A Survey on K-mean Clustering and Particle Swarm Optimization

Pritesh Vora, Bhavesh Oza

***Abstract*— In Data Mining, Clustering is an important research topic and wide range of unsupervised classification application. Clustering is technique which divides a data into meaningful groups. K-mean is one of the popular clustering algorithms. K-mean clustering is widely used to minimize squared distance between features values of two points reside in the same cluster. Particle swarm optimization is an evolutionary computation technique which finds optimum solution in many applications. Using the PSO optimized clustering results in the components, in order to get a more precise clustering efficiency. In this paper, we present the comparison of K-mean clustering and the Particle swarm optimization.**

***Index Terms*— Clustering, K-mean Clustering, Particle Swarm Optimization.**

I. INTRODUCTION

Clustering is a technique which divides data objects into groups based on the information found in data that describes the objects and relationships among them, their feature values which can be used in many applications, such as knowledge discovery, vector quantization, pattern recognition, data mining, data dredging and etc. [1] There are mainly two techniques for clustering: hierarchical clustering and partitioned clustering. Data are not partitioned into a particular cluster in a single step, but a series of partitions takes place in hierarchical clustering, which may run from a single cluster containing all objects to n clusters each containing a single object. And each cluster can have sub clusters, so it can be viewed as a tree, a node in the tree is a cluster, the root of the tree is the cluster containing all the objects, and each node, except the leaf nodes, is the union of its children. But in partitioned clustering, the algorithms typically determine all clusters at once, it divides the set of data objects into non-overlapping clusters, and each data object is in exactly one cluster. Particle swarm optimization (PSO) has gained much attention, and it has been applied in many fields [2]. PSO is a useful stochastic optimization algorithm based on population. The birds in a flock are represented as particles, and particles are considered as simple agents flying through a problem area. And in the multi-dimensional problem space, the particle's location can represent the solution for the problem. But the PSO may lack global search ability at the end of a run due to the utilization of a linearly decreasing inertia weight and PSO may fail to find the required optima when the problem to be solved is too complicated and complex. K-means is the most widely used and studied clustering algorithm. Given a set of n data points in real d-dimensional space (Rd), and an integer k, the clustering problem is to determine a set of k points in Rd, the set of points is called cluster centres, the set of n data points are divided into k groups based on the distance between them and cluster centres. K means algorithm is flexible and simple. But it has some limitation, the cluster result mainly depends on the selection of initial cluster centroids and it may converge to the local optima [3]. However, the same initial cluster centre in a data space can always generate the same cluster results, if a good cluster centre can always be obtained, the K-means will work well.

II. K-MEAN CLUSTERING

James MacQueen, the one who proposed the term "k-means"[4] in 1967. But the standard algorithm was firstly introduced by Stuart Lloyd in 1957 as a technique pulse-code modulation. The K-Means clustering algorithm is a partition-based cluster analysis method [5]. According to the algorithm we firstly select k objects as initial cluster centres, then calculate the distance between each cluster centre and each object and assign it to the nearest cluster, update the averages of all clusters, repeat this process until the criterion function converged. Square error criterion for clustering.

$$E = \sum_{i=1}^{k}\sum_{j=1}^{n_i}\|x_{ij} - m_i\|$$

x_{ij} is the sample j of i-class, mi is the center of i-class, ni is the number of samples i-class, Algorithm step are shown in the fig(1).

K-means clustering algorithm is simply described as follows:

Input: N objects to be cluster {x1, x2........xn}, the number of clusters k;

Output: k clusters and the sum of dissimilarity between each object and its nearest cluster center is the smallest;

- Arbitrarily select k objects as initial cluster centers (m1, m2... mk);
- Calculate the distance between each object xi and each cluster center, then assign each object to the nearest cluster, formula for calculating distance as:

$$d(x_i, m_i) = \sqrt{\sum_{j=1}^{d}(x_{i1} - m_{j1})}$$

i = 1,2........N

j = 1,2........k

d (xi, mj) is the distance between data i and cluster j;

- Calculate the mean of objects in each cluster as the new cluster centers,

Manuscript received on February 2013.

Pritesh Vora, Information Technology, Gujarat Technological University/ L.D. College of Engineering, Ahmedabad, India.

