Application of Particle Swarm Optimization in Data Clustering: A Survey

Sunita Sarkar
Department of Computer Science
Assam university,Silchar

Arindam Roy
Department of Computer Science
Assam university,Silchar

Bipul Shyam Purkayastha
Department of Computer Science
Assam university,Silchar

ABSTRACT
Clustering is the process of organizing similar objects into groups, with its main objective of organizing a collection of data items into some meaningful groups. The problem of Clustering has been approached from different disciplines during the last few year's. Many algorithms have been developed in recent years for solving problems of numerical and combinatorial optimization problems. Most promising among them are swarm intelligence algorithms. Clustering with swarm-based algorithms (PSO) is emerging as an alternative to more conventional clustering techniques. PSO is a population-based stochastic search algorithm that mimics the capability of swarm (cognitive and social behavior). Data clustering with PSO algorithms have recently been shown to produce good results in a wide variety of real-world data. In this paper, a brief survey on PSO application in data clustering is described.

General Terms
Data clustering, K-mean clustering

Keywords
Data mining, Data clustering, , Particle swarm optimization

1. INTRODUCTION
Declining cost of and in data storage cost, rapid advancement in computer networks, data acquisition, improved in computing performance and explosive growth in generation of electronic information, has led to collection and storage of huge amount of data in databases. The amount of data stored in databases continues to grow fast. This large amount of stored data contains valuable knowledge, which could be used to improve the decision-making in an organization. Such large databases have led to the emergence of a field of study called data mining and knowledge discovery in databases[1].

Data Mining is an analytical process exploring data (usually large amounts of it typically business or market related) attempting to find consistent patterns and/or systematic relationships between variables, and then validating the findings by applying the detected patterns to new subsets of data. It aims at making a prediction. Generally speaking, data mining (sometimes also called data or knowledge discovery) is the process of analyzing data from a different prospective and summarizing it into meaningful and usable information. Technically speaking, it is the process of finding correlation and patterns among dozens of fields in relational database by using advanced analytical techniques such as neural network, fuzzy logic and rough set[2],[3]. There are several methods of finding these patterns in a large database. Summarization, association, clustering etc. are some of these methods. Data clustering is the most popular of these methods.

Data clustering is a popular approach of automatically finding classes, concepts, or groups of patterns. It seeks to. partition an unstructured set of objects into clusters (groups). This implies wanting the objects to be as similar to objects in the same cluster and as dissimilar to objects from other clusters as possible. Clustering has been applied in many areas including biology, medicine, anthropology, marketing and economics. Clustering applications include plant and animal classification, disease classification, image processing, pattern recognition and document retrieval. Clustering techniques have been applied to a wide variety of research problems.

Data clustering algorithms can be either hierarchical or partitional [4],[5]. Hierarchical clustering creates a nested set of clusters. Each level in this hierarchy has a separate set of clusters is such a way that at the lowest level, each item is in its unique cluster and at the highest level, all items belong to the same cluster. This hierarchical clustering algorithm can be graphically displayed as a tree, called a dendogram. Such hierarchical algorithms can be agglomerative (bottom-up) or divisive (top-down). Agglomerative algorithms are the ones that begin with each element as a separate cluster and merge them in successively larger clusters. Divisive algorithms, on the other hand, begin with the whole set and proceed to divide it into successively smaller clusters. Hierarchical algorithms have two basic advantages[4]. One is that the number of classes need not be specified a priori, and two, they are independent of the initial conditions. However, the main drawback of hierarchical clustering techniques is that they are static; which is to say that data points assigned to a cluster cannot move to another cluster. Besides this they may fail to separate overlapping clusters due to a lack of information about the global shape or size of the clusters[6]. With partitional clustering the algorithm creates only one set of clusters. These approaches use the desired number of clusters to create final set. The advantages of the hierarchical algorithms happen to be the disadvantages of the partitional algorithms, and vice versa. Jain, Murty and Flynn (1999) presented an extensive survey of various clustering techniques. [6].

It has been recently realized that the partitional clustering technique is well suited for clustering a large dataset on account of their computational requirements being relatively low[7],[8]. The time complexity of this technique is almost linear making it widely usable. The best known partitioning clustering algorithm is the K-means algorithm and its variants [9]. Our understanding shows that this algorithm is simple, straightforward and is based on the firm foundation of the analysis of variances.

The K-mean algorithm seeks to find a partition that minimizes mean square error (MSE) measure. Although it is an extensively useful clustering algorithm, it suffers from

many shortcomings. The objective function of the K-means is not convex[10] and hence it may contain local minima. As a consequence, while minimizing the objective function, there is possibility of getting stuck at local minima as well as at local maxima and at saddle point [11]. The performance of the K-means algorithm depends on the initial choice of the cluster centers. It is also known that the Euclidean norm is sensitive to noise or outliers. It is therefore implied that K-means algorithm should be affected by noise and outliers[12],[13]. Another clustering method Fuzzy C-Mean (FCM) is better when compared to K-mean as cluster boundaries are no more hard boundaries, but it is also dependent upon clustering centre initialization [6]. It is also more complex in computation rather than K-mean. In order to overcome the problem of partitonal clustering various heuristic algorithms have been proposed in the literature surveyed such as Genetic Algorithm (GA), Ant Colony Optimization (ACO), Differential Evolution (DE) and Particle Swarm Optimization (PSO). Literature survey revealed that clustering techniques based on Evolutionary Computing and Swarm Intelligence algorithms outperformed many classical methods of clustering.

