Survey on Particle Swarm Optimization Based Clustering Analysis

Veenu Mangat

University Institute of Engineering and Technology, Panjab University 160014, Chandigarh, India
veenumangat@yahoo.com

Abstract. Clustering analysis is the task of assigning a set of objects to groups such that objects in one group or cluster are more similar to each other than to those in other clusters. Clustering analysis is the major application area of data mining where Particle Swarm Optimisation (PSO) is being widely implemented due to its simplicity and efficiency. When compared with techniques like *K*-means, Fuzzy *C*-means, *K*-Harmonic means and other traditional clustering approaches, in general, the PSO algorithm produces better results with reference to inter-cluster and intra-cluster distances, while having quantization errors comparable to the other algorithms. In recent times, many hybrid algorithms with PSO as one of the techniques have been developed to harness the strong points of PSO and increase its efficiency and accuracy. This paper provides an extensive review of the variants and hybrids of PSO which are being widely used for the purpose of clustering analysis.

Keywords: Clustering Analysis, Particle Swarm Optimization, Hybrid Methods.

1 Introduction

This section gives a brief introduction on clustering analysis and the application of PSO for clustering analysis. The amount of information available and collected nowadays is beyond the human capability of analysing and extracting relevant information or discovering knowledge from it. Such data is heterogeneous, uncertain, dynamic and massive. It is of great significance to explore how to automatically extract the implicit, unknown and potentially helpful information so that it can help in the commercial decision-making activities. This is precisely the task of data mining and knowledge discovery from databases. A dramatic increase in the amount of information requiring in depth analysis has led to the design of new techniques that can perform knowledge extraction efficiently and automatically.

1.1 Clustering Analysis

Clustering analysis is an important technique used in data mining. It involves grouping together similar multi-dimensional data vectors into a number of clusters.

L. Rutkowski et al. (Eds.): SIDE 2012 and EC 2012, LNCS 7269, pp. 301–309, 2012.

The main objective of clustering is to minimize inter-cluster similarity and to maximize intra-cluster similarity [1]. Clustering techniques are basically divided into two types:

- *Hierarchical*: This approach provides a series of nested partitions of the dataset. It divides the data into a nested tree structure where the levels of the tree show similarity or dissimilarity among the clusters at different levels. It is further divided into 'Agglomerative' and 'Divisive' approaches. The divisive approach splits one large cluster into different sub clusters e.g. CHAMELEON [2], BIRCH [3]. In agglomerative approach, the clustering process starts with every data element in individual clusters which are then merged on the basis of their proximity until all data elements are finally in a single cluster e.g. CURE [4] and ROCK [5]. This approach does not need the number of clusters to be specified in advance. It is deterministic and has lower execution time efficiency than partitioning techniques.

- *Partitioning*: In contrast to hierarchical technique which yields a successive level of clusters by iterative fusions or divisions, this technique assigns a set of objects to clusters with no hierarchical structure. These methods try to minimize certain criteria, like square error function. These methods are further divided into 'supervised' and 'unsupervised' algorithms. The supervised algorithms are provided with both the cases (data points) and the concept to be learnt for each case. Common algorithms include K-means and its variants like Fuzzy c-means, Spherical K-Means etc.

1.2 Particle Swarm Optimization

PSO is a technique based upon Swarm Intelligence (SI); an artificial intelligence paradigm for solving optimization problems that originally took its inspiration from the biological examples such as in swarming, flocking and herding phenomena in vertebrates. Particle Swarm Optimization (PSO) incorporates swarming behaviour observed in flocks of birds, schools of fish, or swarms of bees, and even human social behaviour. It is a population-based optimization tool, which could be implemented and applied easily to solve various function optimization problems, or the problems that can be transformed to the function optimization problem. 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. Many evolutionary techniques based on Particle Swarm Optimization have also been developed for unsupervised method.

