Volume 2, No. 1, Jan-Feb 2011

ISSN No. 0976-5697

International Journal of Advanced Research in Computer Science

RESEARCH PAPER

Available Online at www.ijarcs.info

Comparative Analysis of Different Data Clustering Algorithms Based On Swarm Intelligence

Mr. Pankaj K. Bharne
ME Student
Sipna College of Engg. & Technology
Amravati
pankajbharne@gmail.com

Miss. Shweta K. Yewale *
ME Student
Prof. Ram Meghe Institute of Technology & Research,
Badnera
shwetayewale127@gmail.com

Mr. V. S. Gulhane
Assistant Professor
Sipna College of Engg. & Technology,
Amravati
v_gulhane@rediffmail.com

Abstract: For a decade swarm Intelligence is concerned with the design of intelligent systems by taking inspiration from the collective behaviors of social insects. Swarm Intelligence is a successful paradigm for the algorithm with complex problems. This paper focuses on the procedure of most successful methods of optimization techniques inspired by Swarm Intelligence: Ant Colony Optimization (ACO) and Particle Swarm Optimization (PSO). This paper also gives a comparative analysis of PSO and ACO for data clustering.

Keywords: Swarm Intelligence, Data Clustering, Comparative analysis of Data Clustering Algorithms.

I. SWARM INTELLIGENCE

Swarm intelligence (SI) is the collective behavior of decentralized, self-organized systems, natural or artificial. The concept is employed in work on artificial intelligence. The expression was introduced by Gerardo Beni and Jing Wang in 1989, in the context of cellular robotic systems. Swarm Intelligence is an artificial life (AI) technique that focuses on studying the collective behavior of a decentralized system made up by a population of simple agents interacting locally with each other and with the environment. [1]

A swarm has been defined as a set of agents which are liable to communicate directly or indirectly with each other, and which collectively carry out a distributed problem solving. The body can be understood as a swarm of cells and tissues which, unlike the swarms of bees or ants, stick relatively firmly together. However, the swarm of cells constituting a human body is a very different kind of swarm from that of the social insects. The body swarm is not built on ten thousand nearly identical units such as a bee society. [2].

Swarm prompted the design of very efficient optimization and clustering algorithms. The state of the art clustering algorithms based on SI tools are Particle Swarm Optimization (PSO) and Ant Colony Optimization (ACO). [3].

PSO is a new evolutionary computation technique first introduced by Kennedy and Eberhart in 1995 [4]. It is a stochastic optimization approach, modeled on the social behavior of animals such as a flock of birds, a school of fish, or a swarm of bees or a group of people who pursue common goal in their lives [5].

PSO is a population- based search procedure where the individuals, referred to as particles, are grouped into a swarm. Each particle in the swarm represents a candidate solution to the optimization problem. In a PSO system, each particle is “flown” through the multidimensional search space, adjusting its position in search space according to its own experience and that of neighboring particles. A particle therefore makes use of the best position encountered by itself and the best position of its neighbors to position itself toward an optimal solution. The effect is that particles “fly” toward an optimum, while still searching a wide area around the current best solution. The performance of each particle (i.e. the “closeness” of a particle to the global minimum) is measured according to a predefined fitness function which is related to the problem being solved.

The main strength of PSO is its fast convergence, which compares favorably with many global optimization algorithms like Genetic Algorithms (GA), Simulated Annealing (SA) and others. For applying PSO successfully, one of the key issues is finding how to map the problem solution into the PSO particle, which directly affects its feasibility and performance. Applications of PSO include function approximation, clustering, optimization of mechanical structures, and solving systems of equations.

ACO approach was proposed in 1992 by Marco Dorigo et al. to solve several discrete optimization problems [6][7][8]. ACO deals with artificial an system that is inspired from the foraging behavior of real ants, which are used to solve discrete optimization problems [9]. The main idea is the indirect communication between the ants by means of chemical pheromone trials, which enables them to find short paths between their nest and food.

