ISSN:0975-9646

Sagar Tiwari et al, / (IJCSIT) International Journal of Computer Science and Information Technologies, Vol. 4 (4) , 2013, 549 - 552

# Algorithm of Swarm Intelligence Using Data Clustering

Sagar Tiwari, Samta Gajbhiye
*Department of Computer Science & Engineering,*
*SSCET,Bhilai*

*Abstract:* **Problem statement: Data clustering has been applied in numerous fields such as machine learning, data mining, wireless sensor network and pattern recognition. One of the most famous clustering approaches is K-means which effectively has been used in many clustering problems, but this algorithm has some drawback such as local optimal convergence and sensitivity to preliminary points.**
*Approach*: **Particle Swarm Optimization (PSO) algorithm is one of the swarm intelligence algorithms, which is applied in determining the optimal cluster centers. In this study, a cooperative algorithm based on PSO and k-means is presented.**
*Result*: **The proposed algorithm utilizes both global search ability of PSO and local search ability of k-means. The proposed algorithm and also PSO, PSO with Contraction Factor (CF-PSO), k-means algorithms and KPSO hybrid algorithm have been used for clustering six datasets and their efficiencies are compared with each other.**
*Conclusion:* **Experimental results show that the proposed algorithm has an satisfactory efficiency and robustness.**

## 1. INTRODUCTION

Clustering is an unsupervised classification technique in which datasets that are often vectors in multi dimensional space, based on a similarity criterion, are divided into some clusters. Data clustering has vast application in data categorizat- ion (Memarsa deghi and Leary,2003),(Velmuruq- an and Santhanam, 2010), data compression (Celebi, 2011), data mining (Pizzuti and Talia, 2003), pattern recognition (Wong and Li, 2008), compacting (Marr, 2003), machine learning image segmentation and Data clustering importance in various sciences causes the introduction of various methods of data clustering(Hartigan, 1975). When used on a set of objects, which have attributes that characterize them, usually represented as vectors in a multi- dimensional space are grouped into some clusters. When the predefined clusters number is K and there are N m-dimensional data, clustering algorithm would assign each of these data to one of the clusters, such that assigned data to a cluster with respect to a specific criterion are more similar to each other than data in other cluster.

The k-means clustering algorithm was developed by Hartigan (1975) which is one of the earliest and simplest clustering approaches that has been ever widely used. K-means method starts with K cluster centers and divides a set of objects into K subsets. This is one of the most famous and applied clustering techniques since it can be easily under stood and implemented and its time complexity is linear. But k means method has major weaknesses. One of these weaknesses is extra sensitivity to initial values of cluster centre. Objective function of k-means has multiple local optimums and k-means method is not capable to guarantee to pass local optimums. Therefore, if initial position of cluster centers in problem space was chosen in appprivately , this could converge to a local optimum. Data clustering is of NP problems. One of the most Applied methods for finding suitable solution for these kinds of NP problems belongs to swarm Intelligence algorithms. Particle Swarm Optimization (PSO) is one of the most famous swarm intelligence algorithms, which was presented by Kennedy and Eberhart (1995). This algorithm is an effective technique for solving optimization problems that works based on probability rules and population. So far, different PSO-based methods for solving data clustering problem have been presented (Esmin *et al*., 2008; Kao and Lee, 2009; Tsai and Kao, 2010). Presented a hybridized algorithm based on k-means methods and PSO, called KPSO in (Merwe and Engelbrecht, 2003). In KPSO, first, k-means method is executed and then, outcome of k-means is used as one of the particles in initial solution of PSO. Therefore, first in this method, high convergence rate of k-means is used and after k-means converges, PSO is applied for exiting from local minimums and improving the result of k-means. In this study, a cooperative algorithm is proposed based on PSO and k-means. In the proposed algorithm, first, particles perform optimization process in PSO. After particle swarm convergence, obtained cluster centers by particles are used as initial cluster centers of k-means algorithm. After forwarding PSO's output to k-means, particles are reinitialized and performs clustering again. In fact, in the proposed algorithm, PSO is used for a global search and k-means is used for a local search. The proposed algorithm and also k-means, PSO, CF-PSO (Eberhart and Shi, 2000) and KPSO algorithms are applied for clustering 6 real datasets iris, glass, wine, sonar, pima and WDBC. Comparing obtained results from experiments shows an acceptable efficiency of the proposed algorithm.

