IJCSE International Journal of Computer Sciences and Engineering Open Access
Research Paper Volume-4, Special Issue-1 E-ISSN: 2347-2693

# A Shortest Path Similarity Matrix based Spectral Clustering

**Parthajit Roy***
Department of Computer Science,
The University of Burdwan
Burdwan, West Bengal
India-713104
roy.parthajit@gmail.com

**Swati Adhikari**
Department of Computer Science,
The University of Burdwan
Burdwan, West Bengal
India-713104
swatidhkr@gmail.com

**J. K. Mandal**
Dept. of Comp. Sc. &Engg.
The University of Kalyani, Nadia, W.B.
India-741235
jkm.cse@gmail.com

***Abstract*—This paper proposed a new spectral graph clustering model by casting the non-categorical spatial data sets into an undirected graph. Decomposition of the graph to Delaunay graph has been done for computational efficiency. All pair shortest path based model has been adapted for the creation of the underlying Laplacian matrix of the graph. The similarity among the nodes of the graph is measured by a random selection based correlation coefficients. The effectiveness as well as the efficiency of the proposed model has beentasted and measuredwith standard data and the performances are compared with that of existing standard models.**

***Keywords*—**Graph Clustering;Delaunay Triangulation;All-pair Shortest Path Distance; Similarity Matrix; Spectral Clustering.

## 1. INTRODUCTION

Clustering plays a major role in the field of data science and decision support systems. Clustering is a learning technique which is unsupervised, i.e. learning takes place without the help of any training dataset [1].Clustering separates the data objects into meaningful groups or clusters based on proximity measures among the data objects [2]. This technique of grouping objects, based on similarity measures,is used for grouping documents, separating images and for many other such tasks. For this reason, clustering plays an important role in the field of pattern recognition.

There are several standard clustering algorithms exist in literature. A good overview of them has been discussed by Jain and Dubes[3]. A further development in this field has been reported by Xu&Wunsch-II[4]and Everitt*et al*[5]. Out of several branches in clustering, graph based clustering draws much of the attentions of the research communities in recent days. This is because, graph is a natural way of representation of today's real life problems;especially in network based problems. Graph theoretic clustering algorithms are useful for producing clusters where the problem is modeled by using graphs.Graphs are very useful to represent high dimensional unstructured data. So, graph based clustering for large, high dimensional datasets gaining popularity gradually.

A lot of research in the field of graph clustering has been done across the globe. Clustering using structural properties of weighted undirected graph has been used by Newman and Girvan[6]. Their idea is based on the node betweenness property of a graph. A novel betweenness property, based on shortest path, was coined by Freeman[7]. According to him, node betweenness is defined for each nodeas the number of shortest paths in the graph that pass through that node. The

betweenness of an edge as defined by Newman and Girvan[6] is the number of shortest paths connecting any pair of nodes that pass through the edge. They have used this property for clustering.

The proposed algorithm adapts a different approach. Instead of using this betweenness property, in this paper a new approach of all-pair shortest path has been proposed, in order to cluster an undirected graph. Some recent research on shortest path based clustering is done by Nawaz et al[8]. Different standard techniques used in graph clustering has been discussed in a survey paper by Schaeffer[9].

Spectral graph theory deals with the connectivity structures in a graph by casting a graph to an algebraic structure and by analyzing the spectra of the same. The connectivity and Spectral Graph has been discussed by Mohar[10]. A standard clustering method using spectra of graph has been proposed by Shi &Malik[11]. A good survey on spectral clustering has been made by Luxburg[12].The field of spectral clustering is being contributed by Spielman and Teng[13]. The general references on Laplacian matrix can be found in the work of Bapat[14] and Brouwer&Haemers[15]. Some current research in the topic has been done by Li *et al*[16] and Chrysouli et al[17]. Some local distribution based spectral clustering is proposed by Roy *et al*[18][19]. Jia*et al*[20] addressed the latest research progress in the field.

The proposed model uses a computational geometry based graph clustering. Compuational Geometry is the branch of algorithms which deals with geometry of spatial domain and Delaunay triangulation is a branch of computational geometry. Delaunay Triangulation of a set of points is a planar decomposition which maintains the proximity among the points [21].Clustering using Delaunay triangulation is proposed by Jia LV [22] and Deng *et al*[23]. A Fuzzy clustering using Delaunay triangulation is proposed by Roy *et al*[24]. The Spectra of Delaunay triangulation is analyzed by Chen *et al*[25]. This paper proposed to find the all pair shortest path matrix of the adjacency matrix of the Delaunay graph for spatial data clustering.

