ISBN 978-93-84468-20-0
Proceedings of 2015 International Conference on Future Computational Technologies (ICFCT'2015)
Singapore, March 29-30, 2015, pp. 61-67

A Hybrid Approach for Data Clustering using Expectation-Maximization and Parameter Adaptive Harmony Search Algorithm

Vijay Kumar[1], Jitender Kumar Chhabra[2] and Dinesh Kumar[3]

[1]Computer Science and Engineering Department, Manipal University, Jaipur, Rajasthan, India
[2]Computer Engineering Department, National Institute of Technology, Kurukshetra, India
[3]Computer Science and Engineering Department, GJUS&T, Hisar, India

Abstract**:** *This paper presents a novel hybrid data clustering algorithm based on parameter adaptive harmony search algorithm. The recently developed parameter adaptive harmony search algorithm (PAHS) is used to refine the cluster centers, which are further used in initializing Expectation-Maximization clustering algorithm. The optimal number of clusters are determined through four well-known cluster validity indices. The proposed algorithm is evaluated on three real life datasets and compared with the performance of K-Means, Fuzzy C-Means and HS initialize EM (HSEM). Experimental results reveal that the proposed approach provide better results in terms of precision, recall, weighted average, F-Measure and G-Measure.*

Keywords: *Harmony Search Algorithm, Clustering, Expectation-Maximization.*

1. Introduction

The Clustering is a distribution of data into groups based upon similar characteristics of data while minimizing the within group variability and maximizing the between group variability. These techniques have been applied in wide variety of applications such as biology, medicine, engineering and data mining [1].These can be classified into four categories: partition, hierarchical, density-based and grid based clustering. Hierarchical clustering techniques are able to find structures which can be further divided in substructures and so on recursively [1], some of which are: Balanced Iterative Reducing and Clustering using Hierarchies (BIRCH), Clustering using REpresentatives (CURE) [2], and ROCK [3]. Density-based clustering algorithms try to find clusters based on density of data-points in a region [13]. The well known density based clustering techniques are-DBSCAN [14] and DENCLUE [15]. Grid based clustering algorithms quantize the cluster space into a finite number of cells and then perform the required operations on the quantized cluster space [13]. Cells which having more than specified numbers of points are called dense cells. The dense cells are connected to form clusters [13], some of which are: STING [16] and CLIQUE [17].

Partition clustering techniques try to obtain a single partition of data without any other sub-partition and are based on the optimization of an objective function [10]. The most popular partition clustering techniques are: K-Means [2], Fuzzy C-Means, SOM and Expectation-Maximization (EM). Among these, K-Means technique has gained more popularity due to its simplicity and efficiency, but it has weakness of converging to local minima. Another well known approach is EM. EM algorithm uses probability of cluster membership instead of a distance metric [4]. Unlike K-Means, EM is known to be an appropriate optimization algorithm for constructing proper statistical models of the data [4]. Selection of initial cluster centers is crucial for EM. To solve this problem, new techniques have been proposed [5].

In this paper, recently developed parameter adaptive harmony search algorithm (PAHS) is used to improve the initialization of cluster centers for Expectation-Maximization. Our method has two main stages. In the first stage, the PAHS algorithm applied on datasets to find the optimal cluster centers. In second stage, these cluster centers are used in EM as initial cluster centers. After that, EM clustering algorithm is applied to datasets. The rest of the paper is organized as follows. Section 2 presents the basic concepts of Expectation-Maximization and parameter adaptive harmony search algorithm. Section 3 presents proposed clustering approach. Section 4 presents the cluster quality metrics. Section 5 covers the experimental results followed by conclusions in Section 6.

2. Background

2.1. Expectation-Maximization Clustering Algorithm

Expectation Maximization is a statistical technique for maximum likehood estimation using mixture models. It is a model based approach to solve clustering problems [18]. It clusters data in different manner than K-Means. It starts with an initialize estimate for variables and iterates to find the maximum likehood for these variables. The algorithm's inputs are data point, number of clusters and maximum no. of iterations. For each iteration, we execute the Expectation step, which determine the probability of points belong to each cluster. After that, we apply maximization step. The main steps of the EM algorithm are taken from Paul et al. [4].

The major problem in EM is that it is sensitive to the cluster center points selected initially, leading to production of different results based upon different values of initialization. We are using EM as the clusters obtained are more compact and far from other clusters. We initially estimate the cluster centers and number of clusters using some optimization techniques. After that, EM iterates to find optimized clusters.