Particle swarm optimization is a biologically inspired, population-based computational search and optimization method developed in 1995 by Eberhart and Kennedy based on the social behaviors of birds flocking or fish schooling[14]. The concept of particle swarm originated as a simulation of a simplified social system. Originally PSO was designed to graphically simulate the graceful but unpredictable choreography of a bird flock defined as a 'cornfield vector'[15]. It has been successfully applied in several areas of work like clustering problem [16,17],image processing [18] ,function optimization[19] etc. Significantly PSO is simple and requires little memory. It would be noteworthy that, computationally, it is effective and easier to implement when compared to other mathematical algorithms and evolutionary algorithms[20].

2. PARTICLE SWARM OPTIMIZATION

PSO is a population-based search algorithm which is initialized with a population of random solutions, called particles[21]. As against the other evolutionary computation techniques, each particle in this algorithm, called PSO is also associated with a velocity. Particles fly through the search space with velocities that are dynamically adjusted as per their historical behaviors. The particles, therefore have the tendency to fly towards the better and better search area all over the course of the process of search. In PSO a number of simple entities—the particles—are placed in the search space of some problem or function, and each one of these evaluates the objective function at its current location. Thereafter, each particle then determines its movement through the search space by combining some aspect of the history of its own current and best (best-fitness) locations with those of one or more members of the swarm, with some random perturbations. The next iteration takes place after all particles have moved. Eventually the swarm as a whole, like a flock of birds collectively foraging for food, is likely to move close to an optimum of the fitness function.

The particle swarm is actually more than just a collection of particles. A particle by itself has almost does not solve any problem; progress takes place only when they i.e. the particles interact. Populations are organized according to some sort of communication structure or topology. This is often thought of as a social network. The topology typically consists of bidirectional edges connecting pairs of particles. It is like the alphabet j appearing in i's neighborhood, and likewise i in j's neighbour. Each particle communicates with other particles and is affected by the best point found by any member of its topological neighborhood[50].

Each individual in the particle swarm is composed of three D-dimensional vectors, where D is the dimensionality of the search space. These are the current position x_i the previous best position p_i, and the velocity v_i [50]

The i^{th} particle is represented as $X_i = (x_{i1}, x_{i2},. . . , x_{iD})$. At each generation, each particle is updated by the following two 'best' values. The first one is the best previous location (the position giving the best fitness value) a particle has achieved so far. This value is called pBest. The pBest of the i^{th} particle is represented as $P_i = (p_{i1}, p_{i2},. . . , p_{iD})$. At each iteration, the P vector of the particle with the best fitness in the neighborhood, designated l or g, and the P vector of the current particle are combined to adjust the velocity along each dimension, and that velocity is then used to compute a new position for the particle. The portion of the adjustment to the velocity influenced by the individual's previous best position (P) is considered as the cognition component, and the portion influenced by the best in the neighborhood is the social component. With the addition of the inertia factor ω, [22] (brought in for balancing the global and the local search), velocity and position update equations are:

$$v_i = \omega \times v_i + \eta_1 \times rand(\,) \times (p_i - x_i) + \eta_2 \times rand(\,) \times (p_g - x_i)$$

$$x_i = x_i + v_i$$

where rand() and rand() are two random numbers independently generated within the range [0,1] and η1 and η2 are two learning factors which control the influence of the social and cognitive components. In (1.1), if the sum on the right side exceeds a constant value, then the velocity on that dimension is assigned to be $\pm V_{max}$. Thus, particles' velocities are clamped to the range $[-V_{max}, V_{max}]$ which serves as a constraint to control the global exploration ability of particle swarm. Thus, the likelihood of particles leaving the search space is reduced. Grosan et al. [10] presented the basic scheme of PSO algorithm in Fig. 1.

