2019 IEEE International Conference on Systems, Man and Cybernetics (SMC)
Bari, Italy. October 6-9, 2019

A Point Symmetry Distance Based K-Means Algorithm for Distributed Clustering in Peer to Peer Networks

Dinesh Kumar Kotary[1] and Satyasai Jagannath Nanda[1]

Abstract— **In this paper, a distributed K-Means algorithm is proposed based on the point symmetry distance measure which is termed as "Point symmetrical based distributed K-Means (PSDK-Means)" algorithm. Conventional distributed K-Means (DK-Means) clustering is able to detect only spherical shape clusters and it is not suitable for identifying the convex and concave (arbitrary shaped) clusters. The proposed method is implemented to detect spherical, convex and non-convex shape clusters which are distributed over the network at different peers. In the proposed method, cluster centers are shared using the diffusion based cooperation to achieve global clustering of the network. The cluster assignment is carried out using minimum point symmetry distance instead of Euclidean distance. Effectiveness of the proposed PSDK-Means algorithm has been validated on four synthetic and two real life datasets where it is observed to outperform conventional DK-Means algorithm.**

I. Introduction

Peer to Peer (P2P) networks consist of many systems or computers connected over by any medium. P2P networks contain a large volume of data which are distributed across several machines or systems. Wireless sensor network is one of the examples of P2P networks. Clustering is an important task to detect hidden meaning and valuable knowledge in networked data sources. Traditional clustering technique K-Means [1] and DBSCAN [2] require all data to be transferred to a central location which has several limitations like more power consumption, security issue, and large computation complexity. To mitigate these shortcomings, distributed clustering techniques are proposed by the many researchers [3]. In distributed clustering, the global clustering result of whole area is obtained by sharing of significant information to neighboring sites without the requirement of a central site. The distributed clusterings are applicable in segmentation of different files (like music files, world document, image file, etc.) of several connected computers. Some more applications are the clustering of research accomplishments at various institutions, clustering books of libraries located at several sites and clustering of weather monitoring data at different stations.

Distributed clustering using K-Means algorithm is introduced by Bandyopadhyay et al. in 2006 for peer to peer network [3]. Authors termed it as P2P K-means algorithm. In this algorithm every sensor node does clustering using popular K-Means and cluster centers are updated using dimension wise mean. Peer nodes which show a significant change in the cluster center send their cluster center and cluster count to every peer. Each peer node then calculate the weighted mean of the received cluster center which results in the actual cluster center for this iteration. This algorithm faces difficulty in global synchronization as every node moves to the next iteration only after completion of all other nodes present in the network for one iteration. Datta et al. in [4]-[5] proposed a distributed K-means clustering for static and dynamic networks. Thus it works effectively in case of node failure or change in topology. In this approach, K-means clustering is used, and after each iteration, all nodes share cluster count and centroids to their neighbors. Then new centroids are produced at each node using local data and received centroids. This technique does not work well when data is not uniformly distributed over the network. Forero et al. [6]-[7] formulated the distributed K-means clustering as an optimization problem. The help of duality theory is taken to solve the optimization problem. This algorithm faces the difficulty in detecting arbitrary shaped cluster such as convex and nonconvex structures. In [8] a fast and accurate distributed K-means algorithm is reported which uses the consensus method of cooperation among sensor nodes. A distributed version of K-Means for large scale network has been introduced in [9] which is fault tolerant. Zhou et al. [10] proposed a distributed uncertain K-Means clustering for uncertain data (data points are described based on a probability density function).

It is observed from the literature that distributed clustering algorithms based on K-Means do not able to detect the clusters that have arbitrary shape. It is due to the use of conventional Euclidean distance based cluster assignment. The main contribution of this work is to propose a point symmetry based distributed K-Means algorithm which can detect the spherical, convex and nonconvex clusters as long as they show symmetrical structure. In this technique data point, assignment to cluster center is carried out using symmetrical distance. Diffusion method of cooperation is used to share the cluster center which is more robust to link failure or change in network topology [11].

The rest of the paper is organized as follows. Section II begins with basic problem formulation for distributed clustering. Section III describes the original distributed K-means algorithm and proposed point symmetry based distributed clustering. Simulation studies and performance measure for distributed clustering are reported in Section IV. Results and discussions are described in section V. Finally all the analysis are concluded in Section VI.

[1] D. K. Kotary and S. J. Nanda are with Department of Electronics and Communication Engineering, Malaviya National Institute of Technology, Jaipur-302017, Rajasthan, India 2015rec9510@mnit.ac.in, sjnanda.ece@mnit.ac.in

978-1-7281-4569-3/19/$31.00 ©2019 IEEE

II. Distributed Clustering

Consider a P2P network consists of K number of peers, where each peer is represented by Y_k, $\forall k \in [1, K]$. It is assumed that network is fully connected and every peer has a set of (K-1) neighbors. Every peer has a local dataset $W_k = [w_k^1, w_k^2, \dots w_k^n \dots w_k^N]$, $W_k \in R^d$ where d is dimension and N is number of data samples recorded by peer Y_k. The data point of any k^{th} peer is denoted by w_k^n, $\forall n \in (1, N)$.

