Applications of various swarm intelligence techniques in clustering

Tatineni Anusha[1], Garimella Bharathi[2], Tatineni Poojitha[3], Atluri Harshitha sai[4]

[1] *Student, M.Tech, CSE Dept, Gudlavalleru Engineering College, Gudlavalleru, India*

[2] *Sr.Gr.Assistant Professor, CSE Dept, Gudlavalleru Engineering College, Gudlavalleru, India*

[3]*Assistant Professor, CSE Dept, Gudlavalleru Engineering College, Gudlavalleru, India*

[4]*Student, B.Tech, CSE Dept, Gudlavalleru Engineering College, Gudlavalleru, India*

Abstract

In the wake of investigating the drawbacks of the established K-means clustering calculation, this paper consolidates the center thought of K-implies grouping technique with PSO calculation and proposes another grouping strategy which is called grouping calculation dependent on molecule swarm improvement calculation. It utilizes the global optimization of PSO calculation to make up the deficiency of the grouping strategy. Results demonstrate that the calculation is more powerful a molecule swarm is a population of particles, in which every molecule is a moving article which can travel through the search space and can be pulled in to the better positions. PSO must have a wellness assessment capacity to choose the better and best positions; the capacity can take the molecule's position and allots it a wellness fitted value. At that point the goal is to optimize the fitness function.

Keywords: ***Ant colony, Clustering, Honey bee, Particle swam optimization.***

I. Introduction

Clustering is the group of particular set of objects based on their features, aggregating them according to their similarity. Concerning to data mining, this methodology partitions the data implementing a precise join algorithm, most suitable for the desired information analysis. Cluster of data objects are treated as one group. While performing cluster analysis, first partition the set of data into groups based on data similarity and assigns the labels to the groups. The main benefit of clustering over classification is that, it is flexible to changes and helps single out useful features that differentiate different groups. Clustering is one of the key tasks in examining data mining and is also a technique used in statistical data analysis. Popular design of clusters consists of groups with small distances between cluster members, dense areas of the data space and intervals of statistical distributions. Clustering as a result can be formulated as a multi-objective optimization problem. The proper clustering algorithm and parameter settings depend on the individual data set and deliberate use of the results. Cluster analysis is not simply an automatic task, but an iterative process of knowledge discovery or interactive multi-objective optimization that involves trial and failure. It is often essential to adjust data pre processing and model parameters until the result attain the desired properties.

II. Clustering:

I. Partitioning based Clustering:

Partitioning clustering uses iterative process to optimize the cluster centers, as well as the number of clusters. All of the objects are initially considered as a single cluster. Then these objects are divided into number of partitions by iteratively locating the points between the partitions. These algorithms require the analyst to specify the number of clusters to be generated. Some of the partitioning algorithms are K-means, K-medoids.

a) K-MEANS:

Firstly partition of objects into k non-empty subsets. Identify the cluster centroids (mean point) of the present partition. Assign each point to a definite cluster. Calculate the distances from each point and assign points to the nearest cluster, choose the minimum distance from centroid. After re-allotting the points, find the Centroid of the latest cluster formed.

Issues: Handling empty clusters: At assignment step when no points are assigned to cluster this issue occurs. This can be done by choosing the point that contributes most to SSE (sum of squared error) or choose a point from the cluster with the highest SSE.

Outliers: Outliers are the points that are far away from the centriod and they can excessively influence centriod values and in turn this can skew cluster grouping and increase the quantity of time needed to find an optimal solution.

b) K-MEDOIDS

The *k*-medoids algorithm is related to the k-means algorithm. Both the *k*-means and *k*-medoids algorithms are partitioned and both efforts to minimize the distance between points labelled to be in a cluster and a point chosen as the center of that cluster. In distinction to the *k*-means algorithm, *k*-medoids chooses data points as centers and works with a algorithm based on distance metrics. *K*-medoids is a classical partitioning method of clustering that cluster the data set of *n* objects into *k* clusters known *a* priori. Partitioning around Medoids algorithm is common method used in k-medoids. PAM uses a greedy search which may not locate the optimum solution, but it is faster than exhaustive search. A medoids can be defined as the entity of a cluster whose average dissimilarity to all the objects in the cluster is minimal. I.e. it is a mainly centrally located point in the cluster.