Large sets of text documents are now increasingly common. For example, the World-Wide-Web contains nearly 1 billion pages and is growing rapidly (www.alexa.com), the IBM Patent server consists of more than 2 million patents (www.patents.ibm.com), the Lexis-Nexis databases contain more than 2 billion documents (www.lexisnexis.com). Furthermore, an immense amount of text data exists on private corporate intranets, archives of media companies, and in scientifc and technical publishing houses. In this context, applying machine learning and statistical algorithms such as clustering, classification, principal component anal ysis, and discriminant analysis to text data sets is of great practical interest. In this paper, we focus on clustering of text data sets.

### 2.AN INTRODUCTION TO SWARM INTELLIGENCE

The behavior of a single ant, bee, termite and wasp often is too simple, but their collective and social behavior is of paramount significance. A look at National Geographic TV Channel reveals that advanced mammals including lions also enjoy social lives, perhaps for their self-existence at old age and in particular when they are wounded. The collective and social behavior of living creatures motivated researchers to undertake the study of today what is known as *Swarm Intelligence*. Histo- rically, the phrase Swarm Intelligence (SI) was coined by Beny and Wang in late 1980s (Beni and Wang, 1989) in the context of cellular robotics. A group of researchers in different parts of the world started working almost at the same time to study the versatile behavior of different living creatures and especially the social insects. The efforts to mimic such behaviors through computer simulation finally resulted into the fascinating field of SI. SI systems are typically made up of a population of simple agents (an entity capable of perf- orming/executing certain operations) inter- acting locally with one another and with their environment. Although there is normally no centralized control structure dictating how individual agents should behave, local interactions between such agents often lead to the emergence of global behavior. Many biological creatures such as fish schools and bird rocks clearly display structural order, with the behavior of the organisms so integrated that even though they may change shape and direction, they appear to move as a single coherent entity (Couzin *et al*., 2002).

### 3.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 effcients earch and optimization technique. The Particle Swarm Optimization (PSO) (Kennedy and Ebe- rhart, 1995,Kennedy *et al*., 2001), 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. In PSO, a population of conceptual `part- icles' 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 ve-locities 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 particle in PSO may be given as

*Vid*(*t* + 1) = *!:Vid*(*t*) + *C*1*:'*1*:*(*Plid ¡ Xid*(*t*)) +*C*2*:'*2*:* (*Pgd ¡ Xid*(*t*))

*Xid*(*t* + 1) = *Xid*(*t*) + *Vid*(*t* + 1)

The variables *Á*1 and *Á*2 are random positive numbers, drawn from a uni-form distribution and defned by an upper limit *Ámax;*which is a parameter of the system. C1 and C2 are called acceleration constants whereas *!* 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

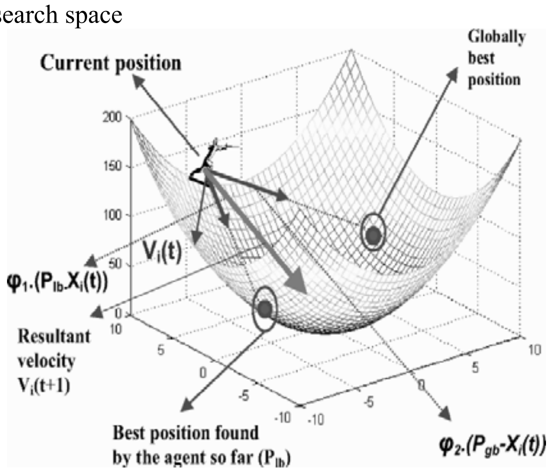


Fig1. Illustrating the velocity updating scheme of basic PSO.