II. DATA CLUSTERING ALGORITHMS BASED ON SWARM INTELLIGENCE

A. Data Clustering

Clustering can be considered the most important *unsupervised learning* problem; so, as every other problem of this kind, it deals with finding a *structure* in a collection of unlabeled data. A loose definition of clustering could be "the process of organizing objects into groups whose members are similar in some way". A *cluster* is therefore a collection of objects which are "similar" between them and are "dissimilar" to the objects belonging to other clusters. [10]We can show this with a simple graphical example in figure 1:

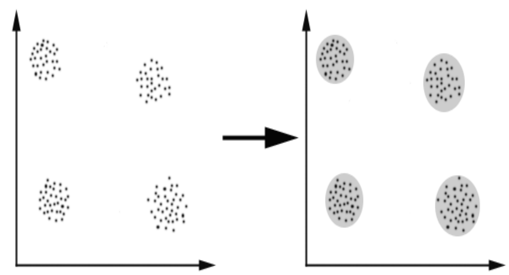

Figure 1. Graphical representation of clusters

Clustering is an important data-mining technique used to find data segmentation and pattern information. Clustering can be defined as the division of data into groups of similar objects. Each group, called cluster, consists of objects that are similar between themselves **and** dissimilar to objects of other groups**.** Clustering technique is widely used in application of financial data classification, spatial data processing, satellite photo analysis, and medical figure auto-detection etc. [11]

The main requirements that a clustering algorithm should satisfy are:

[a] scalability;
[b] Dealing with different types of attributes;
[c] Discovering clusters with arbitrary shape;
[d] Minimal requirements for domain knowledge to determine input parameters;
[e] Ability to deal with noise and outliers;
[f] Insensitivity to order of input records;
[g] High dimensionality;
[h] Interpretability and usability. [10]

Many clustering methods have been proposed. They can be broadly classified into four categories:

[a] partitioning methods,
[b] Hierarchical methods,
[c] density-based methods and
[d] Grid-based methods.

Other clustering techniques that do not fit in these categories have also been developed. These are fuzzy clustering, artificial neural networks and genetic algorithms. [12]

Swarm Intelligence is a relatively new interdisciplinary field of research, which has gained huge popularity in these days. Algorithms belonging to the domain, draw inspiration from the collective intelligence emerging from the behavior of a group of social insects (like bees, termites and wasps). When acting as a community, these insects even with very limited individual capability can jointly (cooperatively) perform many complex tasks necessary for their survival. Problems like finding and storing foods, selecting and picking up materials for future usage require a detailed planning, and are solved by insect colonies without any kind of supervisor or controller. An example of particularly successful research direction in swarm intelligence is Ant Colony Optimization (ACO), which focuses on discrete optimization problems**.** Particle Swarm Optimization (PSO) is another very popular SI algorithm for global optimization over continuous search spaces.

B. Ant Colony Optimization :

An artificial Ant Colony System (ACS) is an agent-based system, which simulates the natural behavior of ants and develops mechanisms of cooperation and learning. The basic idea of a real ant system is illustrated in Fig 2. In the left picture, the ants move in a straight line to the food. The middle picture illustrates the situation soon after an obstacle is inserted between the nest and the food. To avoid the obstacle, initially each ant chooses to turn left or right at random. Ants prefer to follow trails with larger amounts of pheromone, eventually all the ants converge to the shorter path around the obstacle, as shown in Figure 2.

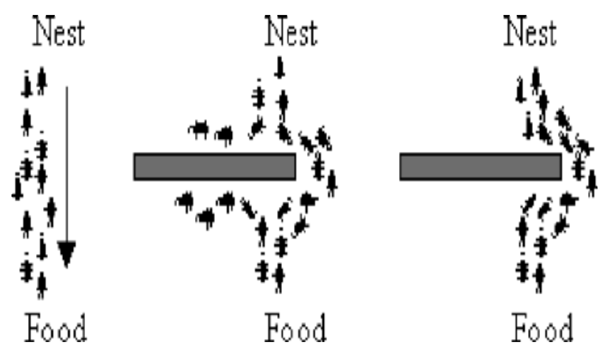

Figure 2. Illustrating the behavior of real ant movements.