K-MEDOIDS Advantages:

1) It is extra robust to noise and outliers as compared to k-means because it minimizes a sum of pair wise dissimilarities instead of a sum of squared Euclidean distances.

2) It can solve K- means problems and create empty clusters and is sensitive to outliers or noise.

3) It can also select the most centred member belonging to the cluster.

K-MEDOIDS Disadvantages:

It requires precision and is complex enough.

II. Hierarchical Clustering

Hierarchical clustering does not state the number of clusters, and the output is independent of the initial condition. However, the hierarchical clustering is static that is the data points assigned to one cluster cannot be reassigned to another cluster. In accumulation, it will fail to separate overlapping clusters due to the lack of information regarding the global shape or size of the clusters. There are mainly two approaches to perform Hierarchical clustering techniques, such as Agglomerative (top-bottom) and Divisive (bottom- top). In Agglomerative hierarchical clustering approach, initially one object is selected and successively merges the neighbour objects based on the distance measure as minimum, maximum and average. The process is constant until a desired cluster is formed. The Divisive hierarchical clustering approach deals with set of objects as single cluster and divides the cluster into further clusters until desired no of clusters are formed. BIRCH, CURE, ROCK, Chameleon, Echidna, Wards, SNN, GRIDCLUST, and CACTUS are some of Hierarchical clustering methods.

Various applications of clustering:

- **Medicine field:** The specialist would identify side effects, such as psychology, tension, anxiety, depression etc. The cluster examination can recognize groups of patients that have similar side effects
- **Biology field:** In this field cluster investigation is mainly used in categorization of species. Analysts can gather a data set of divergent plants and can note diverse characteristics of their phenotypes. A cluster examination can group those observational records into a progression of clusters and assemble scientific classification of gatherings and subgroups of practically equivalent to plants.
- **World Wide Web:** In social network analysis, clustering derives the acquainted results with communities inside extensive groups of individuals. Search result grouping: While performing intelligent grouping of files or documents and websites, clustering can generate a more pertinent and accurate set of search results.

K-Means Clustering

- **Input**: A set, ***V***, consisting of *n* points and a parameter *k*
- **Output**: A set ***X*** consisting of *k* points (*cluster centers*) that minimizes the squared error distortion *d(**V**,**X**)* over all possible choices of ***X***
- **K-means Clustering algorithm**
  1) Pick a number (K) of cluster centers
  2) Assign every data point (e.g., gene) to its nearest cluster center
  3) Move each cluster center to the mean of its assigned data points (e.g., genes)
  4) Repeat 2-3 until convergence

Advantages of K-mean clustering

1. K-mean clustering is straightforward and adaptable.

2. K-mean grouping calculation is straightforward and actualizes.

Disadvantages of K-mean clustering

1. In K-mean clustering client need to determine the quantity of number of cluster in advanced.

2. K-mean clustering calculation execution relies upon an initial centroids that why the calculation doesn't have ensure for optimal solution.

III. Various methods For The Calculation Of Initial Clusters In K Means Algorithm Are Given Below:

1. Forgy's Method:

Forgy's method involves choosing initial centroids randomly from the database. This approach takes advantage of the fact that if we choose points randomly we are more likely to choose a point near a cluster center by virtue of the fact that this is where the highest density of points is located

2. Simple Cluster Seeking Method:

This technique instates the primary seed with the main incentive in the database. It at that point computes the separation between the picked seed and the following point in the database, on the off chance that this separation is more noteworthy than some edge, this point is picked as the second seed, else it will move to the following occurrence in the database and repeat the procedure. When second seed is picked it will move to the following case in the database and ascertain the separation between this case and the two seeds previously picked, on the off chance that both these separations are more prominent than the edge, select the occurrence as the third seed. This procedure is repeated until the point .when K seeds are picked.

