International Journal of Trend in Research and Development, Volume 3(1), ISSN: 2394-9333
www.ijtrd.com

An Overview of Particle Swarm Optimization And Bat Algorithm For Data Clustering

[1]Gunjan Dashora, [2]Payal Awwal
[1] Student ,[2]Assistant Professor,
[1]Computer Science Department,
[1]Govt. Women Engineering College, Ajmer, India

***Abstract*—**Data clustering can be considered as an essential research topic in the field of data mining. Clustering is the field of grouping similar data samples together in some way , according to some criteria. It is the process of standardizing data into meaningful groups, called clusters. A new paradigm Swarm Intelligence is a collective behaviour of social systems like insects such as ants(ant colony optimization), Fish Schooling, Particle swarm optimization, Bat algorithm etc. Advanced study have show that partitional clustering algorithms are more convenient for clustering large datasets. The most commonly used partitional clustering algorithm is K-means due to easily implementation and most efficient in terms of execution time. The drawback of K-means is that it is sensitive to the selection of initial partition and may converge to local optima. This paper consider the de-merits of standard K-means algorithm for data clustering and a comparative analysis of PSO and Bat algorithm is shown with some of their advantages.

Keywords— *Particle Swarm Optimization, K-Means, Bat algorithm, swarm intelligence.*

I. INTRODUCTION

Data Mining is a important field, which includes area like database technology, pattern recognition, neural networks, knowledge processing, high performance computing, artificial intelligence etc. Clustering is an important technology used in data mining to find patterns of unknown dataset. In clustering, groups have similar objects. However, clustering technique suffers from many shortcoming like not work well in scaling with large size of data, finding arbitrary shapes of clusters or dealing effectively with the presence of noise. In clustering, entities are partition into groups based on the feature of each entity. The well formed, separated and homogenous group is called cluster. The two types of clustering algorithms are discriminative and generative. The discriminative algorithms, consider a pair wise similarities between every document and based on these similarities, the objective function is used to produce an optimal clustering. On, the other hand, the generative algorithms consider an underlying distribution of data and cluster centroids are produced by maximize the fit of distribution.

The process of grouping of data into number of clusters are known as data clustering. The aim of data clustering is to keep the objects in a single cluster are similar to each other but they are different from objects in other clusters.[1].

K-means is the most popular and widely used method for clustering. In k-means , we use Euclidean distance for good clustering results.[2].However k-means contains several drawbacks like trapped into local optima and local minima and it is also sensitive to initial cluster centers.[3][4].

A new prototype Swarm Intelligence(SI) is being used in research settings to improve the management and control of large numbers of collaborating entities like computer and sensor networks, communication, satellite constellations and many more. It is a collective behavior of insects. Particle swarm optimization and bat algorithm are used to overcome the problem of local optima in k-means algorithm. The PSO works as the behavior of birds in nature and Bat algorithm works as the echolocation behavior of bats in nature.

II.DATA CLUSTERING USING K-MEANS

Semi structured or unstructured datasets are classify with the help of k-means clustering. K-means clustering is simple and it has the ability to handle voluminous datasets. Therefore, this is the one of the most common and effective method to classify data.

The parameter used in k-mean clustering is the number of clusters and the initial set of centroids. The distance of each item in the dataset is calculated with each of the centroids of the respective cluster. The item is then assigned to the cluster with which the distance of the item is least. The centroid of the cluster to which the item was assigned is recalculated.

The standard k- means algorithm is as follows-

Initial positions of K cluster centers are determined randomly. Following phases are repeated:

a)For each data vector : the vector is allocated to a cluster which its Euclidean distance from its center is less than the other cluster centers. The distance to cluster center is calculated by Eq. (1):

$$Dis(X_p, Z_j) = \sqrt{\sum_{i=1}^{p}(X_{pi} - Z_{ji})^2} \quad (1)$$

In Eq (1), X_p is p^{th} data vector, Z_j is j^{th} cluster center and D is the dimension of data and cluster center.

b) Cluster center are updated by Eq (2):

$$Z_j = \frac{1}{n_j}\left[\sum_{\forall Xp \in Cj} Xp\right] \quad (2)$$

In Eq.(2), n_j is the number of data vectors corresponding to j^{th} cluster and C_j is a subset of the total data vectors which constitute j^{th} cluster and are in it.

Phases (a) and (b) are repeated until stop criterion is satisfied.

The main drawback of K-means algorithm is that the result of cluster is sympathetic to the selection of initial cluster centroids and may trap to the local optima.[5] Therefore, the main processing of K-means is decided by the initial selection of the cluster centroids. Therefore, for generating the initial cluster centroids, we employee some other global optimal searching algorithm.