The main advantage of PSO is that it has less parameters to adjust. Other advantages are that PSO does not have any complicated evolutionary operators such as crossover, mutation as in genetic algorithm[23]. It has shortcomings too. PSO gives good results and accuracy for single objective optimization, but for multi objective problem it stuck into local optima[24]. Another problem in PSO is its nature to a fast and premature convergence in mid optimum points. Several PSO variants have been developed to solve this problem.[25]

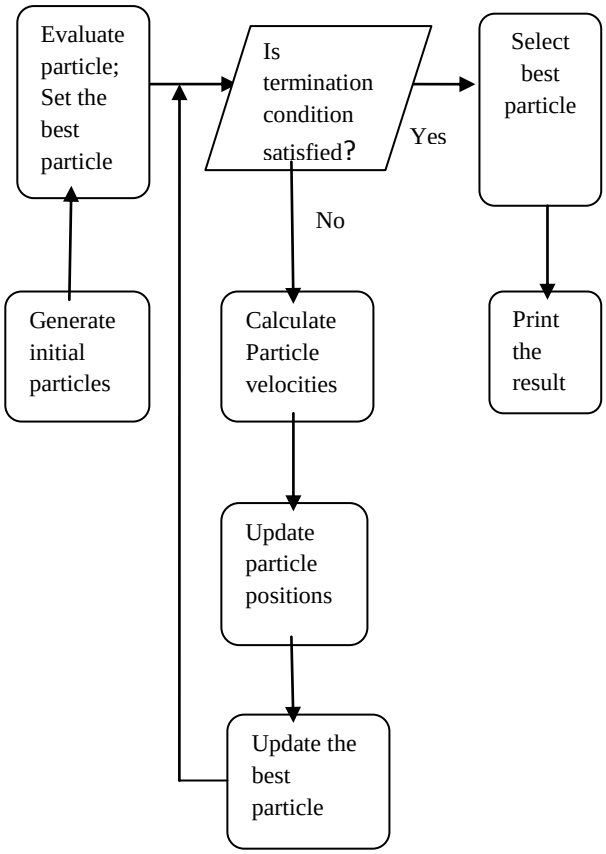

Fig.1: The basic structure of PSO.

3. APPLICATION OF PSO IN DATA CLUSTERING

Van der merwe and Engelhrecht[26] proposed two methods to cluster data using PSO. While in one method standard gbest PSO was used to find the centroid of a user specified number of clusters. In the second method the algorithm is then extended to use K-means clustering to seed the initial swarm. The results of two PSO approaches were compared to K-mean algorithm. This showed that the PSO approaches have better convergence to lower quantization errors, and in general, larger inter-cluster distances and smaller intra cluster distances.

Ahmadyfard and Modares [27] proposed another clustering algorithm, which is a hybrid of PSO and K-mean, named as PSO-KM algorithm. In this PSO algorithm is initially applied to search all space for a global solution. When global solution is found, K-mean clustering algorithm is used for faster convergence to finish the clustering process.

Ghali et al.[28] presented a exponential particle swarm optimization (EPSO) to cluster data. In EPSO exponential inertia weight is used instead of linear inertia weight. A comparison between EPSO clustering algorithm and particle swarm optimization (PSO) was made. it showed that EPSO clustering algorithm has a smaller quantization error than PSO clustering algorithm, i.e. EPSO clustering algorithm more accurate than PSO clustering algorithm.

Chen and Ye[17] proposed a algorithm based on PSO, called PSO-clustering that automatically search cluster centre in the arbitrary data set This proposed algorithm overcomes the shortcomings of K-Means algorithm, performance of which is highly dependent upon its nature of selection of initial cluster centre.

Marinakis et al.[29] found that feature selection is an optimization problem .They proposed a hybrid PSO-GRASP algorithm which used PSO algorithm for a solution to the feature selection problem and GRASP algorithm is for the clustering of data. The proposed algorithm has the potential to overcome the drawbacks of traditional clustering due to the nature of stochastic and population-based search.

Yi et al.[30] proposed a fuzzy PSO clustering method (FPSOC) and its variation FPSOCS for image clustering. In these methods, the particles search the optimal cluster centers in solution space, and the images are classified according to the membership degree of images to cluster centers. In the fuzzy PSO based approaches, feature weights are introduced which are dynamically adjusted during clustering.

Omran [31]proposed a new clustering method based on PSO (DCPSO) for image segmentation. It was proposed to tackle the color image quantization. The method used binary PSO algorithm to automatically determines the ‘‘optimum’’ number of clusters and simultaneously clusters the data set.

Srinoy and Kurutach [32] proposed a novel model for the intrusion detection system, based on hybridization artificial ant cluster algorithm and k-mean particle swarm optimization. In this approach, initially Artificial ant clustering algorithm is used to create raw clusters and then these clusters are refined using K-mean particle swarm optimization (KPSO). This approach is capable of recognizing only the known attacks as well as to detecting suspicious activity that may cause new, unknown attack.

The fuzzy c-means algorithm is sensitive to initialization and is easily trapped in local optima. On the other hand the particle swarm algorithm is a global stochastic tool which could be implemented and applied easily to solve various function optimization problems, or the problems that can be transformed to function optimization problems. Izakian et al. [33] presented a hybrid fuzzy clustering method based on FCM and fuzzy PSO (FPSO) to overcome the shortcomings of the fuzzy c-means. Experimental results over six well known data sets, Iris, Glass, Cancer, Wine, CMC, and Vowel illustrated that the proposed hybrid FCM-FPSO method is efficient and can reveal very encouraging results in term of quality of solution

Mehdizadeh[34] reported that determining suitable suppliers in the supply chain is a key strategic consideration. The nature of these decisions is usually complex and unstructured. In general, many quantitative and qualitative factors, such as quality, price, and flexibility and delivery performance, need to be considered in order to determine suitable suppliers. The author presented a hybrid algorithm namely FPSO integrating the fuzzy c-means (FCM) and the particle swarm optimization to clustering suppliers under fuzzy environments into manageable smaller groups with similar characteristics. The numerical analysis shows that the proposed PSO improves the performance of the fuzzy c-means (FCM) algorithm.

