ISSN No. 0976-5697

Volume 8, No. 3, March – April 2017

International Journal of Advanced Research in Computer Science

RESEARCH PAPER

Available Online at www.ijarcs.info

An Enhanced Bat Algorithm for Data Clustering Problems

Neeraj Dahiya
Department of Computer Science & Engineering
SRM University
Sonipat, Haryana

Surjeet Dalal
Department of Computer Science & Engineering
SRM University
Sonipat, Haryana

Savita Khatri
Department of Computer Science & Engineering
BMIET
Sonipat, Haryana

Abstract: Data Clustering in Data Mining is a domain which never gets out of focus. Clustering a date was always an easy task, but achieving the required accuracy, precision and performance have been never so easy. K means being an archaic clustering algorithm got tested and experimented thousands of times with a variety of datasets due to its robustness and simplicity but what this algorithm proposed was not suggested before. The proposed algorithm uses K means Algorithm for the Evaluation and Validation purposes whereas Optimization of the data is done by the help of Bat Algorithm. The drawbacks of K mean mainly its local convergence and initializing number of clusters at the early stage, which are still an issue has aroused the process of working on this algorithm. So for attaining the global convergence the Swarm Intelligence is preferred over Genetic Algorithm and many other techniques. For the latter one the algorithm combined two functions one of them help in knowing the number of clusters which are optimal for the particular dataset and the other one validates the results using another function and compares the various metrics which will define the goodness and fitness of the algorithm. In one line the complete overview of the proposed algorithm can be described, performing validation with the help of a numerical function of the k means and giving the final touch of Optimizing the data by k means bat algorithm'. The algorithm is tested for over 4 datasets available in UCI Repository and the results were expectedly great.

Keywords: Data mining · Data clustering · Optimization · Bat Algorithm Data Mining, K means Algorithm

I. INTRODUCTION

Clustering in Data Mining is an important field which has resolved many issues in various fields such as Pattern Recognition, Machine Learning, and Evolutionary strategy, Genetic Algorithms, Optimization Algorithms and Swarm Intelligence. The goal of clustering is to cluster the unlabeled data into groups of similar data point or objects. It helps in sorting the unsorted scattered data to a single location according to the similarity factor. The appropriate use of parameter setting such as Distance Function, Density Threshold and Numbers of Clusters expected will also affect the clustering of the dataset. The choice of these factors should be taken into consideration before clustering because otherwise they can lead to problems afterwards like noise, outliers, irregular number and location of clusters thus formed. Clustering algorithms can be divided into 3 categories. Some of them are as:

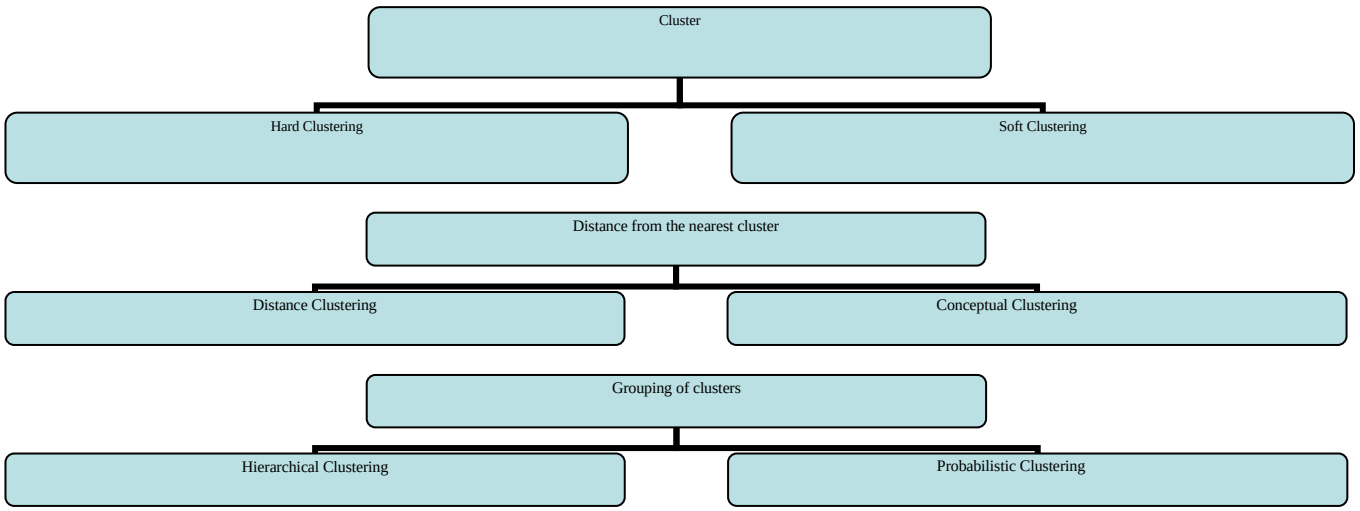


Figure 1

K Means is an oldest Centroid-based algorithm. It is an exploratory data analysis technique. It follows a non-hierarchical method of grouping objects. Being an NP- hard it always converges locally [1-3]. As it is not trivial, it cannot

find the global minimum solution for any problem. K means, according to the Macqueen, who introduced the algorithm said that K means is Clustering of the qualitative and quantitative insights into a large multivariate set rather than just Clustering. Lloyd proposed a filtering algorithm on the basis of the other used techniques which is also known as Lloyd's algorithm. The K means when performed for N dimensional data with k clusters and n objects for some iteration the results were changing throughout. Each time the new outcome appears with different value [4-6]. So, to overcome the basic issues of K means we deduced this algorithm which removed all the problems related to the number of iterations, speed, accuracy, time error and precision.

The Bat algorithm is an algorithm following the Swarm Intelligence in which some randomly generated particles converge globally by updating the particle's position and velocity. The idea of the amalgamation of these two algorithms is not new. It had almost been a decade. Many papers of combination of K means and Bat were introduced like [7], [9], and [10]. But the Bat is a globally converging algorithm sometimes got stuck in local minima. K means Bat and enhanced Bat K means have improved the performance, but the accuracy and the number of iterations were so high. Thus, to completely remove the issue we deduced the algorithm which will always provide a global optimum solution. Both the drawbacks of K means and Bat were taken into consideration and algorithm tried to reduce the time error complexity as well as number of iterations

