BIG DATA DRIVEN GLOWWORM SWARM OPTIMIZATION USING SUBTRACTIVE DATA CLUSTERING ALGORITHM

Shaikh Nafeesa[1], Sindhuja K[2], Pramela Devi[3]
[1]M.Tech, Department of Computer Science & Engineering ,MVJ College Of Engineering, Near ITPB, Whitefield, Bengaluru , Karnataka, India.
[2,3]Assistant professor, Department of Computer Science & Engineering, MVJ College Of Engineering, Near ITPB, Whitefield, Bengaluru , Karnataka, India.

Abstract
Clustering is the process of grouping similar set of data into specified number of groups or clusters. Clustering large amount of data is one of the most challenging that has been used in many areas of application such as bioinformatics, World Wide Web, Social networking and etc. A recent study has shown that PSO with K-means algorithm is the most popular algorithm for large data clustering. The major problem in kPSO is, that at first the number of clusters is not known and clustering result depends on initial centroid selection and lost its accuracy if the dataset is of high dimensions. In this project, a scalable design and implementation of Glowworm Swarm Optimization with Subtractive data clustering along with MapReduce (MR-SGSO) is introduced to handle large amount of data in Big Data area. The Glowwrom has been used, to make use of its solving numerous multimodal problems, which in term of clustering means to search for multiple centroids. The Subtractive clustering algorithm is one-pass, fast algorithm for finding the number of clusters and its centers for any amount of data. MapReduce has been used for the parallelization since it provides data locality, load balancing and fault tolerance. The experimental results show that MR-SGSO scales exceptionally well with expanding data set sizes and accomplishes an optimize speedup while keeping up the efficient clustering quality.

***Keywords*: Data Clustering, Glowworm Swarm Optimization (GSO), Subtractive Clustering Algorithm, Parallel Processing.**

I. INTRODUCTION

Data clustering, it is a collection of similar data into a particular group or clusters. Clustering large amount of data is one of the important and challenging tasks that has been used many application areas such as social media, image segmentation, bio-informatics and etc. If the same cluster is having maximum number of similarity and different cluster which is having minimum number of similarity then the clustering algorithm is said to be more efficient. There are two major clustering techniques: “Hierarchical” and “Partitioning”. In hierarchical clustering, the execution is a tree showing an arrangement of clustering with each cluster being a partition of the dataset. On the other hand, Partitioning clustering the dataset has been partitioned to number of clusters.

In the past year it has been seen that partitioning clustering is well defined for clustering large number of dataset. The partitioning technique time complexity is linear almost and it has been widely used. The K-means algorithm is one of the known partitioning techniques. The most popular K-means clustering suffer from various drawbacks, hear the clustering result depends upon initial number of centroid selection and it loses its accuracy when the dataset is of high dimension.

Particle Swarm Optimization (PSO) is a continuous population based optimization technique. PSO is biologically inspired like genetic algorithm and ant colony optimization.
In PSO, the population of particle or swarm of particle that it starts from objective function space which is placed in random position and

move for searching a local optima. The
The particle of PSO moves in multidimensional solution space. After each instance the position and velocity of the particle is updated and hence try to move to optimum solution point. The PSO also suffer from several drawbacks, it has slow coverage rate. The particle and velocity of PSO should depend upon the previous particle to get updated.

Subtractive clustering method is fast compared to K-means which will give the number of clusters and centroid for each cluster within a particular time frame for any amount of data. The subtractive clustering module predicts initial centroid clusters for the next phase and finds the optimal number of clusters.

In recent years, some researches has described about the idea of swarm for clustering. Swarm intelligent algorithm has developed feature which is self-organized and which share information to get the best solution among the various swarm agent. GSO is influenced by the action of insects know as glowworm. This glowworm consist of luciferin that is which
emits light, which can be used for multiple purposes e.g., the glowworm emits more light when it goes in search of a food and give sign to other worms that it has found the food. The GSO, solve various task of optimization, where each worm of glowworm search for the centroid of cluster which is also known as sub solution. Hence the composition of this sub solution can form a global solution for any clustering problem.