Niknam et al.[35] proposed an efficient hybrid evolutionary optimization algorithm based on a combination of Particle Swarm Optimization (PSO) and Ant Colony Optimization (ACO), so as to be called PSO-ACO, for optimally clustering N object into K clusters. In this algorithm, the decision making process of each particle for

selecting the best guide just before its movement is reinforced with the ACO method. The performance of the new PSO-ACO algorithm was compared with those of ACO, PSO and K-means clustering. The simulation results revealed that the proposed evolutionary optimization algorithm is robust and suitable for handing data clustering.

In standard PSO the non-oscillatory route can quickly cause a particle to stagnate and it may also prematurely converge on suboptimal solutions that are not even guaranteed to local optimal solution. Premalatha and Natarajan[36] proposed a new approach based on discrete binary PSO algorithm with local search for data clustering and applied in the data sets. This approach provides a method for particles to steer clear the local stagnation and the local search is applied to improve the goodness of fit.

Gene clustering methods are essential in the analysis of gene expression data collected over time and under different experimental conditions. However Microarray expression data for thousands of genes can now be collected efficiently and at a relatively low cost. Xiao et al.[37]proposed a hybrid SOM/PSO algorithm for gene clustering. In the hybrid SOM/PSO algorithm, SOM is first used to cluster the dataset. At this stage either regular SOM or SOM with conscience was used. Then PSO was initialized with the weights produced by SOM at the first stage and then PSO was used to refine the clustering process.

Abdul Latiff et al.[38] stated that in clustering of wireless sensor network the number of clusters is one of the key parameters determining the lifetime of the sensor network. They proposed a dynamic multi-objective clustering approach using binary PSO (DCBMPSO) algorithm for wireless sensor networks. This proposed algorithm automatically finds the optimal number of clusters in the network resulting minimum total network energy dissipation. They defined, two clustering metrics namely total network energy consumption and intra-cluster distance for the selection of the best set of network cluster heads.

Rana et al.[39] proposed a hybrid sequential clustering algorithm based on combining the K-Means algorithms and PSO algorithms which uses PSO in sequence with K-Means algorithm for data clustering. This algorithm seeks to overcome drawbacks of both algorithms, improves clustering and avoids being trapped in a local optimal solution. In this algorithm initial process starts by PSO due to its fast convergence and then the result of PSO algorithm is tuned by the K-Means near optimal solutions.

Olesen et al.[40] presented a hybrid approach for clustering based on particle swarm optimization (PSO) and bacteria foraging algorithms (BFA). The proposed method AutoCPB (Auto-Clustering based on particle bacterial foraging) uses autonomous agents to cluster chunks of data by using simplistic collaboration. This algorithm extends the advantages of social influence in PSO with the influence of bacterial foraging behavior.

Hyma et. al.[41] proposed a new method of integrating PSO and GA for document clustering . In this proposed approach two ways namely parallel and transitional are followed to use the integrated algorithm. In the parallel approach each algorithm run for user defined numbers of iterations simultaneously and then fixed numbers of good particles are swapped. In the transitional method the results of one algorithm after user defined numbers of iterations are passed to the other algorithm alternatively.

Premalatha et al.[42] opined that for a large high dimensional dataset, conventional PSO conducts a globalized searching for the optimal clustering, but it may be trapped in a local optimal area. They proposed a hybrid Particle Swarm Optimization (PSO) -Genetic Algorithm (GA) approaches for the document clustering so as to overcome such a problem. This hybrid mechanism of global search models PSO and GA enhances the search process by improving the diversity as well as converging In this method crossover operation of GA is applied for information swapping between two particles and the mutation operation is applied to PSO to increase the diversity of the population

Cui et al.[43] proposed a hybrid PSO based algorithm for document clustering. In this algorithm, they applied the PSO, K-means and a hybrid PSO clustering algorithm on four different text document datasets. The results have shown that the hybrid PSO algorithm can generate more compact clustering results than the K-means algorithm.

Hwang et al.[44] stated that one of the big issue with clustering algorithm was to define the number of clusters at the start of the clustering process by the user. To overcome such a problem, they proposed an algorithm based on particle swarm optimization (PSO) and fuzzy theorem which automatically determines the appropriate number of clusters and their centers. The results revealed that the proposed algorithm is able to determine the number of clusters accurately.

Das et al.[45] worked out a modified PSO based algorithm, called Multi-Elitist PSO (MEPSO) model for clustering complex and linearly non-separable datasets. In this algorithm kernel—induced similarity measure was used instead of Euclidean distance metric. They also reported that for nonlinear and complex data Euclidean distance causes severe misclassifications but it works well when data is hyper spherical and linearly separable.