The rest of the paper is organized as follows. Graph and some of its representations are discussedin Section II.Eigen system of a graph has also been discussed here.This section also discussed about the Delaunay Triangulation, Shortest Path matrix and outlined two well-known clustering algorithms. The proposed methodis given in Section 3.Experimental setup and results, along the comparisons of the same with that of the existing algorithms has been discussed in Section 4. The concluding remarks are given in Section 5 and references are drawn at the end.

## 2. GRAPH,GRAPH MATRICES AND EIGENVALUES AND EIGENVECTORS OF GRAPH

This section discussed the essential graph properties used in the present paper. Graph is problem representation model which is used to represent pairwise relationship between objects. The *nodes* of the graph represent the objects and the *edges* represent the relationship between the nodes which is always a pairwise relationship. A *simple graph* is an undirected graph with no self-loops and no parallel edges between nodes.

There are several algebraic representation of a graph available. Out of which,Adjacency Matrix, Laplacian Matrix and Incidence Matrix are frequently used for solving problems related to clustering. The Adjacency matrix *A*, of an undirected graph *G*, is a symmetric matrix of order |V|×|V| where |V|isthe number of vertices of the graph G. The (*i*, *j*)-th entry $A_{ij}$of A is 1 if nodes *i* and *j* are agrees the relationshipand 0 otherwise.The leading diagonals i.e. $A_{ii}$, $i = 1, 2\ldots|V|$, are all zeroes if G is simple graph.If the graph is an weighted graph, then $A_{ii}=w$, for $i = 1, 2\ldots |V|$, where *w* is the weight of the edge (i,j). That means,

$$A_{ij} = \begin{cases} w, \text{ if } i \text{ and } j \text{ are adjacent} \\ 0, \text{ otherwise} \end{cases}$$

The Laplacianmatrix *L* of the graph *G* is defined as,

$$L_{ij} = \begin{cases} d_i, \text{ if } i = j \\ -A_{ij}, \text{ if } i \neq j \end{cases}$$

where$d_i$ is the sum of values of i-th row of the adjacency matrix A. The Laplacian matrix can be written as,

$$L = D - W$$

where*D* is an|V|×|V| diagonal matrix and whose *i*-th entry $d_i$is equal to *i*-th row-sum of the elements of the adjacency matrix, $i = 1, 2, \ldots\ldots, |V|$and *W* is the weight matrix of the graph *G*. Such type of Laplacian matrix is called the Un-normalized LaplacianMatrix of the graph *G*.

There exist other variants of Graph Laplacians also. These are known as Normalized Symmetric Laplacian and Random WalkLaplacian. Random Walk Laplacian is more popular in the field of clustering though the Symmetric Laplacian is also used. The definitions of the two are shown in equation 1 and 2 respectively.

$$L_{sym} = I - D^{-1/2} W D^{-1/2} \quad (1)$$
$$L_{rw} = I - D^{-1} W \quad (2)$$

$L_{sym}$ is symmetric and $L_{rw}$ is closely related to random walk in a weighted graph.$L$, $L_{sym}$ *and* $L_{rw}$ are all positive semi-definite and have $|V|$ non-negative real-valued eigenvalues. 0 is the smallest eigenvalue of all these threeLaplacians.

Among these $|V|$ Eigenvalues, different eigenvalues hold some specific spectral properties and the corresponding eigenvectors carry different information that can be used to cluster the graph [12].The eigenvector corresponding to the second smallest eigenvalue, for example, ofa $|V| \times |V|$ Laplacian matrix of a graph, is known as the Fiedler vector.

In the present paper a planar graph has been considered for clustering. To keep proximity information intact, a dense planar graph has been chosen. In graph theory, maximal Planar Graphs are known as triangulations. Amaximal planar subdivision can be defined as a subdivision such that no edge connecting two nodes can be added to this subdivision without destroying its planarity. The present paper considers such type of graph for data clustering.

A Delaunay Triangulation of a set of points P in a plane is a special type of triangulation which forms a planar graph where none of the points in P is inside the circle that circumscribes any triangle in that triangulation[21].

The shortest path distance is the distance between two nodes in a graph, where the sum of the weights of its component edges is minimized. The all-pair shortest path problem finds the shortest path between every pair of nodes of a graph. The graph may contain negative edges but no negative cycles.

In this paper, two well-known clustering algorithms — K-means Clustering algorithm and Hierarchical Clustering algorithm — are used for comparison of the result of the proposed model.