ACS was proposed by Dorigo *et al.* (Dorigo and Gambardella, 1997) as a new heuristic to solve combinatorial optimization problems. This new heuristic, called Ant Colony Optimization (ACO) has been found to be both robust and versatile in handling a wide range of combinatorial optimization problems. The main idea of ACO is to model a problem as the search for a minimum cost path in a graph. Artificial ants as if walk on this graph, looking for cheaper paths. Each ant has a rather simple behavior capable of finding relatively costlier paths. Cheaper paths are found as the emergent result of the global cooperation among ants in the colony. The behavior of artificial ants is inspired from real ants. [13]

The ACO algorithm is employed to find an optimal order of traversal of the cities. Let τ be a mathematical entity modeling the pheromone and $\eta_{ij} = 1/r\ (i\ ,\ j)$ is a local heuristic. Also let allowed $k(t)$ be the set of cities that are yet to be visited by ant k located in city i. Then according to the classical ant system (Everitt, 1993) the probability that ant k in city i visits city j is given by:

$$p_{ij}^{k}(t) = \begin{cases} \dfrac{[\tau_{ij}(t)]^{\alpha}[\eta_{ij}]^{\beta}}{\sum_{h \in allowed_k(t)} [\tau_{ih}(t)]^{\alpha}[\eta_{ih}]^{\beta}} & if\ h \in allowed_k(t) \\ 0 & otherwise \end{cases} \quad (1)$$

Then once all ants have built their tours, pheromone is updated on all the ages as,

$$\tau(i,j) = (1-\rho).\tau(i,j) + \sum_{k=1}^{m} \Delta\tau_k(i,j) \quad (2)$$

From Equation (2), we can guess that pheromone updating attempts to accumulate greater amount of pheromone to shorter tours.

The ACO differs from the classical ant system in the sense that here the pheromone trails are updated in two ways. Firstly, when ants construct a tour they locally change the amount of pheromone on the visited edges by a local updating rule. Now if we let γ to be a decay parameter and. $\Delta\tau(\mathrm{i}, \mathrm{j}) = \tau_0$ such that τ_0 is the initial pheromone level, then the local rule may be stated as,

$$\tau(i,j) = (1-\gamma).\tau(i,j) + \gamma.\Delta\tau(i,j) \quad \textbf{(3)}$$

Secondly, after all the ants have built their individual tours, a global updating rule is applied to modify the pheromone level on the edges that belong to the best ant tour found so far. If · be the usual pheromone evaporation constant, D*gb* be the length of the globally best tour from the beginning of the trial and

$\Delta\tau^{/}(\mathrm{i}\ ,\ \mathrm{j}) = 1/\ \mathrm{D}_{gb}$ Only when the edge (*i; j*) belongs to global-best-tour and zero otherwise, then may we express the global rule as follows:

$$\tau(i,j) = (1-\kappa).\tau(i,j) + \kappa.\Delta\tau^{/}(i,j) \quad (4)$$

The main steps of ACO algorithm are presented in Algorithm 1.

[i] Algorithm 1 : Procedure of ACO

[a] Initialize pheromone trails;
[b] repeat *{*at this stage each loop is called an iteration*}*
[c] Each ant is positioned on a starting node
[d] repeat *{*at this level each loop is called a step*}*
[e] Each ant applies a *state transition rule like rule (1)* to incrementally build a solution and a *local pheromone-updating rule like rule (3);*
[f] until all ants have built a complete solution
[g] global pheromone-updating rule like rule (4) is applied.
[h] until terminating condition is reached

C. The Particle Swarm Optimization (PSO)

The concept of Particle Swarms, although initially introduced for simulating human social behaviors, has become very popular these days as an efficient search and optimization technique. The Particle Swarm Optimization as it is called now, does not require any gradient information of the function to be optimized, uses only primitive mathematical operators and is conceptually very simple. [13]