Fun and Chen[46] worked out an evolutionary PSO learning-based method to optimally cluster N data points into K clusters. The hybrid PSO and K-means, with a novel alternative metric algorithm is called Alternative KPSO-clustering (AKPSO) method. It developed to automatically detect the cluster centers of geometrical structure data sets. In AKPSO algorithm, the special alternative metric is considered to improve the traditional K-means clustering algorithm to deal with various structure data sets.

Sridevi and Nagaveni[47] presented a clustering algorithm that employs semantic similarity measure. They have proposed a model by combining ontology and optimization technique to improve the clustering In this model the ontology similarity is used to identify the importance of the concepts in the document and the particle swarm optimization is used to cluster the document.

Johnson and Sahin[48] introduced four methods of PSO, (Interia methods, Inertia with predator prey option, Constriction method and Constriction with predator prey option) to explain the PSO application in data clustering. The four methods were evaluated in a number of well-known benchmark data sets and were compared with K-mean and

fuzzy c-means. The results have shown significant increase in performance and lower quantization error.

Shan et al. [49] proposed an algorithm based on Grid and Density with PSO (HCBGDPSO) to discover clusters with arbitrary-shape . First density of grid cells was computed considering overlapped influence region of data points and then PSO algorithm was applied to find the clusters.

Table 1 summarizes all these methodologies with different parameters like datasets used, evaluation parameters applied, and future work (explicitly mentioned) for easy and quick reference

Table 1. Comparison of various PSO based data clustering methods

Paper referred	**Clustering Algorithm**	**Dataset**	**Evaluation parameters**	**Future Work**
DW van der Merwe AP Engelhrecht [26]	Gbest PSO, Hybrid PSO and K-means algorithm	Iris , Breast Cancer, Wine, Automotives	Quantization error, Inter cluster distance and intra cluster distance	Extend the fitness function to optimize the inter and intra cluster distances, Experiment on higher dimensional problems-and large number of patterns , determination of optimal number of clusters dynamically.
Neveen I. Ghali, Nahed El-Dessouki, Mervat A. N., and Lamiaa Bakrawi [28]	PSO, Exponential Particle Swarm Optimization (EPSO)	Breast cancer, Iris, Yeast, Lences, Glass	Quantization error	-------
Surat Srinoy and Werasak Kurutach [32]	Hybrid artificial ant cluster algorithm and kmean particle swarm optimization	KDD'99 data set	Recognition of known network attacks	-------
Esmaeil Mehdizadeh[34]	Fuzzy PSO alogrithm	Artificial data set, iris, wine and image segmentation	Objective function value and CPU time	-------
Hesam Izakian, Ajith Abraham, Václav Snášel[33]	Hybrid fuzzy c-means fuzzy particle swarm algorithm for clustering	Iris , Cancer, Wine, glass, CMC, vowel	Objective function values	-------
T. Niknam, M. Nayeripour and B.Bahmani Firouzi [35]	Particle swarm optimization - ant colony optimization (PSO-ACO) algorithm	Iris, Wine, Vowel and CMC	Function value, Standard deviation and number of function evaluation	-------
K. Premalatha and A.M. Natarajan[36]	PSO with local search	Iris , Wine, glass	Fitness value , Inter and Intra Cluster similarity	-------
Xiang Xiao, Ernst R. Dow, Russell Eberhart, Zina Ben Miled and Robert J. Oppelt [37]	Hybrid SOM –PSO algorithm	Yeast data set and rat data set	Average merit, execution time	-------
N. M. Abdul Latiff, C. C. Tsimenidis, B. S. Sharif and C. Ladha [38]	Binary PSO with multi-objective clustering approach (DCBMPSO)	100 nodes	Number of cluster, network lifetime and delivery of data messages	To investigate the DC-BMPSO algorithm properties such as the effect of varying algorithm parameter, *init p* on the number of clusters, as well as on network performance
Sandeep Rana, Sanjay Jasola, Rajesh Kumar [39]	PSO in sequence with K-Means	Artificial problem, Iris and wine	Quantization error, Inter and Intra Cluster distance	Variations in PSO algorithm and its hybridization with K-Means algorithm
Jakob R. Olesen, Jorge	*AutoCPB*	Artificial dataset,	QEF metric, ID metric, number of	To identify a rule to minimize local optima, to apply to other domains such

