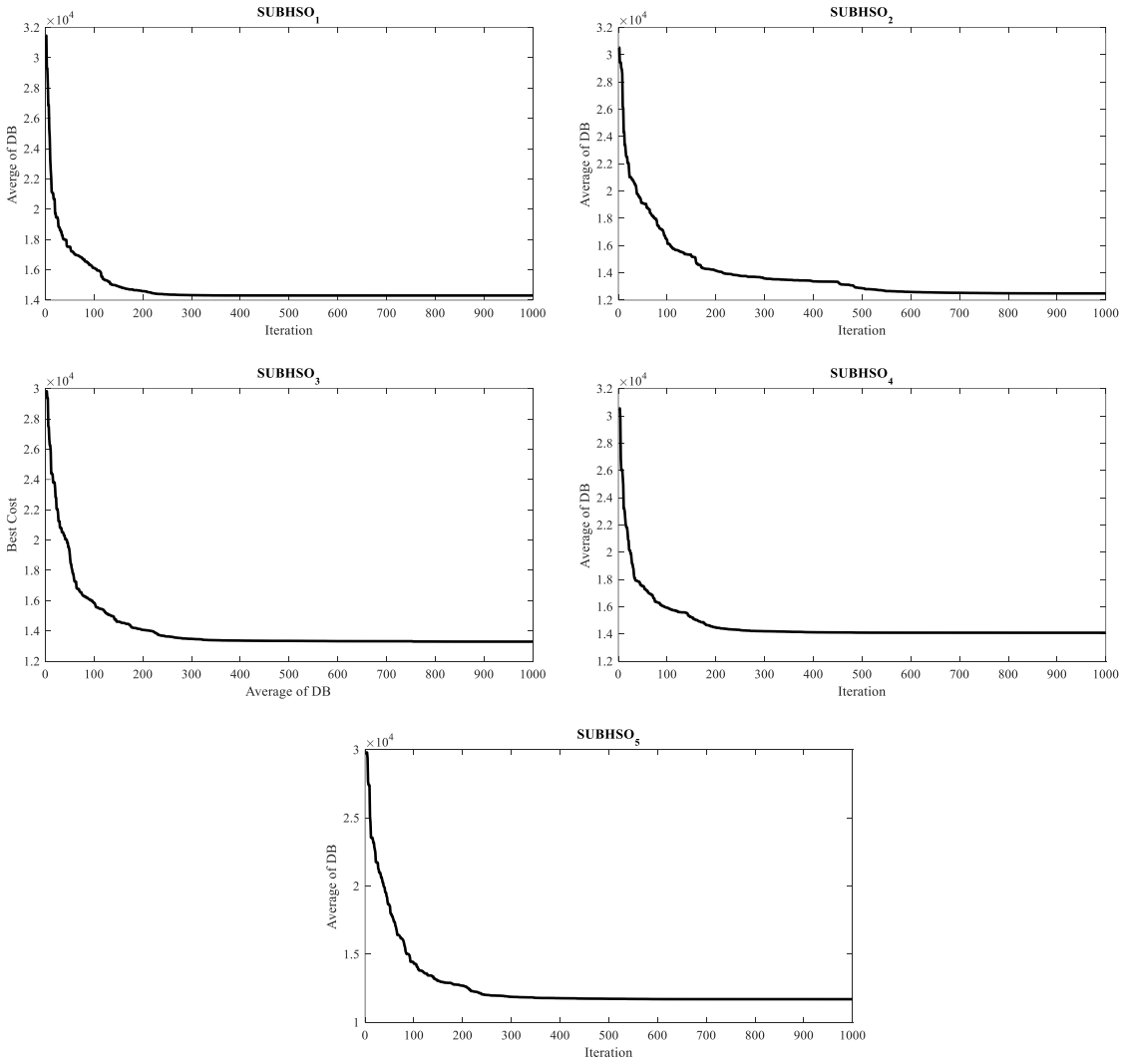

Fig. 10. Graph of optimization of the fitting function values in different branches of the hyper algorithm.