In PSO, a population of conceptual `particles' is initialized with random positions X*i* and velocities V*i*, and a function, *f*, is evaluated, using the particle's positional coordinates as input values. In an n-dimensional search space, X*i* = (x*i*1, x*i*2, x*i*3,...,x*in*) and V*i*= (v*i*1, v*i*2, v*i*3,...,v*in*)*:* Positions and velocities are adjusted, and the function is evaluated with the new coordinates at each time-step. The basic update equations for the d-th dimension of the i-th particle in PSO may be given as;

$$V_{id}(t+1) = \omega.V_{id}(t) + C_1.\varphi_1.(P_{lid} - X_{id}(t)) + C_2.\varphi_2.(P_{gd} - X_{id}(t))$$
$$X_{id}(t+1) = X_{id}(t) + V_{id}(t+1) \quad (4)$$

The variables Á1 and Á2 are random positive numbers, drawn from a uniform distribution and defined by an upper limit *Ámax;* which is a parameter of the system. C1 and C2 are called acceleration constants where as ω is called inertia weight. P*li* is the local best solution found so far by the i-th particle, while P*g* represents the positional coordinates of the fittest particle found so far in the entire community. Once the iterations are terminated, most of the particles are expected to converge to a small radius surrounding the global optima of the search space. The velocity updating scheme has been illustrated in Figure 3 with a humanoid particle.

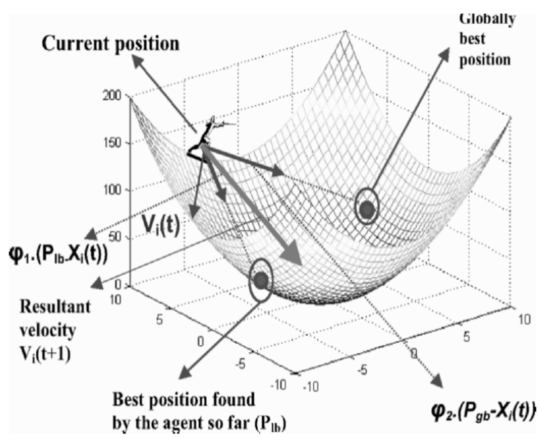

Figure3. Illustrating the velocity updating scheme of basic PSO

A pseudo code for the PSO algorithm is presented in Algorithm 2.

[i] Algorithm 2: Procedure for PSO

Input: Randomly initialized position and velocity of the particles: X*i* (0) and V*i*(0)
Output: Position of the approximate global optima X¤
[a] while terminating condition is not reached do
[b] for *i* = 1 to *numberofparticles* do
[c] Evaluate the ¯tness: =*f*(X*i*(*t*));
[d] Update P(*t*)andg(*t*);
[e] Adapt velocity of the particle using Equation 2;
[f] Update the position of the particle;
[g] end for
[h] end while

D. Procedure of Ant Colony algorithm for Data Clustering

[i] Algorithm 3: Procedure ACA

[a] Place every item X*i* on a random cell of the grid;
[b] Place every ant k on a random cell of the grid unoccupied by ants;
[c] Iteration count *Ã* 1;
[d] While iteration count < maximum iteration do
[e] For *i* = 1 to *no of ants* do
[f] If unladen ant and cell occupied by item X*i* then
[g] Compute *f*(X*i*) and *Ppick¡up*(X*i*);
[h] Else
[i] if ant carrying item xi and cell empty then
[j] Compute *f*(X*i*) and *Pdrop*(X*i*);
[k] Drop item X*i* with probability *Pdrop*(X*i*);
[l] End if
[m] End if
[n] Move to a randomly selected, neighboring and unoccupied cell;
[o] End for
[p] t *Ã* t + 1
[q] End while
[r] Print location of items;

E. Procedure of Particle Swarm Optimization Algorithm for Data Clustering

[i] Algorithm 4: The PSO Clustering Algorithm

1: Initialize each particle with K random cluster centers.
2: **for** iteration count = 1 to maximum iterations do
3: **for** all particles i do
4: **for** all pattern X*p* in the dataset do
5: calculate Euclidean distance of X*p* with all cluster centroids
6: assign X*p* to the cluster that have nearest centroid to X*p*
7: **end for**
8: calculate the fitness function *f*(Z*i;Mi*)
9: **end for**
10: find the personal best and global best position of each particle.
11: updating formula of PSO.
12: **end for**