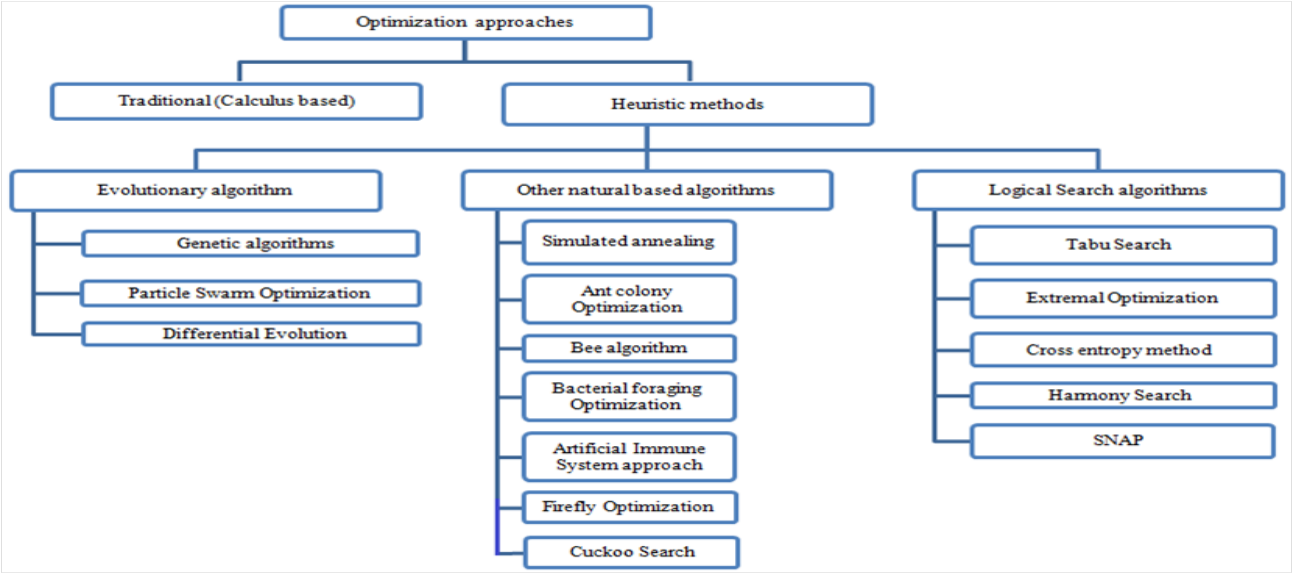


Figure 2

II. PROBLEM STATEMENT

Clustering by K means is still a preference, but the drawbacks, it carries with it cannot be completely removed. While observing a cancer data the number of attributes it attained were more than 8 that is a very large count. If the clusters are to be formed it will be creating a mess. The diagnosis of the type of cancer will be a long procedure. Through a survey, it found that more than 1.5Lac women are having breast cancer and 70% of them die due to inappropriate diagnosis and treatment. Total patient of cancer of all types is more than 10Lac in India. The problem created by these types of data sometimes may cost a life by improper diagnosis or treatment many clustering algorithms and their combination with other optimization algorithms was not that much capable of providing the consistent result in any number of iterations? Though we have not experimented on the cancer data but the proposed algorithm used wine data which also has more than 11 attributes and the results were expectedly good. Thus, the proposed algorithm tried to improve the level of clustering through k means using Bat algorithm [11-15].

The issues that k means faces and was not able to produce the expected results are its local converging property and defining clusters initially. Firstly, the local convergence of k means restricts its path of finding global minimum value. The global optimum value of an algorithm provides the minimum value which is near to the solution which gives the best results. There are many earlier papers which tried to achieve the global optimum with great accuracy but could not perform well. Secondly, the initialization of a number of clusters at the very first step of algorithm restricts it from being performed dynamically well. Thus, an algorithm proposed in this paper will overcome the both issues and provide better performance, accuracy, precision and also reduced time error complexity than the earlier suggested algorithms [16-17].

III. K MEANS AND ITS EXTENSIONS

The K means being an archaic algorithm is robust and preferred till date. Its easiness and handy behavior allowed it to be used in thousands of researches The centers that are defined initially are in each cluster and these centers are to be placed in a manner that different clusters different locations that is 'No two cluster centers overlap each other at any point'. Thus are placed too far from each other. It is an unsupervised iterative stochastic Clustering algorithm which is easy in implementation, uses less memory and has a better computational efficiency [18]. This paper [K means] has sped up the process and accuracy got improved due to which the overall performance of the algorithm got better. But the average function was still too large. Similarly, in paper [k means] a cost function defined and with various values of kit are tested. The cost function defined in the proposed algorithm with k number of clusters performed really well and produced accuracy and cost is far better than the earlier ones.

IV. BAT ALGORITHM VARIANTS

The bat set of rules (BA) became initially brought in, which has been related to benchmark functions, consequently BA performs BA has been correctly implemented to hard optimization problem consisting of motor wheel optimization hassle, clustering problem, in conjunction with famous engineering optimization duties. BA shows within the said literature on attracted the authors to pick this set of rules for attribute reduction assignment. Bats are animals that have wings and consist of the capability of echolocation:

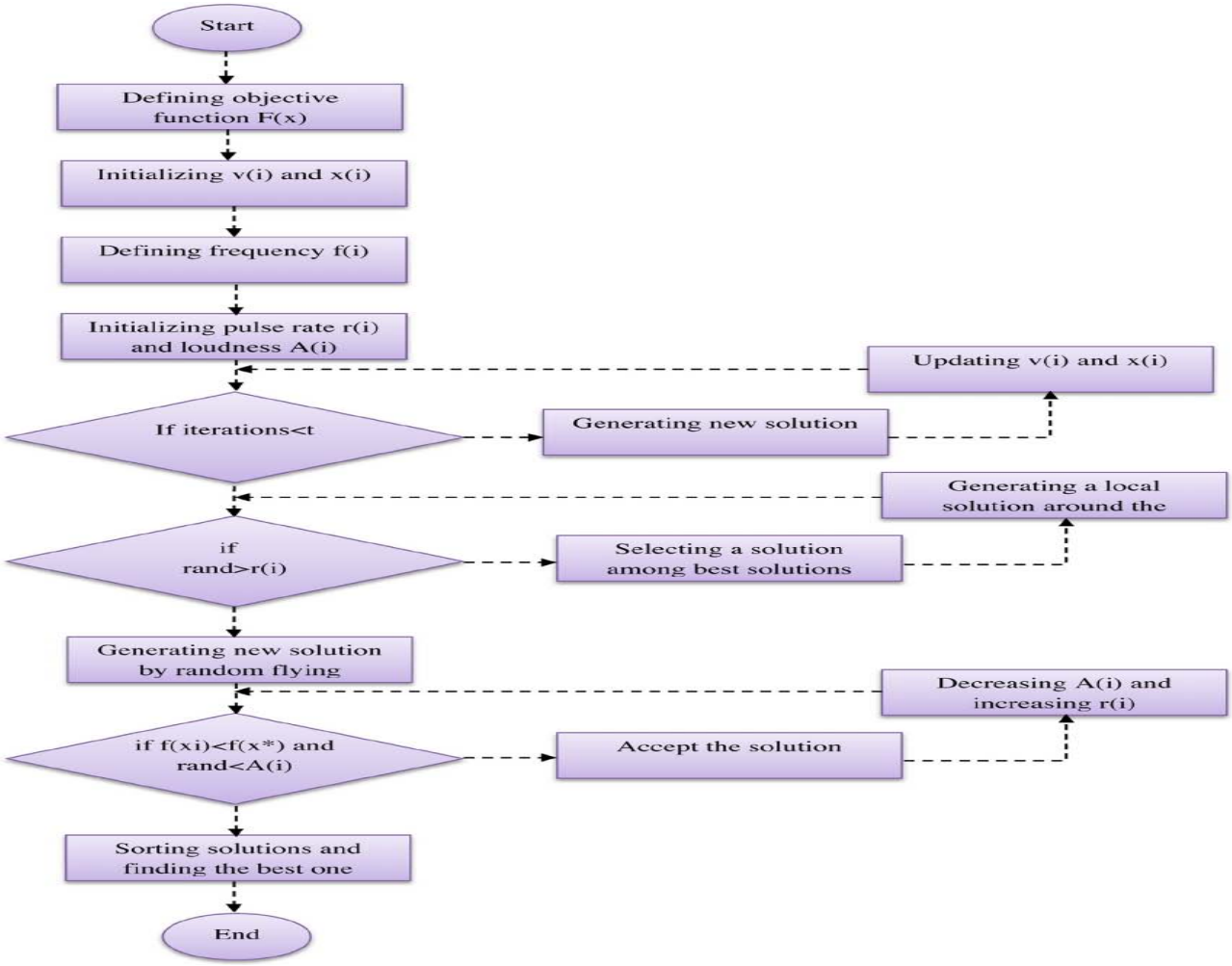


Figure 3

Advantage of Bat algorithm

- Simple, easy and flexible to implement
- Solve a wide range of problems and highly nonlinear problems efficiently.
- Give best solution in quick time.
- It gives promising results.
- Works well with complicated problems

- Fuzzy Logic Bat Algorithm (FLBA): Khan et al. (2011) presented a variant by introducing fuzzy logic into the bat algorithm; they called their variant fuzzy bat algorithm.
- Multi objective bat algorithm (MOBA): Yang (2011) extended BA to deal with multi objective optimization, which has demonstrated its effectiveness in solving a few design benchmarks in engineering.
- K-Means Bat Algorithm (KMBA): Komarasamy and Wahi (2012) presented a combination of K-means and bat algorithm (KMBA) for efficient clustering.
- Chaotic Bat Algorithm (CBA): Lin et al. (2012) presented a chaotic bat algorithm using L´evy flights and chaotic maps to carry out parameter estimation in dynamic biological systems.
- Binary bat algorithm (BBA): Nakamura et al. (2012) developed a discrete version of both algorithms to solve classifications and feature selection problems.
- Differential Operator and Levy flights Bat Algorithm (DLBA): Xie et al. (2013) presented a variant of bat algorithm using differential operator and L´evy flights to solve function optimization problems.
- Improved bat algorithm (IBA): Jamil et al. (2013) extended the bat algorithm with a good combination of L´evy flights and subtle variations of loud and pulse emission rates. They tested the IBA versus over 70 different test functions and proved to be very efficient

