International Journal of Research in Advent Technology, Vol.3, No 4, April 2015
E-ISSN: 2321-9637

Applications of Swarm Intelligence in Data Clustering: A Comprehensive Review

Paramjeet Kaur[1], Dr. Harish Rohil[2]
[1]*Ph.D. Research Scholar, Deptt. of CSA, CDLU, Sirsa*
[2]*Associate Professor, CRSU, Jind*
[1]*paramjeet.nicegirl@gmail.com*, [2]*harishrohil@gmail.com*

Abstract - Data mining is a growing area of research and development. Data mining is the process of identifying the hidden patterns and structures from massive volume of data. Data clustering is most important technique of data mining and widely used in various domains i.e. web, E-com, biology, machine learning. Data Clustering refers to the process in which data objects are grouped into clusters in such a way that objects within a cluster have high similarity in comparison to one another but are different from objects in different clusters. Due to large application area, we need effective and efficient clustering algorithms or methods which are scalable and can work on large databases, can handle complex shaped data. In recent years, many algorithms have been developed for solving clustering as well as numerical and combinatorial optimization problems. Swarm intelligence algorithms are most efficient in solution of such optimization problems. Swarm intelligence is an important area of Artificial intelligence. It is based on the collective or collaborative behaviour of self organized agents like ant's colony, bird's flocks, fish school, bee colony, bats, frogs, bacterium etc. Clustering using different swarm intelligence based metaheuristics is being used as an alternative to more conventional clustering techniques. If we consider 'data clustering' as optimization problem, Data clustering with Swarm Intelligence algorithms are being used to generate better results in a wide variety of real-world data. In this paper, we present a comprehensive survey on applications of various swarm intelligence techniques in data clustering.

Index Terms*:* data clustering, swarm intelligence, ant colony, PSO, k-means, CSO

1. INTRODUCTION

TODAY'S era is digital era. A huge amount of digital data is available around us. As we know that data is generated at very fast speed and also all the generated data is not of much interest to most of users, therefore there is an urgent need for generation of new tools, theories and techniques that can help in extracting the useful information from multidimensional databases. Data mining and KDD are a step toward this extraction process. Data mining is a growing area of research and development. Data mining is the process of identifying the hidden patterns and structures from massive volume of data. It is a interdisciplinary concept. It can extract the useful patterns from huge data sets of different types like numeric, text, images, video, speech and mixed representation. Some of the applications of data mining include various domains like web, ecommerce era, bioinformatics, tracking frauds, identifying marketing strategy, gaming strategy, finding out good and bad customers, find out disease and effectiveness of treatment(medicine area) etc. Various techniques of data mining are useful in such applications like classifications algorithms, clustering algorithms, association rule mining, summarization, sequence analysis. Current research studies show that data mining need not to be constrained to stochastic, combinatorial and classical hard optimization based techniques. We need to dwell on soft computing approaches, artificial intelligence, and swarm intelligence to find the optimized solution.

SECTION 2 gives introduction of different data clustering techniques. Section 3 presents introduction of swarm intelligence techniques. Section 4 presents a comprehensive review of applications of swarm intelligence in data clustering. Section 5 concludes the paper with future remarks.

2. DATA CLUSTERING

Data clustering is most important technique of data mining and widely used in various domains i.e. web, E-com, biology, machine learning. Data Clustering refers to the process in which data objects are grouped into clusters in such a way that objects within a cluster have high similarity in comparison to one another but are different from objects in different clusters. Clustering differs from classification in that there is no target variable for clustering [1]. The clustering task does not try to classify, estimate, or predict the value of a target variable. Therefore clustering is an

unsupervised learning technique versus classification, which belongs to supervised learning.

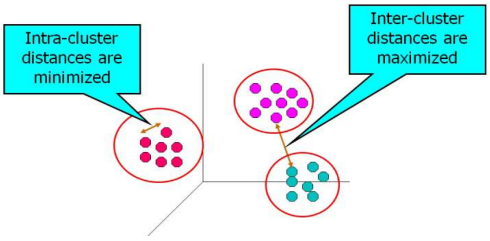

Fig.1- Data Clustering example

Various applications of clustering are market research, pattern recognition, data analysis, image processing. It is widely being used in web document clustering. Due