III. PARTICLE SWARM OPTIMIZATION

A meta-heuristic algorithm, PSO was originally developed by Eberhart and Kennedy in 1995.[6][7].Particle Swarm Optimization is a population based stochastic optimization technique and tries to find the optimal solution using a population of particles. It is inspired by collective behavior of a bird flock for searching of some optimum place in multidimensional space by adjusting their movements and distances for better search.PSO follows two types of approaches: one is called cognitive and another is called social or collective. Particles in a problem space are initialized randomly and then search for optimal solution by updating their generations.PSO is defined as follows:-

A. Elements used in PSO

The following elements are used in PSO:-

a) Particle: We can define the particle as $P_i \varepsilon [a,b]$, where P=1,2,3….D and a,bε R. Here D is for dimension and R is for real numbers.

b) Fitness function: It is also called as objective function and it is used to find the optimal solution.

c) Local Best: It is defined as the best position of the particle among its all positions visited so far.

d) Global Best: It is the position where the best fitness is achieved among all the particles visited so far.

e) Velocity Update: Speed and direction of the particle is determined by the velocity vector.

f) Position Update: For optimal fitness, all the particles try to move toward the best position. Each particle in PSO updates their positions to find the global optima.

Each particle consist of its own position and velocity which are randomly initialized. Among all the particles,each particle have to maintain its local best position P_{best} and global best position G_{best} .The following equations are used to update the velocity and position of a particle-

$$V_{id}= W*Vi_d+ C_1*rand_1*(Pi_d- Xi_d)+C_2*rand_2*(P_{gd}-Xi_d) \quad (3)$$

This equation requires that each particle record its current coordinate X_{id} , its velocity V_{id} that indicates the speed of its movement along the dimensions in a problem space.

$$X_{id}= X_{id}+ V_{id} \quad (4)$$

Here W denites the interia weight factor; P_{id} the location of particle that acquaintance the best fitness value; P_{gd} is the location of particle that acquaintance the global best fitness value. C_1 and C_2 (constants) are denoted as acceleration coefficients; d is the dimension of the problem space; rand1 and rand2 are random values which has a range (0,1).

P_{id} and P_{gd} are the coordinates where the best fitnessvalues were computed. At each generation, the best fitness value are updated by the following equation-

$$P_i(t+1)=\begin{cases} P_i(t) & f(X_i(t+1))\leq f(X_i(t)) \\ X_i(t+1) & f(X_i(t+1))>f(X_i(t)) \end{cases} \quad (5)$$

Here f denotes the fitness function; ‘t’ denotes the generation step and $P_i(t)$ stands for the best fitness value and coordination where the values are calculated. Fig:1 shows the pseudo code of original PSO

```
Initialize the population randomly
While( Population Size)
{
Loop
Calculate fitness
If fitness value is better from the best fitness value(Pbest) in history then
Update Pbest with the new Pbest
End loop
Select the particle with the best fitness value from all particles as Gbest
While maximum iterations or minimum error criteria is not attained
{
For each particle
Calculate particle velocity by equation 3
Update particle position according to equation 4
Next
}
}
```

Fig:1 Pseudo code of original PSO

IV. BAT ALGORITHM

Xin-She Yang developed Bat algorithm in 2010.[8] Echolocation is an important feature of bat behavior. To detect and avoid obstacles , bat uses sonar echoes. It is generally known that the sound pulses which reflects from obstacles are transformed into a frequency. For navigation, the bat uses time delay from emission to reflection. Bat usually emit short, loud sound impulses. After hitting and reflecting to detect the location of prey, the bats transform their own pulse into useful information. The pulse rate is usually defined as 10 to 20 times per second and wavelengths are vary in the range of 0.7 to 17mm or inbound frequencies of 20-500 KHz.

To implement the algorithm, the pulse frequency and rate have to be defined. The pulse rate can be determined in the range from 0 to 1, where 0 means there is no emission and at 1,emission rate is maximum.[9][10][11]. There are three rules which are used when implementing the bat algorithm:
To sense the distance and difference between the food prey and background barriers, echolocation is used by bats.
When bats searching for their prey, they fly randomly with velocity V_i at position X_i with a fixed frequency f_{min} with varying wavelength γ and loudness A_0. Wavelength of their emitted pulses are automatically adjust and adjust the rate of pulse emission rε[0,1],depending on the proximity of their target.
We assume that the loudness varies from a large (positive) A_0 to a minimum constant value A_{min}.
Initialization of the bat population is performed randomly. For generating new solutions, pulse frequency and velocity is calculated as-

$$Q_i^{(t)}=Q_{min}+(Q_{max}-Q_{min})U(0;1) \quad (6)$$

Where U(0,1) is a uniform distribution.