In clustering, each peer dataset W_k gets divided into Z clusters C_k^z, $\forall z \in (1, Z)$ with corresponding cluster center denoted by c_k^z. Distributed clustering is performed based on partitional approach which has the following properties:

1) The local dataset at every peer node comprise of the sum of data points present in all the clusters

$$W_k = \bigcup_{z \in Z} C_k^z \tag{1}$$

2) At every peer node, each cluster C_k^z should possess at least one data point

$$C_k^z \neq \phi, \quad k \in (1, K), \quad z \in (1, Z) \tag{2}$$

3) At every peer node, there should not be any common data points between two different clusters

$$C_k^{z_1} \bigcap C_k^{z_2} = \phi, \quad \forall z_1 \neq z_2 \text{ and } z_1, z_2 \in (1, Z) \tag{3}$$

4) After convergence of clusters using diffusion mode, at all peers of P2P Network the cluster center of same clusters should be nearly same

$$c_{k_1}^z = c_{k_2}^z, \quad \forall k_1, k_2 \in (1, K) \text{ and } z \in (1, Z) \tag{4}$$

In distributed clustering the intra-cluster distance (Euclidean distance) between data element w_k^n and cluster center c_k^z is minimized

$$\text{Minimize } \zeta = \sum_{k \in K} \sum_{z \in Z} \sum_{n \in C_z^k} \| w_k^n - c_k^z \|^2 \tag{5}$$

where $n \in [1, n_1]$, and n_1 is the number of data points present in cluster C_k^z.

III. Proposed Point Symmetry Based Distributed K-means Clustering

A. Background on Distributed K-Means clustering

Distributed K-Means (DK-Means) clustering algorithm is proposed by Forero et al. [6], where author called it as method of multiplier (MoM) K-Means algorithm. In this method distributed clustering problem is solved as an optimization problem using Lagrangian multipliers.

DK-Means clustering algorithm is used to cluster the data that are collected by different peers. The main steps for the algorithm are as follows:

Step 1: Every peer $k \in K$ randomly select the cluster centers c_k^z from the dataset W_k and initialize Lagrangian multipliers λ_k^z. Appropriate vale of $\mu > 0$ is chosen.

Step 2: Assign each data points w_k^n to cluster center c_k^z from which it has minimum Euclidean distance.

Step 3: Broadcast the initial cluster centers c_k^z to all the peers and Lagrangian multipliers λ_k^z to next peer $k1 \in K$

Step 4: Peer k update the cluster center c_k^z using following:

$$c_k^z = \frac{\sum_{x_{k,n} \in c_k^z} (x_{k,n}) + \sum_{k \in K} (2 \cdot \mu \cdot c_k^z - \lambda_k^z + \lambda_{k1}^z)}{2 \cdot \mu \cdot \mid W_k \mid + \mid C_k^z \mid} \tag{6}$$

Step 5: Peer $k \in K$ update the value of its Lagrange multipliers $\lambda_{k,k1}^z$, $\forall z \in Z, k1 \in K$ as:

$$\lambda_{k,k1}^z = \lambda_{k,k1}^z + \mu (c_k^z - c_{k1}^z) \tag{7}$$

Repeat **Step 2-5** until convergence criteria is satisfied.

B. Point Symmetry Based Distributed K-means Clustering

Modified point symmetry based distance $d_{ps}(w_k, c_k)$ associated with data point w_k from the cluster center c_k of node k is described in [12],[13]. It shows improved performance over the earlier PS (point symmetry) distance [14]. Let a data point of peer k be w_k. The symmetrical (reflected) data point of w_k associated with cluster center c_k is $2 \times c_k - w_k$. This point is denoted by $w_k^{'}$. If $knear$ represent unique nearest neighbors of $w_k^{'}$ located at Euclidean distances of d^i, $i = 1, 2, ...knear$. Then

$$d_{ps}(w_k, c_k) = d_{sym}(w_k, c_k) \times d_e(w_k, c_k) \tag{8}$$

$$= \frac{\sum_{i=1}^{knear} d^i}{knear} \times d_e(w_k, c_k) \tag{9}$$

where $d_e(w_k, c_k)$ denotes the Euclidean distance between the point w_k and cluster center c_k of peer k and $d_{sym}(w_k, c_k)$ is symmetrical distance of a point to particular cluster center. The value of d_{sym} is small if $w_k^{'}$ situated in the dataset of node k. Whereas d_{sym} is large when the reflected point is not symmetrical to associated cluster. The $d_{ps}(w_k, c_k)$ is composition of symmetrical based distance and Euclidean distance denoted as point symmetry distance. If point symmetry distance is chosen for cluster assignment rather than Euclidean distance then clustering algorithm can detect the cluster of arbitrary shape if they have symmetrical clusters.

Basic steps of point symmetry based K-Means are as follows:

Step 1. Initialization: Every peer $k \in K$, randomly select the cluster centers c_k^z from the dataset W_k and initialize Lagrangian multipliers λ_k^z. Choose appropriate vale of $\mu > 0$.

Step 2. Coarse tuning: Now original DK-Means algorithm is used to cluster the data using minimum Euclidean distance until the convergence criteria is satisfied (80% of total iteration). After the 80% of iteration the algorithm goes to fine tuning operation.