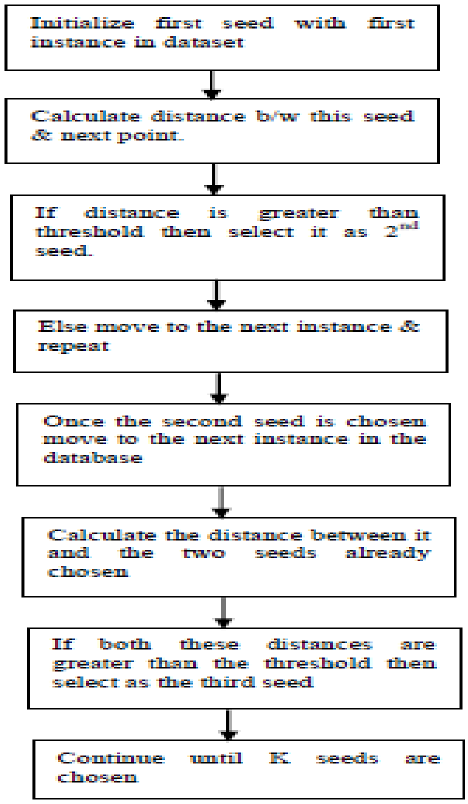


Fig 2: Simple Cluster Seeking Method

1. The upside of this technique is that it enables the client to control the separation between various clusters focuses.
2. As it may, the strategy additionally experiences a few restrictions which incorporate, the reliance of the technique.
3. Request of the focuses in the database, and, all the more fundamentally, the client must settle on the limit esteem.

3. KKZ Method:

In the underlying stage a point x is picked as the essential seed, this point is preferably at the edge of the data. By then the procedure finds a point most distant from x and this point will be the second seed. By then the system finds out the division of all concentrations in the dataset to the nearest of first and second seed. The third seed is the point which is the most far from its nearest seed. The path toward picking the furthest point from its nearest seed is repeated until the point that the moment that K seeds are picked.

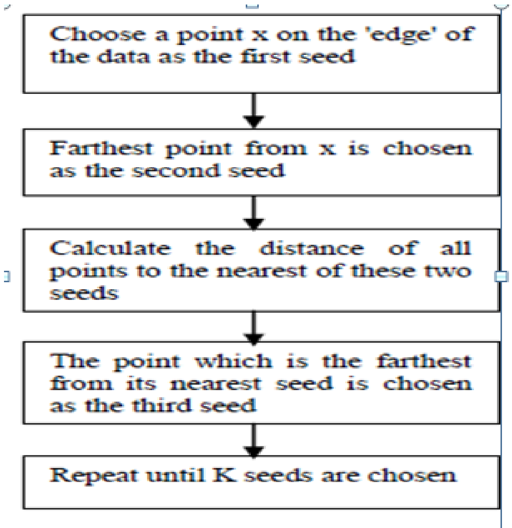


Fig 3: KKZ Method

4. Bradley and Fayyad's Method:

Bradley and Fayyad recommended another procedure for discovering introductory bunch centroids in K-implies calculation. In the initial step the information is separated into 10 subsets. In the second step K-implies calculation is connected on every one of the 10 subsets, the underlying centroids for these are picked utilizing Forgy's technique. The consequence of the 10 keeps running of the K-implies calculation is 10K focus focuses. These 10K focuses are then given as contribution to the K-implies calculation and the calculation run multiple times, every one of the 10 runs instated utilizing the K last centroid areas from one of the 10 subset runs. The outcome subsequently acquired is introductory group centroids for the K-implies calculation The primary preferred standpoint of the strategy is that it expands the proficiency of the outcome by the undeniable certainty that underlying centroids are gotten by various keeps running of the K-implies calculation. The real downside of this instatement strategy is that it requires a considerable measure of computational exertion In this strategy we need to run the K-implies calculation various quantities of times which builds the time taken by the technique to deliver the coveted outcome. Likewise, this technique requires more memory to store transitional consequences of various keeps running of K-implies. This makes the utilization of this strategy restricted to circumstances where computational time, space and speed does not make a difference.

Table 1: Comparison among existing methods

Method Name	Strengths	Weaknesses
Forgy's Method	• Simplest method. • Give quick results. • User do not have to supply any threshold value	• Randomness in choosing initial clusters gives extreme results
Simple Cluster Seeking Method	• Allow the user to control the distance b/w cluster centers.	• Dependency on the order of data points in the database. • The user has to supply a threshold value.
KKZ Method	• Simple method for choosing unique initial clusters. • Does not depend on any threshold value.	• Outliers pose a challenge to this method.
Bradley & Fayyad's Method	• Increases the efficacy of the result by running the K-means algorithm many times.	• It requires a lot of computational effort.