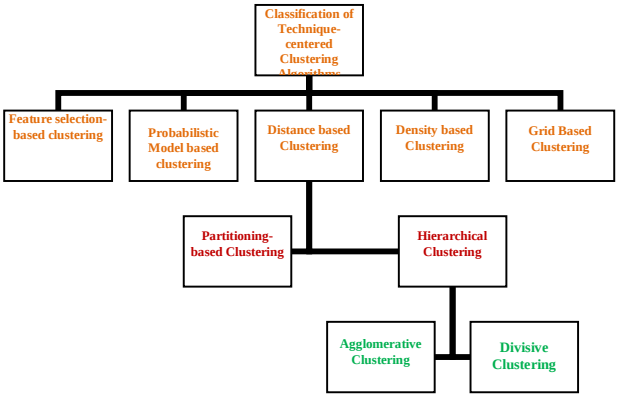

Fig. 2: Classification of Technique-centered Clustering Algorithms

to large application area, we need effective and efficient clustering algorithms or methods which are scalable and can work on large databases, can handle complex shaped data. Many clustering algorithms exist in the literature. Different researchers categorize them based on different criteria. Some categorize these based on underlying technique such as distance based, density based, probabilistic techniques etc. Some algorithms are classified as per type of data involved in clustering. For example, some algorithms can cluster only numerical data whereas others are applicable for categorical data or time series data.

The nature of the data greatly impacts the choice of methodology used for the clustering process. Furthermore, some data types are more difficult than others because of the separation between different kinds of attributes such as behavior or contextual attributes.

We will describe, in the following sections, one of the categorization, i.e. Technique-centered clustering algorithms.

a. Feature selection based Clustering

Dimensionality reduction [2] is a technique to remove irrelevant and redundant features. Dimensionality reduction techniques can be categorized mainly into feature extraction and feature selection. The quality of learning may degrade due to the existence of irrelevant features and it may also consume more memory and computational time that could be saved if these features were removed. From the clustering point of view, irrelevant features' removal will not negatively affect clustering accuracy. Methods of feature selection for clustering are categorized into filter [3] wrapper [4], and hybrid models [5]. A wrapper model evaluates the candidate feature subsets by the quality of clustering while a filter model is independent of clustering algorithm. Thus, the filter model is preferable in terms of computational time and as unbiased toward any clustering method, while the wrapper model produces better clustering if we have prior information about the clustering method. In order to reduce the computational cost incurred in the wrapper model, filtering criteria are utilized to select the candidate feature subsets in the hybrid model.

b. Probabilistic Model based Clustering

Probabilistic model-based clustering methods have been used in many applications, ranging from image segmentation, handwriting recognition, document clustering, topic modeling to information retrieval [6]. Model-based clustering approaches attempt to optimize the fit between the observed data and some mathematical model using a probabilistic approach. These methods are based on the assumption that the data are generated by a mixture of underlying probability distributions. Each cluster can be represented mathematically by a parametric probability distribution, such as a Gaussian or a Poisson distribution. Thus, the clustering problem is transformed into a parameter estimation problem since the entire data can be modeled by a mixture of *K* component distributions.

Typical examples of algorithms of this category include mixture models, EM algorithm and its variations and probabilistic topic models.

c. Distance based clustering

Distance based clustering methods are those which measure similarity between objects by computing the distance between each pair. There are a number of methods for computing the distance in a multidimensional environment such as Euclidean distance, Manhattan distance, Chebychev distance and some others. These methods are simple and easy to implement. These can be generally divided into two

types- Partitioning based clustering and Hierarchical Clustering.

1. Partitioning based Clustering algorithms

As the name suggests, Partition based clustering partitions the total n no. of elements into k no. of segments, where each segment represents a cluster. A partitioning based clustering algorithm obtains a single partition of the data instead of a clustering structure, such as the dendrogram produced by a hierarchical technique. Partition-based methods have advantages in applications involving large data sets for which the construction of a dendrogram is computationally expensive.

The *k*-means [7] is the simplest and most commonly used partitioning based clustering algorithm employing a squared error criterion. It starts with a random initial partition and keeps reassigning the patterns to clusters based on the similarity between the pattern and the cluster centers until a convergence criterion is met (e.g., there is no reassignment of any pattern from one cluster to another, or the squared error ceases to decrease significantly after some number of iterations). The *k*-means algorithm is popular because it is easy to implement, and its time complexity is linear in *n* where *n* is the number of patterns. A major problem with this algorithm is that it is sensitive to the selection of the initial partition and may converge to a local minimum of the criterion function value if the initial partition is not properly chosen.

Typical examples of algorithms of this category include K-means and its variants, PAM (K-Mediods), K-Modes, CLARA, CLARANS.