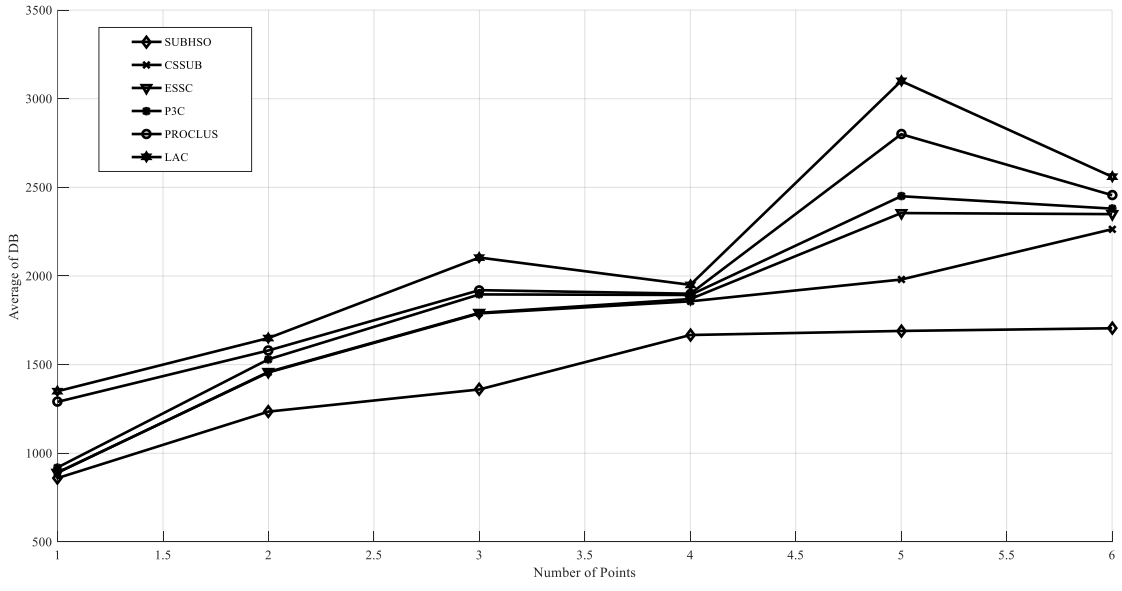

Fig. 11. Comparison of mean fitness function for different number of points.

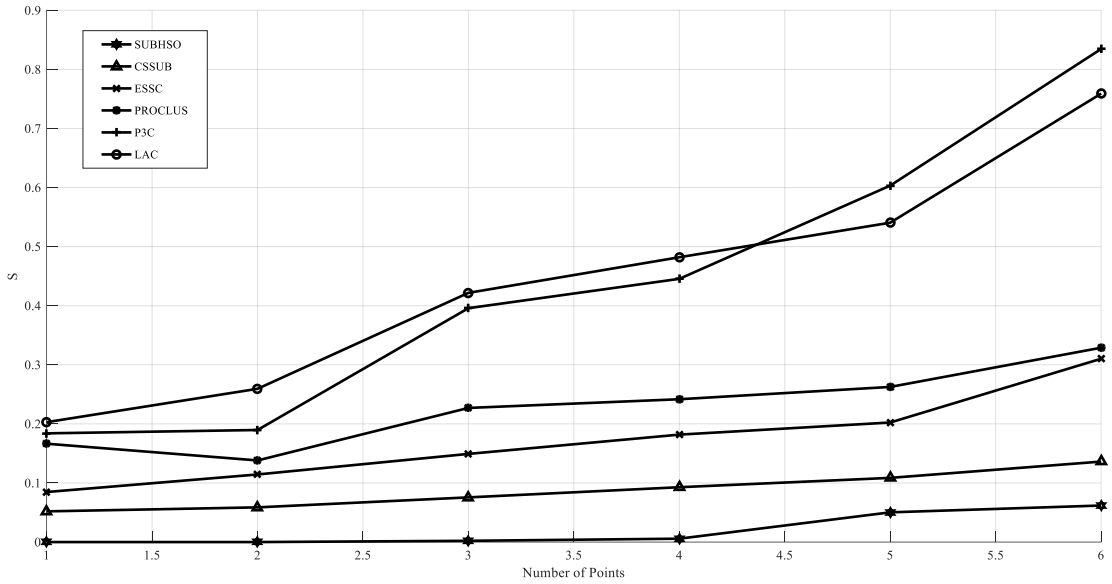

Fig. 12. Changes of variable *S* with increasing point number.