Cordero H., and Yifeng Zeng [40]		Iris, Wine, Pima, Haberman, Breast Cancer, Glass and Yeast	clusters and elapsed times	as attribute clustering, more specific analysis of parameter setting
J.Hyma, Y.Jhansi and S.Anuradha [41]	Hybrid PSO and Genetic algorithm	Document dataset	Intra Cluster distance	To extend this work to deal with other sorts of documents.
Swagatam Das, Ajith Abraham, Amit Konar [45]	Kernel_MEPSO (Multi-Elitist PSO) algorithm	Synthetic dataset, Glass, Wine ,Breast cancer, Image , segmentation and Japanese vowel	Mean and standard deviation of the clustering accuracy, Mean and standard deviation of the number of clusters, unpaired t-tests, execution time, Mean and standard deviations of the number of fitness function evaluation	Improve the performance of the algorithm over high dimensional datasets by incorporating some feature selection mechanism in it. Automatic clustering in lower dimensional subspaces with MEPSO may also be a worthy topic of further research
Fun Ye and Ching-Yi Chen [46]	Alternative KPSO-clustering (AKPSO)	Artificial datasets and iris dataset	Cluster center location, distortion measure	-------
K. Premalatha and A.M. Natarajan[42]	Hybrid PSO and Genetic algorithm	Library Science, Information Science and Aeronautics	Fitness value	-------
Alireza Ahmadyfard and Hamidreza Modares[[27]	PSO-Kmeans	synthetic data sets (SET I, SET II and SET III), Iris and Cancer	Error rate and Mean square error	-------
Yannis Marinakis, Magdalene Marinaki, and Nikolaos Matsatsinis[29]	Hybrid PSO-GRASP(Greedy Randomized Adaptive Search Procedure)	Australian Credit Breast Cancer Wisconsin 1 (BCW1) Breast Cancer Wisconsin 2 (BCW2) Heart Disease (HD) Hepatitis 1 (Hep1) Ionosphere (Ion) Spambase(spam) Iris Wine Olive Oil	Feature selection	Use of different algorithms for both feature selection phase and clustering algorithm phase.
Wensheng Yi1, Min Yao and Zhiwei Jiang [30]	fuzzy particle swarm optimization clustering algorithm.	Iris, Wine, Ionosphere, sunset/sunrise images, beach and grassland images	Entropy and cluster purity	To extend the clustering algorithm to video stream, Improvement of the speed of the algorithms
Ryan K. Johnson, Ferat Sahin[48]	PSO (Interia methods, Inertia with predator prey option, Constriction method	Iris, Breast Cancer, Wine,E. Coli, glass and	Quantization Error (mean), Quantization Error (std.	Clustering of dynamic data

	and Constriction with predator prey option)	Segmentation	dev.)	
Jen-Ing G. Hwang, Chia-Jung Huang [44]	Hybrid scheme of differential evolution based PSO and fuzzy *c*-means(EDPSO)	Iris, Breast Cancer, Wine	Effect of perturbed velocity, determine an appropriate number of Clusters, Jaccard index	-------
Mahamed G. H. Omran , Ayed Salman and Andries P. Engelbrecht[31]	Dynamic clustering algorithm based on PSO (DCPSO)	Synthetic images,Lenna, mandrill, jet, peppers, MRI and Lake Tahoe	Mean and Standard deviation	Application of the DCPSO algorithm to general data, to investigate the effect of high dimensionality on the performance of the DCPSO, use of other clustering algorithms such as FCM and KHM to refine the cluster centroids, incorporation of spatial information into the DCPSO algorithm
Xiaohui Cui, Thomas E. Potok [43]	Hybrid Particle Swarm Optimization (PSO) and K-means document clustering	TREC-5, TREC-6, and TREC-7	Average distance between documents and the cluster centroid(ADVDC)	-------
Sridevi.U. K., Nagaveni. N. [47]	PSO clustering using ontology similarity	NewsGroups	Sum of squared error, Precision, Recall, F-measure, Time in minutes	Fuzzy ontology based methodology for clustering knowledge and personalized searching method
Ching-Yi Chen and Fun Ye[17]	PSO clustering algorithm	Artificial data set	Object function and Cluster centre	
Shi M. Shan, Gui S. Deng, Ying H. He[49]	Hybridization of Clustering Based on Grid and Density with PSO (HCBGDPSO	Artificial dataset	Shape of clusters	To devise an application and finding a way of adaptively tuning the parameters in HCBGDPSO

4. CONCLUSIONS

This paper has presented a review of previous researches conducted in the areas of Particle Swarm Optimization (PSO), PSO hybrids and their application to data clustering. Researches in this field shows that if PSO is hybridized with other clustering algorithms, then it yields better results in various optimization problems in terms of efficiency and accuracy when compared with other evolutionary algorithms such as GA,SA etc. The implementation of hybrid PSO algorithms for data clustering yields optimal number of clusters which results in better prediction and analysis of data. A comprehensive survey of literature in this area has therefore been given to help provide more insight in this subject. After having done the survey we would like to do the following:

- We would like to analyse and evaluate existing PSO based approaches in data clustering to know about the strengths and shortcomings of the existing systems.
- Improvement in earlier proposed solutions, if possible.
- We would also like to develop a semantic based text document clustering approach using Universal Networking Language (UNL) and hybrid PSO - SOM algorithm. Universal Networking Language (UNL) would be used to identify the importance of the concepts in the document.

5. REFERENCES

[1] Han,J. ; Kamber, M (2001). "Data Mining: Concepts and Techniques", Morgan Kaufmann, San Francisco.