Step 3. Fine tuning: After coarse tuning phase now cluster assignment takes place using point symmetry distance. A data point w_k^i, $1 < i < N$ is joins cluster z if and only if $d_{ps}(w_k^i, c_k^z) \leq d_{ps}(w_k^i, c_k^y), \forall y = 1, ..., Z, z \neq y$ and $(d_{ps}(w_k^i, c_k^z))/(d_e(w_k^i, c_k^z)) \leq \alpha$. For $(d_{ps}(w_k^i, c_k^z))/(d_e(w_k^i, c_k^z)) > \alpha$ data point w_k^i is allocated to cluster s if $(d_e(w_k^i, c_k^s)) \leq (d_e(w_k^i, c_k^y))$, $y = 1, ..., Z, y \neq s$. The value of α is taken as maximum nearest neighbor

distance considering all data point in any peer as in [12].
Step 4. Updating: Update the cluster centers and Lagrangian multipliers of each peer using (6) and (7).
Step 5. Sharing: Share the cluster centers and Lagrange multipliers using diffusion method of cooperation.
Step 6. Continuation: If there is no change in cluster center or maximum number of iterations has reached, then stop. Otherwise go to Step 3.
The pseudo code of proposed PSDK-Means is shown in Algorithm 1.

IV. Simulation Studies

The simulation studies of proposed PSDK-Means and DK-Means [6] algorithms are carried out in MATLAB R2015a in an Intel Core i3 processor with 4 GB RAM and 500GB hard disk on Windows 8 (64-bit) platform. Both algorithms are allowed to run for 50 iterations, and the obtained performance is recorded. Four number of peers are considered for every six experiments. The value of μ is taken as 6 for each algorithm, and the value of $knear$ is taken as 4 for PSDK-Means algorithm. Three performance evaluation indices are used for accessing the quality of distributed clustering as follows.

A. Minkowski Score

Minkowski Score (MS) is a crucial criterion for evaluating cluster quality subject to the condition when the true partition is known apriori. MS score signifies the deviation of clustering solution obtained from true clustering. In other words, it describes the amount of incorrectly clustered data items.

Let R be the solution obtained after applying the clustering algorithm and T be the true clustering solution. The following parameters for defining MS can be considered:
1. u_{11} be the number of pairs of data element that is present in both R and T clusters.
2. u_{01} be the number of data elements present in only R.
3. u_{10} be the count of data elements in only T.
Then MS is defined as:

$$MS = \sqrt{\frac{u_{01} + u_{10}}{u_{11} + u_{10}}} \tag{10}$$

In P2P network for distributed clustering the Minkowski Score is given by:

$$MS_{dc} = \left(\frac{1}{K}\right) \sum_{k=1}^{K} MS_k \tag{11}$$

where MS_{dc} is average MS of all peer nodes and MS_k is MS of k^{th} peer.

Algorithm 1: Pseudo code of PSDK-Means

```
Define P2P network, number of peers and its neighbors.
Collect data at each peer.
Define number of iterations for course tuning (t_c) and fine tunning (t_f).
Input number of clusters Z, μ, c_k^z, and λ_k^z .
for each iteration until t_c do
    for every peer node k ∈ K do
        Find d_e(w_k, c_k)
        Find C_k^k using minimum d_e(w_k, c_k)
        Share c_k^z and λ_k^z to every peer.
        Update c_k^z and λ_k^z using (6) and (7)
    end
end
for each iteration until t_f do
    for every peer node k ∈ K do
        Find d_ps(w_k, c_k)
        Find C_k^k using minimum d_ps(w_k, c_k)
        Share c_k^z and λ_k^z to every peer.
        Update c_k^z and λ_k^z using (6) and (7)
    end
end
Return C_k^z
```

B. Silhouette index

The Silhouette index (SI) [15] is one of the popularly used performance indices in cluster analysis. A higher value of SI signifies that data points have better matching to its cluster and less similarity from other clusters. Silhouette index for distributed clustering is defined as:

$$SI_{dc} = \left(\frac{1}{NK}\right) \sum_{k=1}^{K} \sum_{n=1}^{N} SI(w_k^n) \tag{12}$$

where N is number of data points in a peer, K is number of peers. The Silhouette at any k^{th} peer associated with n^{th} data is given by

$$SI(w_k^n) = \frac{b(w_k^n) - a(w_k^n)}{max\{a(w_k^n), b(w_k^n)\}} \tag{13}$$

where $a(w_k^n)$ be the average distance between w_k^n and all other data within the same cluster of k^{th} sensor node, $b(w_k^n)$ is the lowest average distance of w_k^n to all points in any other cluster, of which w_k^n is not a member.

C. Dunn index

Dunn index [15] is another performance index to validate cluster quality. It is defined as the ratio of minimum inter-cluster distance to maximum cluster size. Thus larger value of inter-cluster distances and smaller cluster size results in higher Dunn index. The high value of Dunn index leads to better clustering. Mathematically the Dunn index for Z clusters for k^{th} sensor node is defined as:

$$DI(Z_k) = \overset{min}{i \in Z_k} \left\{ \overset{min}{j \in Z_k, j \neq i} \left\{ \frac{\delta(C_k^i, C_k^j)}{\overset{max}{z \in Z_k} \{\Delta(C_k^z)\}} \right\} \right\} \tag{14}$$

where $\delta(C_k^i, C_k^j) = min\left\{d(p_i, p_j) : p_i \in C_k^i, p_j \in C_k^j\right\}$, $\Delta(C_k^z) = max\left\{d(p_i, p_j) : p_i, p_j \in C_k^i\right\}$, d is distance. In

P2P network for distributed clustering the Dunn index is given by :

$$DI_{dc} = \left(\frac{1}{K}\right) \sum_{k=1}^{K} DI(Z_k) \tag{15}$$

V. Result Analysis and Discussions

A. Synthetic distributed datasets analysis

1) Circle_3_2 dataset : This dataset used in [13] and contains three clusters one ring shape, one rectangle shape, and one line shape cluster. Two dimensional 400 data points having three clusters as reported in Fig. 1(a). The same dataset is provided to all four peers, after that distributed clustering is performed using PSDK-Means and DK-Means algorithms. The clustering results of each peer at this dataset for PSDK-Means and DK-Means are reported in Fig. 2 and Fig. 3 respectively. It is clear from Fig. 2 PSDK-Means correctly clustered this dataset and having zero Mikowski score validated 100% accuracy for this dataset. It can be observed from Fig. 3 that many data points are clustered wrongly using DK-Means that's why Minkowski score is high as reported in Table I.

