Performance Analysis of BAT K-Means Clustering Algorithm Using Gene Expression Data set

P. Neelavathi,
Department of Computer Science,
Periyar University,
Salem-636011
neelapalani15@gmail.com

K. Thangavel,
Department of Computer Science,
Periyar University,
Salem-636011
drktvelu@yahoo.com

E.N.Sathish Kumar,
Department of Computer Science,
Periyar University,
Salem 636011
en.sathishkumar@yahoo.in

***Abstract*-- Clustering is one of the popular data mining methods aiming at representing large dataset by a collection of cluster. Clustering gene expression data sets is a difficult task because of seed selection. Hence to group the gene expression data set we proposed hybrid Bat K-Means algorithm (BKM). The proposed system have been utilizes the concepts of Bat optimization techniques and K-Means Algorithms. In this approach initial cluster centre are computed by using Bat optimization technique which gives nearly the exact centre of the cluster based on the echolocation behaviour of Bat location and velocity. This method is tested upon the "cell-free RNA across pregnancy time course from pregnant women" gene expression data set. The performance of the BKM is evaluated using confusion matrix validation and compared. Further, experimental studies show that BKM based clustering outperforms.**

Keywords - Gene Expression, Microarray Dataset, Bat algorithm, K-Means clustering, Bat K-Means clustering algorithm.

I. INTRODUCTION

Mainly Data mining algorithms for microarray gene expression data deal by the difficulty of clustering. Cluster study of gene expression data have to be a useful tool for identifying co-expressed genes. DNA microarrays are emerged as the important technology to evaluate gene expression levels primarily, because of their high throughput. Results starting this test generally exist in the form of a data matrix in which rows represent genes and columns represent conditions or samples. Each entry in the matrix is a measure of the expression level of a particular gene in an exact condition. Analysis of this data set reveals genes of unknown functions and the discovery of functional relationships between genes. Co-expressed genes can able to be grouped into clusters based on their expression patterns of gene. Clustering can be performed based on genes and sample. Into gene based clustering, the genes are treated as objects and the samples are features. within sample based clustering, the sample can be partitioned into homogeneous groups anywhere the genes be consider as features among the samples as objects [1, 2, 3, 4, 5].

Optimization algorithms have expansively been applied on diverse and numerous area over the last few decades. Various natural-world optimization problems are especially complicated, where better optimization algorithms are needed. The aim of an optimization algorithm is to locate a set of values for the parameter, i.e., the independent variables to maximizes or minimizes the importance of one or more dependent variables. Various complex optimization problems cannot be solved within enclosed computation time. Thus the algorithms capable of finding near-optimal or at least practically good solution within logical computation time include drawn the attention of the scientific community. During the last a small number of decades, the scientific community have observed the emergence of a number of nature-inspired optimization algorithms. Swarm Intelligence (SI) be subfield of Computational Intelligence to is dedicated to mimic the behavior of natural swarms to locate solution for complex optimization problems which be not easily tackled by other approaches. Swarm intelligence (SI) is base on collective performance of self-organized systems. Typical swarm intelligence algorithms consist of Particle Swarm Optimization (PSO). Recently some new swarm based algorithm namely Bat Algorithm (BA) emerges in the field of optimization algorithms. The algorithm proposed in this paper is based on the Bat optimization algorithm, which established on K-Means clustering [8, 9, and 10].

The structure of this paper be as follows. Sections II outline the fundamentals of the bat algorithm and K-Means algorithm, while Sections III describes the confusion matrix validation and its measures. Section IV summarizes the data set, and Section V describes the experimental results. Finally, Section VI concludes briefly.