In this paper, we present a Glowworn swarm optimization (GSO) with subtractive data clustering algorithm along with MapReduce to get the better coverage rate compared to kPSO algorithm. The GSO with subtractive clustering algorithm that performs the fast clustering. The subtractive algorithm is one pass, fast algorithm for finding the number of clusters and its centers for any amount of data. The MapReduce is used for parallel processing, so that the time taken for large amount of data will be the efficient one with best cluster quality.

The rest of the paper is arranged in the following manner: Section II describe about the various research on clustering algorithm using PSO & GSO. Section III describe about the Subtractive clustering algorithm. Section IV describe about the Glowworm swarm optimization. Section V describe about MapReduce technique. Section VI describe about the proposed system. Section VII gives the experimental result of the project. Section VIII gives the conclusion.

II. RELATED WORK

In this section, we present the survey of various clustering algorithm which has been used in big data. Clustering algorithm such as k-means, particle swarm optimization, and so on has been discussed. Due to sheer volume of big data and heterogeneous structure the process of clustering has become a great challenge. Especially, when there is diverse amount of data present, getting an efficient and best quality cluster in a less computation time is a challenging task in a recent research on big data.

In [1], Hua Fang discuss about the significant challenges of big data in terms of analytics, management, ethics, storage and networking. It is a survey of big data research which gives the multiple advantages of big data in terms of cost reduction, development of new services and time improvement in computing a task.

In [2], Ling LIU discussed about the computing infrastructure which has experienced a major changeover from single processing device to all over networking device. The paper gives an overview of computing infrastructure for big data processing, storage and networking challenges.

In [3], Mariam El-Tarabily discussed about the clustering technique and explains about the drawbacks of K-means clustering algorithm. The hybrid Subtractive + PSO clustering gives the efficient and fast clustering approach. Experiment results of Subtractive + PSO clustering algorithm gives the optimal solution compared to ordinary PSO after nearly 50 iterations. The proposed approach in this generates the most compact result for clustering. The Subtractive clustering algorithm gives the initial centroid cluster and also generates number of clusters for PSO.

In [4], Chetna Sethi proposed a Linear PCA based hybrid K-means PSO algorithm for clustering large dataset. The proposed algorithm overcomes the drawbacks of K-means, by combining the K-means along with the linear PCA (Principal Component Analysis) which is used for reducing dimension. The PCA-K-PSO algorithm, this there algorithm is been collaborated to cluster large amount of dataset. Where the K-means is been used for its fast

convergence rate and PSO for global search ability.

In [5], S.Rana proposed a new algorithm for clustering large dataset that is Boundary restricted adaptive particle swarm optimization (BRA-PSO) which has been tested on nine dataset and hence this proposed algorithm has been compared with those of NM-PSO, K-PSO, and K-means algorithm. The proposed system is more robust, fast and show accurate coverage speed compared to other algorithm.

In [6], Bara'a Ali Attea proposed a fuzzy multi-objective particle swarm optimization algorithm to give the effective clustering result for large amount of data. The proposed algorithm checks the effectiveness and efficiency by comparing with existing clustering algorithm. The proposed framework resolve the confusion of cluster assignment which have single or more than one belongingness in one cluster.

In [7], Tongguang Zhang proposed the Qos-aware web service selection based on PSO. The proposed algorithm is been used to deal with large number of non-linear, complex and non-differentiable problem which has been widely used in science and engineering.

In [8], Ibrahim Aljarah proposed a new clustering approach based on Glowworm Swarm Optimization (GSO). The GSO is based on the lighting of worm and hence which is inspired by the nature. The GSO has been used to solve various multimodal problem which in terms of clustering means finding centroid for multiple clusters. It proposed algorithm experiment results shows more efficient clustering technique.

In [9], Dr.R.N.Kulkani proposed a design of parallel Glowworm swarm optimization tool using Mapreduce. The GSO is been used to optimize the cluster, it uses

MapReduce function to get the work done within a less time frame.

III. SUBTRACTIVE CLUSTERING ALGORITHM

In Subtractive data clustering each point of data is considered as participants for center of cluster. Here data points are independent of each other. Subtractive clustering predicts the optimal number of clusters and also find the initial centroid cluster for the next phase.