2) Sphere_3_2 dataset: This dataset has two dimensional 715 points as shown in Fig. 1(b). This dataset has one sphere shape cluster and two elliptical shaped clusters. The clustering result for this dataset is presented in Fig. 4 and Fig 5 for PSDK-Means and DK-Means respectively. It can be observed that most of the points are correctly classified using PSDK-Means. On the other hand in DK-Means the two elliptical shaped clusters are wrongly clustered. The Minkowski score of PSDK-Means is lower than DK-Means as reported in Table I which proves better accuracy of the proposed algorithm. The Dunn index also shows a better performance of PSDK-Means.

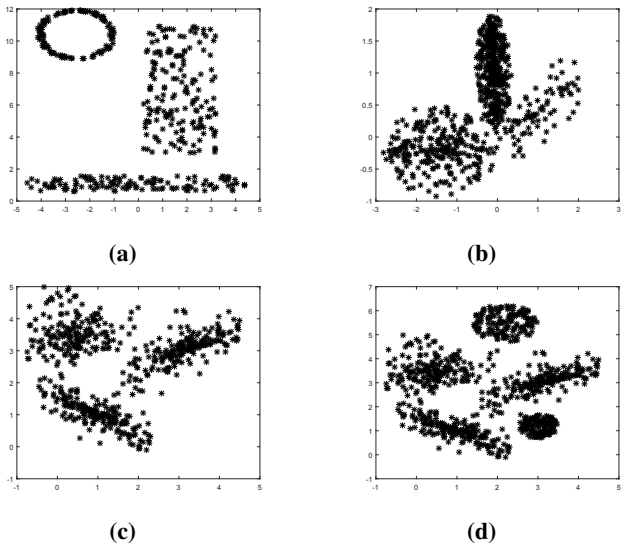

Fig. 1: Datasets given to different peers: (a) Circle_3_2, (b) Sphere_3_2, (c) Ellip_3_2, (d) Mixed_5_2

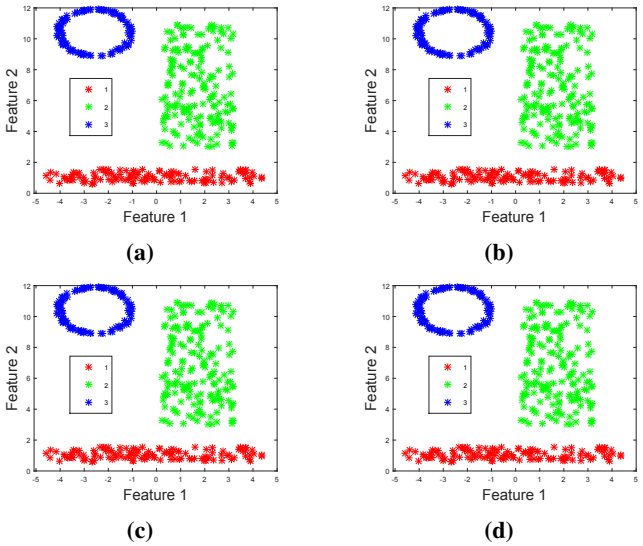

Fig. 2: Clustering result using PSDK-Means for Circle_3_2 dataset: (a) Peer 1, (b) Peer 2, (c) Peer 3 and (d) Peer 4

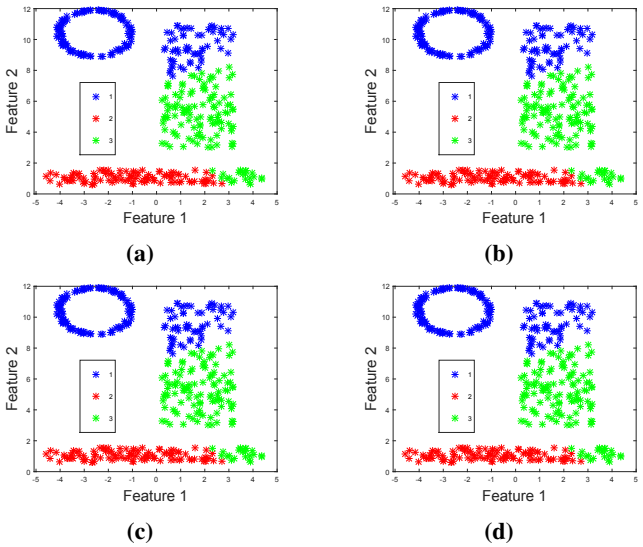

Fig. 3: Clustering result using DK-Means for Circle_3_2 dataset: (a) Peer 1, (b) Peer 2, (c) Peer 3 and (d) Peer 4

3) Ellip_3_2 dataset: This dataset is used in [14] having two dimensional 577 elements. This dataset contains one random and two elliptical shaped clusters as shown in Fig. 1(c) Distributed clustering result of each four peers is reported in Fig. 6 and Fig. 7 for PSDK-means and DK-Means respectively. Most of the points are correctly clustered using PSDK-Means while DK-means clustering shows poor performance. The Minkowski score of PSDK-Means is less which shows improved performance as compared to DK-Means. The Dunn index is better for PSDK-Means as compared to DK-Means.