2. Hierarchical Clustering

Hierarchical clustering algorithms solve the problem of clustering by developing a binary tree-based data structure called the dendrogram. Once the dendrogram is constructed, one can automatically choose the right number of clusters by splitting the tree at different levels to have different clustering solutions for the same dataset without rerunning the clustering algorithm again. Hierarchical clustering algorithms [8] were developed to overcome some of the disadvantages of partitioning-based clustering methods. Hierarchical clustering can be achieved in two different ways, such as, Agglomerative and divisive clustering. Agglomerative methods start by taking singleton clusters (that contain only one data object per cluster) at the bottom level and continue merging two clusters at a time to build a *bottom-up* hierarchy of the clusters. Divisive methods, on the other hand, start with all the data objects in single cluster and split it continuously into two groups generating a *top-down* hierarchy of clusters. These are ineffective at capturing arbitrarily shaped clusters. So, to capture arbitrarily shaped clusters, algorithms such as CURE [9] and CHAMELEON [10], COBWEB, Self-Organizing Maps [11] are proposed as extensions of hierarchical algorithm.

d. Density-based Clustering

Density-based clustering [12] is a nonparametric method having no assumptions about the number of clusters or their distribution.] Partitioning based algorithms cannot find any arbitrary shaped clusters. So in that case Density based clustering approaches can be used. Density-based clusters are connected, dense areas in the data space separated from each other by sparser areas. Further, the density within the areas of noise is assumed to be lower than the density in any of the clusters [12]. The idea behind this approach is to continue growing the given cluster as long as the density in the neighborhood exceeds some threshold. The advantage of this method is that it can find the cluster of any shape and can also be used to filter out the noise (outliers) [13].

DBSCAN and its extension, OPTICS, DENCLUE are typical representative algorithms of this category.

e. Grid-based Clustering

Grid-based clustering [14] algorithms partition the data space into a finite number of cells to form a grid structure and then form clusters from the cells in the grid structure. Clusters correspond to regions that are denser in data points than their surroundings. Grids were initially proposed by Warnekar and Krishna [15] to organize the feature space, e.g., in GRIDCLUS [16], and increased in popularity after STING [17], CLIQUE and WaveCluster were introduced. The great advantage of grid-based clustering is a significant reduction in time complexity, especially for very large data sets. Rather than clustering the data points directly, grid-based approaches cluster the neighborhood surrounding the data points represented by cells. In most applications since the number of cells is significantly smaller than the number of data points, the performance of grid-based approaches is significantly improved.

3. SWARM INTELLIGENCE

Swarm intelligence is an important area of Artificial intelligence. The term ‘Swarm Intelligence’ was coined by Beny and Wang in early 1980’s. The term ‘Swarm’ refers to group of agents and ‘Intelligence’ refers to collective behaviour i.e. Swarm Intelligence is based on the collective or collaborative behaviour of self organized agents like ant’s colony, bird’s flocks, fish school, bee colony etc. [18]. Thus swarm is used to represent an aggregated term used for fishes, birds,

insects such as ants, termites, bees, etc. performing collective behaviour. According to Kennedy and Eberhart (2001), swarm intelligence is most recent paradigm in bio-inspired computing because it includes the nature inspired methods for solving various optimization problems. It is based on the collective social behaviour of organisms. This collective behaviour results from the interaction of individual components locally with each other and with their environment. Various popular swarm intelligence based methods are discussed as below.

a. Ant Colony Optimization

Ants are social insects i.e. insects that live in colonies and whose behaviour is responsible for survival of whole colony than that of the individual component of colony [20]. The most important and interesting behaviour of ant colony is its food foraging behaviour and how ants can find the paths between source of food and nest of ants. Another important behaviour which can be sought in real ants is brood sorting and corpse filtering [19].

b. Particle Swarm Optimization

This algorithm [21] is based on the swarming behaviour of bird's flocks. Bird flocking behaviour is first simulated by Craig Reynolds. Bird flocking can be referred as the social collective motion of a large no. of interacting birds with common group objective. The underlying principle for such interactions is the nearest neighbour principle where birds adjust their motion by following some flocking rules depending upon their nearest neighbours. To implement the flocking behaviour of birds, Reynolds proposed three simple rules: flock centering, collision avoidance, velocity matching. PSO is a population based search strategy that finds the optimal solutions using a set of flying particles (birds) with velocities that are dynamically adjusted according to their historical performance as well as their neighbours in search space. Each particle remembers its historically best position. Various advancements have been done in field of PSO. Application area of PSO includes Image and video analysis, clustering which is the biggest application area. Image analysis covers iris recognition, fruit quality grading, face detection and recognition, traffic stop sign detection, image segmentation, image/pixel classification, scene matching, object detection. Video analysis applications include MPEG optimization, motion estimation, object tracking, and body posture tracking.