2.2. Parameter Adaptive Harmony Search Algorithm

The concept of Harmony Search (HS) algorithm was first presented by Geem et al. [11]. It is a metaheuristic algorithm that imitates the music improvisation process where the musician improvise their instruments' pitch by searching the prefect state of harmony. It has been successfully applied in a wide variety of optimization problems such as timetabling, structure design, vehicle routing, sudoku puzzle solving, tour planning, etc.[11, 12].

To enrich the searching behaviour and to avoid being trapped in a local optimum, Kumar et al. [18] proposed a parameter adaptive harmony search (PAHS) algorithm. In PAHS algorithm, the two control parameter named as Harmony Memory Consideration Rate (HMCR) and Pitch Adjustment Rate (PAR), were being allowed to change dynamically. The computational procedure of PAHS algorithm can be summarized as follows [18, 19]:

Step 1. Initialization of the optimization problem and algorithm parameters: The optimization problem can be defined as Minimize (or Maximize) $f(x)$ such that $x_i \in [LB_i, UB_i], \quad i = 1, 2, ..., n$. Where $f(x)$ is the objective function, $x = (x_1, x_2, ..., x_n)$ is set of decision variables, n is the number of decision variables. LB_i and UB_i are the lower and upper bounds of decision variable x_i respectively. The parameters of the PAHS are harmony memory size (HMS), range of harmony memory consideration rate ($HMCR_{\min}$, $HMCR_{\max}$), range of pitch adjustment rate ($PAR_{\min}$, $PAR_{\max}$), range of distance bandwidth ($BW_{\min}$, $BW_{\max}$), and number of improvisation (NI).

Step 2. Initialization of Harmony Memory (HM): The HM consists of HMS harmony vectors. It is filled with randomly generated solution vectors and sorted by the values of the objective function $f(x)$.

Step 3. Improvisation of New Harmony: A new harmony vector $x' = (x'_1, x'_2, ..., x'_n)$ is generated using three rules: memory consideration, pitch adjustment and random selection as follows:

For each $i \in [1,n]$ do

$$\text{HMCR} = \text{HMCR}_{\min} + \frac{(\text{HMCR}_{\max} - \text{HMCR}_{\min})}{\text{NI}} \times \text{gn}$$

$$PAR = PAR_{\min} + \frac{(PAR_{\max} - PAR_{\min})}{NI} \times (NI - gn)$$

$$BW = BW_{max} \times e^{\left(\frac{ln(BW_{min}/BW_{max})}{NI} \times gn\right)}$$

if $U(0,1) \le HMCR$ **then** /* memory consideration */

begin

$x_i' = x_i^l$ **where** $l \square U(1,2,....,HMS)$

if $U(0,1) \le PAR$ **then** /* pitch adjustment */

begin

$x_i' = x_i' \pm BW \times Rand, \quad Rand \square U(0,1)$

endif

else /* random selection ***/**

$x_i' = LB_i + (UB_i - LB_i) \times Rand$

endif

done

Step 4. Updation in HM: If newly generated harmony vector $x' = (x_1', x_2', ..., x_n')$, evaluated in term of objective function value, is better than the worst harmony vector in HM, it is replaced with worst harmony vector. This is the step of algorithm where a decision should be taken whether the new harmony vector is to included in HM or not.

Step 5. Checking the termination criterion: If the maximum number of improvisation step is reached then computation is terminated and the algorithm returns the best harmony vector. Otherwise, Steps 3 and 4 are repeated.

3. Proposed Clustering Approach

The proposed cluster refinement procedure initializes Expectation Maximization using PAHS (PAHSEM). In EMPAHS, PAHS is executed on the dataset to provide the cluster centers. These are used to initialize cluster centers for EM algorithm.

In the proposed method, the clustering has two stages. At the first stage, the harmonies are initialized with random values. Each harmony vector is a sequence of real numbers representing the K cluster centers. For d dimensional space, the length of agent is $K \times d$. The first d positions represent the d dimensions of first cluster center, the next d positions represent the second cluster center and so on. The K cluster centers are encoded in the each harmony vector are initialized. The objective function is Euclidean distance, which must be minimum. The computation procedure of PAHS must do till predefined iteration.

In the second stage, the EM algorithm initialized with position of best harmony. The EM clustering algorithm refine the cluster centers.

4. Cluster Quality Metrics

4.1. Cluster Validity Indices

The three different validity indexes are used for finding the optimal number of clusters. These validity indexes are: Rand Index [6], Jaccard Index [7], and Mirkin Index [8]. Rand, and Jaccard indexes give largest

value for optimal number of clusters whereas Mirkin index gives smallest value, when number of clusters attains optimal value.