4) Mixed_3_2 dataset: This Dataset has 850 points having two dimensions. This data has three spherical shaped clusters and two elliptical shaped clusters as shown in Fig. 1(d) The

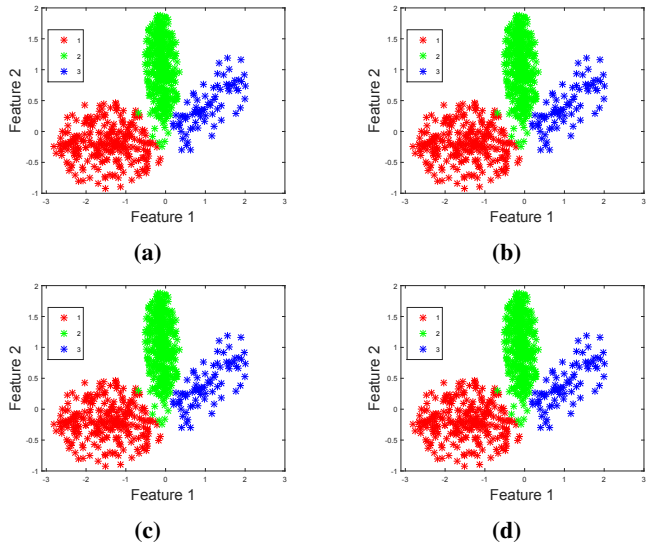

Fig. 4: Clustering result using PSDK-Means for Sphere_3_2 dataset: (a) Peer 1, (b) Peer 2, (c) Peer 3 and (d) Peer 4

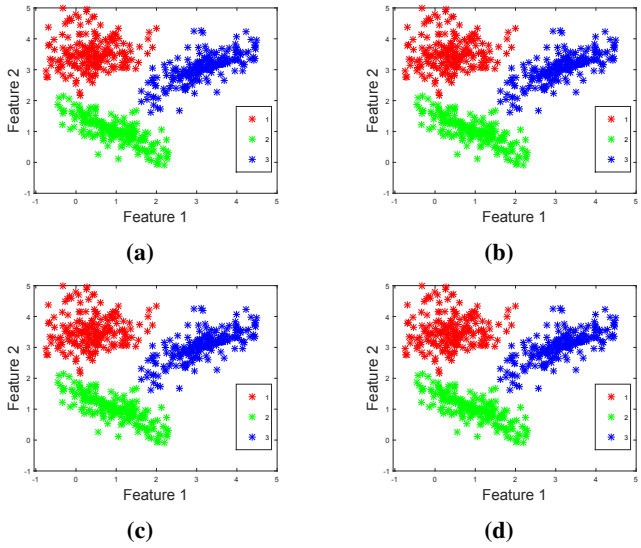

Fig. 6: Clustering result using PSDK-Means Ellip_3_2 dataset: (a) Node 1, (b) Node 2, (c) Node 3 and (d) Node 4

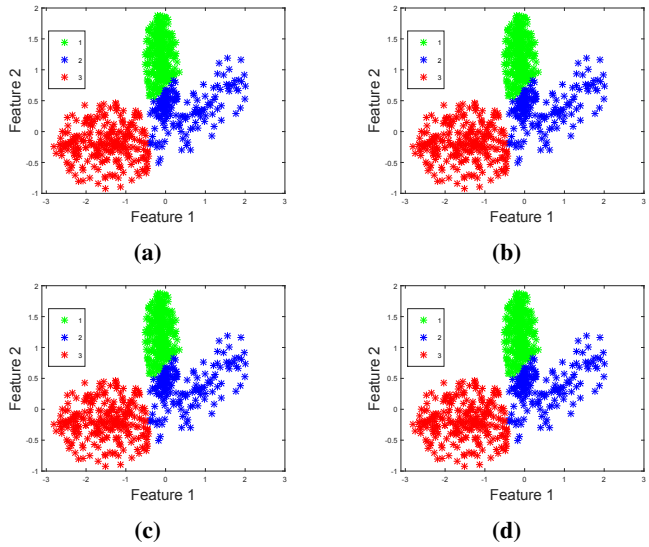

Fig. 5: Clustering result using DK-Means for Sphere_3_2 dataset: (a) Peer 1, (b) Peer 2, (c) Peer 3 and (d) Peer 4

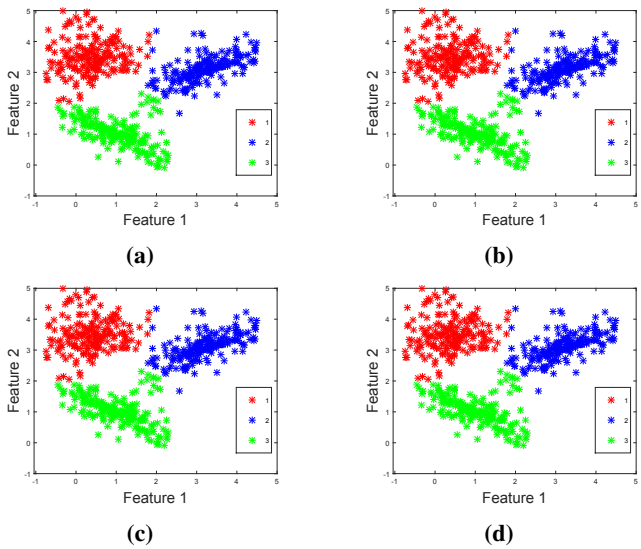

Fig. 7: Clustering result using PSDK-Means for Ellip_3_2 dataset: (a) Peer 1, (b) Peer 2, (c) Peer 3 and (d) Peer 4

clustering results are reported in Fig. 8 and Fig. 9 using PSDK-Means and DK-Means respectively. The Minkowski score and Dunn index is better in PSDK-Means for this dataset.