The K-Means algorithm worksin an iterative manner. It assumes the number of clusters $K$ as a *priory* and minimizes the error function $E = \sum_{i=1}^{K} \sum_{j=1}^{N} \| X_j - K_i \|^2$ ,where $K$ is the number of clusters, and $K_i$ is the center of $i^{th}$ cluster.

Hierarchical clustering algorithm groups the data in two ways — either by first considering all data objects as a single cluster (root) and next dividing the root into a number of small clusters (Divisive Clustering algorithm) or works in a reverse order. i.e. forms clusters of singleton nodes and merges them up (Agglomerative Clustering algorithm).

## 3. THE PROPOSED METHOD

The proposed model is a shortest path similarity based Laplacian clustering model. The motivation behind the model comes from the random walk nature of graphs. The random walk has been applied to Delaunay Triangulation of the original graph and on its Laplacian.

A random walk on a graph is a stochastic process which randomly jumps from node to node [9]. If a cluster is dense, then it stays for long time within that cluster and hardly jumps out of it. The transition matrix $T$ of the random walk is defined by,

$$T = D^{-1} W$$

where$D$ is the degree diagonal matrix and $W$ is the weight matrix. Random walk $T$ has connection with Normalized Laplacian$L_{rw} = I - T$.

If there exists shortest path between the data vectors $u$ and $v$ and that distance be $x$, then the data vectors whose shortest path to the data vector $v$ are almost same as $x$ will be in the same cluster, otherwise they will belong to the different cluster. To achieve the above mentioned goal, first the Delaunay Triangulation, of the graph representing the original dataset, has been computed. After this, the adjacency matrix of the Delaunay graph has been created. On this adjacency matrix, all-pair shortest path distance matrix has been computed. Then, some percentage (20% to 50%) of the original data points in the dataset are chosen randomly. The correlation of these data points is measured.If the correlations areless than a

predetermined threshold value then theyare set to zero. The correlation of a data vector to itself is also assigned to 0. Thus, a new adjacency matrix is formed with all leading diagonals set to zero.

Next the Normalized Laplacian matrix (Random Walk) is formed from the newly created adjacency matrix. Eigenvalues and eigenvectors of this Laplacian matrix is calculated and the eigenvectors related to K smallest eigenvalues have been considered for clustering using known clustering algorithms that have already discussed in Section 2. Following is the outline of the proposed model in algorithmic form.

*Proposed Algorithm*:

[The algorithm assumes that the number of clusters *K*,is known in advance.]

Input:Data Vector Set D, cut-off threshold value for selecting correlation of data vectors, percentage P for selecting random data vectors from the data set D and Number of Clusters K
Output: K numbers of clusters

Step 1: Compute Delaunay graph DG of the graph represented by Input data vector D.
Step 2: Compute the Adjacency matrix of this Delaunay graph DG.
Step 3: Compute the All-pair Shortest Path matrix SP from DG. If there is no edge between two data vectors, then set weight corresponding to that edge to a large amount.
Step 4: Select P percent of random data vectorsfrom the All-pair Shortest Path matrix SP.
Step 5: Evaluate the correlation matrix CM of this random data set.Set the correlation less than the threshold value to zero; also make the correlation of a data vector to itself to zero.
Step 6: Compute the Laplacian matrix LM from this correlation matrix CM using Random-Walk Laplacian$L_{rw}$.
Step 7: Calculate the Eigenvalues and Eigenvectors of the Laplacian matrix LM.
Step 8: Choose eigenvectors associated with the K smallest eigenvalues and apply K-means Clustering or Hierarchical Clustering algorithm to this vector to produce K numbers of clusters.

*Correctness of the Model*:

This subsection deals with the correctness of the proposed model.Firstly, the original graph has been decomposed to Delaunay graph. In this process the shortest path will slightly differ. But this is not less than a certain factor. Secondly, the shortest path proximity will remain invariant. i.e., the points which were close in the original graph will remain close in the Delaunay graph and the vice-versa because Delaunay graph ensures the proximity.

Third important point is, shortest path distances are metric. i.e. they follow the triangular inequality. Because of this property of shortest path, the correlation of the shortest path distances from the two close points is high. i.e., if a point C is far from a point A and B is close to A, then C is also far from B. If shortest path would not be a metric, then this assumption would not be true.