2 Related Work

2.1 Original Version (Early Developments)

PSO was first introduced by J. Kennedy and Eberhart in 1995 [6]. They developed this method for optimization of continuous non-linear functions. A '*swarm*' refers to a collection of a number of potential solutions where each potential solution is known

as a '*particle*'. In the standard PSO method, each particle is initialized with random positions X_i and velocities V_i, and a function, *f* (*fitness function*) is evaluated. The aim of PSO is to find the particle's position that gives the best evaluation of a given fitness function using the particle's positional coordinates as input values. In a k-dimensional search space, $X_i = (x_i1, x_i2, x_i3,...,x_ik)$ and $V_i = (v_i1, v_i2, v_i3,...,v_ik)$. Positions and velocities are adjusted, and the function is evaluated with the new coordinates at each step. In each generation, each particle updates itself continuously by following two extreme values: the best position of the particle in its neighbourhood (*lbest or localbest or personalbest*) and the best position in the *swarm* at that time (*gbest or globalbest*) [7]. After finding the above values, each particle updates its position and velocity as follows:

$$v_{i.k}(t + 1) = wv_{i.k}(t) + c_1r_{1.k}(t) (y_{i.k}(t) - x_{i\,.\,k}(t))+ c_2r_{2.k}(t) (y`_k(t) – x_{i.k}(t)) \quad (1)$$

$$x_i (t+1) = x_i(t) + v_i(t+1) \quad (2)$$

Where: $v_{i.k}$ is the velocity of the *i-th* particle in the *t-th* iteration of the *k-th* dimension; $x_{i.k}$ is the position of the *i-th* particle in the *t-th* iteration of the *k-th* dimension; r_1 and r_2 are random numbers in the interval [0, 1]; c_1 and c_2 are learning factors, in general, $c_1=c_2=2$. An improvement in these parameters and their optimized values have been done in recent papers by the researchers which will be discussed later. '*w*' is the inertia weight factor generally selected in the range (0.1, 0. 9). This parameter was introduced in [8] which illustrated its significance in the particle swarm optimizer. Equation (1) is used to calculate the particle's new velocity according to its previous velocity and the distances of its current position from its own best experience and the group's best experience. The velocity is thus calculated based on three contributions:

- A fraction of the previous velocity.
- The cognitive component which is a function of the distance of the particle from its personal best position.
- The social component which is a function of the distance of the particle from the best particle found thus far (i.e. the best of the personal bests).

The personal best position y_i of particle 'i' can be computed as:

$$y_i (t+1) = y_i (t) \quad \text{if } f(x_i(t+1)) >= f(y_i(t))$$

$$\text{or } y_i (t+1) = x_i(t+1) \quad \text{if } f(x_i(t+1) < f(y_i(t)) \quad (3)$$

Equation (1) reflects the *gbest* version of PSO whereas in the lbest version the swarm is further divided into overlapping neighbourhoods and the best particle in each neighbourhood is determined. For the *lbest* version the social component of (1) changes to:

$$C_2r_{2.k}(t)(y`_{j.k}(t)-x_{i.k}(t)) \quad (4)$$

Where: $y`_j$ is the best in the neighbourhood of *i-th* particle. The particle flies towards a new position according to equation (2). The PSO is usually executed with repeated application of equations (1) and (2) until a specified number of iterations have been exceeded or when the velocity updates are close to zero over a number of iterations.

2.2 PSO Clustering

In [7] the authors have used the first kind of hybrid PSO technique for data clustering by hybridizing PSO with the popular K-means algorithm. The k-means algorithm has been used to seed the initial swarm and PSO is then used to refine the clusters formed by K-means. In this algorithm a single particle represents the N_c cluster centroid vectors. That is, each particle x**,** is constructed as follows:

$$x_i = (m_{i1} \ldots : m_{ij}, .., m_{iNc}) \tag{5}$$

Where: m_{ij} is the *j-th* cluster centroid vector of the *i-th* particle in cluster C_{ij}. The fitness of particles can be easily measured as the quantization error,

$$J_e = \sum_{j=1}^{Nc} \frac{[\sum_{\forall Zp \in Cij} d(z_p.m_j)/|C_{ij}|]}{N_c} \tag{6}$$

where d is the distance to the centroid given by equation:

$$d(z_p.m_j) = \sqrt{\sum_{k=1}^{Nd} (z_{pk} - m_{jk})^2} \tag{7}$$

'k' subscripts the dimension and $|C_i|$ in equation (7) is the number of data vectors belonging to cluster C_{ij} i.e. the frequency of that cluster. The authors also proposed the standard *gbest* PSO clustering algorithm using this hybrid technique. Results proved that this hybrid technique is more efficient than the standard PSO technique or traditional K-means algorithm alone. In [9] the authors have used PSO to decide the vector of the cluster centre. The following fitness function in equation (4) is evaluated and then compared with the particle's best solution. The updating of position and velocities vectors are carried out according to (1) and (2) till the algorithm meets its stopping criterion.