**Algorithm : The PSO Algorithm**

**Input**: Randomly initialized position and velocity of the particles: X*i*(0) and V*i*(0)

**Output**: Position of the approximate global optima X¤

1: while terminating condition is not reached do
2: for *i* = 1 to *number of particles* do
3: Evaluate the fitness: =*f*(X*i*(*t*));
4: Update P(*t*) and g(*t*);
5: Adapt velocity of the particle using Equation 3;
6: Update the position of the particle;
7: end for
8: end while

Let P = *f*P1, P2... P*ng* be a set of n patterns or data points, each having d features. These patterns can also be represented by a profile data matrix X*n£d* having n d-dimensional row vectors. The i-th row vector X*i* characterizes the i-th object from the set P and each element X*i;j* in X*i* corresponds to the j-th real value feature (*j* = 1*;* 2*; :::::; d*) of the i-th pattern ( i =1,2,...., n). Given such an X*n£d;a* partitional clustering algorithm tries to find a partition C = *f*C1, C2,......, C*Kg* of *K* classes, such that the similarity of the patterns in the same cluster is maximum and patterns from different clusters differ as far as possible. The partitions should maintain the following properties:

1. Each cluster should have at least one pattern assigned i.e. *Ci 6= c#*
2. Two different clusters should have no pattern in common. i.e. *Ci \ Cj = c,8i 6= j* and *i; j 2 f*1*;* 2*; :::;Kg*. This property is

required for crisp (hard)clustering. In Fuzzy clustering this property doesn't exist.
3. Each pattern should definitely be attached to a cluster i.e.$K$S$i$=1$Ci = P$.

## 4. MATERIALS AND METHODS

**K-means algorithm:** Clustering in D-dimensional Euclidean space is a process in which a set of N members, based on a similarity criterion, is divided into K groups or clusters. Various clustering methods are represented so far. The base of clustering algorithms is measuring the similarity between data and it is determined how much similar these two data vectors are, by a function. K-means algorithm is one of the oldest and most famous clustering methods. This method sorts data vectors in D-dimensional space in clusters, which their number was determined before, this clustering is based on Euclidean distance between data and cluster center which are considered as similarity criterion. Euclidean distance between data vectors of a cluster with the center of that cluster is less than their Euclidean distance with other cluster centers. Standard k-means algorithm is as below: Initial positions of K cluster centers are determined randomly. Following phases are repeated:
For each data vector: the vector is allocated to a cluster which its Euclidean distance from its center is less than the other cluster centers. The distance to cluster center is calculated by Eq. 1

$$\text{Dis(X,Z)}= \sqrt{\sum_{i=1}^{D}(X-z)^2}$$

In Eq. (1), Xp is pth data vector, Zj is jth cluster center and D is the dimension of data and cluster center. b) Cluster center are updated by Eq.

$$Z_{j=}\frac{1}{n}\sum_{x\neq c_j} x$$

## 5. PARTICLE SWARM OPTIMIZATION ALGORITHM:

PSO is one of the swarm intelligence methods and evolutionary optimization techniques, which was proposed by Kennedy and Eberhart (1995). PSO was presented according to animals social inter- actions such as bird folk and fish swarm. In this method, there is a swarm of particles that each of particles shows a feasible solution for optimization problem. Every particle tries to move toward final solution by adjusting its path and moving toward the best personal experience and also the best swarm experience.
Suppose that the population size is N. For particle i (1□□i □□N) in D-dimension space, current position is xi = (xi1 , xi2 , … , xiD)
and velocity is xi = (vi1 , vi2 , … , viD).
During optimization process, velocity and position of each particle at each step is updated by Eq.3 and 4

$$v_{i,j}(t+1)=Wv_{i,j}(t)+c_1R_{i,j}^1(Pbest_{i,j}(t)-X_{i,j}(t)) +C_2R_{i,j}^2(Gbest_j(t) - X_{i,j}(t))\ldots\ldots(3)$$

$$X_{i,j}(t+1)=X_i(t)+V_i(t+1)\ldots\ldots \quad (4)$$

where, xi,j is the component j of particle i, c1 and c2 are acceleration coefficients and w is inertia weight that can be a constant number or a positive function (Shi and Eberhart, 1998). R is a random number with uniform distribution in interval [0, 1]. Pbesti (t) is the best position that is found by particle i until time t (the best individual experience of particle i) and Gbest (t) is the best position that until time t is found by whole swarm's members (the best swarm experience). At each iteration, the best individual experience of particle i is given by Eq. 5:

$$Pbeat_i(t+1)= \{Pbest_i(t), if\ f(x_i(t+1)\geq f(pbest_i(t))$$
$$\{Pbest_i(t), if\ f(x_i(t+1)< f(pbest_i(t))\}$$

where, f(x) is the fitness value of vector x. The best swarm experience is given by Eq. 6

$$\text{Gbest(t+1)} = \arg \min f(Pbest_i(t+1), 1\leq i\leq N)$$

Clerc presented another version of PSO in which by using construction factor (CF-PSO), PSO conv- ergence rate has been improved. In this version of PSO, particles velocity is updated by Eq. 7:

$$V_{i,j}(t+1)=\chi\{v_{i,j}(t)+c_1r_{i,j}^1(Pbest_{i,j}(t)- \quad x_L+c_2R_{i,j}^2 (Gbest_j(t) - X_{i,j}(t))\}$$

the appropriate value of χ is 0.729843788 and c1=c2=2.05 (Eberhart and Shi, 2000). According to how particles move in PSO, particles may leave search space, which leads to decrease efficiency and algorithm convergence rate. To remove this problem, some constraints are considered for velocity components' values. For this reason, in each of iterations, after computing velocity by Eq. 3, all of its components' values would be considered in various dimensions. The value of each velocity vector component can be clamped to the range [-Vmax,Vmax] to reduce the likelihood of particles leaving the search space. The value of Vmax is usually chosen to be K□Xmax (Here, Xmax is the length of changes interval in search space dimensions), with 0.1□K□1 (Bergh and Engelbrecht, 2004). To find the optimal cluster centers, PSO algorithm applies Eq. 8 as the fitness function (Tsai and Kao, 2010). Eq. 8 shows generating function of Sum of Intra Cluster Distances (SICD) which is one of the most known evaluating criteria for clustering data. Less value of SICD is higher quality the clustering is performed. Therefore, for data clustering, PSO algorithm should minimize the fitness function

## 6. PROPOSED ALGORITHM

In this section, a new cooperative algorithm based on PSO and k-means algorithms is described. The purpose of designing the proposed algorithm is to take advantages of both algorithms and remove their weaknesses. K-means is of high convergence rate, but it's very sensitive to initializing the cluster centers and in the case of selecting inappropriate initial cluster centers, it could converge to a local optimum. PSO can pass local optima to some extent but cannot guarantee reaching to global optima. However, PSO's computational comp- lexity for data clustering is much more than k-means. How the proposed algorithm functions remove weaknesses of these two algorithms and apply their advantages is as following:

In the proposed algorithm, first, the particles are initialized in PSO. Each of particles contains K cluster centers which are displaced in the problem space by performing PSO algorithm. PSO continues to perform until the particles converge. After convergence of PSO, Gbest position including the best cluster centers which have found by particles so far is considered as the input of k-means. Then, k-means algorithm starts working and while it is not converged, it continues working. Therefore, PSO searches globally and as far as it can, it passes local optima. After convergence of PSO's particles, PSO's output would have an appropriate initial cluster centers for k-means

```
Coop-KPSO:
for each Particle i
    initialize x_i , v_i
    P_i = x_i
endfor
G = arg max_{P_i} J(P_i)
repeat
    for each Particle i
        update v_i using Eq. (3)
        Check the velocity boundaries.
        update x_i using Eq. (4)
        if J(x_i) > J(Pbest_i) then
            P_i=x_i
        endif
        if J(Pbest_i) > J(G) then
            Gbest=Pbest_i
        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
```