B. Real life distributed datasets analysis

1) Weather station dataset: This dataset has been taken from Canada government official site which keeps records of different weather parameters like temperature, relative humidity, wind speed, wind direction, pressure, etc. by installing many sensors at different stations [16]. Four nearby stations PA Toronto Hyundai, PA Downsview, PA Toronto North Yark Motors and PA Concord Ryder of Ontario, Canada, USA, have been considered here for validation of distributed clustering. Three parameter, temperature (degree Celsius), relative humidity (%) and wind speed (km/h) have been considered. The data has been recorded from October month at every hour and have 744 (31×24) samples. The data of four nodes has been shown in Fig. 10. The clustering result of each peer is reported in Fig. 11 using PSDK-Means. The better Silhouette index and Dunn index shows improved performance of the proposed method as compared to DK-Means.

2) Intel laboratory dataset:: It is a practical data set acquired by 54 sensors deployed in Intel Berkeley Research lab, USA over the period 28^{th} February and 5^{th} April 2004 [17]. Mica2Dot sensors are used to record the temperature (degrees Celsius), humidity (%), light intensity (Lux) and

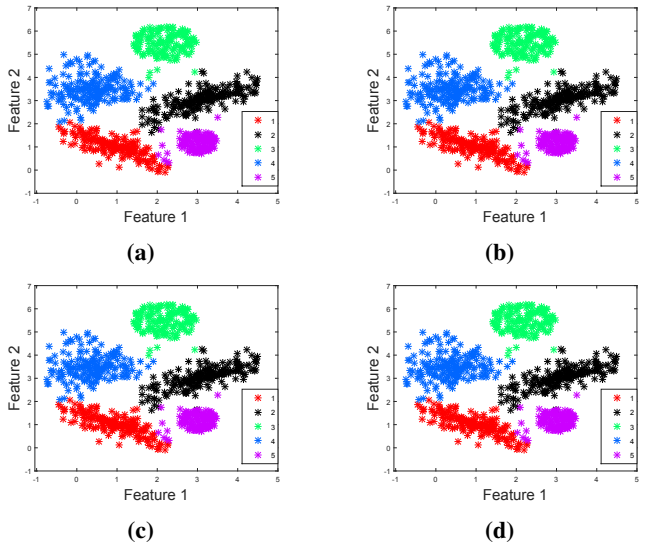

Fig. 8: Clustering result using PSDK-Means for Mixed_5_2 dataset: (a) Peer 1, (b) Peer 2, (c) Peer 3 and (d) Peer 4

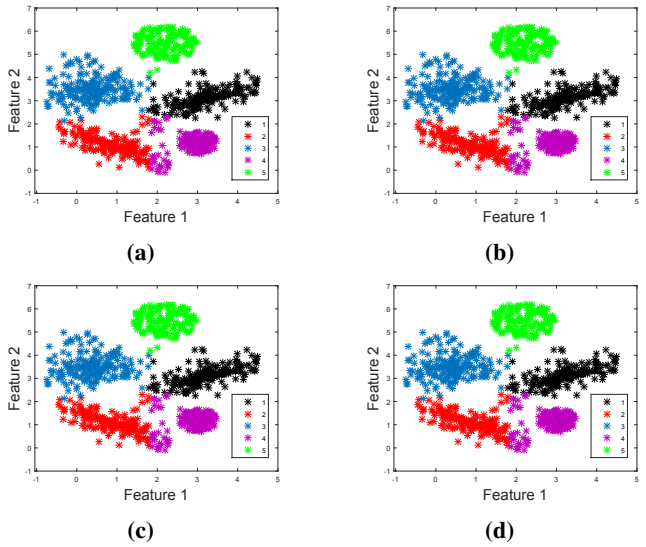

Fig. 9: Clustering result using DK-Means for Mixed_5_2 dataset: (a) Peer 1, (b) Peer 2, (c) Peer 3 and (d) Peer 4

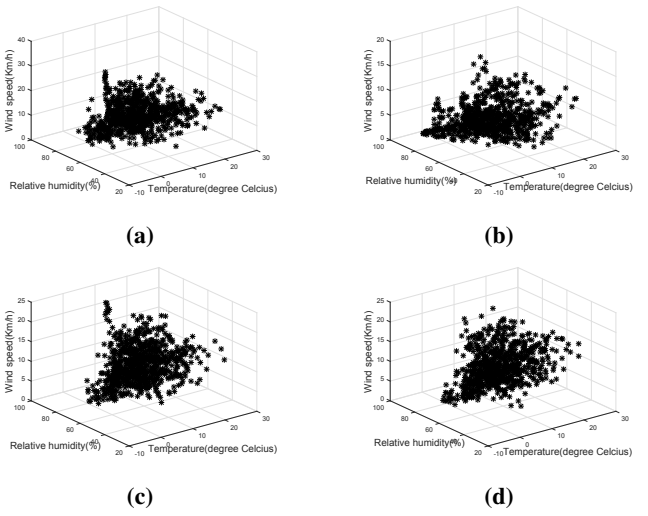

Fig. 10: Weather station data collected by different peers: (a) Peer 1, (b) Peer 2, (c) Peer 3 and (d) Peer 4