II. METHODOLOGY

A. Bat Algorithm

Bat Algorithm is a meta-heuristic approach and it is working based on echolocation behavior of bats. The bat has the capability to locate its prey in complete darkness. This algorithm is developed based on this hunting

behavior of bats. Bats are mammals by wings. Bats are natural with the advanced capability of echolocation. Microbats are insectivores. Echolocation is an individual type of sonar, used by the microbats to avoid obstacles, detect prey, and pinpoint their roosting crevices into the dark. Bats release a high sound frequency to listen the echo that bounces back from the neighboring objects. Bats radiate frequency differs in qualifications. The frequency is associated with their food gathering strategies. Bats use small frequency-modulated signals to sweep during about an octave. Signal bandwidth of bats varies depends on the species. The echolocation characteristics of microbats emphasize some approximate or idealized rules, by which the difference of Bat Algorithm may be developed [8, 9, and 10]. Pseudo code for Bat algorithm is described in Algorithm 1.

Algorithm 1: Bat Algorithm

1. Objective function $f(x), x = (x_1, \ldots, x_d)$
2. Initialize the bat population $x_i = 1,2, \ldots . n$ and v_i
3. Set the pulse frequency f_i at x_i
4. Initialize the rates r_i and the loudness A_i
5. While (t < Maximum number of iterations)
6. Generate latest solutions by updating the pulse frequency, Velocities and locations
7. If $(rand > r_i)$
8. Select a solution among the best solutions
9. Generate a local solution among the select best solution
10. End if
11. Generate a new solution by flying randomly
12. If$\left(rand < A_i \; and \; f(x_i) < f(x^*)\right)$
13. Accept the new solutions
14. Increase x_i and reduce $A_i$values
15. End if
16. Rank the bats and find the current bestx^*
17. End while
18. Position process results and visualization

B. K-Means Clustering

One of the mainly popular clustering methods is K-Means clustering algorithm. It generate K objects as initial centroid arbitrarily, where K is a user specified parameter. Each point be then assigned to the cluster by the closest centroid. After that the centroid of each cluster is updated by taking the mean of the data points of each cluster. Various data points may move from one cluster to other cluster. Again we calculate new centroid and assign the data points to the suitable clusters. We repeat the assignment and update the centroid, until convergence criteria is met i.e., no point changes clusters, or equivalently, until the centroid remain the same. In this algorithm mostly Euclidean distance is used to find distance between data points and centroid [6, 11, and 12]. Pseudo code for the K-means clustering algorithm is described in Algorithm 2.

Algorithm 2: K-Means Clustering Algorithm

Input: D = {d_1, d_2, d_3, .., d_n } // Set of n data points.
K - Number of desired clusters
Output: A set of K clusters.
Steps:
1. Arbitrarily choose k data points from D as initial centroids;
2. Repeat
Assign each point di to the cluster which has the closest centroid;
Calculate the new mean for each cluster;
Until convergence criteria is met.

A. Confusion Matrix

A confusion matrix contains information on actual and predicted classifications through by a classification system. Presentation of such systems is normally evaluated using the data in the matrix.

TABLE I. CONFUSION MATRIX FOR A TWO-CLASS PROBLEM

	Positive prediction	**Negative prediction**
Positive class	True positive (TP)	False negative (FN)
Negative class	False positive (FP)	True negative (TN)

Table I shows a confusion matrix for a two-class difficulty by positive and negative class value. Since such a matrix, it is possible to extract a number of widely used metrics to measure the performance of a classifier.

Accuracy:

Overall accuracy is the proportion of the total number of predictions that were correct, it is defined as in the equation.

$$Accuracy = \frac{TP + TN}{TP + FN + TN + FP} \tag{1}$$

It is possible to derive five performance metrics as of Table 1 to measure the classification performance on the positive and negative classes independently.

Sensitivity or Recall (True positive rate):

It is the proportion of positive cases that were correctly identified, as calculated using the equation:

$$Recall = \frac{TP}{TP + FN} \tag{2}$$

Specificity (True negative rate):

It is defined as the proportion of negatives cases that were classified correctly, as calculated using the equation:

$$Specificity = \frac{TN}{TN + FP} \tag{3}$$

False Positive Rate (FPR):

The false positive rate (FPR) is the amount of negatives cases to be incorrectly classified as positive, as calculated using the equation:

$$\mathrm{FPR} = \frac{\mathrm{FP}}{\mathrm{FP} + \mathrm{TN}} \tag{4}$$

Precision:

Precision is the proportion of the predicted positive cases to be correct, while calculated using the equation:

$$Precision = \frac{TP}{TP + FP} \tag{5}$$

F-measure:

A measure that combines precision and recollect is the harmonic mean of precision and recall. The traditional F-measure or balanced F-score is defined since in the following equation,

$$Fmeasure = 2 * \left(\frac{Precision * Recall}{Precision + Recall}\right) \quad (6)$$

The advantages of the above five performance measures are of individual independent of class costs and a priori probabilities. The aim of a classifier is to minimize false positive and negative rates, or differently to maximize true negative and positive rates.