V. ENHANCED BAT ALGORITHM APPLICATION IN DATA CLUSTERING PROBLEMS

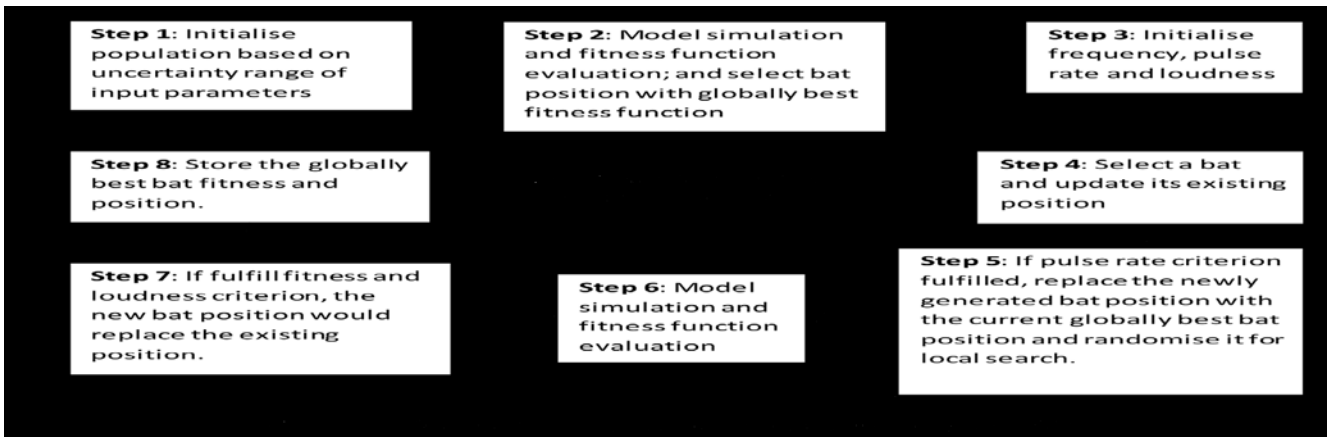
Methodologies Used: - The various algorithms and methodologies applied in the proposed algorithm are K means Clustering, Genetic Algorithm, Cluster analysis, Internal and External Evaluation indices and Bat algorithm along with clustering which are already discussed above. The formulation of the algorithm by using various functions and concepts is done.

Proposed Algorithm: - The proposed algorithm tries to focus on the basic drawbacks of k means which were an issue for a long time and even after experimenting it many other the algorithm was run on a dell laptop with ram 4gb and Matlab r2015a and over 4 datasets that were imported from the UCI machine learning repository for the evaluation purpose. It contains iris dataset with 5 classes and 150 instances in 4 dimensions, Portugese wine data with 2 types: red and white. The wine red having 1599 instances and wine white with 4898 instances along with more than 11 attributes each. At last bean small dataset with 47 instances and 35 attributes. The algorithm got implemented and results were recorded in table 1. It clearly shows the improved sum of the mean square error distance and the optimal number of clusters to be formed after running the function. The Optimal clusters can be cross verified with the cluster originally formed by the dataset used before running K means with Bat. In most of the cases the output was correct except for wine white dataset. The Sum of Mean Square distances got improved in all cases in all datasets with a little margin, but the values get upgraded even up to 4 places of decimals. The Accuracy defines how accurate the algorithm's output is and more the accuracy rate better are the outputs. In case of IRIS dataset has been the best of all and it has proved to be far better than the accuracy rate produced in the results of paper [19-23].

This algorithm consists of the following main components:

- ✓ Initialization,
- ✓ Variation in operation,
- ✓ Local search,
- ✓ Evaluation of a solution,
- ✓ Replacement.