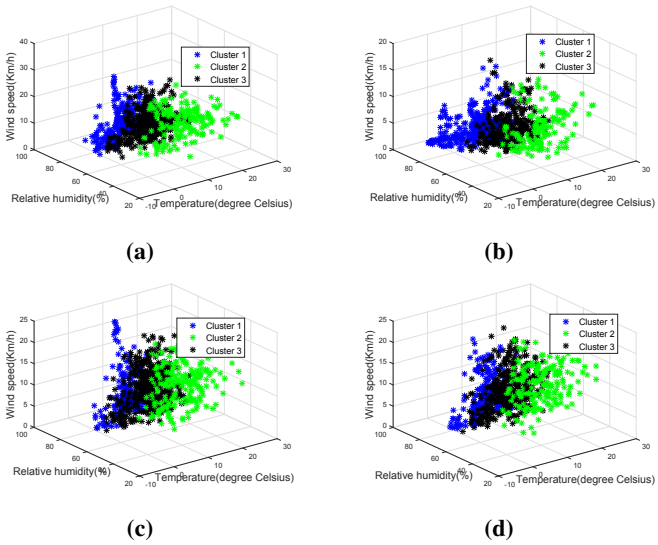

Fig. 11: Clustering result using PSDK-Means for weather station data: (a) Peer 1, (b) Peer 2, (c) Peer 3 and (d) Peer 4

voltage (V) reading at every 31 seconds. Here all the four parameters are considered for distributed clustering. In the simulation study, four peers are taken into consideration; each possesses 800 data points collected on 28^{th} February 2004 from midnight 12:00 to 23:59hrs. Clustering result of each peer is reported in Fig. 13 using PSDK-Means. Similarly, DK-Means algorithm is also simulated, and tabular results are shown in Table I.

C. Results of performance indices

The Minkowski Score, Silhouette and Dunn indices of both PSDK-Means and DK-Means algorithm is presented in Table I. The best results obtained are highlighted in bold letters. The Minkowski Score is evaluated only for the four labeled synthetic datasets. Both the real life datasets do not have labels thus for them Silhouette and Dunn indices are computed. It is observed that the results of proposed PSDK-Means is superior considering Minkowski Score and Dunn index. While considering Silhouette index as performance measure three out of six datasets have better results.

VI. Conclusion

In this paper, a point symmetry based distributed K-Means is proposed for P2P networks. A global clustering is achieved for network with the sharing of cluster centers without requirement of any central station. Distributed clustering with point symmetry distance measure achieves better performance compare to Euclidean distance based approach. The comparative analysis suggests that the proposed method

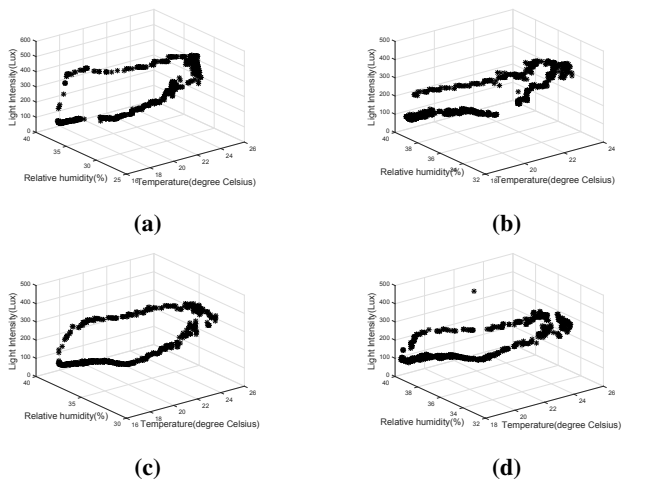

Fig. 12: Intel laboratory dataset collected by different peers: (a) Peer 1, (b) Peer 2, (c) Peer 3 and (d) Peer 4

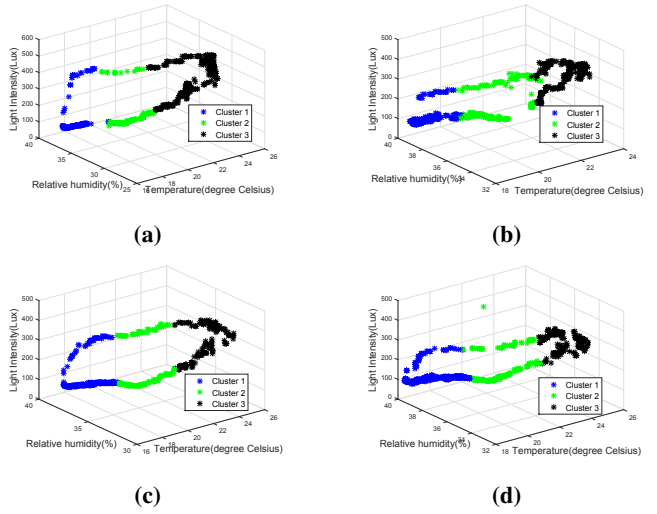

Fig. 13: Clustering results using PSDK-Means for Intel lab dataset: (a) Peer 1, (b) Peer 2, (c) Peer 3 and (d) Peer 4

is able to detect spherical, convex and non-convex clusters in effective manner. The simulation results on four synthetic and two real-life datasets validate the robust performance of the proposed approach. Minkowski Score and Dunn validation indices reveals the better performance of the proposed method over conventional DK-Means algorithm. .

