International Journal of Computational Intelligence and Informatics, Vol. 3: No. 1, April - June 2013

Enhanced Binary Small World Optimization Algorithm for High Dimensional Datasets

K Thangadurai
P.G. and Research Department of Computer Science, Government Arts College (Autonomous), Karur, India.
ktramprasad04@yahoo.com

N Kurinjivendhan
P.G. and Research Department of Computer Science, Government Arts College (Autonomous), Karur, India.
vendhancs1489@gmail.com

***Abstract*-Large scale databases with high dimensional datasets can be mined and used for making decisions which may be unknown information but effective and will be used in the related fields like bio-informatics, medical, business, etc. Clustering is an unsupervised method that creates group of objects or clusters such that objects in the same group are very similar and objects in different group are very distinct. It allows users to analyze data from many different dimensions and categorize it, and summarize the relationships. Technically, binary small world optimization algorithm (BSWOA) is newly implied technique. In this paper, we review and compare these clustering algorithms to identify their efficiency and differences among them.**

Keywords-Clustering, Knowledge discovery, Clustering methods, Dynamic clustering problems.

I. INTRODUCTION

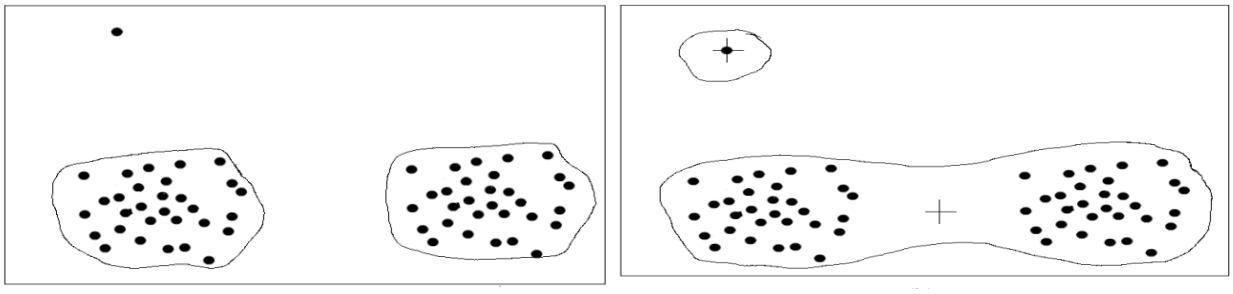
This paper Survey the publications with related to clustering related releases in this last 5 to 10 years. Clustering techniques used for knowledge discovery, understand the types of datasets and patterns. Cluster analysis is not only a specific algorithm, it is a general task and it can be achieved by various algorithms that they differ significantly what constitutes a cluster and how to efficiently find them [1]. Clustering methodology use some predefined criteria and find small distances among the cluster members, intervals or some statistical distributions.

Clustering can be a multi-objective optimization problem when used to denote the process of extracting appropriate knowledge from large data sets.

II. Existing Data Clustering Methods

A. K-means clustering

Clustering is the study of techniques for automatically finding classes without any prior knowledge. An effective Clustering algorithm may be any one of the following usually for forming the clusters. Prototype based partitioning method for clustering is K-means algorithm for low dimension category datasets. A set of n objects O = { o_1, o_2 … o_n} in a D dimensional metric space to be clustered into k number of sought clusters. In the initial step the cluster represented by C= (c_1,c_2,…c_k). In second step the algorithm randomly select k objects, in that each of them represents a cluster center or mean [2].



(a) (b)

Figure 1. Dead point problem (k-means)

New centroid of a group calculated by averaging the positions of the each point and each moving dimension position of each centroid. Random initialization of initial center leads to dead point problem in k-means algorithm. Initial centroids selection by systematic selection approach (ICSS) is used to solve the dead point problem [3].

In the Figure 1(b) the mean of each cluster is marked by "+". K-means algorithm is sensitive to outliers because an object with an extremely large value may substantially distort the distribution of data [14]. In the Figure.1(b) the mean of each cluster is marked by "+". K-means algorithm is sensitive to outliers because an object with an extremely large value may substantially distort the distribution of data [14].