Begin
Initialization. Set the generation counter $t = 1$; Initialize the population of NP bats P randomly and each bat corresponding to a potential solution to the given problems; define loudnes A_i, pulse frequency Q_i and the initial velocities v_i $(i = 1, 2, \ldots, NP)$; set pulse rate r_i.
While the termination criterion is not satisfied ***or*** $t < MaxGeneration$ ***do***
Generate new solutions by adjusting frequency, and updating velocities and location Solutions (2),
if (rand $> r_i$) **then**
Select a solution among the best solutions;
Generate a location solution around the selected best solution
endif
Generate a new solution by flying randomly
if (rand $< A_i$ && $f(x_i) < f(x_*)$)
Accept the new solution
Increase r_i and reduce A_i
endif
Rank the bats and the find the current best x_*
$t = t + 1$;
endwhile
Post-processing the results and visualization.
end.



The difference in the accuracy rate of both the paper differs from 7 points. The rate in the paper of IRIS data with highest performance was 88.30% and the accuracy rate in a proposed algorithm comes out to be 95%. The Accuracy rate of the other datasets has also increased with some difference in K means with Bat and in K means without Bat. The Precision is a parameter which denotes the exactness of the algorithm. It represents how useful the results are after the number of iterations being performed. Higher the precision value more precise and perfect the algorithm is for the task. Here the precision values of IRIS data got increased after the amalgamation of K means Bat instead it was not so good when K means tested without Bat the Recall parameter defines the sensitivity. It tells how sensitive the algorithm is to the irregularities if happen during the procedure. The completeness of an algorithm can be denoted by Recall. The better recall means the value should be near to 1. The more it is near to 1 better the value of Recall. The other best thing produced by the proposed algorithm is the reduced number of iterations drastically can be observed in K means with Bat in comparison with K mean without Bat. This reduction has helped in reducing the time error complexity of the algorithm. It has

helped in enhancing the performance of the algorithm by many times from the previously defined algorithms. At, last the optimal number of Centroid and the Best Cost obtained after running the simple Bat algorithm can be compared with the number of clusters obtained in the starting of the algorithm, which proves to be approximately equal in all cases except wine white. The Best Cost of all the datasets cannot be represented completely in a table format being irregular in number for each dataset the average of Best Cost can be seen here, for

- ✓ Bean Small Dataset: 20.34
- ✓ Wine Red Dataset: 2.62
- ✓ Wine, White Dataset: 2.38
- ✓ Iris Dataset: 1.66

Table 1

Parameters	Bean		Wine White		Wine Red		Iris	
	K Means Without	K Means	K Means Without Bat	K Means	K Means Without Bat	K Means	K Means Without Bat	K Means
Optimal k	4	5	1	8	5	5	4	4
No of Observations	49	46	4798	4998	1699	1599	160	190
Sum Of Distance	1.3395	1.3439	6.8888	7.308	14.5178	14.687	0.9394	0.3396
Accuracy	62.340	64.468	10.94	12.94	20.35	25.2	52.6667	86
Precision	33.478	35.454	0.1162	0	0	0	33.2105	89.285
Recall	1	1	0.25	0	0	0	0.54	1
No Of Iterations	3	5	81	42	37	29	14	5

VI. CONCLUSION

The proposed algorithm has accomplished the task of removing the issues of basic K means Algorithm and has produced the results which perform better than the earlier suggested algorithms in terms of Accuracy, Recall, and Precision and minimized the number of iterations with a great difference that will affect the time error complexity of the algorithm and reduced it. Firstly, the problem of initialization of a number of clusters at the starting is resolved by using the function which will produce the optimal number of clusters required for the particular data set. The function has helped in validating the predicted outputs with the actual ones which cross verifies the algorithm with the help of supervised learning and proves to be the best. Secondly the issue of local convergence, which got settled when the Bat algorithm was applied over the improved K means and the Consistency it achieved in various numbers of iterations on a single data set has made it perfect for large datasets. Thus, deduced algorithm has performed its best and the results were expectedly great. The Global convergence is achieved and proves to be consistent in any number of iterations without getting trapped in any local minima. The consistency, reduced time error complexity and improved accuracy rate is a major outcome of the proposed algorithm.

VII. REFERENCES

[1]. W. Du, B. Li, Multi-strategy ensemble particle swarm optimization for dynamic optimization, Information Sciences, 2008, 178, pp 3096–3109.

[2]. Anil.K.Jain, ”Data Clustering: 50 years beyond K means”, Elsevier, Pattern Recognition Letters, DOI: 10.1016/j.patrec 2009.09.011, 2009.