The final important point is, that the random walk Laplacian simulates random walk in a graph. i.e. a random walk will take longer time to come out from a dense sub-graph and relatively shorter time to come out from a sparse sub-graph. From the above mentioned discussions, it is clear that the proposed algorithm will work correctly to identify the clusters.

## 4. EXPERIMENTS AND RESULTS

The performance of the proposed model has been tasted with two standard datasets namely *iris* and *flame*. The benchmark dataset*iris*is due to UCI Machine Learning Unit [26] and the *flame* dataset is due to [27].

Iris is a dataset of 150 instances of three types of flowers namely Iris Setosa, Iris Versicolor and Iris Virginica. All of these three types of flowers have 50 instances in each. Among these three types of flowers, only one type of flower is linearly separable from the other two and no missing attributes exist there. Each instance has four attributes. These attributes are Sepal length, Sepal width, Petal length and Petal width.
Flame dataset has 240 instances and two attributes. Flame dataset has two clusters. These clusters are not well separated. The clusters are not linearly separable also. One of the clusters is not even convex shaped. The Iris data set is selected because it contains both linearly separable and non-separable classes. The flame dataset is chosen because it consists of clusters of non-convex shape. The performances of the proposed model will thus be tasted in case of non-separable clusters as well as arbitrary shaped clusters.

To analyse the performances, external clusters validity indexes are used. These indices are used to measure the similarity between results produced by two algorithms. In the present paper, the similarity between the computed result and the actual result has been considered for all the models. External indices which have been used for comparison are Czekanowski-Dice, Folkes_Mallows, Hubert, Jaccard, Rogers-Tanimotoand Russel-Rao.A detail discussion on such and other indexes can be found in the work of Saha*et al*[28] and Roy *et al* [29].

The performance of the proposed model is compared with that of two standard models. These are K-Means model and Hierarchical Model (Average Linkage).

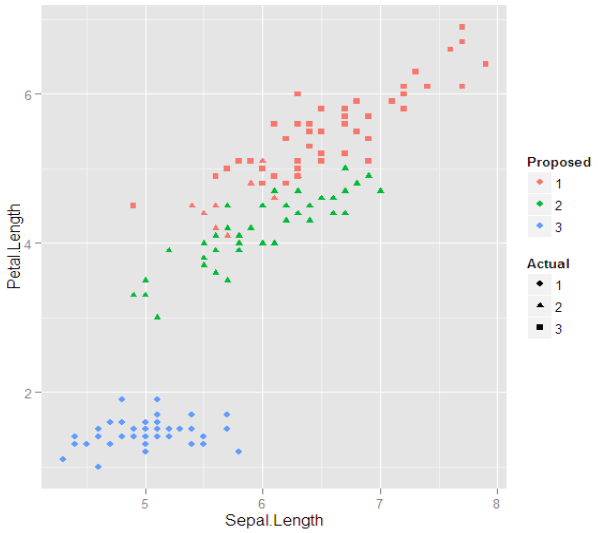


Fig. 1.Clustering of *iris* data using Proposed Algorithm

The result obtained by the proposed algorithm on *iris* dataset is shown in Figure 1.From this figure it is clear that our proposed algorithm can successfully cluster the linearly non-separable classes of *iris* data.

TABLE I
COMPARATIVE STUDY OF THE RESULTS OF K-MEANS, HIERARCHICAL AND PROPOSED ALGORITHM ON IRIS DATASET.

| External Indices | K-Means Algorithm | Hierarchical Algorithm | **Proposed Algorithm** |
|---|---|---|---|
| CZ-Dice | 0.8206 | 0.8400 | **0.8890** |
| Folkes Mallows | 0.8208 | 0.8407 | **0.8898** |
| Hubert | 0.7305 | 0.7597 | **0.8350** |
| Jaccard | 0.6958 | 0.7248 | **0.8014** |
| Rogers Tanimoto | 0.8037 | 0.8318 | **0.8752** |
| Russel Rao | 0.2751 | 0.2837 | **0.2958** |

The comparative study between Hierarchical(average linkage), K-Means and the proposed algorithm on iris dataset in different external indices is shown in Table I. Observing the results given in Table I, it is clear that the proposed algorithm performs better on *iris* dataset. In all the indexes, the proposed model gains highest value.