In M-dimensional space, consider the collection of n data points $\{x_1, x_2, \ldots\ldots, x_n\}$. Since each data points is a participant for center cluster, at point x_i the density measured by

$$Di = \sum_{j=1}^{n} \exp\left(-\frac{||x_i - x_j||^2}{(r_a/2)^2}\right)$$

Where r_a is considered as positive constant. Hence, if there are have many neighborhood data points then the data point value is of high density. The neighborhood for the data point is radius r_a, which is used to measure the highest density.

After the density measure is calculated of all data points, the first cluster is selected which has the highest density measure. x_i is recalculated for each data point which is given by:

$$D_i = D_i - D_{c_1} \exp\left(-\frac{||x_i - x_{c_1}||^2}{(r_b/2)^2}\right)$$

Where r_b is considered as positive constant. The data points which are not selected as the next cluster are the one whose density measure is reduced, which was close to first cluster center. The positive constant r_a is normally smaller then r_b, so that the cluster are not closely placed.

After recalculating the density measure for each data point, the x_{c2} is selected as a next cluster and hence density measure is recalculated for each data point. Until a sufficient number of clusters are generated the iterative process is repeated. Hence these cluster centers are used in GSO algorithm for initial clustering centers.

IV. GLOWWORM SWARM OPTIMIZATION ALGORITHM

In recent years, some researches has described about the idea of swarm for clustering. Swarm intelligent algorithm has developed feature which is self-organized and which share information to get the best solution among the various swarm agent. GSO is influenced by the action of insects know as glowworm. This glowworm consist of luciferin that is which

emits light, which can be used for multiple purposes e.g., the glowworm emits more light when it goes in search of a food and give sign to other worms that it has found the food. The GSO, solve various task of optimization, where each worm of glowworm search for the centroid of cluster which is also known as sub solution. Hence the composition of this sub solution can form a global solution for any clustering problem.

The basic working of the GSO algorithm is the result of transaction between the given three

mechanisms which are:

1. **Fitness broadcast***:* Glowworm consist of emission of light pigment called luciferin, the highest value of luciferin calculates the fitness value of glowworm. This allows the glowworm to glow at certain rate which is directly proportional to the optimized function value. There is no reduction in the luciferin level, if it is sensed by the neighbour due to distance.
2. **Positive Axis**: In a search space of multiple glowworm, each glowworm moves to the neighbour whose glow is brighter then itself. The probabilistic mechanis has been used to select the best from them.
3. **Adaptive neighbourhood***:* To identify neighbours, glowworm uses adaptive neighbourhood range.

The step by step process of GSO algorithm is given by:

Set number of dimensions $= m$
Set number of glowworms $= n$
Let s be the step size
Let $x_i(t)$ be the location of glowworm i at time t
deploy_agents_randomly;
for $i = 1$ to n do $\ell_i(0) = \ell_0$
$r_d^i(0) = r_0$
set maximum iteration number $= iter\_max$;
set $t = 1$;
while $(t \leq iter\_max)$ do:
{
for each glowworm i do: % Luciferin-update phase
$\ell_i(t) = (1-\rho)\ell_i(t-1) + \gamma J(x_i(t))$;

for each glowworm i do: % Movement-phase
{
$N_i(t) = \{j : d_{ij}(t) < r_d^i(t); \ell_i(t) < \ell_j(t)\}$;
for each glowworm $j \in N_i(t)$ do:
$p_{ij}(t) = \frac{\ell_j(t)-\ell_i(t)}{\sum_{k\in N_i(t)} \ell_k(t)-\ell_i(t)}$;
$j = select\_glowworm(\vec{p})$;
$x_i(t+1) = x_i(t) + s\left(\frac{x_j(t)-x_i(t)}{\|x_j(t)-x_i(t)\|}\right)$
$r_d^i(t+1) = \min\{r_s, \max\{0, r_d^i(t) + \beta(n_t - |N_i(t)|)\}\}$;

}
$t \leftarrow t+1$;
}

The above GSO algorithm with subtractive clustering algorithm has been used in our project for processing large amount of dataset within less computation time along with MapReduce to process the large dataset.