c. Bee Swarm Intelligence

Bees are social insects living in colonies. There are three kinds of bees in a colony: drones, queen and workers [22]. A very interesting swarm in nature is honey bee swarm that allocates the tasks dynamically and adapts itself in response to changes in the environment in a collective intelligent manner. These bees have photographic memories, space-age sensory and navigation systems, group decision making process during selection of their new nest sites, brood storing, retrieving and distributing honey and pollen, communication and foraging. These characteristics can be utilized by researchers to model the intelligent behaviors. Some of the most important features that are being utilized include bee dance (communication), bee foraging, queen bee, task selection, collective decision making, nest site selection, mating, floral/pheromone laying, navigation systems.

d. Bat Algorithm

Bat algorithm [23] is swarm intelligence based algorithm which is worked on the echolocation of bats. This algorithm was developed by Xin-She Yang in 2010. It is a new metaheuristic algorithm for solving the many optimization problems. Bats are based on the echolocation behaviour of microbats. They can find their prey o food and also they can know the different type of insects even in a complete darkness. Since we know that microbats are the insectivore who have the quality of fascinating. These bats use a type of sonar namely as echolocation. They emit a loud sound pulse and detect a echo that is comes back from their surrounding objects. Their pulse varying in properties and will be depend on the species. Their loudness is also varying. When they are searching for their prey, their loudness is loudest if they are far away from the prey and they will become slow when they are nearer to the prey. Now for emission and detection of echo which are generated by them, they use time delay. And this time delay is between their two ears and the loudness variation of echoes.

e. Other Swarm Intelligence based techniques

The Bacteria foraging optimization is inspired by the chemotaxis behavior of bacteria that recognizes chemical gradients in the environment (such as nutrients) and move toward or away from specific signals. Bacterial foraging (BFO), applied the algorithm to the optimization of a benchmark function. In reproduction, the health of each bacterium represents its fitness value. All bacteria are sorted according to their health status and only the first half of population survives. The surviving bacteria are split into two identical ones in order to form a new population. Thus, the population of bacteria is kept constant [24].

Lampyridae is a family of insects that are capable to produce natural light (bioluminescence) to attract a mate or a prey, which are commonly called as fireflies or lightning bugs. The firefly algorithm (FA) was proposed

by Yang [25] and the algorithm was applied to the optimization of benchmark functions.

Similar to ants, cockroaches leave chemical trails in their faces as well as emitting airborne pheromones for swarming and mating. Other cockroaches will follow these trails to discover sources of food and water, and also discover where other cockroaches are hiding. Thus, cockroaches can exhibit emergent behaviour, in which group or swarm behaviour emerges from a simple set of individual interactions. Cockroaches' agents are defined using three simple behaviors: cockroaches search for the darkest location in the search space and the fitness value is directly proportional to the level of darkness (find darkness phase); cockroaches socialize with nearby cockroaches (find friend phase) and third behavior is cockroaches periodically become hungry and leave the friendship to search for food (find food phase) [26].

The shuffled frog-leaping algorithm is a memetic metaheuristic that is designed to seek a global optimal solution by performing a heuristic search. It is based on the evolution of memes carried by individuals and a global exchange of information among the population. it combines the benefits of the local search tool of the particle swarm optimization and the idea of mixing information from parallel local searches to move toward a global solution [27].

4. APPLICATIONS OF SWARM INTELLIGENCE IN DATA CLUSTERING

A vast variety of approaches for data clustering have been proposed in the Literature. Some of them are based on simple data clustering approaches described in section 2. A family of bio-inspired algorithms, especially Swarm Intelligence (SI) has recently emerged that meets the clustering requirements and has successfully been applied to a number of real world clustering problems. Here we present a review of applications of different approaches of swarm intelligence in clustering of different types of data.

a) Applications of Ant Colony System in data clustering

Ant-based clustering sorting was first introduced by **Deneubourg et al. (1991)** to explain phenomena of corpse clustering and larval sorting in ants [28]. They proposed the distributed sorting algorithm, inspired by brood sorting behavior, for use by robot teams. As robots move randomly with no communication, have no global representation and can only perceive the objects in front of them but can differentiate between objects with certain degree of errors. The probability of picking up or dropping down the objects is given as a function of how many same objects the ants have met in recent past. This generate a positive feedback that is sufficient to coordinate the robot activities which result in sorting of objects in clusters.

Lumer and Faieta (1994) modified Deneubourg et al.'s basic model using a dissimilarity-based evaluation of the local density, in order to make it suitable for clustering of data [29]. They have introduced the notion of short-term memory within each agent. Each ant can memorize a small number of locations where it has successfully dropped an item and from where to pick.