VII. Acknowledgement

This research work is supported by Ph.D. fellowship to Mr. D. K. Kotary under Visvesvaraya Ph.D. Scheme for Electronics and IT, Miety, Govt. of India.

References

[1] J. A. Hartigan and M. A. Wong, "Algorithm as 136: A k-means clustering algorithm," *Journal of the Royal Statistical Society. Series C*, vol. 28, no. 1, pp. 100–108, 1979.

[2] M. Ester, H.-P. Kriegel, J. Sander, X. Xu *et al.*, "A density-based algorithm for discovering clusters in large spatial databases with noise." in *Kdd*, vol. 96, no. 34, 1996, pp. 226–231.

TABLE I: Comparative analysis of Minkoski Score (MS_{dc}), Silhouette (SI_{dc}) and Dunn index (DI_{dc}) for all datasets.

Index	Dataset	PSDK-Means	DK-Means
MS_{dc}	Circle_3_2	**0.0000**	0.5568
	Sphere_3_2	**0.1496**	0.5127
	Ellip_3_2	**0.1766**	0.2427
	Mixed_5_2	**0.1940**	0.3662
SI_{dc}	Circle_3_2	0.6422	**0.6623**
	Sphere_3_2	**0.7265**	0.6447
	Ellip_3_2	0.7874	**0.8012**
	Mixed_5_2	0.7466	**0.7600**
	Weather station	**0.5596**	0.5532
	Intel Lab data	**0.8764**	0.8511
DI_{dc}	Circle_3_2	**0.1613**	0.0365
	Sphere_3_2	**0.0403**	0.0153
	Ellip_3_2	**0.0859**	0.0649
	Mixed_5_2	**0.0152**	0.0143
	Weather station	**0.0289**	0.0257
	Intel lab data	**0.0785**	0.0609

[3] S. Bandyopadhyay, C. Giannella, U. Maulik, H. Kargupta, K. Liu, and S. Datta, "Clustering distributed data streams in peer-to-peer environments," *Inf. Sci.*, vol. 176, pp. 1952–1985, 2006.

[4] S. Datta, C. Giannella, and H. Kargupta, "K-means clustering over a large, dynamic network," in *Proceedings of the 2006 SIAM International Conference on Data Mining*. SIAM, 2006, pp. 153–164.

[5] G. C. Datta, Souptik and H. Kargupta, "Approximate distributed k-means clustering over a peer-to-peer network," *IEEE Transactions on Knowledge and Data Engineering*, vol. 21, no. 10, pp. 1372–1388, 2009.

[6] P. A. Forero, A. Cano, and G. B. Giannakis, "Consensus-based k-means algorithm for distributed learning using wireless sensor networks," in *Proceedings of the Workshop on Sensors, Signal and Info. Process., Sedona, AZ*, 2008, pp. 11–14.

[7] C. A. Forero, Pedro A and G. B. Giannakis, "Distributed clustering using wireless sensor networks," *IEEE Journal of Selected Topics in Signal Processing*, vol. 5, no. 4, pp. 707–724, 2011.

[8] J. Qin, W. Fu, H. Gao, and W. X. Zheng, "Distributed k-means algorithm and fuzzy c-means algorithm for sensor networks based on multiagent consensus theory," *IEEE transactions on cybernetics*, vol. 47, no. 3, pp. 772–783, 2017.

[9] G. Di Fatta, F. Blasa, S. Cafiero, and G. Fortino, "Fault tolerant decentralised k-means clustering for asynchronous large-scale networks," *Journal of Parallel and Distributed Computing*, vol. 73, no. 3, pp. 317–329, 2013.

[10] J. Zhou, L. Chen, C. P. Chen, Y. Wang, and H.-X. Li, "Uncertain data clustering in distributed peer-to-peer networks," *IEEE transactions on neural networks and learning systems*, vol. 29, no. 6, pp. 2392–2406, 2018.

[11] K. Dimple, D. K. Kotary, and S. J. Nanda, "Diffusion least mean square algorithm for identification of iir system present in each node of a wireless sensor networks," in *Computational Intelligence in Data Mining*. Springer, 2019, pp. 709–720.

[12] S. Saha and S. Bandyopadhyay, "A new multiobjective simulated annealing based clustering technique using symmetry," *Pattern Recognition Letters*, vol. 30, no. 15, pp. 1392–1403, 2009.

[13] S. Bandyopadhyay and S. Saha, "Gaps: A clustering method using a new point symmetry-based distance measure," *Pattern recognition*, vol. 40, no. 12, pp. 3430–3451, 2007.

[14] M.-C. Su and C.-H. Chou, "A modified version of the k-means algorithm with a distance based on cluster symmetry," *IEEE Transactions on Pattern Analysis & Machine Intelligence*, no. 6, pp. 674–680, 2001.

[15] S. J. Nanda and G. Panda, "A survey on nature inspired metaheuristic algorithms for partitional clustering," *Swarm and Evolutionary computation*, vol. 16, pp. 1–18, 2014.

[16] "Government of canada weather station dataset." [Online]. Available: http://climate.weather.gc.ca/historical_data/search_historic_data_e.html,

[17] "Intel berkely reseach lab IBRL dataset." [Online]. Available: http://db.csail.mit.edu/labdata/labdata.html,