VI. Validation of the Proposed Algorithm

Student's t-test is a method of testing the difference, if any, between the mean of a sample and the mean of a population when the standard deviation of the population is unknown [29]. To use this test, the investigated variable should be expressed by the distance scale and distributed normally. Equations (18) and (19) show the distribution, mean values, and standard deviation of both populations under investigation. One of them is the mean fitness function of the proposed algorithm (X) and the other is the mean fitness function of existing algorithm (Y). Equation (20) is the test conditions, (21) is the test statistic value, and (22) is the confidence interval of 1−α % for the difference between the mean values of samples and fitness function value of the proposed algorithm and its predecessors.

$$DB_{X1}, DB_{X2}, \ldots, DB_{Xn} \sim N\left(\mu_{DB_X} \cdot {\sigma^2}_{DB_X}\right) \quad (18)$$

$$DB_{Y1}, DB_{Y2}, \ldots, DB_{Yn} \sim N\left(\mu_{DB_Y} \cdot {\sigma^2}_{DB_Y}\right) \quad (19)$$

$$H_0 : \mu_{DB_X} \leq \mu_{DB_Y} \quad (20)$$

$$H_1 : \mu_{DB_X} \geq \mu_{DB_Y}$$

$$\frac{\bar{X}-\bar{Y}-(\mu_x-\mu_y)}{\sqrt{\frac{s_x^2}{n_x}+\frac{s_y^2}{n_y}}} \sim t \sim t_v \quad v = n_x + n_y - 2 \quad (21)$$

$$\bar{X} - \bar{Y} \pm t_{\frac{\alpha}{2}v} \sqrt{\frac{s_x^2}{n_x} + \frac{s_y^2}{n_y}} \quad (22)$$

Results from Table V confirm the hypothesized improvement of the mean fitness function of the proposed algorithm relative to its predecessors.

TABLE V
Results From Statistical Hypothesis Testing for Comparison of Mean Value of Objective Function of the Proposed Algorithm and Its Predecessors

New method	Available methods	Accepted area for t_v (α = 0.05) (n_x, n_y = 15)	t_v	Accepted condition
SUBHSO	PROCLUS	$(-\infty . t_{\frac{\alpha}{2}v}]$ = $(-\infty = 2.0484]$	1.980	$\mu_{DB_X} \leq \mu_{DB_Y}$
	LAC		1.431	
	P3C		1.076	
	FSC		1.988	
	CSSUB		2.001	
	ESSC		1.567	

VII. Conclusion and Recommendations

The conventional data mining methods are not suitable for big data clustering. As a result, a research set, entitled dimension reduction methods, has been developed. The subspace research method is a dimension reduction technique. In these methods, the concentration is on the calculation of subspaces of data, which avoids computational complications without affecting the clustering accuracy. This article has proposed a new algorithm, SUBHSO, in which the cluster centres generated by CCSUB are changed and optimized in a simulation-optimization loop and the Davies–Bouldin index, regarded as the fitness function in this study, has been

calculated at each stage. Comparison of the fitness function of the proposed algorithm with that of its predecessors for different dataset points suggests the improvement of clustering, which enhances with an increase in the dataset points. This article has used high-volume of data of the electricity market, including high-volume of attributes and stored data, to help the EMM predict the demand based on the existing parameters and clustering results. In order to advance the subject of this paper, we suggest the following:

1. It is suggested that the variables be broken up into components before entering the analysis mechanism by signal analysis methods such as the wavelet.
2. It is suggested using methods such as PCA to decrease the dimensions of input data to CSSUB.
3. It is suggested using more advanced algorithms such as the parallel k-means algorithm for clustering.
4. It is suggested using the roulette cycle mechanism to select better subsets in CSSUB.

REFERENCES

[1] H. Chen, and Z. Mao, "Study on the failure probability of occupant evacuation with the method of Monte Carlo sampling," *Procedia Engineering*, vol. 211, 2018, pp. 55–62. https://doi.org/10.1016/j.proeng.2017.12.137

[2] T. G. Penkova, "Principal component analysis and cluster analysis for evaluating the natural andanthropogenic territory safety," *Procedia Computer Science*, vol. 112, 2017, pp. 99–108. https://doi.org/10.1016/j.procs.2017.08.179

[3] E. Vera, D. Lucio, L. A. F. Fernandes, and L. Velho, "Hough transform for real-time plane detection in depth images," *Pattern Recognition Letters*, vol. 103, 2018, pp. 8–15. https://doi.org/10.1016/j.patrec.2017.12.027

[4] M. H. Yang, J. H. Li, and B. X. Liu, "Fractal analysis on the cluster network in metallic liquid and glass," *Journal of Alloys and Compounds*, vol. 757, 2018, pp. 228–232. https://doi.org/10.1016/j.jallcom.2018.05.069