Monmarche (1999) combined the stochastic and exploratory principles of clustering ants with the deterministic and heuristic of the popular k-means algorithm in order to improve the convergence of the ant-based clustering algorithm [30]. The proposed method is called AntClass and is based on the work of Lumer and Faieta [29]. The AntClass algorithm allows an ant to drop more than one object in the same cell, thus forming heaps of objects.

V. Ramos and J.J. Merelo (2002) developed a new strategy, called ACLUSTER, for unsupervised clustering as well as for dealing with data retrieval problems. This algorithm was employed for textual document clustering [31]. The authors proposed the use of bio-inspired spatial transition probabilities, avoiding randomly moving agents, which may explore non-interesting regions.

Labroche et al.(2002) proposed a clustering algorithm, called ANTCLUST, based on a modeling of the chemical recognition system of ants [32] . This system allows the construction of a colonial odor used for determining the ants' nest membership, such that ants can discriminate between nest mates and intruders. Thus it helps to create homogeneous groups of individuals sharing a similar odor by exchanging the chemical cues. The results are compared with k-means and AntClass algorithms. This algorithm proved best in extracting the knowledge from real web sessions.

Handl et al. (2003) proposed a scheme that enables an unbiased interpretation of the clustering solutions obtained by ant-based clustering algorithms [33]. They performed an analytical evaluation of clustering results on real and artificial data sets using four different evaluation measures.

Azzag et al. (2003) presented a new algorithm for document clustering in hierarchical manner and automatic generation of portals sites [34]. This model is based on the self-assembling behavior observed in real ants where ants progressively get attached to an existing

support and successively to other attached ants. The artificial ants that we have defined will similarly build a tree. Each ant represents a document. The results are compared with AHC algorithm.

Vizine et al. (2005) proposed an Adaptive Ant-Clustering Algorithm (A^2CA) [35]. They have made a series of improvements to Lumer and Faieta's system. A^2CA is more robust in terms of the number of clusters found and tends to converge into good solution as the clustering process evolves.

A hybrid ant-based clustering method [36] is presented by **Omar et al (2013)** with new modifications to the original ant colony clustering model (ACC) to enhance the operations of ants, picking up and dropping off data items. Ants' decisions are adapted from two classical cluster analysis methods: Agglomerative Hierarchical Clustering (AHC) and density-based clustering.

J. Chircop and C. D. Buckingham (2013) applied the Multiple Pheromone Ant Clustering Algorithm (MPACA) model to real-world data from two domains [37]. The task for the MPACA in each domain was to predict class membership where the classes for the logistics domain were the levels of demand on haulage company resources and the mental-health classes were levels of suicide risk. Results on these noisy real-world data were promising, demonstrating the ability of the MPACA to find patterns in the data with accuracy comparable to more traditional linear regression models.

K. Aparna and M.K.Nair (2014) proposed a new method to improve the cluster quality on high dimensional data set by ant based refinement algorithm [38]. The proposed algorithm is tested with data from different domains. The results show that refined initial starting points and post processing refinement of clusters based on ACO can lead to improved solutions in terms of entropy, time taken and accuracy of clusters.

ii) Applications of PSO in data clustering

Van der Merwe and Engelbrecht (2003) applied PSO for data clustering. Each solution is represented by the coordinates of a user pre-defined number of cluster centroids [39]. Each data instance is assigned to the nearest cluster centroid, using Euclidean distance. The fitness function used is the quantization error, which can be seen as the average distance from a data point to its cluster centroid, averaged over the different clusters. A drawback is that, as with k-means clustering, the number of clusters needs to be predefined.

Well-known PSO algorithms reported in the literature for solving continuous function optimization problems were comparatively evaluated by **Karthi et al. (2008)** by considering real world data clustering problems. Data clustering problems are solved, by considering three performance clustering metrics [40] such as TRace Within criteria(TRW), Variance Ratio Criteria (VRC) and Marriott Criteria (MC).The results obtained by the PSO variants were compared with the basic PSO algorithm, Genetic algorithm and Differential evolution algorithms.

Abdul Latiff et al. (2008) 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 [41]. 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.

Sridevi and Nagaveni (2005) presented a clustering algorithm on the basis of semantic similarity measure. They have proposed a model by combining ontology and optimization technique to improve the clustering [42]. 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 documents.

Omran (2006) proposed a new clustering method based on PSO for image segmentation [43]. The method was proposed to tackle the problem of color image quantization. The method used binary PSO algorithm to automatically determine the 'optimum' number of clusters.