B. Fuzzy K-means and c-means clustering

In [6] Fuzzy clustering methods allowing the objects placed in one cluster can be gradual members of two or more clusters or classes. In[5] it consists a set X with n objects along with described by m attributes as well as set C with k classes , $C=\{c_1,c_2,\ldots,c_k\}$ is the special key feature of c-means algorithm for each single object x in X.

C. K- Mediods clustering

Instead of taking mean value for cluster centroids as a reference point of the clusters, we can select one representative object per cluster. K-Mediods method is then performed based on the principle of minimizing the sum of the dissimilarities between objects and representative point of the class. The absolute-error criterion is defined as

$$E = \sum_{J=1}^{k} \sum_{p=C_j} | p = O_j | \quad (1)$$

E is the sum of the absolute error for all objects in the dataset. p is the point in space representing a given object in cluster Cj and oj is the representative object of cj.

D. Weighted fuzzy ant colony clustering

In[10] Distance With Connectivity model(DWC) estimates the distance between transactions with the local consistency and global connectivity information.(ACO)ant colony optimization is used for the data clustering process. Combined model of ACO and DWC is to find spherical shaped cluster. The main purpose of the Ant Colony Clustering algorithm with fuzzy are

i) Un-even data distribution handling
ii) Accurate distance measure
iii) Cluster accuracy

E. Fuzzy tabu search method

This method is a metaheuristic global optimization method in which the large combinatorial optimization tasks were performed. Tabu search problem solving depends on the given parameters are tabu list size and the selection of trial solutions. This method also may require iterative procedures for improving the efficiency, speed, accuracy. MEP (Maximum Entropy Principle framework) used for dynamic setting and minimizing.

III. PROPOSED BSWOA APPROACH FOR THE CLUSTERING PROBLEM

BSWOA is newly implied technique to simulate the small world phenomenon and originated by "The research on tracking the shortest path in American social networks" by Stanley Milligram, a social-psychologist [14].

A. Approach mechanism

This algorithm works by the two types of operator: Local searching operator Ψ, random Long-range searching operator Γ. The BSWOA method divides the given dataset into k number of classes. In [12] assumed that $X=\{X_1,X_2,\ldots,X_{n-1},X_n\}$is a dataset that contains *n* objects, X_{ij} represent the j_{th} attribute of the object X_i(*i*=1,2,…,*n* *j*=1,2,…,*d*). In Figure 2, the integer string $a_1, a_2, \ldots a_n$ on behalf of every clustering encoded in to binary string $b_1, b_2, \ldots b_{n-1}, b_n$ as a node in the solution space of BSWOA clustering algorithm [1].

B. Fitness function

The fitness function is the indicator of individual performance, for an essential survival of the fittest in BSWOA iteration. The same as is known, the objective of clustering is to make the distance between objects as undersized as possible, which is too better; In addition makes the distance between different clusters as large as possible. We could redefine the new fitness function as follow

Designate G_k as the k^{th} cluster, Z_k is the number of objects of cluster G_k. Finally the cluster center of the cluster G_k is specified, $C_k = (C_{k1},C_{k2},\ldots,C_{kd})$ must be obtained through the subsequent rule:

$$C_k = \sum_{X_n \in G_K} X_n / Z_k \quad (2)$$

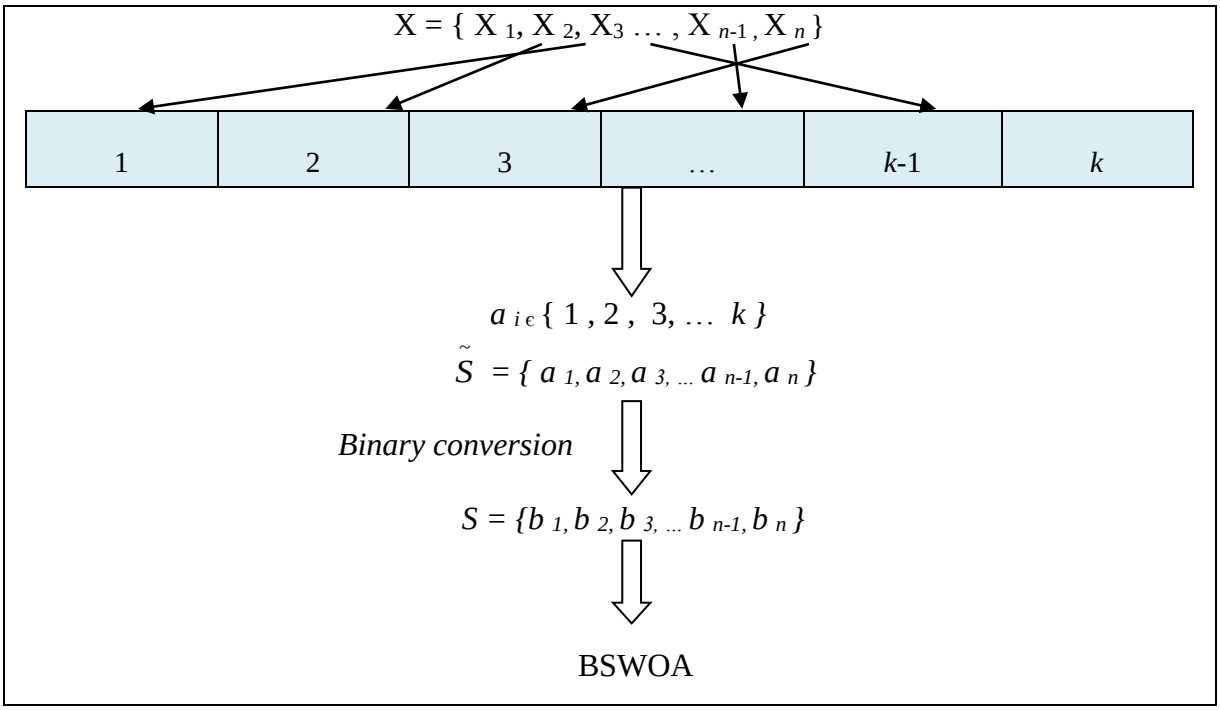


Figure 2: BSWOA binary conversion

Then the with-in cluster variation (WCV) of the k[th] cluster might be calculated by the subsequent rule:

$$WCV(k) = \sum_{X_n \in G_k} \sum_{i=1}^{d} (X_{ni} - C_{ki})^2 \quad (3)$$

C. Description of the algorithm

BSWOA to clustering problem firstly define the coding rules of clustering partition, after that randomly initialize the solution population, refresh solution population by means of integration and do not end until discovering the optimum solution or summit the termination state. The BSWOA algorithm works as follow

Step1: Construct initial node set S (0), subsequently distribute the given data objects into k clusters.
Step2: Calculate the fitness value of every node in node set S (0) subsequent to modification to get the optimal node S*(0).
Step3: Perform small world searching iteration until end criterion has been fulfilled.
Step 4: Initial stage iterations t=0.
Step 5: Initial parameter value i=0.
Step 6: Perform small world searching of the node set S(t).
Step 7: t=t+1.
Step 8: If t > t max, end procedure.

IV. PERFORMANCE ANALYSIS

In [4] the issues of numerous local minima, quantifying coverage cost and cluster resolution, spatio-temporal smoothening, achieving trade-offs between computational cost and resolution. DME, the general framework to identify and track cluster of moving objects successively. Since it require multiple iteration at each time step.

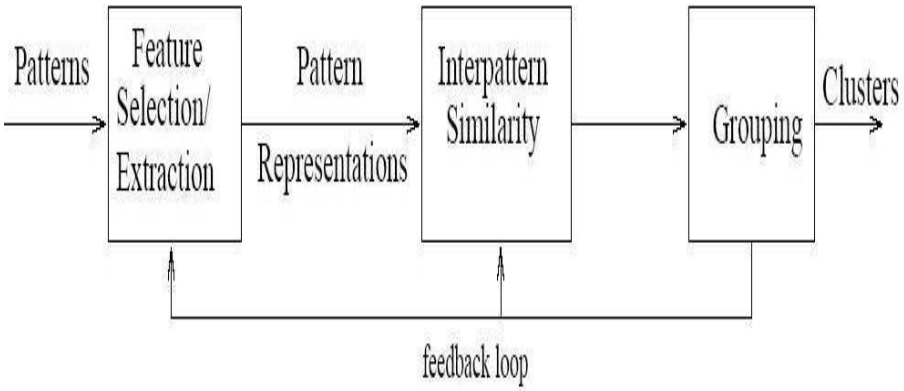


Figure 3. Clustering procedure with a feedback path

V. DYNAMIC CLUSTERING PROBLEMS

A. *Monitoring Cluster Centers*

Clustering-monitoring based on the following decision choices:

(i) Decision to split:

A decision split can be taken when the parameter β satisfies the condition $\beta^{-1}=2\ \lambda_{max=}(c_{x\ (t_c)}|y_{j\ (t_c)})$.

