Krill Herd Clustering Algorithm using DBSCAN Technique

Varinder Singh
Department of CSE/IT
Lovely Professional University
Phagwara, Punjab, India
varindersaini1990@gmail.com

Mrs. Monica Sood
Department of CSE/IT
Lovely Professional University
Phagwara, Punjab, India
monica.14858@lpu.co.in

Abstract— The hybrid approach is proposed to show that the clusters also show the swarm behavior. Krill herd algorithm is used to show the simulation of the herding behavior of the krill individuals. Density based approach is used for discovering the clusters and to show the region with sufficiently high density into clusters of krill individuals that of the arbitrary shape in environment. The minimum distance of each individual krill from food and from high density of herds are considered as the objective function of the krill movement. A density based cluster is a set of density connected objects that is maximal with respect to density-reach ability and noisy objects. The movement of the krill individuals is considered by the foraging movement and random diffusion.

Keywords- Hybrid Krill herd clustering algorithm, DBSCAN, Genetic operators.

I. INTRODUCTION

Swarm intelligence is the self organized behavior that can be of any type of artificial or natural. Krill is the one of the species which shows the swarm behavior found mostly in Antarctic. Clustering is the process of grouping a set of physical or abstract objects into classes of similar objects. The density based clustering technique can be used to discover clusters with arbitrary shapes. DBSCAN searches for clusters by checking the neighborhood of each point in the given space or environment. DBSCAN then collects density reachable objects from the core point object, which may involve the few density reachable clusters. There can be the different numbers of clusters in any given region. These outliers may be the group of the other cluster which can also create the cluster in different shapes. [7][4]

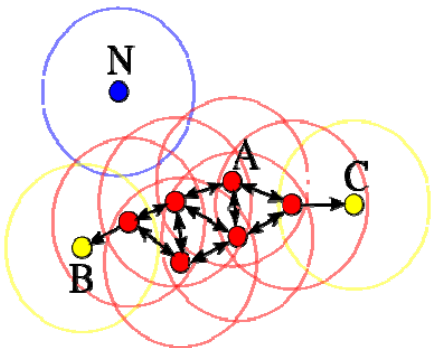


Figure 1: Density-reachable Cluster with noise[5]

As seen in figure, there are the different clusters and that shows the border and core points of the different objects. Points at A are core points. Points B and C are density-reachable from A and thus density-connected and belong to the same cluster. Point N is a noise point that is neither a core point nor density-reachable. Clustering is used for outliers detection where these outliers are very interesting than any other single objects. These outliers may be the group of the other cluster which can also create the cluster in different shapes. The approach for using this technique is to find the best optimal path by the krills and reaching the food. The core krill is selected on the basis of the density and the fitness value of the each individual and can be calculate the best result for the reaching of the food. [1]

There is requirement of two different parameters for DBSCAN ε (eps) and the minimum number of points required to form a cluster (minPts). The formation of clusters starts with an arbitrary starting point that has not

been visited. This point's ε-neighborhood is retrieved, and if it contains sufficiently many points, a cluster is started. Otherwise, the point is labeled as noise. Note that this point might later be found in a sufficiently sized ε-environment of a different point and hence be made part of a cluster.

The main motive of using this technique is to show that the clustering can also be applied to the natural intelligence techniques and the clustering also shows the swarm behavior and thus used with the nature inspired algorithms to help to find the food for the krill individuals. Motion of the krill individuals is due to the core point of the cluster, they try to maintain the high density of cluster and move due to their mutual effects. The direction of the motion induced by them is measured by the local density of clusters and by the movements of the core krill of the cluster. In this the genetic operator's crossover and mutation are also used to increase the effectiveness of the results and to find the best possible result. [4][6][1]

In section 2, we have defined the literature respective of the presented work. In section 3, the proposed work is defined.

II. LITERATURE SURVY

Many of the earlier researchers did a lot of work for finding the best optimal path to reach the destination. In this section, the work done by the earlier researchers is discussed. In year 2012, Gandomi A and Alavi A performed a work to expose and explore the different paradigms of finding the best optimal path up to the food location. The KH algorithm is based on the herding behavior of the krill individuals. The position of the krill individuals is time dependent and can be calculated the three different kinds of main factors. Movement induced by the presence of the other krills, foraging activity and random diffusion. Mutation tells about using the best value as compared to any other value which is obtained from the different number of iterations. These steps are performed again until the best result is found.[1]

Another work performed by Ester M, Sander J, et al for the main technique which can be used for the topic is density based clustering approach. It used for discovering the arbitrary shaped clustered data. DBSCAN technique is used for efficiently for calculating the large databases. It represents the notion of the clusters which based on the density of the different clusters of data. It details about the retrieving all the points that density reachable from the given points using the correct parameters. DBSCAN uses global values for Eps and MinPts i.e. the same value for all the clusters. The DBSCAN algorithm explains about the finding the new neighborhood from the randomly selected core points. On the basis the core points, the new epsilon-neighbors are estimated. The neighbors are estimated on the basis of the particular distance from the core point around that point and the new arbitrary based clusters are formed.[6][4]