[5] T. Cui, F. Caravelli, and C. Ududec, "Correlations and clustering in wholesale electricity markets," *Physica A: Statistical Mechanics and its Applications*, vol. 492, 2018, pp. 1507–1522. https://doi.org/10.1016/j.physa.2017.11.077

[6] G. Zhu, J. Wang, and H. Lu, "Clustering based ensemble correlation tracking," *Computer Vision and Image Understanding*, vol. 153, 2016, pp. 55–63. https://doi.org/10.1016/j.cviu.2016.05.006

[7] S. Chormunge, and S. Jena, "Correlation based feature selection with clustering for high dimensional data," *Journal of Electrical Systems and Information Technology*, vol. 5, no. 3, 2018, pp. 542–549. https://doi.org/10.1016/j.jesit.2017.06.004

[8] K. Fujiwara, M. Kano, and S. Hasebe, "Development of correlation-based clustering method and its application to software sensing," *Chemometrics and Intelligent Laboratory Systems*, vol. 101, no. 2, 2010, pp. 130–138. https://doi.org/10.1016/j.chemolab.2010.02.006

[9] R. Veroneze, A. Banerjee, and F. J. von Zuben, "Enumerating all maximal biclusters in numerical datasets," *Information Sciences*, vol. 379, 2017, pp. 288–309. https://doi.org/10.1016/j.ins.2016.10.029

[10] S. Chen, J. Liu, and T. Zeng, "Measuring the quality of linear patterns inbiclusters," *Methods*, vol. 83, 2015, pp. 18–27. https://doi.org/10.1016/j.ymeth.2015.04.005

[11] G. F. de Sousa Filho, L. dos A. F. Cabral, L. S. Ochi, and F. Protti, "Hybrid metaheuristic for bicluster editing problem," *Electronic Notes in Discrete Mathematics*, vol. 39, 2012, pp. 35–42. https://doi.org/10.1016/j.endm.2012.10.006

[12] M. Wang, X. Shang, X. Li, W. Liu, and Z. Li, "Efficient mining differential co-expression biclusters in microarray datasets," *Gene*, vol. 518, no. 1, 2013, pp. 59–69. https://doi.org/10.1016/j.gene.2012.11.085

[13] Y. Lee, J. Lee, and C. H. Jun, "Stability-based validation of bicluster solutions," *Pattern Recognition*, vol. 44, no. 2, 2011, pp. 252–264. https://doi.org/10.1016/j.patcog.2010.08.029

[14] F. Divina, B. Pontes, R. Giráldez, and J. S. Aguilar-Ruiz, "An effective measure for assessing the quality of biclusters," *Computers in Biology and Medicine*, vol. 42, no. 2, 2012, pp. 245–256. https://doi.org/10.1016/j.compbiomed.2011.11.015

[15] C. C. Aggarwal, J. L. Wolf, P. S. Yu, C. Procopiuc, and J. S. Park, "Fast algorithms for projected clustering," *Proceedings of the 1999 ACM SIGMOD International Conference on Management of Data*, SIGMOD, ACM, New York, NY, USA, 1999, pp. 61–72. https://doi.org/10.1145/304181.304188

[16] G. Moise, J. Sander, and M. Ester, "Robust projected clustering," *Knowledge and Information Systems*, vol. 14, no. 3, 2008, pp. 273–298. https://doi.org/10.1007/s10115-007-0090-6

[17] G. Gan, and J. Wu, "A convergence theorem for the fuzzy subspace clustering (fsc) algorithm," *Pattern Recognition*, vol. 6, no. 2, 2008, pp. 1939–1947. https://doi.org/10.1016/j.patcog.2007.11.011

[18] Z. Deng, K. S. Choi, F. L. Chung, and S. Wang, "Enhanced soft subspace clustering integrating within-cluster and between-cluster information," *Pattern Recognition*, vol. 43, no. 3, 2010, pp. 767–781. https://doi.org/10.1016/j.patcog.2009.09.010

[19] X. Chen, Y. Ye, X. Xu, and J. Z. Huang, "A feature group weighting method for subspace clustering of high-dimensional data," *Pattern Recognition*, vol. 45, no. 1, 2012, pp. 434–446. https://doi.org/10.1016/j.patcog.2011.06.004