[2] Guoyin, W; Jun, H; Qinghua, Z; Xiangquan, L; Jiaqing, Z (2008). "Granular computing based data mining in the view of rough set and fuzzy set". In International conference on Granular computing. Proceedings in IEEE GRC. pp 67–67

[3] Shanli.W(2008) Research on a new effective data mining method based on neural networks. In

[4] International symposium on electronic commerce and security. pp 195–198

[5] Frigui, H, Krishnapuram, R (1999). "A robust competitive clustering algorithm with applications in computer vision," IEEE Trans. Pattern Anal. Mach. Intell., vol. 21, no. 5, pp. 450–465.

[6] Leung, Y; Zhang, J; Xu,Z (2000). "Clustering by scale-space filtering," IEEE Trans. Pattern Anal. Mach. Intell., vol. 22, no. 12, pp. 1396–1410.

[7] Jain, A. K ; Murty, M. N. Flynn, P. J. (1999). "Data clustering: a review. ACM Computing Survey 31(3):264–323

[8] Steinbach, M; Karypis, G; Kumar, V. (2000). A Comparison of Document Clustering Techniques. TextMining Workshop, KDD.

[9] Zhao. Y; Karypis G(2004). Empirical and Theoretical Comparisons of Selected Criterion Functions for Document Clustering, Machine Learning, 55 (3): pp. 311-331.

[10] Hartigan, J. A (1975). Clustering Algorithms. John Wiley and Sons, Inc., New York, NY.

[11] Grosan, C; Abraham, A; Chis,M (2006). Swarm Intelligence in Data Mining, Studies in Computational Intelligence (SCI) 34, 1–20

[12] Selim, S.Z.; Ismail, MA (1984) K-means Type Algorithms: A Generalized Convergence Theorem and Characterization of Local Optimality, IEEE Transaction on Pattern Analysis and Machine Intelligence, 6, 81-87

[13] Wu, K.L, Yang, M.S (2002) Alternative C-means Clustering Algorithms. Pattern Recognition, 35, 2267-2278

[14] Jones, G; Robertson, A; Santimetvirul, C; Willett, P (1995) Non-hierarchic document clustering using a genetic algorithm. Information Research, 1(1)

[15] Kennedy, J; Eberhart, RC (1995) Particle swarm optimization. In: Proceedings of IEEE conference on neural networks, Perth, Australia, pp 1942–1948.

[16] Kennedy, J (1997). Minds and cultures: Particle swarm implications. Socially Intelligent Agents. AAAI Fall Symposium. Technical Report FS-97-02, Menlo Park, CA: AAAI Press, 67-72.

[17] Paterlini, S; Krink, T (2006) Differential evolution and particle swarm optimization in partitional clustering. Comput Stat Data Anal 50:1220–1247

[18] Chen, CY; Ye, F (2004). Particle swarm optimization algorithm and its application to clustering analysis. In: Proceedings of the IEEE international conference on networking, sensing and control. Taipei, Taiwan, pp 789–794

[19] Niu, Y; Shen, L (2006) An adaptive multi-objective particle swarm optimization for color image fusion. Lecture notes in computer science, LNCS. pp 473–480

[20] Silva, A; Neves, A; Costa, E (2002). Chasing the swarm: a predator pray appoach to function optimization. In: Proceedinge of the MENDEL, international conference on soft computing.

[21] Senthil, MA; Rao, MVC; Chandramohan, A (2005). Competitive approaches to PSO algorithms via new acceleration co-efficient variant with mutation operators. In: Proceedings of the fifth international conference on computational intelligence and multimedia applications

[22] Hu, X; Shi, Y; Eberhart, RC (2004) Recent Advances in Particle Swarm, In Proceedings of Congress on evolutionary Computation (CEC), Portland, Oregon, 90-97

[23] Shi, Y; Eberhart, RC (1998). A modified particle swarm optimizer. In Proceedings of the IEEE Congress on Evolutionary Computation (CEC), Piscataway, NJ. 69-73

[24] Boeringer, D-W; Werner, DH(2004).”Particle swarm optimization versus genetic algorithm for phased array synthesis”. IEEE Trans Antennas Propag 52(3):771-779

[25] Junliang, L; Xinping, X (2008). Multi-swarm and multi-best particle swarm optimization algorithm. In: IEEE world congress on intelligent control and automation. pp 6281–6286

[26] Rana, S; Jasola, S; Kumar, R, “A review on Particle Swarm Optimization Algorithms and Applications to data clustering”. Springer Link Artificial Intelligence Review vol.35, issue 3:211–222.2011

[27] Van der Merwe ,DW; Engelhrecht, AP (2003) Data clustering using particle swarm optimization. In: Conference of evolutionary computation CEC’03, vol 1. pp 215–220

[28] Ahmadyfard, A; Modares, H (2008) Combining PSO and k-means to enhance data clustering. In: International symposium on telecommunications. pp 688–691

[29] Ghali, NI; Dessouki,NE ; Mervat A. N; Bakrawi, L(2008) Exponential Particle Swarm Optimization Approach for Improving Data Clustering. World Academy of Science, Engineering and Technology 42 .