Jian li, Yu Wei, et.al explained in their publication about the DBSCAN technique. DBSCAN algorithm is computationally expensive, limiting its performance in large-scale data sets, especially in high dimensional data sets. The high complexity is rooted from the region queries, a very common operation in density based algorithms, which brings the complexity of the algorithms to O (n2), where n is the number of database objects. With the help of index structure the complexity can be reduced to O (nlogn), however it is inefficient to create the index structure especially for high dimensional data sets or large-scale databases. In this paper they explained a new concept named memory effect (ME). ME can be used to shrink the scope of region queries to neighboring objects. Based on ME we have improved DBSCAN algorithm evidently, and empirical experiments have shown the improvement in both effectiveness and efficiency. At last, they gave the theoretical analysis of MEDBSCAN algorithm and talk about the influence of parameters.[2]

Ram A, Sharma A, et.al explained about the clustering technique. It can find out the clusters of different shapes and sizes from the large amount of data which is containing noise and outliers. But the clusters detected by it contain large amount of density variation within them. It cannot handle the local density variation that exists within the cluster. For good clustering a significant density variation may be allowed within the cluster because if we go for homogeneous clustering, a large number of smaller unimportant clusters may be generated. In this paper we propose an Enhanced DBSCAN algorithm which keeps track of local density variation within the cluster. It calculates the density variance for any core object with respect to its e -neighborhood. If density variance of a core object is less than or equal to a threshold value and also satisfying the homogeneity index with respect to its e -neighborhood then it will allow the core object for expansion. The experimental results show that the proposed clustering algorithm gives optimized results.[3]

Hua Z, Zhenxing W, et.al Density-based clustering algorithms, which are important algorithms for the task of class identification in spatial database, have many advantages such as no dependence on the number of clusters, ability to discover clusters with arbitrary shapes and handle noise. However, clustering quality of most density-based clustering algorithms degrades when the clusters are of different densities. To address this issue, this paper brings forward a clustering algorithm based on characteristics of density distribution--CCDD algorithm. Firstly, it divides data space into a number of grids. Secondly, it re-divides data space into many smaller partitions, according to each grid's one-dimensional or multi-dimensional characteristics of density distribution. Finally, it uses an improved DBSCAN algorithm, which chooses different parameters according to

each partition's local density, to cluster respectively. The experimental results show that CCDD algorithm, which is superior in quality and efficiency to DBSCAN algorithm, can find clusters with arbitrary shapes and different densities in spatial databases with noise.[8]

III. PROPOSED WORK

The presented work is combination of different approaches for the clustering of the krill herds. As the krills follows the density based approach for finding the shortest path for reaching the food. Thus the DBSCAN technique is used here to implement the shortest path finding for the krills by discarding the extra krills which are not in density reachable area.

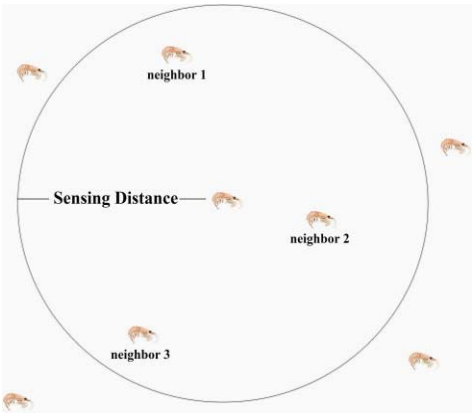


Figure 2: Krills with in Cluster[1]

A new hybrid algorithm is proposed for the formation of the clusters and movement induced by the clusters due to the movement induced by the single core point in the clusters that core point is the krill that have the best fitness value computed by using the different genetic operators. The fitness value is compared with the different krills that are within the particular cluster that is under consideration. Density is equals to the number of the points within the particular radius (EPS). A point is a core point if it has a more than specified points (MinPts) with Eps. These are the points that are the Interior of the cluster. A border point has fewer than MinPts than within Eps, but it is in the neighborhood of the core point. Noise is any point which is not within any point that is not any core point or any border point.

Then after getting the result from the iteration, the positions of the clusters are updated. Once the positions of the clusters are updated, then check if the result found is best solution around the neighborhood positions particles, then the terminate the operations and then the optimal best solution is found and the probability of reaching the food increases.