Prof. Bhavesh Oza, Computer Engineering Department Gujarat Technological University/ L.D. College of Engineering, Ahmedabad, India.

$$m_i = \frac{1}{N_i}\sum_{j=1}^{N_i} x_{ij}$$

i=l, 2 . . . k; N_i is the number of samples of current cluster i;

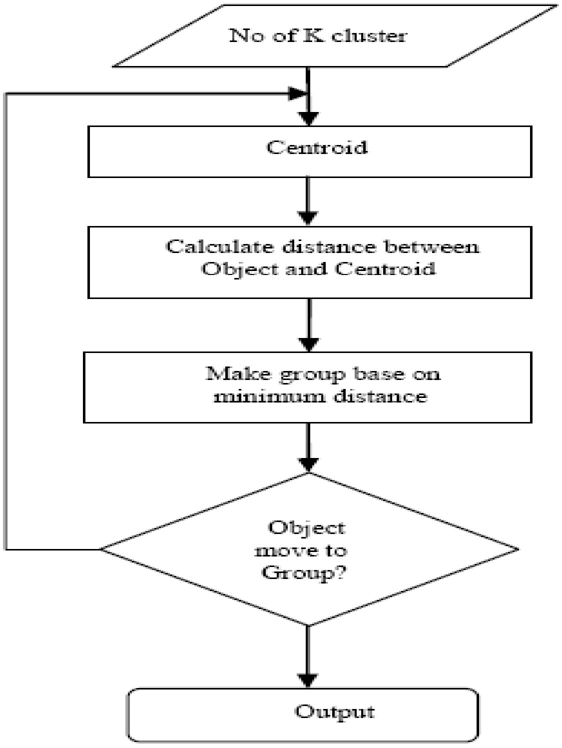

Fig.1 Flowchart of K-mean

III. PARTICLE SWARM OPTIMIZATION

PSO was introduced by Kennedy and Eberhart[6], it was based on the swarming behaviour of animals and human social behaviour. A particle swarm is a population of particles, in which each particle is a moving object which can move through the search space and can be attracted to the better positions. PSO must have a fitness evaluation function to decide the better and best positions, the function can take the particle's position and assigns it a fitness value. Then the objective is to optimize the fitness function. In general, the fitness function is pre-defined and is depend on the problem.

Each particle has own coordinate and velocity to change the flying direction in the search space. And all particles move through the search space by following the current optimum particles. Each particle consists of a position vector z, which can represent the candidate solution to the problem, a velocity vector v, and a memory vector pid, which is the better candidate solution encountered by a particle. Suppose the search space is n-dimensional, then the ith individual can be represented as:

Zi = {Zi1, Zi2,…….Zin}

Vi = {Vi1, Vi2,…….Vin}

i=1,2,3…………..n.

Where n is the size of swarm. The best previous experience of the ith particle is represented as:

pid_i = {pid_i1, pid_i2………..pid_in}

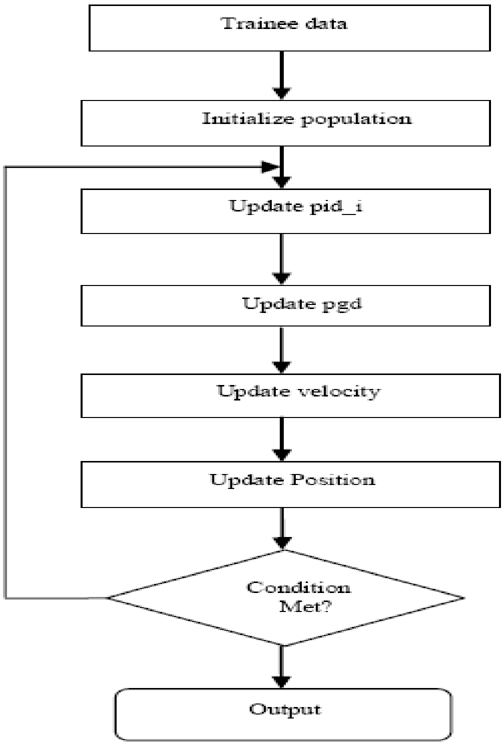

Fig.2 flowchart of PSO

Another memory vector pgd is used, which is the best candidate solutions encountered by all particles. The particles are then manipulated according to the following equations:

V_id(t+1)=wv_id(t)+η1rand(pid_i-Z_id(t))+η2rand(pgd-Z_id (t)),

Z_id (t+1)=Z_id (t)+V_id (t+1),

d= 1,2….n

Where w is an inertia weight, which is used for controlling the effect of previous history of velocities on current velocity, and controls the trade off between the local and global exploration abilities of the swarm. A small inertia weight facilitates local exploration, while a big one tends to facilitate global exploration. In order to get a better global exploration, w can be gradually decreased to get a better solution.η1 and η2 are two positive constants, rand is a uniformly generated random number. The equation shows that in calculating the next velocity for a particle, the previous velocity of the particle, the best location in the neighbourhood about the particle, the global best location all contribute some influence to the next velocity. Particle's velocities in each dimension can arrive to a maximum velocity vmax, which is defined to the range of the search space in each dimension. [3]The process of the PSO can be described as follows: First, It will initialize a population of particles with velocities and random positions in search space.

Secondly, for each particle i, update the position and velocity according to ,compute the fitness value according to the fitness function, update pid_i and pgd if necessary, repeat this process until termination conditions are met. Flowchart shown in Fig-2.

IV. ADVANTAGE AND DISADVANTAGE OF K-MEAN AND PSO

Advantages of K-mean clustering

- K-mean clustering is simple and flexible.
- K-mean clustering algorithm is easy to understand and implements.

Disadvantages of K-mean clustering

- In K-mean clustering user need to specify the number of cluster in advanced [7].
- K-mean clustering algorithm performance depends on a initial centroids that why the algorithm doesn't have guarantee for optimal solution [7].

Advantages of PSO

- PSO based on the intelligence and it is applied on both scientific research and engineering.
- PSO have no mutation and overlapping calculation. The search can be take place by the speed of the particle. Most optimist particle can able to transmit the information onto the other particles during the development of several generations, and the speed of researching is faster.[8]
- PSO accepts the real number code, and that is decided directly by the solution. Calculation in PSO is simpler and efficient in global search [8]

Disadvantages of PSO

- It is slow convergence in refined search stage and weak local search ability.
- The method cannot work on the problems of non-coordinate systems like the solution of energy field and the moving rules for the particles in the energy field.

V. CONCLUSION

Study of the k-mean clustering and Particle swam optimization we say that the k-mean which is depend on initial condition, which cause the algorithm may converge to suboptimal solution. On the other side Particle swarm optimization is less sensitive for initial condition due to its population based nature. So Particle swarm optimization is more likely to find near optimal solution.

ACKNOWLEDGMENT

Author of this paper is thankful to Asst. Prof. Bhvesh Oza, for providing his invaluable time to review the ideas and also answer the queries promptly. Author would also thank to Head of the Department, Prof. D. A. Parikh, their colleagues, friends, classmates, teachers and other guides for providing their immense support and helpful comments to improve this paper.

REFERENCES

1. A. Jain, M. Murty and P. Flynn, "Data Clustering: A Review", ACM Computing Surveys, Vol.31, No. 3, Sep 1999, pp. 264–323.
2. H. M. Feng, C.Y. chen and F. Ye, "Evolutionay fuzzy particle swarm optimization vector quantization learning scheme in image compression", Expert Systems with Applications. Vol. 32, No. 1, 2007, pp. 213-222.
3. Jinxin D. And Minyong Q., "A new Algorithm for clustering based on particle swarm optimization and k-Means", International Conference Intelligence,2009,pp 264-268.
4. Shalove Agarwal, Shashank Yadav and Kanchan Singh, "K-mean versus k-mean++ clustering Techniques", in IEEE 2012
5. Juntao Wang and Xiaolong Su, "An improved k-mean clustering algorithm", in IEEE, 2011, pp 44-46.
6. R. Eberhart and J. Kennedy, " Particle swarm optimization ", Proc. of the IEEE Int. Conf. on Neurad l Networks, Piscataway, NJ., 1995, pp. 1942–1948.
7. Garbriela derban and Grigoreta sofia moldovan, "A comparison of clustering techniques in aspect mining", Studia University, Vol LI, Number1, 2006, pp 69-78.
8. Qinghai B., "The Analysis of Particle Swarm Optimization Algorithm", in CCSE, February 2010, vol.3.

AUTHORS PROFILE

Pritesh Vora ,Information Technology, Computer Engineering Department, L.D. college of Engineering, Gujarat Technological University, Ahmedabad, Gujarat.

Prof. BHAVESH A. OZA, MTech in CE, BE in IT, works as Assistant Professor in Computer Engineering, Computer Engineering Department, L.D. college of Engineering, Gujarat Technological University, Ahmedabad, Gujarat.