4.2. Cluster Quality Analysis

Weighted average, precision, recall, F-Measure and G-Measure [9] have been used for goodness/quality of clustering. Precision is a measure of the ability of system to present only relevant items. Recall is a measure of the ability of the system to present all the relevant items. Weighted average is the ratio of correctly claimed classes to the total number of classes.

5. Experimental Results and Discussions

5.1. Real -Life datasets

All the clustering techniques used in this paper have been tested over three real-life datasets of UCI database [20]. The real life datasets are: “Iris", “Wine” and “Haberman” datasets. Table I presents the details of these datasets.

TABLE I: UCI Datsets

Dataset	No. of Datapoints	No. of Features	No. of Classes
Iris	150	4	3
Wine	178	13	3
Haberman	306	3	2

5.2. Parameter setting for the algorithms

For Harmony Search, we use HMCR, BW and PAR is 0.85, 0.0005 and 0.5 respectively. For PAHS, the range of PAR and HMCR is set [0.2, 0.5] and [0.5, 0.85] respectively. The maximum number of iteration is 100. HMS for both algorithm is set to 25. For EM algorithm, we use Gaussian mixture model.

5.3. Experiment 1 and Results

First, we calculate the optimal number of clusters required for *Iris*, *Wine* and *Haberman* datasets. For this, we use four validity indexes for K-Means, FCM and EM clustering techniques. Figures 1(a-e) show the effect of varying number of clusters on validity indexes for iris dataset. The results reveal that the best values for all cluster validity indexes are achieved when number of cluster is 3.

From Figs. 1(a), 1(c) and 1(d), Rand, Jaccard and FM indexes attain largest value when the number of clusters is 3 for all above said methods. Mirkin index gives optimal number of clusters when index attains minimum value. We observed from Fig. 1(b) that Mirkin index having lowest value at number of clusters is 3. So, the optimal number of clusters for all above said techniques is 3 for Iris dataset.

Figs. 2 (a-e) show validity indexes for Wine dataset. Fig. 2(a) shows the Rand Index for wine dataset. It attains maximum value at number of clusters is 3. Similarly, we observed from Figs. 2(c) and 2(d) that, Jaccard and EM indexes having maximum value at number of clusters is 3. Mirkin Index shown in figure 2(b), which indicates the optimal number of clusters is 3.

For Haberman dataset, validity indexes are shown in Figs. 3(a-e). We observed from all figures that the optimal number of clusters is 2.

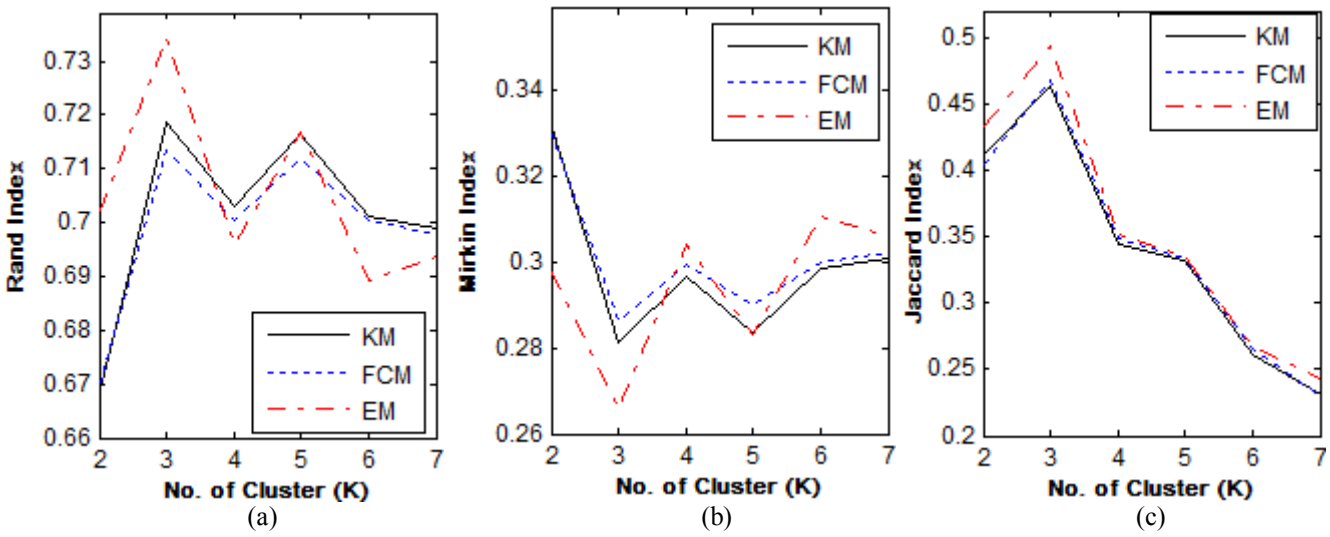

Fig. 1: Cluster Validity Indices for Iris dataset; (a) Rand (b) Mirkin (c) Jaccard.