IV. Various swarm Intelligence (SI) Models:

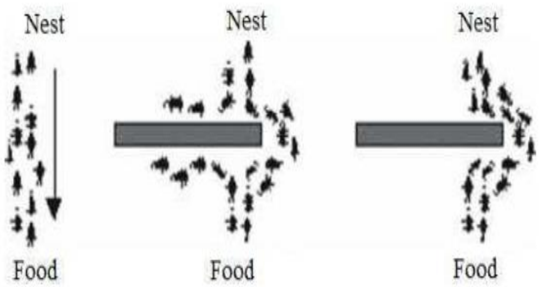
Swarm intelligence models are referred to as computational models inspired by natural swarm systems. To date, several swarm intelligence models based on different natural swarm systems have been proposed in the literature, and successfully applied in many real-life applications. Examples of swarm intelligence models are: Ant Colony Optimization Particle Swarm Optimization Artificial Bee Colony primarily focus on the most popular swarm intelligences models, namely, Particle Swarm Optimization.

1. Ant Colony Optimization (ACO) Model:

The first example of a successful swarm intelligence model is Ant Colony Optimization (ACO), which has been originally used to solve discrete optimization problems in the late 1980s. ACO draws inspiration from the social behavior of ant colonies. It is a natural observation that a group of **'almost blind'** ants can jointly figure out the shortest route between their food and their nest without any visual information. The following section presents some details about ants in nature, and shows how these relatively unsophisticated insects can cooperatively interact together to perform complex tasks necessary for their survival.

Foraging behaviour of ants is as follows:

1. Ants can't straightforwardly speak with one another; circuitous correspondence is called as stigmergy.
2. At the point when the ants discover their nourishment source they quickly returned close to the home on its way back it leaves a concoction substances called as pheromone. These pheromones are unstable in nature they continue dissipating.
3. Ants are fit for detecting this pheromone and the course is pulled in by different ants, they proceed onward a similar track. What's more, every subterranean insect leaves their concoction substances and thickness the track so that on the off chance that some other ants are in the source, they can pursue the pheromone thickness and discover their sustenance source.



: Ants' stigmergic behaviour in finding the shortest route between food and nest

Advantages of the Ant Colony Optimization

a. Have guaranteed covergence

b. Positive Feedback accounts for rapid discovery of good solutions.

Disadvantages of the Ant Colony Optimization

a. Sequences of random decisions (not independent)

b. Probability distribution changes by iteration.

c. Research is experimental rather than theoretical

d. Time to convergence uncertain (but convergence is guaranteed!).

2. Particle Swarm Optimization.

PSO is one of the swarm evolutionary optimization systems, PSO was introduced by creature's social connections, for example, winged creature people and fish swarm. In this technique, there is a swarm of particles that every one of particles demonstrates a practical answer for advancement issue. Each molecule endeavors to push toward definite arrangement by changing its way and advancing toward the best close to home involvement. In this segment, another agreeable calculation dependent on PSO and k-implies calculations are depicted.

PSO Strengths:

a. PSO is not only characterized by its fast convergence behaviour, but also by its simplicity.
b. PSO has an acquire potential to adjust to an evolving situation.

PSO has the following advantages and disadvantages

Advantages:

a. The algorithm can easily be implemented;
b. The global search of the algorithm is efficient,
c. The dependency on the initial solution is smaller,
d. It is a fast algorithm,
e. The algorithm has less parameter for tuning.

Disadvantages:

a. The algorithm has a weakness regarding local search,

b. It has a slow convergence rate,

c. It may get trapped in local minima for hard optimization problems.

```
for each Particle i
      initialize xi , vi
      Pi = xi
endfor
G = arg max_Pi J(Pi)
repeat
      for each Particle i
            update vi using Eq. (3)
            Check the velocity boundaries.
            update xi using Eq. (4)
            if J(xi) > J(Pbesti) then
                 Pi=xi
            endif
            if J(Pbesti) > J(G) then
                  Gbest=Pbesti
           endif
      endfor
      if PSO is converged then
           Execute k-means on Gbest
           if outcome of k-means is better than bulletin then
             bulletin = outcome of k-means
           endif
           Reinitialize PSO
           endif
until stopping criterion is met
```