A. Proposed KHC Algorithm

The methodology used in algorithm is simply finding the best and easily finding the core point and finding the path to the food by finding the different number of iterations. This algorithm is proposed for best solution and for the complete results of all the methodology. The algorithm explained below:

- Select any point A from the given area.
- Retrieve all points density-reachable from *A* w.r.t. ε and MinPts…… *(*)*
- Calculate the fitness value using genetic operators.
- Calculate the motion induced by other objects w.r.t. A.
- Use genetic operators such as mutation and crossover for the comparison of best result.
- Update the position of krill objects.
- Perform the steps from (*) for checking again.
- End

The selection of the core point from the cluster is depends on ε and MinPts parameters. These are parameters which are required for the selection on the basis of density reachable objects. The fitness function is defined as the distance from food and highest density of krills. The time dependent actions that are to be considers

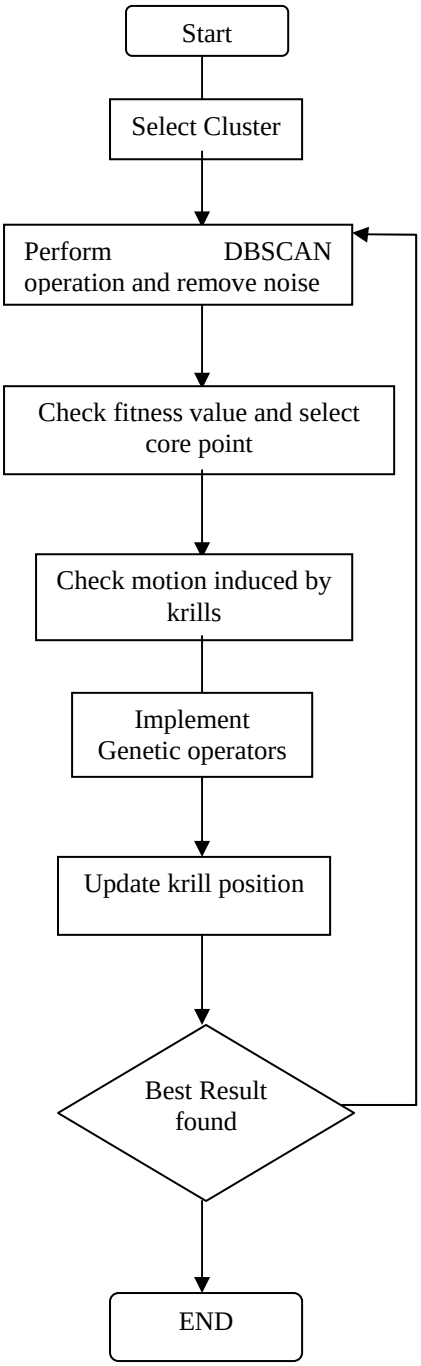
- Motion induced by the krills
- Random diffusion

And then check if the best result is found or not. This all depends on the behavior of the krills.

B. Density based clustering technique

DBSCAN is a density based clustering algorithm. Here the density is calculated by the number of points with in particular radius ε (Eps). A core point is point if it has more than the specified number of points (MinPts). Core point is in the interior of the cluster. A border point has fewer than *MinPts* within *Eps* but is in neighborhood of a core point. A noise point is any point that is neither a core point nor a border point. So we are considering only the point which holds the boundary condition.

The flowchart for the proposed methodology:



IV. CONCLUSION

This work is about to define the simple approach for calculating the best path by the krills. Based on the density based technique the algorithm is proposed for the optimization. This gives the easiest approach for the grouping of different objects and this becomes easy to define them in groupings and that can be used for easy differentiable than others. The presented work is the model for implementing the best possible path by using the DBSCAN technique.

V. REFERENCE

[1] Gandomi.A, Alavi.A "Krill herd algorithm: new bio-inspired optimization algorithm" Elsevier Publications, Communications in Nonlinear Science and Numerical Simulation, 2012.

[2] Jian.L, Wei.Y, Bao-Ping.Y, "Memory effect in DBSCAN algorithm," 4th International Conference on Computer Science & Education, 25-28 July 2009.

[3] Ram .A, Sharma.A, Jalal.A.S, Agrawal.A, Singh.R "An Enhanced Density Based Spatial Clustering of Applications with Noise," IEEE International Advance Computing Conference, 2009.

[4] http://en.wikipedia.org/wiki/DBSCAN.

[5] http://en.wikipedia.org/wiki/File:DBSCAN-Illustration.svg.

[6] Kamber.M and Han. J, Data Mining: Concepts and Techniques, The Morgan Kaufmann publications, San Francisco, Pp 418-423.

[7] Lai L.L, Nieh.T "Swarm Intelligence" published in transmission and Distribution Conference and Exhibition: Asia and Pacific, 2005 IEEE/PES, Pp 1-5.

[8] Hua.Z, Zhenxing, Liancheng.Z, Qian.W, "Clustering algorithm based on characteristics of density distribution," 2nd International Conference on Advanced Computer Control 27-29 March 2010.