B. Data of cell-free RNA across pregnancy

Analysis of cell-free plasma from pregnant women data set, during the first, second, third trimesters and immediately post-partum. This is downloaded as of Gene Expression Omnibus (http://www.ncbi.nlm.nih.gov/geo/) [7]. In this Dataset, the total number of genes is: 33297, with sample 48. Results provide insight into a noninvasive means to monitor the expression status of many tissues and measure temporal expression of genes longitudinally during development. Suffix and prefix should be used more properly.

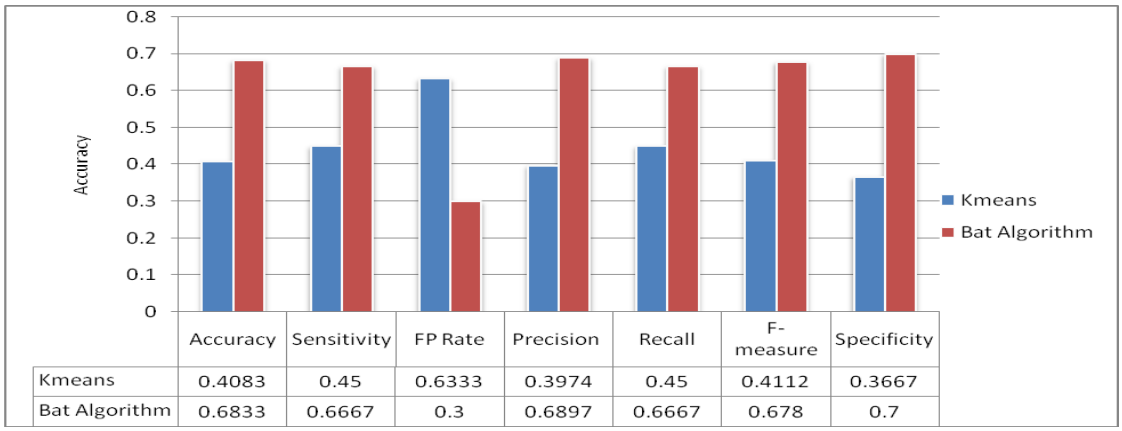
IV. Experimental Analysis And Results

In this section, the performance of the BKM is evaluated by using confusion matrix and their measures such as Accuracy, Sensitivity, False Positive Rate, Precision, Recall, F-Measure, and Specificity. Experiments are conducted to evaluate the robustness of the BKM algorithm for initial centroid selection of the cluster. The comparison results of proposed with KM and BKM algorithms are tabulated in Table I.

The experimental results show that the highest correctly classified accuracy 68.33% by BATKM and the lowest correctly classified accuracy 58.33% by K-Means fifth run. The Minimum false positive rate is decreased to 0.3% in BKM and highest false positive rate 0.8% is recognized from K-Means third, fifth and sixth run. The average false positive rate of K-Means is obtained as 0.633%. Experimental results shows that BKM based clustering outperformed compared with the existing K-Means clustering. Figure 1 illustrates the accuracy versus datasets comparison results. Based on the initial centroid of each cluster the BKMA gives better accuracy for all the datasets, but KM and BATKM gives less accuracy.