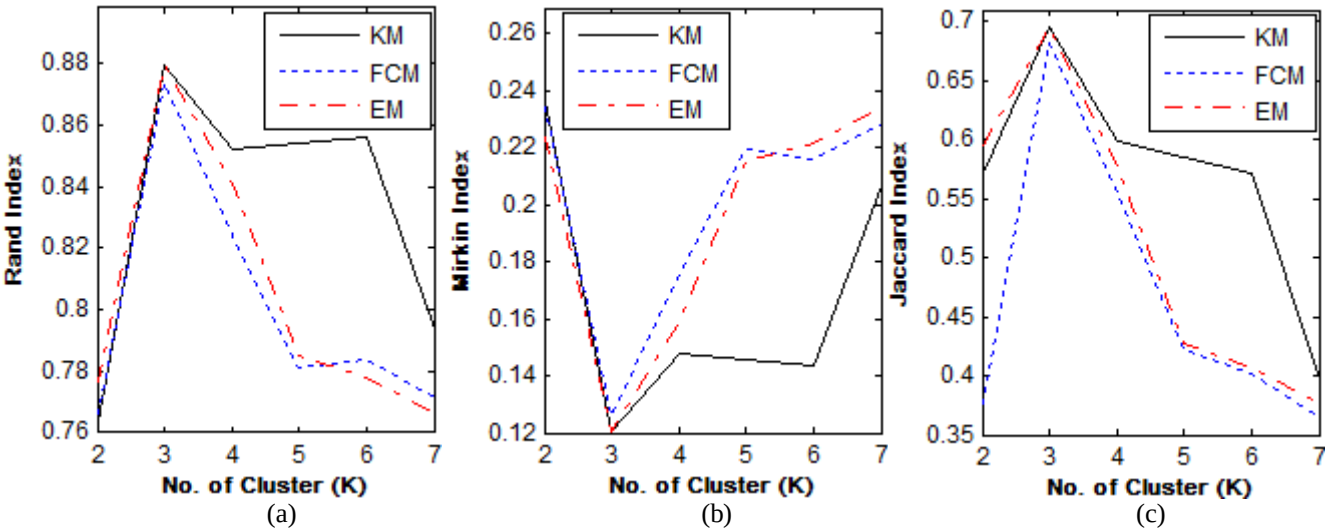

Fig. 2: Cluster Validity Indices for Wine dataset; (a) Rand (b) Mirkin (c) Jaccard.

5.4. Experiment 2 and Results

To evaluate the performance of Initialized EM based on Parameter Adaptive Harmony Search (PAHSEM) technique, we compared it with K-Means (KM), FCM, Expectation-Maximization (EM) and Initialized EM based on Harmony Search given by Geem (HSEM) Clustering techniques. These were applied on three datasets. These datasets also have a class label for classification purpose. We use optimal number of clusters for Iris, Wine and Haberman datasets. Tables 2, 3 and 4 show the comparison between proposed PAHSEM approach and above said techniques for iris, wine and haberman datasets respectively. The results reveal that PAHSEM outperforms K-Means, FCM, EM and HSEM clustering in terms of precision, recall, weighted average, F-Measure and G-measure.