[20] D. S. Modha, and W. S. Spangler, "Feature weighting in k-means clustering," *Machine Learning*, vol. 52, no. 3, 2003, pp. 217–237. https://doi.org/10.1023/A:1024016609528

[21] C. Domeniconi, D. Gunopulos, S. Ma, B. Yan, M. Al-Razgan, and D. Papadopoulos, "Locally adaptive metrics for clustering high dimensional data," *Data Mining and Knowledge Discovery*, vol. 14, no. 1, 2007, pp. 63–97. https://doi.org/10.1007/s10618-006-0060-8

[22] Y. Zhu, K. M. Ting, and M. J. Carman, "Grouping points by shared subspaces for effective subspace clustering," *Pattern Recognition*, vol. 83, 2018, pp. 230–244. https://doi.org/10.1016/j.patcog.2018.05.027

[23] H. Chen, W. Wang, and X. Feng, "Structured sparse subspace clustering with within-cluster grouping," *Pattern Recognition*, vol. 83, 2018, pp. 107–118. https://doi.org/10.1016/j.patcog.2018.05.020

[24] W. Zhu, J. Lu, and J. Zhou, "Nonlinear subspace clustering for image clustering," *Pattern Recognition Letters*, vol. 107, 2018, pp. 131–136. https://doi.org/10.1016/j.patrec.2017.08.023

[25] X. Wang, Z. Lei, X. Guo, C. Zhang, H. Shi, and S. Z. Li, "Multi-view subspace clustering with intactness-aware similarity," *Pattern Recognition*, vol. 6, no. 2, 2018, pp. 50–63. https://doi.org/10.1016/j.patcog.2018.09.009

[26] Y. Chen, and Z. Yi, "Locality-constrained least squares regression for subspace clustering," *Knowledge-Based Systems*, vol. 163, 2019, pp. 51–56. https://doi.org/10.1016/j.knosys.2018.08.014

[27] Ł. Struski, J. Tabor, and P. Spurek, "Lossy compression approach to subspace clustering," *Information Sciences*, vol. 435, 2018, pp. 161–183. https://doi.org/10.1016/j.ins.2017.12.056

[28] D. L. Davies, and D. W. Bouldin, "A Cluster Separation Measure," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. PAMI-1, no. 2, 1979, pp. 224–227. https://doi.org/10.1109/TPAMI.1979.4766909

[29] N. Amjady, F. Keynia, and H. Zareipour, "Wind power prediction by a new forecast engine composed of modified hybrid neural network and enhanced particle swarm optimization," *Sustainable Energy*, vol. 2, no. 3, 2011, pp. 265–276. https://doi.org/10.1109/TSTE.2011.2114680

[30] T. P. Latchoumi, K. Balamurugan, K. Dinesh, and T. P. Ezhilarasi, "Particle swarm optimization approach for waterjet cavitation peening," *Measurement*, vol. 141, 2019, pp. 184–189. https://doi.org/10.1016/j.measurement.2019.04.040

[31] F. Korner-Nievergelt, T. Roth, S. von Felten, J. Guélat, B. Almasi, and P. Korner-Nievergelt, "Chapter 12: Markov chain Monte Carlo simulation," in *Bayesian Data Analysis in Ecology Using Linear Models with R, BUGS, and STAN*, Academic Press, 2015, pp. 197–212. https://doi.org/10.1016/B978-0-12-801370-0.00012-5

[32] IGMC. [Online] Available from: https://www.igmc.ir

Mebaholdin Salami is a student at Central Tehran Branch of Islamic Azad University. He also obtained Bachelor's and Master's degrees in industrial engineering in Iran. Now he is working on big data, especially in energy demand & price forecasting. In addition, he is interested in multi-objective model and metaheuristic solving method.
E-mail: mesbah.salami@yahoo.com

Farzad Movahedi Sobhani received his Ph. D. in Industrial Engineering from the Department of Industrial and Systems Engineering, Tarbiat Modares University, Tehran, Iran. He is currently a faculty member at the Department of Industrial Engineering, Science and Research Branch, Islamic Azad University, Tehran, Iran. His interests include business process management, knowledge management, and data mining.
E-mail: farzadmovahedi@gmail.com
ORCID iD: https://orcid.org/0000-0002-4602-2710

Mohammad Sadegh Ghazizadeh is an Associate Professor at Shahid Beheashti University (Power and Water University of Technology). He is interested in power system operation and control, electricity markets and smart grid.
E-mail: ghazizadeh.ms@gmail.com