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. **Izakian et al. (2009)** presented a hybrid fuzzy clustering method based on FCM and fuzzy PSO (FPSO) [44] 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

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. (2003)** proposed a hybrid SOM/PSO algorithm [45] for gene clustering. In the hybrid SOM/PSO algorithm, SOM is first used to cluster the dataset. Then PSO was initialized with the weights produced by SOM at the first stage and then PSO was used to refine the clustering process.

Niknam et al. (2008) proposed an efficient hybrid evolutionary optimization algorithm [46] 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.

Hyma et al.(2010) proposed a new method of integrating PSO and GA for document clustering [47]. 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.

Rana et al. (2010) 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 [48]. 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. (2009) 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) [49] 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.

Premalatha et al. (2010) 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) [50] 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. (2005) proposed a hybrid PSO based algorithm for document clustering [51]. 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. Das et al. (2008) 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.

Johnson and Sahin (2009) 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 [52]. 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. (2006) proposed an algorithm based on Grid and Density with PSO (HCBGDPSO) to discover clusters with arbitrary-shape [53]. First density of grid cells was computed considering overlapped influence region of data points and then PSO algorithm was applied to find the clusters.

iii) Applications of Artificial Bee Colony algorithms in data clustering

An interesting application area of ABC is data mining. Particularly clustering, feature selection and rule discovery.

Karaboga and Ozturk (2011) proposed a novel clustering approach based on ABC and tested it on thirteen of typical test data sets from the UCI Machine Learning Repository [54].
Zhang et al. (2010) presented an ABC clustering algorithm [55] to optimally partition n objects into k clusters where Deb's rules are used to direct the search direction of each candidate.

Karaboga and Ozturk (2010) tested the performance of ABC on fuzzy clustering [56] and showed that ABC algorithm is also successful in fuzzy clustering.

Hsieh and Yeh (2011) proposed a concept [57] for machine learning that integrates a grid scheme into a least squares support vector machine (called GS-LSSVM) for classification problems, where ABC is used to optimize parameters for LSSVM learning.

Zhang et al. (2011) described a methodology for automatically extracting a convenient version of T-S fuzzy models from data using a novel clustering technique [58], called variable string length ABC algorithm based fuzzy c-means clustering approach.

Li et al. (2011) studied on study the risk of dams in the perspective of clustering analysis [59] and to improve the performance of fuzzy c-means clustering they proposed an ABC with fuzzy c-means.

Zhao and Zhang (2011) proposed an improved kernel fuzzy c-means clustering algorithm based on ABC [60] which integrates the advantages of kernel fuzzy c-means and ABC algorithm.

Shukran et al. (2011) proposed the use of the ABC algorithm as a new tool [61] for data mining particularly in classification tasks and indicated that ABC algorithm is competitive, not only with other evolutionary techniques, but also to industry standard algorithms such as PART, SOM, naive bayes, classification tree and nearest neighbour (kNN).

iv) Applications of other swarm intelligence based techniques in data clustering

B. Santosa and M. K. Ningrum (2009) proposed a new Cat Swarm Optimization algorithm for clustering problem [62]. The new CSO clustering algorithm was tested on four different datasets. The modification is made on the CSO formula to obtain better results. Then, the accuracy level of proposed algorithm was compared to those of K-means and PSO clustering. The modification of CSO formula can improve the performance of CSO Clustering. The comparison indicates that CSO clustering can be considered as a sufficiently accurate clustering method.

Komarasamy and Wahi (2012) studied K-means clustering using bat algorithm [63] and they concluded that the combination of both K-means and BA can achieve higher efficiency and thus performs better than other algorithms.

Xiujuan Lei et al. (2011) proposed a novel algorithm using Bacteria Foraging Optimization (BFO) algorithm to avoid the influence of cluster number on experimental result of clustering PPI networks. The initial position of the bacterium was considered to be the cluster center and the positions that the bacterium moved were regarded as the adjacent nodes of cluster center. The simulation result [64] showed that the algorithm not only improved the accuracy of cluster result, but also automatically determined the cluster number.

B. Pitchaimanickam and S. Radhakrishnan (2013) proposed a new approach [65] for clustering wireless sensor network into k optimal clusters. They have introduced the Bacteria Foraging Algorithm that forms optimal clusters by identifying the cluster head that have energy more than the average energy of the k-optimal clusters. The approach is implemented in ns2 and simulation results show the improvement of the life time of the network by increase in the number of alive nodes, reduction in the energy consumption.