- The algorithm keeps track of three global variables:
- Target value or condition.
- Global best (gBest) value indicating which particle's data is currently closest to the Target.
- Stopping value indicating when the algorithm should stop if the Target isn't found

TABLE III. COMPARISON OF K-MEANS ALGORITHM AND PSO ALGORITHMS BASED ON CLUSTERING CRITERIA

Criteria	K-means	PSO	Reason
Time/Space complexity	Best	Worst	In k-means there is only a simple initialization of cluster center and its similarity to other data points need to measure. Therefore, no need to perform complex operations (need to save Pbest, Gbest etc.). So time and memory requirement is less in k-means.
Efficiency	Not efficient	Efficient	There is a chance that K-means gets stuck in a local optimum and solution provide by K-means is not always an optimal solution.
Efforts required in implementation	Less	More	Because we need to handle lots of parameters such as inertia weight, acceleration coefficient, velocity components in PSO.
Sensitivity to outliers	Sensitive	Not sensitive	In K-means, there is a need to define the number of clusters in advance, so there is a possibility that these selected clusters contain an outlier
A prior specification of number of cluster	Need to specify	No need to specify	In K-means we have to first define a cluster center, and based on defined cluster center K-means find the similarity of each data point to the cluster center, and iterate until stopping criteria are met.
Effect of initial partition on final result	Sensitive to the initial partition	Not sensitive to initial partition	If we select different initial partition, then result is different in case of k-means, because it finds similarity based on initial selected cluster center, whichever data instances are more similar to the cluster's center, assign data instances to that particular cluster center. Whereas, PSO finds the position of optimum cluster centers automatically and then assign data to cluster based on similarity.

Genetic Algorithm

GA is a powerful technique of optimization area. it is an optimization tools used widely in solving problems based on natural selection & genetics. It is an adaptive heuristic search algorithm applied to solve the optimization problems.

Simple Genetic Algorithm

1. Selection of initial population of individuals
2. Fitness function evaluation: every individuals in the population is evaluated.
3. Iterate on this generation until termination criteria meet (noise population,fitness value)

a. Choose the best-fit individuals for reproduction
b. Produce new offspring through crossover k mutation operations
c. Evaluate the individual fitness of new individuals
d. Select new individuals as a replacement for least fit population

Advantages of Genetic Algorithm

1. GA provides a great flexibility to make an efficient implementation for a specific problem.
2. The greatest advantage of genetic algorithms is that the fitness function can be altered to change the behaviour of the algorithm.

PSO Advantages over GA:

1. The key advantage of PSO over GA is that it is algorithmically simpler, yet more robust and generally converges faster than GA In fact, the simplicity of PSO allowed scientists from different backgrounds, not necessarily related to computer science or programming skills, to use PSO as an efficient optimization tool to a wide-range of application domains.
2. PSO is more able to control convergence than GA.
3. Research is experimental rather than theoretical
4. Time to convergence uncertain (but convergence is guaranteed!)

2 HONEY BEES

Honey bees are one of the instances of swarm. Honey bee swarms are dynamic and insightful; they are equipped for separating different assignments among different honey bees. The exercises that the honey bees perform are rummaging, putting away, recovering and dispersing nectar, gathering dust, correspondence and adjusting to the adjustments in the earth in the aggregate way with no focal control. Scrounging in honey bees is not the same as that of ants. Honey bees arrange their provinces extremely well that there is no need of hibernation for it. Honey bees are constantly social and live respectively in states. Honey bees create nectar by gathering the dust and nectar from the blooms and store it in the honeycombs. Honey bees in nature are exceptionally sorted out, there are three gatherings of honey bees they are utilized honey bees and jobless honey bees. Jobless honey bees are passerby and scouts honey bees. Utilized honey bees are typically the accomplished honey bees and they go looking for nourishment source. These honey bees' moves arbitrarily starting with one bloom then onto the next and continue investigating different blossoms with the end goal to locate the best nourishment until the point that they are worn out. When the utilized honey bees discover the nourishment source with the rich of sustenance, they return and impart to the passerby honey bees through waggle move inside the hive Scout honey bees search for various sources or focuses on, this honey bees look through their source which are blooms dependent on different limitation and in the wake of finding the proper sources they need to show it to alternate honey bees in the hive, so they come back to the hive and there is something many refer to as move floor in the hive where these scout honey bees do waggle move, by this move they can convey to different honey bees about the area of the sustenance source. The working drones choose the best way of the source dependent on the move. And all the working drones go to the fitting area which is blossom and they gather the nectar from it andthen they store it inthe brush, the determination of the nourishment source depends on the amount of the sustenance and can likewise be the separation. This conduct of honey bees can be utilized for some streamlining issues and furthermore for the best way strategy.