[3]. J Kennedy, RC Eberhart, “Particle Swarm Optimization”, Proceedings of the IEEE International Joint Conference on Neural Networks, Vol. 4, pp 1942–1948, 1995.

[4]. Bahaman Bahamani, Sergei Vissilvitskii, Ravi Kumar, Andrea Pattani,”Scalable K means++”, 38th International Conference on Very Large Database,2012.

[5]. Sarafrazi, H. P. Nezamabadi and S. Saryazdi, Disruption: A new operator in gravitational search algorithm, Scien-Tia Iranica D, Vol. 18, No. 3, pp. 539–548, 2011.

[6]. R. V. Rao, V. J. Savsani and D. P. Vakharia, Teaching–learning-based optimization: A novel method for constrained mechanical design optimization problems. Computer-aided Design, Vol. 43, No. 3, pp. 303-315, 2011.

[7]. A. Hatamlou, Black hole: A new heuristic optimization approach for data clustering. Information Sciences, Vol. 222, pp. 175-184, 2013.

[8]. K. Arun Prabha, N. Karthikayini. Visalakshi, “Improved Particle Swarm Optimization based K means Algorithm”, 978-1-4799-3966- 4/14, DOI 10.1109/ICICA.2014.21, 2014.

[9]. Clara Pizzuti, Nicolo Procorpio,” A K-means Based Genetic Algorithm for Data Clustering”, Springer, Advances in Intelligent Systems and Computing 527, DOI 10.1007/978-3-319-47364-2 21.

[10]. K. Premalatha, A.M. Natarajan,” A New Approach for Data Clustering Based on PSO with Local Search”, Computer and Information Science, Vol. 1, No.4, November, 2008.

[11]. David Arthur and Sergei Vassilvitskii, “K means++: Advantage of careful seeding”,

[12]. M. R. Ackerman, C. Lammersen, M. Martens, C. Raupach, C, Sohler and K. Swierkot, “Stream KM++: A Clustering Algorithm for data streams”, ALENEX, pages 173-187, 2010.

[13]. O. K. Erol and I. Eksin, A new optimization method: big bang–big crunch. Advances in Engineering Software, Vol. 37, No. 2, pp. 106-111, 2006.

[14]. A. Kaveh and S. Talatahari, A novel heuristic optimization method: charged system search. Acta Mechanica, vol. 213, No. 3-4, pp. 267-289, 2010.

[15]. A. Kaveh, AMAM. Share and M. Moslehi, Magnetic charged system search: a new meta-heuristic algorithm for optimization, Acta Mechanica, Vol. 224, No. 1, pp. 85-07, 2013.

[16]. S. C. Chu, P. W. Tsai and J. S. Pan, Cat swarm optimization, In PRICAI 2006: Trends in Artificial Intelligence, pp. 854-858, 2006.

[17]. T. Kanaugho, D. M. Mount, NS. Netanyahu, C. D. Piatko, R. Silverman and A.Y.Wu, "An Efficient K means Clustering Algorithm: Analysis and Implementation",

[18]. Krista Rizman Zalik, "An Efficient K means Clustering algorithm", Pattern Recognition Letters, July 2008, DOI: 10.1016/j.patrec.2008.02.014.

[19]. T. Kanaugho, D. M. Mount, NS. Netanyahu, C. D. Piatko, R. Silverman and A.Y.Wu," A Local search approximation algorithm for K mean Clustering", Computational Geometry, pages 28 (2-3), 89-112,2004.

[20]. S. Kalyani and K. S. Swarup, "Particle swarm optimization based K Means Clustering approach for security assessment in power systems", Expert systems with applications 38 (2011) 10839-10846.

[21]. DW, Van der Merwe, AP Engelbrecht, "Data Clustering Using Particle Swarm Optimization".

[22]. K. Krishna, M. N. Murthy, "Genetic K means Algorithm", IEEE Transaction system, man, Cybernetics, Vol. 29, No.3, June, 1999.

[23]. Long Tan, "Clustering K means Algorithm based on Improved PSO Algorithm", 2015 Fifth International Conference on Communication Systems and Network Technologies". ISBN:978-1-4799-1797- 6/IEEE DOI 10.1109/CSNT.2015.223.