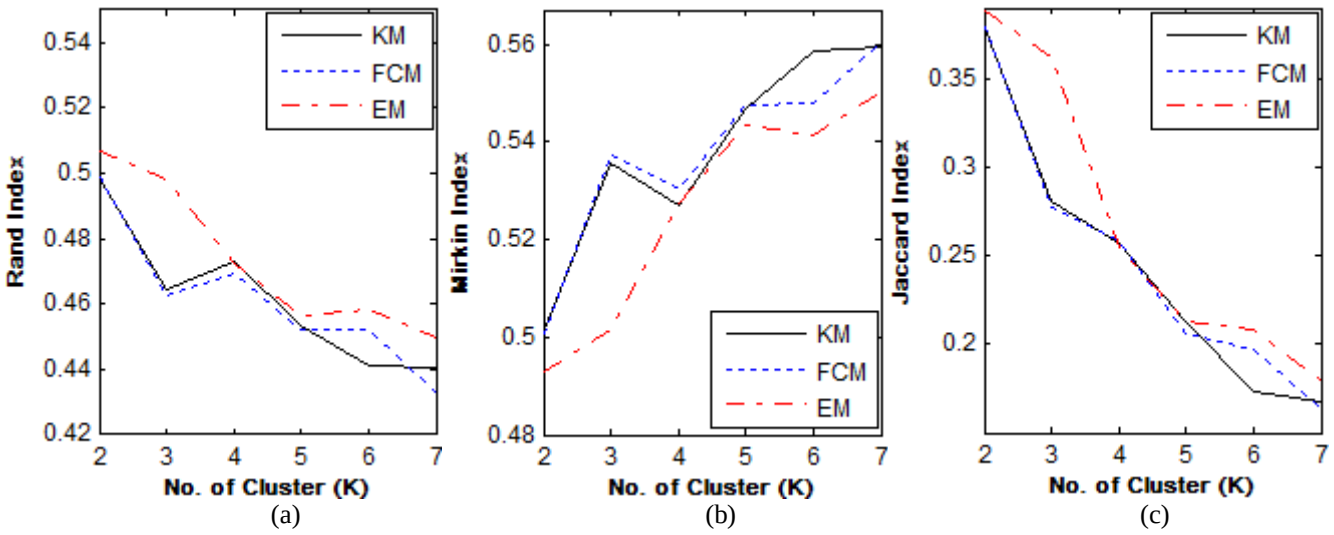

Fig. 3: Cluster Validity Indices for Haberman dataset; (a) Rand (b) Mirkin (c) Jaccard.

TABLE 2: Cluster Quality Metrics for Iris Dataset

	Precision	Recall	Weighted Average	F-Measure	G-Measure
K-Means	0.43181	0.44000	0.44000	0.11111	0.25596
FCM	0.54361	0.53330	0.53330	0.19277	0.35276
EM	0.90718	0.89333	0.89333	0.88525	0.89295
HSEM	0.92708	0.90666	0.90666	0.89970	0.90856
PAHSEM	**0.94276**	**0.94000**	**0.94000**	**0.93797**	**0.93969**

TABLE 3: Cluster Quality Metrics for Wine Dataset

	Precision	Recall	Weighted Average	F-Measure	G-Measure
K-Means	0.16872	0.21005	0.18539	0.04495	0.10710
FCM	0.17311	0.21574	0.19101	0.04549	0.10988
EM	0.31435	0.29866	0.33708	0.13966	0.19352
HSEM	0.94143	0.93507	0.93258	0.93473	0.93650
PAHSEM	**0.97087**	**0.97462**	**0.97191**	**0.97227**	**0.97251**

TABLE 4: Cluster Quality Metrics for Haberman Dataset

	Precision	Recall	Weighted Average	F-Measure	G-Measure
K-Means	0.49338	0.49161	0.51961	0.42619	0.45981
FCM	0.50038	0.50049	0.48039	0.43681	0.46927
EM	0.46652	0.46296	0.55882	0.35309	0.40538
HSEM	**0.63303**	**0.66667**	**0.66667**	**0.60893**	**0.63014**
PAHSEM	**0.63303**	**0.66667**	**0.66667**	**0.60893**	**0.63014**

6. Conclusions

In this paper, a new hybridized method based on the parameter adaptive harmony algorithm and Expectation-Maximization clustering method is proposed to cluster the dataset. In the proposed method, the parameter adaptive harmony algorithm is used to find the optimal cluster centers and then initialized the Expectation-Maximization clustering method with these cluster centers to refine the centers. The proposed algorithm is implemented and tested on three real life datasets. The optimal number of clusters has been calculated for three datasets using three different validity indexes. Experimental results demonstrated that the optimal number of clusters is three for Iris and Wine and two for Haberman datasets. On comparing the results of proposed technique with the others, it has been found that PAHSEM performs better than K-Means, FCM and HSEM Clustering techniques.