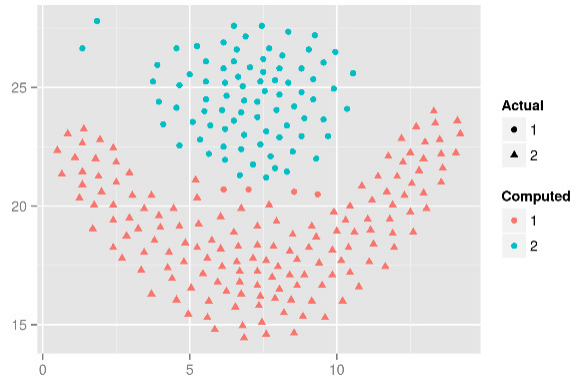


Fig. 2.Clustering of *flame* data using Proposed Algorithm

The performance of the proposed model on *flame* dataset has been shown in figure 2. It is clear from the figure that the performance of the algorithm is almost accurate except at the touching point of the two clusters. Out of 240 points, only four points are identified wrongly.

TABLE II
COMPARATIVE STUDY OF THE RESULTS OF K-MEANS, HIERARCHICAL AND PROPOSED MODEL ON FLAME DATASET.

| External Indices | K-Means Algorithm | Hierarchical Algorithm | **Proposed Algorithm** |
|---|---|---|---|
| Czek-Dice | 0.7581 | 0.7306 | **0.9696** |
| Folkes Mallows | 0.7586 | 0.7311 | **0.9696** |
| Hubert | 0.5012 | 0.4433 | **0.9339** |
| Jaccard | 0.6105 | 0.5756 | **0.9409** |
| Rogers Tanimoto | 0.5998 | 0.5638 | **0.9363** |
| Russel Rao | 0.3921 | 0.3783 | **0.5243** |

The comparative study of the performances of the standard models and that of proposed model is shown in Table II. It is clear from the table that the proposed model's performance is far better than the existing models.

## 5. CONCLUSION AND FUTURE SCOPE

This paper proposes a new data clustering method by using all-pair shortest path distances of graph and its Laplacian spectra. The performance of the proposed model is quite satisfactory in various situations. It can handle non-separability as well as non-convexity of the clusters. Though the performance of the proposed model is satisfactory, nevertheless, there are scopes of further improvements. Instead of using adjacency matrix, other graph matrices can be used and instead of using laplacian spectra, spectra of other types can also be considered.Similarity measures using commute distance of the graph can also be used for further tuning of the performances.

## ACKNOWLEDGEMENT

Authors express theirgratitude to the Department of Computer Science, The University of Burdwan and Department of Computer Science & Engineering for their necessary support and also to UCI Machine Learning Centre for their online standard datasets.

## REFERENCES

[1] A. K. Jain, M. N. Murty and P. J. Flynn, "Data Clustering: A Review," *ACM Computing Surveys*, vol. 31, no. 3, pp. 264-323, Sept. 1999
[2] Duda, R. O., P. E. Hart and D. G. Stork, "*Pattern Classification*," 2nd ed.,John Wiley&Sons,UK, 2008.
[3] A. K. Jain, and R. C. Dubes, "*Algorithms for clustering data*," Prentice Hall, , March 1988.