1. The Foraging Behaviour of Honey Bees:A colony of honey bees can exploit a large number of food sources in big fields and they can fly up to 11 km to exploit food sources .

2. The Waggle Dance of Honey Bees:

The waggle dance is named based on the wagging run (in which the dancers produce a loud buzzing sound by moving their bodies from side to side), which is used by the scout bees to communicate information about the food source to the rest of the colony. The scout bees provide the following information by means of the waggle dance: the quality of the food source, the distance of the source from the hive and the direction of the source

The BA also has advantages and disadvantages

Advantages:

a. The algorithm has local search and global search ability

b. Implemented with several optimization problems

c .Easy to use

d. Available for hybridization combination with other algorithms.

Disadvantages:

A. Random initialization,

B. The algorithm has several parameters,

C. Parameters need to be tuned

Optimization Techniques in Improvement of K-means clustering

Technique	Type	Concept	Pros	Cons
PSO	Swarm Intelligence method	Optimization of non linear functions using methodology of particle swarms (i.e. bird flocking behavior).	1. Simple, easy and derivative free algorithm. 2. Efficient global search ability and efficient to handle complex nonlinear optimization problems.	1. For large search space, premature convergence to local optima. 2. Weak local search. SA
ACO	Swarm Intelligence Method	It used to discover the shortest trail to the source of food from the colony and return back by means of an indirect communication through pheromone.	1. Better performance (when compared to GA and SA). 2. Retains memory of entire colony instead of previous generation only. 3. Robust and also easy to accommodate with other algorithms.	1. For large number of nodes, computationally very difficult to solve. 2. Convergence is guaranteed, but time to convergence is uncertain. 3. Complicated coding. 4. Tradeoffs in evaluating convergence. Genetic
Genetic Algorithm	Aritifical Intelligence	Genetic algorithm is search heuristic usually applied in Optimization problems.	1. Concept is easy to understand 2. Supports multi-objective 3. Optimization Good for “noisy” environments	1. Over-fitting 2. Optimization Time

Table: Optimization Techniques in Improvement of K-means clustering

V. **Conclusion**:

We addressed the clustering problem in this paper. We proposed a method based on combination of the particle swarm optimization (PSO) and the k-mean algorithm. We showed that the combined method has the advantage of both PSO and k-means methods while does not inherent their drawbacks. As the PSO algorithm successfully searches all space during the initial stages of a global search .we have listed the advantages and disadvantages of applications of swarm models .to provide scope for future research.

References:

1. T. Velmurugan,and T. Santhanam, “A Survey of Partition based Clustering Algorithms in Data Mining: An Experimental Approach” Anexperimental approach Information. Technology. Journal, Vol, 10,No .3 , pp478-484,2011.
2. Ahmadyfard A, Modares H (2008) Combining PSO and k-meansto enhance data clustering. In: International symposium on telecommunications,pp 688–691
3. Angeline P J. (1999). Using selection to improve Particle Swarm Optimization.Proceedings of the 1999 Congress on Evolutionary Computation.Piscataway. NJ:IEEE Press,1999:84-89.
4. B. K. Panigrahi, Y. Shi, and M.-H. Lim (eds.): Handbook of Swarm Intelligence.Series: Adaptation, Learning, and Optimization, Vol 7, Springer-Verlag BerlinHeidelberg, 2011. ISBN 978-3-642-17389-9.
5. Ant Colony Optimization Algorithm for Solving Travelling Salesman Problem Krishna H. Hingrajiya, Ravindra Kumar Gupta, Gajendra Singh Chandel , University of Rajiv Gandhi Proudyogiki Vishwavidyalaya, Bhopal (M..)