(ii) Decision to track cluster centers with or without increasing resolution:

At time t if $y_{i,} \leq j \leq M$ are at the cluster centers and the parameter β does not satisfy the splitting condition, then to improve coverage-resolution by increasing cooling rates eventually leading to the splitting condition being satisfied with higher resolution [4].

VI. EVALUATION OF CLUSTERING ALGORITHMS

The main criterion of clustering quality that is the purity and percentage of incorrectly clustered objects [13]. The results of the experiments are mentioned in Figures 4 and 5.

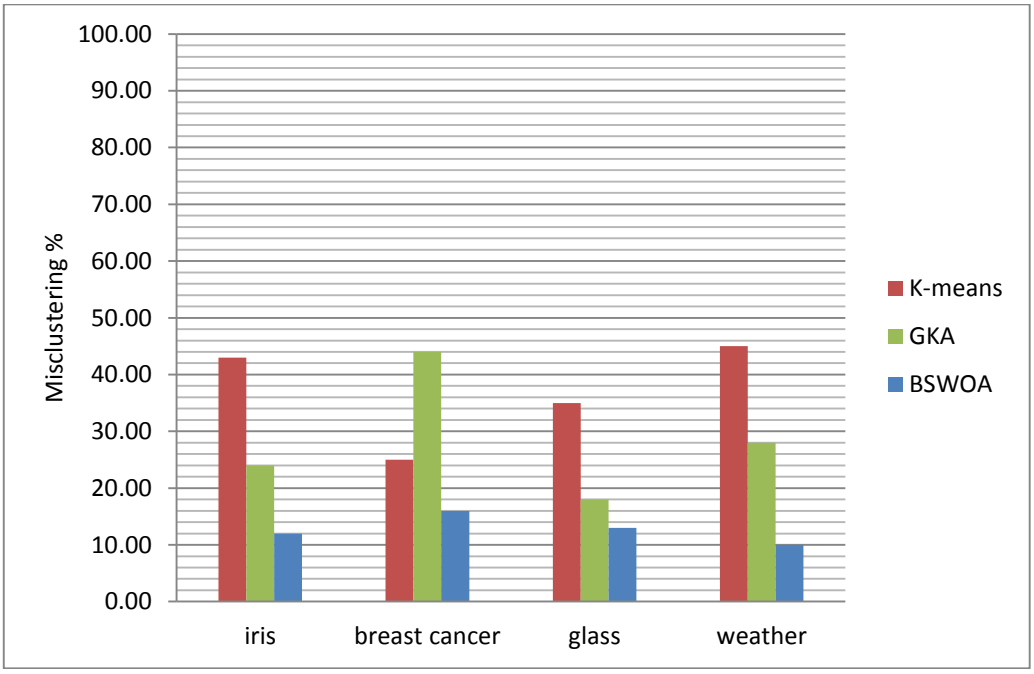


Figure 4. Evaluation of clustering algorithms on mini-newsgroup datasets with misclustering %

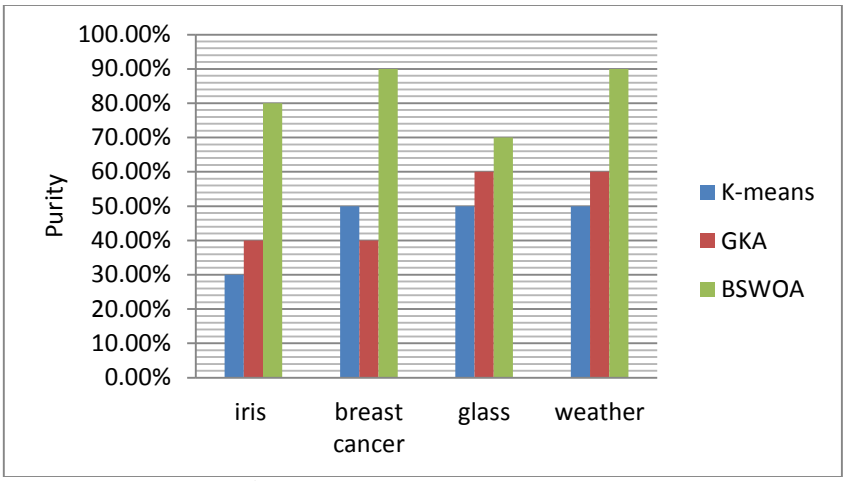


Figure 5. Evaluation of clustering algorithms on mini-newsgroup datasets with purity