$$J = \sum_{i=1}^{K} \sum_{i=1}^{N} \| xi - zj \|^2 \tag{8}$$

In [10] the authors have applied hybrid of K-means and PSO for the purpose of fast and high-quality document clustering to effectively navigate, summarize and organize information. C.M Cohen and Castro [11] used PSO for data clustering by adapting this algorithm to position prototypes (particles) in regions of the space that represent natural clusters of input data set. They proposed the *PSC (particle swarm clustering)* which behaves more like a self-organizing neural network that aids the positioning of the particles (prototypes) in the space following the spatial distribution of the input data. Other than the introduction of inertia weight parameter; another improvement in original PSO proposed by Kennedy and Eberhart was the introduction of the constriction factor [12] that too resulted in fast convergence of PSO algorithm. In equation (1) of the original version of PSO, $v_{i.k}$ is limited to the range $(-V_{max}, +V_{max})$ where V_{max} parameter was introduced to limit the step size or the velocity to prevent explosion that results due to the random weighting of the control parameters in the algorithm. Constriction Coefficients are used to prevent such an explosion. Specifically, the application of constriction coefficients allows control over the dynamic characteristics of the particle swarm, including its exploration versus exploitation properties. Eberhart and Shi [13] have also concluded that constriction factor does not alone guarantee fast convergence and that the fastest convergence can

be achieved by a combined approach of V_{max} parameter clamping strategy and constriction coefficients. Engelbrecht and Bergh proposed a new locally convergent particle swarm optimizer [14] called the *Guaranteed Convergence Particle Swarm Optimizer (GCPSO)*. This new algorithm significantly resulted in faster convergence compared to the original PSO, especially when smaller swarm sizes are used. Yet another paper by Bergh and Engelbrecht [15] employs heuristic approach for initialization of the inertia weight and acceleration of coefficient values of PSO to guarantee convergent trajectories. In [16] an improved PSO method has been proposed in order to solve the problem of easy fall into local optimal solutions, lower convergent precision, slower convergence rates and the poor population diversity. The simulation results of this improved PSO indicated that its performance in terms of optimal precision, efficiency and the stability are much better than that of traditional PSO.

In another version of PSO [17] based on initial population of clustering, the diversity of the population was analyzed according to discrepancy in the solution space and objective function space. The clustering algorithm is used to grab the information of the initial population in order to generate the representative individuals, and then the size of the initial population composed by these representative individuals is reduced. This method provides an effective method to generate initial populations and offers the basis for assessment and regulation of the population diversity in the process of running the algorithm. In a paper [18] by Dai, Lui, and Li, an intelligent method for optimum parameter selection is proposed. Firstly it analyzes the effect of each parameter on algorithm performance in detail. Tests to the benchmark function show that these parameters are better than the experience parameters and results in the optimal fitness and convergence rate. A discrete binary version of the improved PSO has been discussed in [19].On one hand, to improve the convergence rate, the improved algorithm combines the traditional binary particle swarm algorithm with the simulated annealing in order to guide the evolution of the optimal solution, and on the other hand, to simplify the structure of algorithm, the cross-operation of the genetic algorithm is used to replace the update operation of the speed and location.

3 Survey of Hybrid Techniques Based on PSO

Evolutionary algorithms are used nowadays for clustering. Hybridization is a method of combining two or more techniques in a judicious way so that the resulting algorithm contains the positive features of all the combined algorithms. Many hybrids of PSO have been developed so far in order to harness the strong points of the PSO algorithm and further improve its efficiency and accuracy.