Fig2. Pseudo code of proposed algorithm

algorithm starts searching locally. Consequently, in the proposed algorithm, global search ability of PSO has been used and after converging, a great part of optimization process will be given to k-means to utilize high capability of local search of this algorithm and its high convergence rate. Since initial cluster centers for k-means are obtained by PSO and k-means is used for local search, k-means weakness of sensitivity to initial cluster centers is removed. But, PSO capability may not be enough for preventing from being trapped in local optima. If this algorithm is trapped in local optima, it cannot present proper initial cluster values to k-means. Thereafter, according to low ability of k-means in passing local optima, the obtained result cannot be acceptable. To raise this problem, after convergence of PSO, the output of this algorithm is sent to k-means. Simultaneously with starting of k-means, PSO's particles are initialized and start global search again. In fact, in one time of executing the proposed algorithm, PSO has many times of chance to perform an acceptable global search. It should be noted that in the proposed algorithm, in each time of executing PSO, particles just search globally and converge after a short time and k-means undertakes the remaining of optimization process which is local search. Therefore, with respect to low computational complexity of k-means, huge amount of computations for local search is prevented.

In the proposed algorithm, it has been tried to utilize this conserved computation load for giving new opportunities to PSO in order to perform an acceptable global search in at least one of given opportunities to it. Hence, for each execution of global search by PSO, k-means is also performed once.

## CONCLUSION

In this study, a new cooperative algorithm based on k-means and PSO is presented. In the proposed algorithm, PSO performs global search and k-means is responsible for local search. The process of the proposed algorithm is such that the strength and ability of preventing from being trapped in local optimums is improved. The proposed algorithm along with four other algorithms is used for clustering 6 standard datasets and obtained results are compared with each other. Experimental results show that the proposed algorithm is of higher robustness and better efficiency to other tested algorithms. To improve the obtained results of the proposed algorithm, it can increase local search ability around the best found position by the algorithm. This is issue that merits further research.

## REFERENCES


[1]Bergh, F.V.D. and A.P. Engelbrecht, 2004. cooperative approach to particle swarmoptimization. IEEE Trans. Evolutionary Comput., 8: 225-239. DOI: 10.1109/TEVC.2004.826069

[2]Celebi, M.E. 2011. Improving the performance of kmeansfor color quantization. J. Image VisionComput. 29: 260-271. DOI:10.1016/j.imavis.2010.10.002

[3]Eberhart, R.C. and Y. Shi, 2000. Comparing inertiaweights and constriction factors in particle swarmoptimization. Proceedings of the Congress Evolutionary Computation, July 16-19, IEEE Xplore Press, La Jolla, CA , USA., pp: 84-88. DOI: 10.1109/CEC.2000.870279

[4]Esmin, A.A.A., D.L. Pereira and F. de Araujo, 2008.Study of different approach to clustering data by using the particle swarm optimization algorithm.Proceedings of the IEEE Congress on Evolutionary Computation, June 1-6, IEEE Xplore Press, Hong Kong, pp: 1817-1822. DOI: 10.1109/CEC.2008.4631035

[5] Shi, Y. and R. Eberhart, 1998. A modified particle swarm optimizer. Proceedings of the IEEE International Conference Evolutionary Computation, May, 4-9, IEEE Xplore Press, Anchorage, AK , USA., pp: 69-73. DOI: 10.1109/ICEC.1998.699146

[6]Tsai, C.Y. and I.W. Kao, 2010. Particle Swarm Optimization with Selective Particle Regeneration for Data Clustering. Expert Syst. Appl. 38: 6565- 6576. DOI: 10.1016/j.eswa.2010.11.082

[7] Velmuruqan, T. and T. Santhanam, 2010.Computational complexity between K-Means and K-medoids clustering algorithms for normal and uniform distributions of data points. J. Comput. Sci., 6: 363-368. DOI:10.3844/jcssp.2010.363.368

[8]Vannoorenberghe, P. and G. Flouzat, 2006. A beliefbased pixel labeling strategy for medical and satellite image segmentation. Proceedings of the IEEE International Conference on Fuzzy System, Juley 16-21, IEEE Xplore Press, Vancouver, BC., pp: 1093-1098. DOI: 10.1109/FUZZY.2006.1681846

[9]Wong, A.K.C. and G.C.L. Li, 2008. Simultaneous pattern and data clustering for pattern cluster analysis. IEEE Trans. Knowledge Data Eng., 20: 911-923. DOI: 10.1109/TKDE.2008.38