$$V_i^{(t+1)}=V_i^{(t)}+(X_i^{(t)}-\text{best})\,Q_i^{(t)} \quad (7)$$

Local search is done by random walk with direct exploitation. To modify the current best solution we use the equation:

$$X^{(t)}=\text{best}+\varepsilon A_i^{(t)}\,(2U(0,1)-1) \quad (8)$$

Where ε is the scaling factor and $A_i^{(t)}$ is the loudness. The local Search is launched with the proximity (nearness) depending on pulse rate r_i. When bat finds its prey ,the rate of pulse emission r_i increases and the loudness A_i decreases. The bat is moving towards optimal solution according to –

$$A_i^{(t+1)}=\alpha A_i^{(t}\,;\; r_i^{(t)}=r_i^{(0)}[1-\exp(-\gamma\varepsilon)] \quad (9)$$

Here α and γ are constants. α parameter controls the convergence rate of this algorithm. Fig:2 shows the pseudo code of bat algorithm.

1: Objective function f(x), x = (x1; :::; xd)^T
2: Initialize the bat population xi and vi for i = 1…n
3: Define pulse frequency Q_i ε $[Q_{min};Q_{max}]$
4:Initialize pulse rates r_i and loudness A_i
5: while (t < Tmax) // number of iterations
6: Generate new solutions by adjusting frequency and
7: update velocities and locations/solutions
8: if(rand(0; 1) > r_i)
9: Select a solution among the best solutions
10: Generate a local solution around the best solution
11: end if
12: Generate a new solution byfl ying randomly
13: if(rand(0; 1) < A_i and $f(x_i) < f(x)$)
14: Accept the new solutions
15: Increase r_i and reduce A_i
16: end if
17: Rank the bats and find the current best
18: end
19: Postprocess results and visualization

Fig:2 Pseudo code of bat algorithm

CONCLUSION

For inter- disciplinary research , cluster analysis still remains an active field. No single algorithm is known which can group all real world datasets efficiently and without error. However , in recent study, it has been observed that the integrated swarm algorithms consist of an additional optimization function which searches the unexplored part of the search space and improved the current best solution. In this paper, We study the particle swarm optimization and bat algorithm which are used to improve the de-merits of K-means clustering and find the optimum solution.

References

[1] Jiawei Han, Micheline Kamber, “Data Mining: Concepts and Technique”.
[2] K. Jain, M. N. Murty, P. J. Flynn, “Data Clustering: A Review”, *ACM Computing Surveys, Vol. 31, No. 3, September 1999.*
[3] Xiong, H., J. Wu and J. Chen, 2009. K-Means clustering versus validation measures: A data distributionPerspective.IEEETrans.Syst.,Man,Cybernet.PartB,39: 318331.http://www.ncbi.nlm.nih.gov/pubmed/19095536.R. Nicole, “Title of paper with only first word capitalized,” J. Name Stand. Abbrev., in press.
[4] M.N.Joshi. Parallel K-Means Algorithm Distributed Memory Multiprocessors[J].Computer, 2003, 9:3-15.
[5] Cui X., Potok T. E., 2005. Document Clustering using Particle Swarm Optimization, IEEE Swarm Intelligence Symposium 2005, Pasadena, California
[6] Omran, M., Salman, A. and Engelbrecht, A. P.,2002. Image classification using particle swarm optimization. Proceedings of the 4th Asia-Pacific Conference on Simulated Evolution and Learning 2002 (SEAL 2002), Singapore. pp. 370-374.
[7] J. Kennedy and R. Eberhart .”Particle Swarm Optimization”, In Proceedings of IEEE International Conference on Neural Networks, 1995, PP.1942-1948.
[8] X.S. Yang. A new metaheuristic bat-inspired algorithm. Nature Inspired Cooperative Strategies for Optimization (NICSO 2010), pages 65-74, 2010.
[9] A.H. Gandomi, X.S. Yang, A.H. Alavi, and S. Talatahari. Bat algorithm for constrained optimization tasks. Neural Computing & Applications, pages 1-17, 2012.
[10] P.W. Tsai, J.S. Pan, B.Y. Liao, M.J. Tsai, and V. Istanda. Bat algorithm inspired algorithm for solving numerical optimization problems. Applied Mechanics and Materials, 148:134-137, 2012.
[11] X.S. Yang. Review of meta-heuristics and generalised evolutionary walk algorithm. International Journal of Bio-Inspired Computation, 3(2):77-84, 2011.