3.1 Hybridization Perspective of Clustering of Multi-objective and High-Dimensional Problems

For Multi-objective optimization problems (MOPs), the objectives to be optimized are normally in conflict with respect to each other, which means that there is no single

solution for these problems. A research paper [20] published in 2009 proposed a PSO method for MOP using Fuzzy Clustering technique named *Fuzzy Clustering Multi-objective Particle Swarm Optimizer (FC-MOPSO)*. Fuzzy clustering technique provides a better distribution of solutions in decision variable space by dividing the whole swarm into sub-swarms. In FC-MOPSO, the migration concept is used to exchange information between different sub-swarms and to ensure their diversity. The actual data sets used in data mining are high-dimensional and very complex, hence, effective hybrid techniques are required to efficiently cluster them. To reduce dimensionality of datasets PSO along with the *Principal Component Analysis technique (PCA)* [21] is used. PSO has been proved to be effective in clustering data under static environments. In 2010, Serkan, Jenn and Moncef proposed a PSO technique [22] for multidimensional search in dynamic environment by introducing the *Fractional Global Best Formation (FGBF)* technique. This technique exhibits a significant performance for multi-modal and non-stationary environments.

3.2 PSO and Genetic Algorithm (GA) Hybridization

The hybrid of PSO and Genetic Algorithm (GA) is one of the most widely used and efficient technique for clustering data [23]. GA is a randomized global search technique that solves problems by imitating processes observed from natural evolution. Based on the survival and reproduction of the fittest, GA continually exploits new and better solutions. For a specific problem, the GA codes a solution as a binary string called a chromosome (individual). A set of chromosomes is randomly chosen from the search space to form the initial population that represents a part of the solution space of the problem. Next, through computations, the individuals are selected in a competitive manner, based on their fitness measured by a specific objective function. The genetic search operators such as selection, mutation and crossover are then applied one after another to obtain a new generation of chromosomes in which the expected quality over all the chromosomes is better than that of the previous generation. The major problem with the traditional K-means algorithm is that it is sensitive to the selection of the initial partitions and it may converge to local optima. The hybrid of GA and PSO and K-means [23] avoids premature convergence and provides fast data clustering. This hybrid combines the ability of the globalized searching of the evolutionary algorithms and the fast convergence of the k-means algorithm and can avoid the drawbacks of both.

3.3 PSO and DE (Differential Evolution) Hybridization

DE algorithm was proposed by Storn and Price [24] in 1995. DE involves the same operators as GA (selection, mutation and crossover) but differs in the way it operates. PSO-DE hybrids usually combine the evolutionary schemes of both algorithms to propose a new evolutionary position scheme. A modified PSO with differential evolution operator mutations is introduced in [25] to eliminate stagnation and premature convergence of standard PSO. In [26] the authors have used this hybrid

technique in an attempt to efficiently guide the evolution and enhance the convergence. They have evolved the personal experience of the swarm with the DE algorithm [27].

3.4 Other Variants of PSO

A novel technique named *Selective Regeneration Particle Swarm Optimization (SRPSO)* [29] suggests parameter setting and the mechanism of selective particle regeneration. The suggested unbalanced setting of *c1* and *c2* in equation (1) accelerates the convergence of the algorithm while the particle regeneration operation enables the search to escape from local optima and explore other areas for better solutions. A comprehensive review of the various hybridization techniques of PSO and K-means has been discussed in [30]. An improved particle swam optimization algorithm with *synthetic update mechanism* is presented. The synthetic update is made up of three parts: the first is disturbance operation, the second is mutation operation and the last is *gbest* value distribution. Multi-objective PSO algorithm [32] has been used for clustering and feature selection. Features are assigned weights automatically by an algorithm and the features with low weights are then omitted which helps in omitting irrelevant features. Experimental results show that the proposed algorithm performs clustering independently for the shape of clusters and it can have good accuracy on dataset of any shape or distribution. In [33], the authors have developed a new PSO technique which can be applied both when the number of clusters is known as well as when this number is unknown. The authors have proposed a fitness function in case where the number of clusters is known:

$$f^t_p = \sigma^t_p = \sum^{Kp}_{k=1} \sum^{n}_{i=1} w^{pt}_{ik} D(o_i.z^{pt}_k) \quad (9)$$

Where f^t_p is the fitness value of particle p at iteration t. If the number of clusters is unknown the following fitness function is proposed:

$$f^t_p = \sigma^t_p - \min_{k \neq l} D(z^{pt}_k.z^{pt}_l) \quad (10)$$

When the partitioning is compact and satisfactory, the value of σ^t_p should be low, while $min_{k \neq l} D(z^{pt}_k.z^{pt}_l)$ should be high, thereby yielding lower values from the fitness function. Nowadays meta heuristic optimization algorithms have become popular choice for solving complex and intricate problems which are otherwise difficult to solve by traditional methods [34].