V. MAP REDUCE TECHNIQUE

MapReduce is a profoundly adaptable model and can be utilized across many nodes of computer, and it is been used for larger dataset applications, when there are restrictions on multiprocessing also, extensive shared-memory machines. MapReduce influences utilization of two primary activities: Map and Reduce. Both Map and Reduce activities take input and generates output in the form of <key, value>. The Map task goes over a vast number of records and extracts important information from each record, and after that all values with a similar key are sent to the same Reduce activity. The reduce task generates the final result by aggregating the result provided by map function.

A notable and usually utilized execution of MapReduce is Apache Hadoop. It empowers applications to work with petabytes of data utilizing a huge number of independent processors. One of the primary components of Hadoop is, Hadoop Distributed File System (HDFS).HDFS gives high-throughput access to the data and keeps up multiple replicas of the target data block. MapReduce and HDFS work together to compute large amount of data.

VI. PROPOSED SYSTEM

In our proposed system, we take the MAGIC (Major Atmospheric Gamma Imaging Cherenkov Telescope) dataset as an input, which contains no. of instances, attributes, classes. The proposed system performs the following three steps to get the optimize cluster:

1. **Glowworm Swarm Initialization:** For clustering, the clustered data has been taken. In this technique, initially the agent of swarm has been distributed in the search space. The GSO agents carry a quantity of luminescence known as luciferin along with them. The glowworms transmit a light whose force is relative to the related luciferin and communicate with different agents inside a variable neighbourhood.
2. **Centroid Selection:** This module tells the swarm agents are selected in centroid manner to reduce the tasks. Here map and reduce tasks has been performed.
3. **Clustering Movement:** In this module the subtractive clustering technique has been used, the various agents are clustered and

moved. The output of map task is given to reduce task where the reduce class calculates the fitness of each glowworm.

The above modules are performed to implement MR-SGSO. Where, GSO is used to solve multimodal problem, which in terms of clustering is finding multiple centroids. The Subtractive clustering algorithm is used to get the initial number of cluster and which is fast compares to kPSO. The MapReduce technique is been used for parallel processing.

The following figure (1) gives the block diagram for the proposed system:

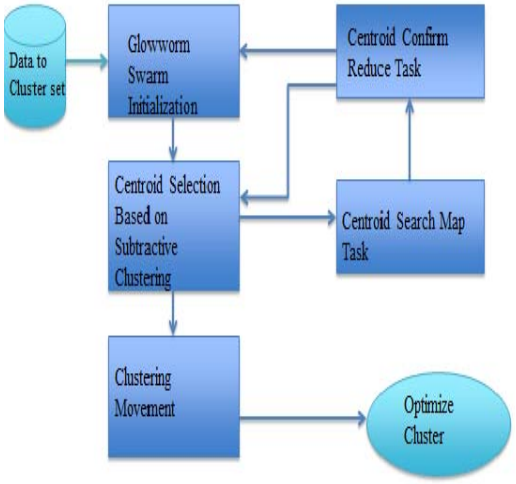

Figure 1: Block Diagram

VII. EXPERIMENTAL RESULT

In this section, we describe about the implementation of our proposed work.

The real data set which is taken as input for our project is MAGIC dataset. This Dataset is MC generated which consist of high energy gamma particle collected from Cherenkov gamma telescope. Under these more than 10000 Cherenkov photons has been collected. The dataset consist of 11 attribute which also include class. It is obtained from UCI machine learning repository.

In figure (2), it describe about the computation time for the kPSO. The graph of kPSO shows the execution time for the different amount of data set. The time taken by kPSO is much compared to our proposed approach. The time taken to execute 55,500 size of data is 10,000 milli sec, and it also the initial number of cluster is not known. The clustering result depends upon the initial number of clusters.