[4] RuiXu and Donald C. Wunsch, II, "*Clustering*," Wiley-IEEE Press, October 24, 2008
[5] B. Everitt, S. Landau S., M. Leese and D. Stahl, "*Cluster Analysis*," Wiley, 5thedn. February, 2011
[6] M.E.J. Newman and M. Girvan, Mixing patterns and community structure in networks, in: R. PastorSatorras, M. Rubi, A. DíazGuilera (Eds.), Statistical Mechanics of Complex Networks: Proceedings of the XVIII Sitges Conference on Statistical Mechanics, in: Lecture Notes in Physics, vol. 625, SpringerVerlag GmbH, Berlin, Germany, 2003.
[7] L.C. Freeman, "A set of measures of centrality based on betweenness," *Sociometry*, vol. 40, no. 1, pp. 35–41, 1977.
[8] Waqas Nawaz, Kifayat-Ullah Khan and Young-Koo Lee, "SPORE: shortest path overlapped regions and confined traversals towards graph clustering", *Applied Intelligence*, vol. 43,no. 1, pp. 208-232, 2015, DOI 10.1007/s10489-014-0637-7
[9] Satu Elisa Schaeffer, "Graph Clustering," *Computer Science Review*, vol. I, pp. 27-64, 2007.DOI 10.1016/j.cosrev.2007.05.001.
[10] B. Mohar, "Some Applications of Laplace eigenvalues of graphs", *Graph Symmetry: Algebraic Methods and Applications*, NATO ASI, Series.C-497, pp.225-275, pub.Kluwer, Editor. G. Hahn and Sabidussi, 1997.
[11] Jianbo Shi and Jitendra Malik, "Normalized Cuts and Image Segmentation", *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol.22, no.8, pp.888-905, August, 2000.
[12] Ulrike von Luxburg, "A Tutorial on Spectral Clustering," *Statistics and Computing*, vol. 17, no. 4, pp. 1-32, 2007.
[13] Daniel A. Spielman and Shang-HuaTeng, "Spectral Partitioning Works: Planar Graphs and Finite Element Meshes", *Linear Algebra and its Applications*, vol.421, no. 2-3, pp.284-305, March, 2007.
[14] R. B. Bapat, "The Laplacian Matrix of A Graph," *The Mathematics Student*, vol. 65, nos. 1-4, pp. 214-223, 1996.
[15] A. E. Brouwer and W. H. Haemers, "Spectra of Graphs," Springer, February, 2011.
[16] Jianyuan Li, Yingjie Xia andYuncai Liu, "Scalable Constrained Spectral Clustering", *IEEE Transactions on Knowledge and Data Engineering*,vol. 27, no. 2, pp. 589-593, September, 2014.
[17] Christina Chrysouli andAnastasiosTefas, "Spectral clustering and semi-supervised learning using evolving similarity graphs", *Applied Soft Computing,* vol.34, No. C, pp. 625-637, 2015.
[18] Parthajit Roy and J. K. Mandal, "A Novel Spectral Clustering based on Local Distribution", *International Journal of Electrical and Computer Engineering(IJECE)*, vol.5, No.2, pp.361-370, April, 2015.
[19] Parthajit Roy, Swati Adhikari and J.K. Mandal, A Novel Similarity Matrix based Spectral Clustering for Two Class Problems, in *Proceedings of the National Conference on Computing, Communication and Information Processing (NCCCIP-2015)*, ISBN: 978-93-84935-27-6, pp:149-157, May, 2015.
[20] HongjieJia, Shifei Ding, XinzhengXuand RuNie, "The latest research progress on spectral clustering", *Neural Computing & Applications*, vol.24, , no. 7-8 pp.1477-1486, June, 2014 , DOI 10.1007/s00521-013-1439-2.
[21] M. D. Berg, O. Cheong, M. V. Kreveld and M. Overmars, "Computational Geometry: Algorithms and Applications" 3rd ed., Springer-verlag, 2008.
[22] Jia LV, "Clustering Algorithm Based on Delaunay Triangulation Density Metric", *inProceedings of the Seventh International Conference on Fuzzy Systems and Knowledge Discovery (FSKD 2010)*,vol.4, pp.1621-1624, 2010.
[23] Min Deng, Qiliang Liu, Tao Cheng and Yan Shi, "An adaptive spatial clustering algorithm based on delaunay triangulation", *Computers, Environment and Urban Systems*, vol.35, no. 4, pp.320-332, July, 2011.
[24] Parthajit Roy and J. K. Mandal, "A Delaunay Triangulation Preprocessing Based Fuzzy-Encroachment Graph Clustering for Large Scale GIS Data", *in Proceedings of the International Symposium on Electronic System Design, 2012*, pp.300-305, December, 2012.
[25] Renjie Chen, Yin Xu, Craig Gotsman, Ligang Liu, "A spectral characterization of the Delaunay triangulation", *Computer Aided Geometric Design*, vol.27, no. 4, pp.295-300, May, 2010.
[26] R. A. Fisher, "UCI machine learning repository," 1936. [Online]. Available: http://archive.ics.uci.edu/ml
[27] L. Fu and E. Medico, " FLAME, a novelfuzzyclusteringmethod for the analysis of DNA microarray data," *BMC Bioinformatics*, vol. 8, no. 3, January 4, 2007, DOI: 10.1186/1471-2105-8-3.
[28] SriparnaSaha and SanghamitraBandyopadhyay, "Performance Evaluation of Some Symmetry-Based Cluster Validity Indexes", *IEEE Transaction on Systems, Man, and Cybernetics-Part C: Applications and Reviews*, Vol-39, No-4, pp-420-425, July 2009.
[29] Parthajit Roy and J.K. Mandal, Performance Evaluation of Some Clustering Indices, *presented in the conference Computational Intelligence in Data Mining - 2014*, Volume 3, Published in the Springer online in Smart Innovation, Systems and Technologies, Volume 33, ISSN:2190-3018, ISBN(print): 978-81-322-2201-9, ISBN(online):978-81-322-2202-6 , pp:509-517, 2015.