4 Conclusion

One of the major reasons for the wide use of Particle Swarm Optimization is that there are very few parameters to adjust. A single version, with very slight variations works well in a wide variety of applications. PSO has been used for approaches that can be used across a wide range of applications such as clustering of web usage data, image segmentation, system design, multi-objective optimization, classification, pattern recognition, biological system modelling, scheduling, signal processing and robotic applications. The hybridisation of PSO with other evolutionary algorithms like

GA and DE has been very effective in improving its efficiency and accuracy. Due to its simplicity and efficiency, PSO is gaining a lot of attention from the researchers and the recent developments show that hybrid PSO methods will emerge as a successful optimization technique in diverse applications.

References

[1] Han, J., Kamber, M.: Data Mining: Concepts and Techniques. Morgan Kaufmann Publishers (2002)
[2] Karypis, G., Han, E.-H., Kumar, V.: CHAMELEON: A Hierarchical Clustering Algorithm using Dynamic Modelling. Computer 32, 68–75 (1999)
[3] Zhang, T., Ramakrishnan, R., Livny, M.: BIRCH: An Efficient Data Clustering Method for very Large Databases. In: Widom, J. (ed.) Proceedings of the 1996 ACM SIGMOD International Conference on Management of Data, SIGMOD 1996, Montreal, Quebec, Canada, pp. 103–114. ACM Press, New York (1996)
[4] Guha, S., Rastogi, R., Shim, K.: Cure: An efficient clustering algorithm for large databases. In: Proceedings of the 1998 ACM SIGMOD International Conference on Management of Data, Seattle, WA, pp. 73–84 (1998)
[5] Guha, S., Rastogi, R., Shim, K.: ROCK: A robust clustering algorithm for categorical attributes. In: Proceedings of the International Conference on Data Engineering, pp. 512–521 (1999)
[6] Kennedy, J., Eberhart, R.C.: Particle Swarm Optimization. In: Proceedings of the IEEE International Conference on Neural Networks (ICNN), Australia, vol. IV, pp. 1942–1948 (1995)
[7] Van Der Merwe, D.W., Engelbrecht, A.P.: Data clustering using particle swarm optimization. In: Proceedings of the IEEE Congress on Evolutionary Computation, Canberra, Australia (2003)
[8] Shi, Y., Eberhart, R.C.: A modified particle swarm optimizer. In: Proceedings of the IEEE International Conference on Evolutionary Computation, Alaska (1998)
[9] Cheo, C.Y., Ye, F.: Particle Swarm Optimization Algorithm and Its Application to Clustering Analysis. In: Proceedings of the 2004 IEEE International Conference on Networking, Sensing Control, Taiwan (2004)
[10] Cui, Potok, Palathingal: Document clustering using Particle Swarm Optimization. In: IEEE Swarm Intelligence Symposium, SIS (2005)
[11] Cohen, S.C.M., de Castro, L.N.: Data Clustering with Particle Swarms. In: Proceedings of the IEEE Congress on Evolutionary Computation (2006)
[12] Clerc, M., Kennedy, J.: The Particle Swarm- Explosion, Stability and Convergence in a Multidimensional Complex Space. IEEE Transactions on Evolutionary Computation 6(1) (February 2002)
[13] Eberhart, R.C., Shi, Y.: Comparing inertia weights and constriction factors in particle swarm optimization. In: Proceedings of the 2000 Congress on Evolutionary Computation, pp. 84–89 (2000)
[14] van den Bergh, F., Engelbrecht, A.P.: A New locally convergent Particle Swarm optimizer. In: Proceedings of the IEEE Conference on Systems, Man and Cybernetics (2002)
[15] van den Bergh, F., Engelbrecht, A.P.: A study of particle swarm optimization particle trajectories. Information Sciences 176, 937–971 (2006)
[16] Yang, J., Xue, L.: Adaptive Population Differentiation PSO Algorithm. In: Third International Symposium on Intelligent Information Technology Application (2009)