III. APPLICATIONS OF DATA CLUSTERING ALGORITHMS

A. Applications of ACO

Ant colony optimization algorithms have been applied to many combinatorial optimization problems, ranging from quadratic assignment to fold protein or routing vehicles and a lot of derived methods have been adapted to dynamic problems in real variables, stochastic problems, multi-targets and parallel implementations.

It has also been used to produce near-optimal solutions to the travelling salesman problem. As a very good example, ant colony optimization algorithms have been used to produce near-optimal solutions to the travelling salesman problem.

ACO algorithms have been applied for following problems:
[a] Scheduling problem
[b] Vehicle routing problem
[c] Assignment problem
[d] Set problem
[e] Classification
[f] Connection-oriented network routing
[g] Connectionless network routing
[h] Data mining
[i] Discounted cash flows in project scheduling
[j] Distributed Information Retrieval
[k] Grid Workflow Scheduling Problem
[l] Image processing
[m] Intelligent testing system
[n] System identification
[o] Protein Folding
[p] Power Electronic Circuit Design

B. Applications of PCO

The first practical application of PSO was in the field of neural network training and was reported together with the algorithm itself (Kennedy and Eberhart 1995). To date, there are hundreds of publications reporting applications of particle swarm optimization algorithms.
Although PSO has been used mainly to solve:
[a] Unconstrained problem
[b] single-objective optimization problem
[c] Multi-objective optimization problems
[d] Problems with dynamically changing landscapes.
[e] Telecommunications
[f] Control
[g] Data mining
[h] Design,
[i] Combinatorial optimization
[j] Power systems
[k] Signal processing, and many others.

IV. COMPARATIVE ANALYSIS OF DATA CLUSTERING ALGORITHMS

Ant colony optimization algorithms have been applied to many combinatorial optimization problems, ranging from quadratic assignment to fold protein or routing vehicles and a lot of derived methods have been adapted to dynamic problems in real variables, stochastic problems, multi-targets and parallel implementations.

The practical application of PSO was in the field of neural network training and was reported together with the algorithm itself (Kennedy and Eberhart 1995). Many more areas of application have been explored ever since, including telecommunications, control, data mining, design, combinatorial optimization, power systems, signal processing, and many others.

The ACO and PSO can be analyzed for future enhancements such that new research could be focused to produce better solution by implementing the effectiveness and reducing the limitations of PSO. Plans to endow PSO with fitness sharing, aiming to investigate whether this helps in improving performance can be implemented in the evolutionary algorithms.

Both the ACO and PSO algorithm are the data clustering algorithms by implementing swarm behavior. Whereas the ACO is more applicable for problems where source and destination are predefined and specific. At the same time PSO is a clustering algorithm in the areas of mutli-objective, dynamic optimization and constraint handling. The ACO is more applicable for problems that require crisps results and PSO is applicable for problems that are fuzzy is nature. All these characteristic of the ACO and PSO are implicitly evident in the following applications.

A. Advantages and Disadvantages of Particle Swarm Optimization algorithm

[i] Advantages

[a] PSO is based on the intelligence. It can be applied into both scientific research and engineering use.
[b] PSO has no overlapping and mutation calculation. The search can be carried out by the speed of the particle. During the development of several generations, only the most optimist particle can transmit information onto the other particles, and the speed of the researching is very fast.
[c] The calculation in PSO is very simple. Compared with the other developing calculations, it occupies the bigger optimization ability and it can be completed easily.
[d] PSO adopts the real number code, and it is decided directly by the solution. The number of the dimension is equal to the constant of the solution.