VII. Conclusion And Future Work

In this paper, BSWOA algorithm is compared with various clustering algorithms especially with k-means, Genetic k-means to identify BSWOA's purity and low misclustering %. These algorithms have been implemented in MATLAB R 8A. All experiments with BSWOA, K-means and GKA were run in Windows 7 on laptop computer with Intel(R) Core(TM) i3, 2.27 GHz processors and 2GB RAM. The simplicity of clustering algorithms makes it the choice for many clustering tasks. The experimental results illustrate the advantages of the proposed BSWOA method which are (1) Purity (2) Fast convergence (3) Fast execution and (4) low misclustering percentage. This proposed BSWOA method needs to be improved to perform with less iteration that will be a next task in future.

Acknowledgement

We would like to thanks to the Principal, HOD and faculty members of P.G and Research Department of Computer Science and Research Scholars Government Arts College (Autonomous), Karur for their encouragement to publish this work.

References

[1] Shiwei Wu, Shaohong Yin and Min Li, "A New approach for clustering problem based on binary small world optimization algorithms", IEEE International Conference on Computer Science and Automation Engineering, Vol. 3, 2012, pp. 412 – 416.

[2] M. Aouf, L. Lyanage and S. Hansen, " Review of data mining clustering techniques to analyze data with high dimensionality as applied in gene expression data", IEEE International Conference on Service Systems and Service Management, July 2008, pp. 1- 5.

[3] Syed Sibte Raza Abidi, Jason ong, "A Data mining Strategy for Inductive Data Clustering: A Synergy between Self-Organizing Neural Networks and K-means clustering techniques", IEEE TENCON Proceedings, Vol. 2, 2000, pp. 568 – 573.

[4] Puneet Sharma, Srinivasa M. Salapaka, Carolyn L. Beck, "Entropy-based framework for dynamic coverage and clustering problems", IEEE Transcation on Automatic Control, Vol. 57, Issue 1, Jan 2012, pp.135 – 150.

[5] Parsons, L., E. Haque, and H. Liu, "Subspace Clustering for High Dimensional Data: A review", CEINT, Vol. 6, Issue 1, 2004, pp.90 – 105.

[6] Yilei Wang Hui pan and Tao Li, "The data mining of the E-government on the basis on fuzzy logic", IEEE International Conference on Integration Technology, 2007, pp.774-777.

[7] Md. Zahidul Islam and Ljiljana Brankovic, "DETECTIVE: A decision tree based categorical value clustering and perturbation technique for preserving privacy in data mining", IEEE International Conference on Industrial Informatics, 2005, pp.700-707.

[8] Li Yanhong, Liu Peng, Qin Zheng, "An analysis model of financial statements based on data mining", IEEE International conference on Intelligent Systems, 2006, pp.847-850.

[9] S. Nithya and R. Manavalan, "Breast Cancer Data Analysis Using Weighted Fuzzy Ant Colony Clustering", IJCII Vol. 1, No. 3, Dec 2012.

[10] R.Barani and M.Sumathi, "Cluster based evaluation of image fusion algorithms", IJCII, Vol. 1, No. 4, March 2012.

[11] J. Han, M. Kamber, "Data Mining: Concepts and Techniques", 2nd edition, Morgan Kaufmann Publishers, March 2006.

[12] M. Hall, E. Frank, G. Holmes, et.al., "The WEKA Data Mining Software: An Update", SIGKDD Explorations, Vol 11, Issue 1. 2009, pp. 10-18.

[13] T. Jeffrey and S. Milligram, "An experimental study of the small world problem", Sociometry, Vol. 32, No. 4, pp. 425–443.

[14] K.Thangadurai, M.Uma, M.Punithavalli, "A study on rough clustering", Global journal of computer science and technology, Vol. 10, Issue 5, July 2010, pp 55-58.