5. CONCLUSION

Clustering plays important role in the process of knowledge discovery and data mining. A vast variety of approaches for data clustering have been proposed in the Literature. Some of them are based on simple data clustering approaches described in section 2. Bio-inspired algorithms, especially Swarm Intelligence (SI), meet the clustering requirements and have successfully been applied to a number of real world clustering problems. This paper has presented a review of swarm intelligence techniques like ant colony system, particle swarm optimization, bee colony optimization, bat algorithms, cat swarm optimization and their application in data clustering. If SI based techniques are hybridized with other clustering algorithms, then they provides better results in various optimization problems in terms of efficiency and accuracy when compared with other evolutionary algorithms and classical clustering algorithms.

REFERENCES

[1]. K.P.Soman, S. Diwakar and V.Ajay. Data Mining : Theory and Practice, PHI, 2006
[2]. C.C. Aggrawal, C.K. Reddy. Data Clustering : Algorithms and Applications, Chapter 2, CRC, 2014
[3]. M. Dash, K. Choi, P. Scheuermann, and H. Liu. Feature selection for clustering—A filter solution. In Proceedings of the Second International Conference on Data Mining, pages 115–122, 2002
[4]. Roth and T. Lange. Feature selection in clustering problems. *Advances in Neural Information Processing Systems*, 16, 2003.
[5]. J.G. Dy. Unsupervised feature selection. *Computational Methods of Feature Selection*, pages 19–39, 2008.
[6]. C.C. Aggrawal, C.K. Reddy. Data Clustering : Algorithms and Applications, Chapter 3, CRC, 2014
[7]. J. MacQueen, Some methods for classification and analysis of multivariate observations," Proc. of the Fifth Berkeley Symp. OnMath. Stat. and Prob., vol. 1, pp. 281-296, 1967
[8]. C.C. Aggrawal, C.K. Reddy. Data Clustering : Algorithms and Applications, Chapter 4, CRC, 2014
[9]. S. Guha, R. Rastogi, and K. Shim. Cure: An efficient clustering algorithm for large databases. In *ACM SIGMOD Record*, volume 27, pages 73–84. ACM, 1998.
[10]. G. Karypis, E. H. Han, and V. Kumar. CHAMELEON: Hierarchical clustering using dynamic modeling. *Computer*, 32(8):68–75, 1999.
[11]. T. Kohonen. The self-organizing map. *Proceedings of the IEEE*, 78(9):1464–1480, 1990.
[12]. C.C. Aggrawal, C.K. Reddy. Data Clustering : Algorithms and Applications, Chapter 5, CRC, 2014
[13]. J. Han and M. Kamber. Data Mining: Concepts and Techniques, Elsevier, Morgan Kauffman Publishers, 2006
[14]. C.C. Aggrawal, C.K. Reddy. Data Clustering : Algorithms and Applications, Chapter 6, CRC, 2014
[15]. C. S. Warnekar and G. Krishna. A heuristic clustering algorithm using union of overlapping pattern-cells. *Pattern Recognition*, 11(2):85–93, 1979.
[16]. E. Schikuta. Grid-clustering: an efficient hierarchical clustering method for very large data sets. *Proceedings of the 13th International Conference on Pattern Recognition*, 2:101–105, 1996
[17]. WeiWang, Jiong Yang, and Richard R. Muntz. STING: A statistical information grid approach to spatial data mining. In *VLDB'97, Proceedings of 23rd International Conference on Very Large Data Bases*, Athens, Greece, pages 186–195. Morgan Kaufmann, 1997.
[18]. Parpinelli, R.S., and Lopes, H.S. 2011 "New inspirations in swarm intelligence: a survey", International Journal of Bio-Inspired Computation, Vol. 3, No. 1, pp. 1-16.
[19]. Dorigo, M., Maniezzo, V. & Colorni, A. 1996 " Ant System: Optimization by a Colony of Cooperating Agents", IEEE Transactions on Systems, Man, and Cybernetics-Part B, Vol. 2, No. 1, pp. 29-41.
[20]. Dorigo, M., and Stützle, T. "Ant Colony Optimization", MIT Press, 2004Tavel, P. 2007 Modeling and Simulation Design. AK Peters Ltd.
[21]. R.Karthi, S.Arumugam, K. Rameshkumar, Comparative evaluation of Particle Swarm Optimization Algorithms for Data Clustering using real world data sets, IJCSNS International Journal of Computer Science and Network Security, VOL.8 No.1, PP 203-212, January 2008
[22]. Karaboga, D. 2005 "An idea based on honey bee swarm for numerical optimization', Technical report, Erciyes University, Engineering Faculty, Computer Engineering Department.
[23]. Komarasamy, G., and Wahi, A., (2012). An optimized K-means clustering technique using bat algorithm, European J. Scientific Research, Vol. 84, No. 2, pp. 263-273.
[24]. Passino, K. 2002 "Biomimicry of bacterial foraging for distributed optimization and control", IEEE Control Systems Magazine, pp. 52–67.
[25]. Yang, X.S. 2009 "Firefly algorithms for multimodal optimization", SAGA 2009, Lecture Notes in Computer Science, 5792, pp. 169-178.
[26]. Havens, T., Spain, C., Salmon, N. and Keller, J. 2008 "Roach infestation optimization", IEEE Swarm Intelligence Symposium, September, pp.1–7.
[27]. Eusuff, M.M. and Lansey, K.E. 2003 "Optimization of water distribution network design using the shuffled frog leaping algorithm", Journal of Water Resource Planning Management, , 210 – 225.Forman, G. 2003. An extensive empirical study of feature