[ii] Disadvantages

[a] The method easily suffers from the partial optimism, which causes the less exact at the regulation of its speed and the direction.
[b] The method cannot work out the problems of scattering and optimization.
[c] The method cannot work out the problems of non-coordinate system, such as the solution to the energy field and the moving rules of the particles in the energy field.

B. Advantages and Disadvantages of Ant colony Optimization algorithm

[i] Advantages

[a] Inherent parallelism
[b] Positive Feedback accounts for rapid discovery of good solutions
[c] Efficient for Traveling Salesman Problem and similar problems
[d] Can be used in dynamic applications (adapts to changes such as new distances, etc)

[ii] Disadvantages

[a] Theoretical analysis is difficult
[b] Sequences of random decisions (not independent)
[c] Probability distribution changes by iteration
[d] Research is experimental rather than theoretical
[e] Time to convergence uncertain (but convergence is guaranteed!)

V. CONCLUSION

Clustering remains an active field of interdisciplinary research till date. No single algorithm is known, which can group all real world datasets efficiently and without error.

The ACO differs from the classical ant system in the sense that here the pheromone trails are updated in two ways. Firstly, when ants construct a tour they locally change the amount of pheromone on the visited edges by a local updating role. Secondly, after all the ants have built their individual tours, a global updating rule is applied to modify the pheromone level on the edges that belong to the best ant tour found so far.

The ability of Particle Swarm Optimization (PSO), heuristic technique for search of optimal solutions based on the concept of swarm, to efficiently face classification of multiclass database instances. PSO is also tied to Evolutionary Computation, namely to Genetic Algorithms (GA) and to Evolutionary Programming.

VI. REFERENCES

[1] http://en.wikipedia.org/wiki/Swarm_intelligence
[2]http://www.molbio.ku.dk/MolBioPages/abk/PersonalPages/Jesper/Swarm.html
[3] O.A. Mohamed Jafar and R. Sivakumar, “Ant-based Clustering Algorithms: A Brief Survey”, International Journal of Computer Theory and Engineering, Vol. 2, No. 5, October, 2010.
[4] J. Kennedy and R. Eberhart, “Particle Swarm Optimization”, IEEE International Conf. on Neural Networks, Piscataway, NJ, Vol. 4, 1995, pp. 1942-1948.
[5] J. Kennedy and R. Eberhart, “Swarm Intelligence”, Morgan Kaufmann Publishers, Inc., San Francisco, CA, 2001.
[6] Marco Dorigo and Thomas Stutzle, “Ant Colony Optimization”, Prentice-Hall of India Private Ltd., New Delhi, 2006.
[7] G. DI Caro and M. Dorigo, “Extending AntNet for Best-effort Quality-of-Services Routing”, Ant Workshop on Ant Colony Optimization, http://iridia.ulb.ac.be/ants98/ants98.html, 1998, pp. 15-16.
[8] Marco Dorigo and Thomas Stutzle, “The Ant Colony Optimizaion Metaheuristic: Algorithms, Applications, and Advances”, In: F. Glover and G.Kochenberger (Eds.), Handbook of Metaheuristics, Kluwer Academic Publishers
[9] M. Dorigo, V. Maniezzo and A. Colorni, “Ant system: optimization by a colony of cooperating agents”, IEEE Transaction on Systems, Man and Cybernetics- Part B, 26(1), 1996, pp. 29-41.
[10]http://home.dei.polimi.it/matteucc/Clustering/tutorial_html/
[11] Zhuo Chen, Qing-Chun Meng, “AN INCREMENTAL CLUSTERING ALGORITHM BASED ON SWARM INTELLIGENCE THEORY”, Third International Conference on Machine Learning and Cybernetics, Shanghai, 26-29 August 2004.
[12] D.T. Pham, S. Otri, A. Afify, M. Mahmuddin and H. Al-Jabbouli, “Data Clustering Using the Bees Algorithm”, Proc. 40th CIRP International Manufacturing Systems Seminar 2007.
[13] Ajith Abraham, He Guo, and Hongbo Liu “Swarm Intelligence: Foundations, Perspectives and Applications”, Springer-Verlag Berlin Heidelberg2006.