[30] Marinakis, Y; Marinaki, M; and Matsatsinis, N (2007). A Hybrid Particle Swarm Optimization Algorithm for Clustering Analysis .DaWaK 2007, Lecture notes in computer science, LNCS 4654, pp. 241–250

[31] Yi, W; Yaoand, M; Jiang, Z(2006).Fuzzy Particle Swarm Optimization Clustering and Its Application to Image Clustering.

[32] Omran, M; Salman, A; Engelbrecht AP (2006). Dynamic clustering using particle swarm optimization with application in image segmentation. Pattern Anal Appl 8:332–344

[33] Srinoy, S; Kurutach, W (2006).Combination Artificial Ant Clustering and K-PSO Clustering Approach to Network Security Model. International Conference on Hybrid Information Technology (ICHIT'06)

[34] Izakian,H ; Abraham, A; Snášel V(2009)Fuzzy Clustering Using Hybrid Fuzzy c-means and Fuzzy Particle SwarmOptimization. World Congress on Nature & Biologically Inspired Computing (NaBIC 2009)

[35] Mehdizadeh, E (2009) A fuzzy clustering PSO algorithm for supplier base management. International Journal of Management Science and Engineering Management Vol. 4 (2009) No. 4, pp. 311-320

[36] Niknam,T; Nayeripour, M; Firouzi, BB(2008). Application of a New Hybrid optimization Algorithm on Cluster Analysis Data clustering. World Academy of Science, Engineering and Technology 46

[37] Premalatha, K and Natarajan, AM(2008) A New Approach for Data Clustering Based on PSO with Local Search. International Journal of Computer and Information Science , vol 1, No. 4, 139

[38] Xiao, X; Dow,ER; Eberhart, R; Miled , ZB ;Oppelt, RJ (2003). Gene clustering using self-organizing maps and particle swarm optimization. Proceedings of International Symposium on Parallel and Distributed Processing.

[39] Abdul Latiff, N.M.; Tsimenidis, C.C.; Sharif, B.S.; Ladha, C.(2008). Dynamic clustering using binary multi-objective Particle Swarm Optimization for wireless sensor networks. IEEE 19th International Symposium on Personal, Indoor and Mobile Radio Communications, 2008. PIMRC 2008. IEEE 19th International Symposium pp 1 - 5

[40] Rana,S; Jasola,S; Kumar, R(2010). A hybrid sequential approach for data clustering using K-Means and particle swarm optimization algorithm.International Journal of Engineering, Science and Technology.Vol. 2, No. 6, pp. 167-176

[41] Olesen,J.R.; Cordero H.,J; Zeng, Y(2009). Auto-Clustering Using Particle Swarm Optimization and Bacterial Foraging.Lecture Notes in Computer Science, LNCS 5680, pp. 69–83

[42] Hyma, J; Jhansi, Y; Anuradha, S(2010). A new hybridized approach of PSO & GA for document clustering. International Journal of Engineering Science and Technology Vol. 2(5), 1221-1226

[43] Premalatha, K and Natarajan, AM(2010). Hybrid PSO and GA Models for Document Clustering. Int. J. Advance. Soft Comput. Appl., Vol. 2, No. 3,

[44] Cui, X ; Potok, TE, (2005), Document Clustering Analysis Based on Hybrid PSO+Kmeans Algorithm, Journal of Computer Sciences (Special Issue), ISSN 1549-3636, pp. 27-33.

[45] Hwang, J.-I.G.; Huang, C-J.(2010) Evolutionary dynamic particle swarm optimization for data clustering. In: International Conference on Machine Learning and Cybernetics (ICMLC)

[46] Das S, Abraham A, Konar A (2008) Automatic kernel clustering with a multi-elitist particle swarm optimization algorithm. Pattern Recognit Lett 29:688–699

[47] Fun, Y. and Chen, C. Y.(2005). Alternative KPSO-clustering algorithm. Tamkang J. Sci. Eng., 8, 165–174.

[48] Sridevi.U. K. and Nagaveni. N.(2011) Semantically Enhanced Document Clustering Based on PSO Algorithm. European Journal of Scientific Research Vol.57 No.3 (2011), pp.485-493

[49] Johnson Ryan, K; Sachin, Ferat (2009) Particle swarm optimization methods for data clustering. In: IEEE fifth international conference soft computing computing with words and perceptions in system analysis, decision and control. Pp 1-6

[50] Shan, SM; Deng, GS; He, YH(2006). Data Clustering using Hybridization of Clustering Based on Grid and Density with PSO. In: IEEE International Conference on Service Operations and Logistics, and Informatics.

[51] Poli R; Kennedy, J; Blackwell, T. Particle Swarm Optimization An Overview. Springer Link, Swarm Intelligence, vol. 1, issue 1: 33–57, 2007