[17] He, D., Chang, H., Chang, Q., Liu, Y.: Particle Swarm Optimization Based on the Initial Population of Clustering. In: Proceedings of the Sixth International Conference on Natural Computation, ICNC (2010)
[18] Dai, Y., Liu, L., Li, Y.: An Intelligent Parameter Selection Method for Particle Swarm Optimization Algorithm. In: Proceedings of the Fourth International Joint Conference on Computational Sciences and Optimization (2011)
[19] Jun, X., Chang, H.: The Discrete Binary Version Of The Improved Particle Swarm Optimization Algorithm. In: Proceedings of the IEEE International Conference on Management and Service Science MASS (2009)
[20] Benameur, L., Alami, J., El Imrani, A.: A New Hybrid Particle Swarm Optimization Algorithm for Handling Multiobjective Problem Using Fuzzy Clustering Technique. In: Proceedings of the International Conference on Computational Intelligence, Modelling and Simulation (2009)
[21] Qian, X.-D., Li-Wie.: Date Clustering using Principal Component Analysis and Particle Swarm Optimization. In: Proceedings of the 5th International Conference on Computer Science & Education Hefei, China (2010)
[22] Kiranyaz, S., Pulkkinen, J., Gabbouj, M.: Multi-dimensional particle swarm optimization in dynamic environments. Expert Systems with Applications 38, 2212–2223 (2011)
[23] Abdel-Kader, R.F.: Genetically Improved PSO Algorithm for Efficient Data Clustering. In: Proceedings of the IEEE Second International Conference on Machine Learning and Computing (2010)
[24] Price, K., Storn, R., Lampinen, J.: Differential Evolution: A Practical Approach to Global Optimization. Springer, Berlin (2005)
[25] Zheng, X.: Modified Particle Swarm Optimization with Differential Evolution Mutation. In: Proceedings of the Sixth International Conference on Natural Computation, ICNC (2010)
[26] Epitropakis, M.G., Plagianakos, V.P., Vrahatis, M.N.: Evolving cognitive and social experience in Particle Swarm Optimization through Differential Evolution. In: IEEE Congress on Evolutionary Computation, CEC (2010)
[27] Xu, R., Xu, J., Wunsch: Clustering with Differential Evolution Particle Swarm Optimization. In: IEEE Congress on Evolutionary Computation CEC (2010)
[28] Kuo, R.J., Lin, L.M.: Application of a hybrid of genetic algorithm and particle swarm optimization algorithm for order clustering. Decision Support Systems 49(4), 451–462 (2010)
[29] Tsai, C.-Y., Kao, I.-W.: Particle swarm optimization with selective particle regeneration for data clustering. Expert Systems with Applications 38, 6565–6576 (2011)
[30] Shen, Jin, Zhu, Zhu: Hybridization of Particle Swarm Optimization with the K-Means Algorithm for Clustering Analysis. In: Proceedings of the IEEE 5th International Conference on Bio-Inspired Computing: Theories and Applications BIC-TA (2010)
[31] Li, F.: An Improved Particle Swarm Optimization Algorithm with Synthetic Update Mechanism. In: IEEE Third International Symposium on Intelligent Information Technology and Security Informatics (2010)
[32] Javani, M., Faez, K., Aghlmandi, D.: Clustering and feature selection via PSO algorithm. In: IEEE International Symposium on Artificial Intelligence and Signal Processing (2011)
[33] Cura, T.: A particle swarm optimization approach to clustering. Expert Systems with Applications 39, 1582–1588 (2012)
[34] Thangaraj, R., Pant, M., Abraham, A., Bouvry, P.: Particle swarm optimization: Hybridization perspectives and experimental illustrations. Applied Mathematics and Computation 217, 5208–5226 (2011)