References

[1] A. K. Jain, M. N. Murty, and P. J. Flynn, "Data Clustering: A review," *Int. J. ACM Computing Surveys*, vol. 31, pp. 264-323, 1999

http://dx.doi.org/10.1145/331499.331504

[2] S. Guha, R. Rastogi, and K. Shim, "CURE: An Efficient Clustering Algorithm for Large Data Sets," in *Proc. ACM SIGMOD Conference*, 1998

[3] S. Guha, R. Rastogi, and K. Shim, "ROCK: A Robust Clustering Algorithm for Categorical Attributes," in *Proc. IEEE Conf. Data Engg.*, 1999

http://dx.doi.org/10.1109/ICDE.1999.754967

[4] P.S. Bradley, U. Fayyad, and C. Reina, "Efficient Probabilistic Data Clustering: Scaling to Large Databases," Microsoft Research Report, MSR-TR-98-35, pp. 1-25, 1999

[5] W. Abd-Almageed, A. El-Osery, and C. E. Smith, "Non-Parametric Expectation Maximization: A Learning Automata Approach," in *Proc. IEEE Conf. Systems, Man and Cybernetics*, 2003

http://dx.doi.org/10.1109/ICSMC.2003.1244347

[6] W. M. Rand, "Objective Criterion for Evolution of Clustering Methods,". *J. American Stat. Assoc.* vol. 66, pp. 846–850, 1971

http://dx.doi.org/10.1080/01621459.1971.10482356

[7] P. Jaccard, "The distribution of flora in the alpine zone," *J. New Physiologist.* vol. 11, pp. 37–50, 1912

http://dx.doi.org/10.1111/j.1469-8137.1912.tb05611.x

[8] B. G. Mirkin, and L. B. Cherny, "Deriving a distance between paritions of a finite set," *J. Auto. Remote Control*, vol. 31, pp. 91–98, 1970

[9] G. Kowalski, *Information Retrieval Systems- Theory and Implementation*, Kluwer Academic Publishers, 1997

[10] M. Filippone, F. Camastra, F. Masulli, and S. Rovetta, "A survey of kernel and spectral methods for clustering," *J. Pattern Recogn.* vol. 41, pp. 175–190, 2008

http://dx.doi.org/10.1016/j.patcog.2007.05.018

[11] Z. W. Geem, J. Kim, and G. V. Loganathan, "A New Heuristic Optimization Algorithm: Harmony Search," *J. Simulation*, vol. 76, pp. 60–68, 2001

http://dx.doi.org/10.1177/003754970107600201

[12] M. Mahdavi, M. Fesanghary, and E. Damangir, "An improved harmony search algorithm for solving optimization problems," *J. Appl. Math. Comput.*, vol. 188, pp. 1567–1579, 2007

http://dx.doi.org/10.1016/j.amc.2006.11.033

[13] S. B. Kotsiantis, and P. E. Pintelos, "Recent Advances in Clustering: A brief Survey," WSEAS Trans. Inform. Sci. Appl., pp-73-81, 2004

[14] M. Ester, H. P. Kriegel, J. Sander, and X. Xu, "A density-based algorithm for discovering clusters in large datasets with noise," *Proc. Int. Conf. on Knowledge Discovery and Data Mining*, Potland, pp. 226-231, 1996

[15] A. Hinneburg, and B. Keim, "An efficient approach to clustering in large multimedia datasets with noise," in *Proc. Int. Conf. on Knowledge Discovery and Data Mining*, Potland, pp. 58-65, 1998

[16] W. Wang, J. Yang, and R. Muntz, "STING: a Statistical Information Grid Approach to Spatial Data Mining," in *Proc. VLDB Conf.*, Greece, 1999

[17] R. Agarwal, J. Gehrke, D. Gunopulas, and P. Raghavan, "Automatic subspace clustering of high dimensional data for data mining applications," in *Proc. ACM SIGMOD Conf. on the Management of Data*, pp. 94-105, 1998

http://dx.doi.org/10.1145/276304.276314

[18] V. Kumar, J. K. Chhabra and D. Kumar, "Parameter adaptive harmony search for unimodal and multimodal optimization problems," *Journal of Computational Science*, vol. 5, no. 2, pp. 144-155, 2014.

http://dx.doi.org/10.1016/j.jocs.2013.12.001

[19] V. Kumar, J. K. Chhabra and D. Kumar, "Clustering Using Modified Harmony Search Algorithm," *Int. J. Comput. Intell. Studies*, vol. 3, no. 2/3, pp. 113-133, 2014.

http://dx.doi.org/10.1504/IJCISTUDIES.2014.062726

[20] C. L. Blake, and C. J. Merz, UCI Repository of Machine Learning, 1998

http:/www.ics.uci.edu/~mlearn/databases/.