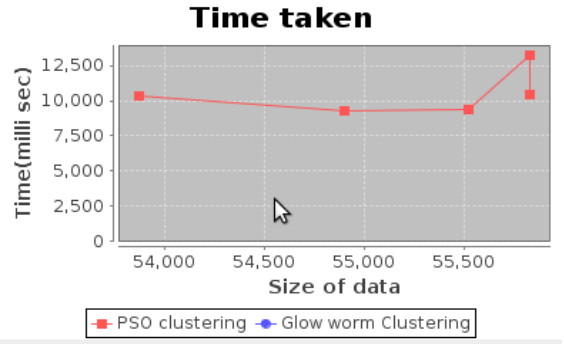

Figure 2: Computation time of k-PSO

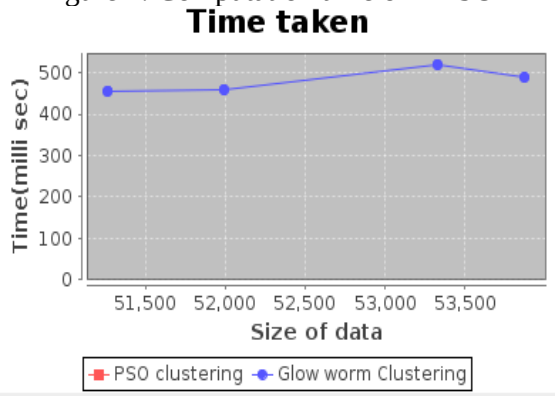

Figure 3: Computation time of MR-SGSO

In figure (3), it describe about the computation time for MR-SGSO. The graph of MR-SGSO shows the execution time for different amount of data. The time taken for MR-SGSO is 460milli sec for 53,500 size of data. Which is more efficient compared to time taken kPSO. The proposed algorithm is fast and more efficient compared to existing one.

It solves multiple multimodal problem and gives the best cluster quality within less computation time.

VIII. CONCLUSION

A scalable design and implementation of Glowworm Swarm Optimization (GSO) with Subtractive data clustering along with MapReduce is proposed. The PSO with K-means is an existing approach, it is an effective method however it take a long time to process large amount of dataset. Therefore, MR-SGSO was proposed to overcome the inefficiency of k-PSO for big datasets. MRSGSO shows that SGSO can parallelized efficiently with MapReduce to process large data set. Experiments were conducted to compare the efficiency of kPSO and MR-SGSO, hence MR-SGSO gives better cluster quality and solve various multimodal problem within less computation time.

REFERENCES

[1]. Ling Liu, "Computing infrastructure for big data processing. Frontiers of Computer Science, vol. 7, no. 2, pp. 165-170, 2013.

[2]. H. Fang, Z. Zhang, C. J. Wang, M. Daneshmand, C. Wang, and H. Wang, "A Survey of Big Data Research," IEEE Network, pp. 6-9, September/October 2015

[3]. Mariam El-Tarabily, Rehab Abdel-Kader, Mahmoud Marie3, Gamal Abdel-Azeem,"A PSO-Based Subtractive data", IJORCS, Volume 3,Issue 2, (2013).

[4]. Chetna Sethi, Garima Mishra,"A Linear PCA based hybrid K-means PSO algorihm for clustering large dataset", IJSER, Volume 4, Issue 6, June 2013

[5]. S.Rana, S.Jasola , R.Kumar, "A boundary Restricted adaptive particle swarm optimization for data clustering", Springer, 8 June 2012

[6]. Bara'a Ali Attea, "A fuzzy multi-objective particle swarm optimization for effective data clustering", Springer, 24 July 2010

[7]. Tongguang Zhang," Qos-aware web service selection based on particle swarm optimization", Journal of Network, Vol.9, March 2014

[8]. Ibrahim Aljarah and S. A. Ludwig, "A new clustering approach based on glowworm swarm optimization." in IEEE Congress on Evolutionary Computation. Mexico, Cancun: IEEE, pp. 2642–2649, 2013

[9]. Dr.R.N.Kulkarni, Aishwarya H.M, Arun Kumar, Deeksha Patil, Deeksha Jain.P,"Design of Parallel Glowworm Swarm Optimization Tool Using Map Reduce,IJCSIT, Vol.7, 2016

[10]. M. Shamim Hossain, Senior Member, IEEE, M. Moniruzzaman, and G. Muahammad, Member, IEEE, Ahmed Al Ghoneim, Member IEEE, Atif Alamri, Member IEEE,"Big Data-Driven Service Composition Using Parallel Clustered Particle Swarm Optimization in Mobile Environment", IEEE Transaction, 2016