TABLE II. EXPERIMENTAL RESULT OF EACH CLASSIFIER USING PREGNANCY DATASET

S. No	Runs	Accuracy	Sensitivity	FP Rate	Precision	Recall	F-measure	Specificity
1	**KM_R1**	0.3	0.3	0.7	0.3	0.3	0.3	0.3
2	**KM_R2**	0.7	0.7	0.3	0.7	0.7	0.7	0.7
3	**KM_R3**	0.2	0.2	0.8	0.2	0.2	0.2	0.2
4	**KM_R4**	0.45	0.3	0.4	0.42	0.3	0.35	0.6
5	**KM_R5**	0.2	0.2	0.8	0.2	0.2	0.2	0.2
6	**KM_R6**	0.6	1	0.8	0.5556	1	0.714	0.2
7	**KM_AVG**	0.4083	0.45	0.6333	0.3974	0.45	0.4112	0.3667
8	**BKM**	0.6833	0.6667	0.3	0.6897	0.6667	0.678	0.7



	Accuracy	Sensitivity	FP Rate	Precision	Recall	F-measure	Specificity
Kmeans	0.4083	0.45	0.6333	0.3974	0.45	0.4112	0.3667
Bat Algorithm	0.6833	0.6667	0.3	0.6897	0.6667	0.678	0.7

Figure 1. Evaluation Accuracy of Existing and Proposed Method

V. CONCLUSION

In this paper, the performance of standard and proposed versions of the Bat K-Means clustering algorithm are investigated and compares their performance on the constant optimization problems. The basic Bat K-Means clustering algorithm gained improved results in optimizing the lower-dimensional function but shows poor performance on the higher dimensional multimodal function. In order to improve the performance on high dimensional problems, we have proposed a simple Bat K-Means clustering algorithm method for selection of initial centroid to achieve good quality of clusters. The performance of the K-Means and Bat K-Means algorithms have been compared using Gene Expression data and observed that the proposed Bat K-Means performed well.

ACKNOWLEDGMENT

The authors immensely acknowledge the UGC, New Delhi for partial financial assistance under UGC-SAP (DRS) Grant No. F.3-50/2011.

REFERENCES

[1] Jain A. K., Murthy M. N., Flynn P. J., "Data clustering: A Review", ACM Computing Surveys, pp. 265-323, Vol. 31(3), 1999.

[2] Jinyan Li., Limson Wong., Qiang Yang., "Data Mining in Bioinformatics", IEEE Computer Society, pp. 16-18, 2005.

[3] Ruchi Singh., Richa Sharma., "Bioinformatics: Basics, Algorithms and Applications", University Press, 2010.

[4] W. Frawley, G. Piatetsky-Shapiro, C. Matheus, "Knowledge discovery in databases: an overview", AI, Magazine, 213-228, 1992.

[5] Shital A Raut and S.R. Sathe "A modified Fast-map K-means clustering algorithms for Large Scale Gene Expression Datasets", International Journal of Bioscience, Bioinformatics vol.1, No- 4, November 2011.

[6] Kohei Arai and Ali Ridho Barakbah, "Hierarchical K-means algorithms for the centred initialization for K-means",Report of faculty of Science and Engineering, Saga University volume 36, No-1, 2007.

[7] Gene Expression Omnibus. http://www.ncbi.nlm.nih.gov/geo/

[8] Monica Sood and Shilpi Bansal "K-Medoids Clustering Technique using Bat Algorithm", International Journal of Applied Information Systems (IJAIS), 5(8):20-22, 2013.

[9] Taher N, Babak A, "An efficient hybrid approach based on PSO, ACO and k-means for cluster analysis", Appl Soft Comput 10(1):183-197. 2010.

[10] Xin-She Yang, "A new metaheuristic Bat inspired Algorithm. Studies in Computational Intelligence", Springer, 2010.

[11] E.N.Sathishkumar, K.Thangavel and T.Chandrasekhar, "Hybrid KMean-QuickReduct Algorithm for Gene Selection," IEEE International Conference On Research And Development Prospects On Engineering And Technology -2013 (ICRDPET-2013), on 29th-30th, March 2013.

[12] E.N.Sathishkumar, K.Thangavel and D. Arul Pon Daniel, "Efficacious Clustering Algorithm for Gas Sensor Array Drift Dataset," International Journal of Computational Intelligence and Informatics. Vol.3, No.3, ISSN:2349-6363, 2014.