selection metrics for text classification. J. Mach. Learn. Res. 3 (Mar. 2003), pp. 1289-1305.

[28]. J. L. Deneubourg, S. Gross, N. Franks, A. S. Franks, C. Detrain and L. Chretien, " The dynamics of collective sorting: Robot-like ants and ant-like robots", In Proceedings of the First International Conference on Simulation of Adaptive Behavior: From Animals to Animats, Cambridge, MA, MIT Press, 1991, pp. 356-363.

[29]. E.D. Lumer and B. Faieta, "Diversity and adaptation in populations of clustering ants", Cambridge: MIT Press, In D. Cliff, P. Husbands, J.-A. Meyer, & S.W. Wilson (Eds.), From animals to animats: Proceedings of the Third International Conference on Simulation of Adaptive Behavior, 1994, pp. 501-508.

[30]. N. Monmarche, "On Data Clustering with Artificial Ants", In: Freitas AA, (ed.), Data Mining with Evolutionary Algorithms: Research Directions – Papers from the AAAI Workshop, AAAI Press, pp. 23-26.

[31]. Ramos and J.J. Merelo, "Self-Organized Stigmergic Document Maps: Environment as a Mechanism for Context Learning", In E. Alba, F. Herrera, J.J. Merelo et al. Eds., AEB'2002, First Spanish Conference on Evolutionary and Bio-inspired Algorithms, Spain, 2002, pp. 284-293.

[32]. N. Labroche, N. Monmarche and G. Venturini, " A new clustering algorithm based on the chemical recognition system of ants", Proc. Of the 15th European Conference on Artificial Intelligence, IOS Press, France, 2002

[33]. J. Handl, J. Knowles and M. Dorigo, "On the performance of ant-based clustering", Proc. of the Third International Conference on Hybrid Intelligent Systems Frontiers in Artificial Intelligence and Applications, IOS Press, Vol. 104, 2003, pp. 204-213.

[34]. H. Azzag, N. Monmarche, M. Slimane and G. Venturini, "AntTree: a new model for clustering with artificial ants", Evolutionary Computation, CEC'03, Vol. 4, 2003, 2642-2647

[35]. A.Vizine, L.N. de Castro, E.R. Hruschka and R.R. Gudwin, "Towards improving clustering ants: An adaptive clustering algorithm", Informatica Journal, Vol. 29, 2005.

[36]. W.Omar, A. Badr and A. E. Hegazy, "Hybrid Ant-Based Clustering Algorithm With Cluster Analysis Techniques", Journal of Computer Science 9 (6): 780-793, 2013 , ISSN: 1549-3636

[37]. J. Chircop and C. D. Buckingham, The Multiple Pheromone Ant Clustering Algorithm and its application to real world domains, IEEE, Proceedings of the 2013 Federated Conference on Computer Science and Information Systems pp. 27–34

[38]. Aparna, K.; Nair, M.K.(2014). 'Enhancement of K-Means algorithm using ACO as an optimization technique on high dimensional data', 2014 International Conference on Electronics and Communication Systems (ICECS), IEEE, pp. 1-5

[39]. V. der Merwe & Engelbrecht, A. P. (2003). Data clustering using particle swarm optimization. In IEEE congress on evolutionary computation (1) (pp. 215–220). New York: IEEE.

[40]. R.Karthi, S.Arumugam, K. Rameshkumar, Comparative evaluation of Particle Swarm Optimization Algorithms for Data Clustering using real world data sets, IJCSNS International Journal of Computer Science and Network Security, VOL.8 No.1, PP 203-212, January 2008

[41]. 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

[42]. 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

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

[44]. 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)

[45]. 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.

[46]. 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

[47]. 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

[48]. 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

[49]. 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

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

[51]. 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.

[52]. 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

[53]. 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.

[54]. Karaboga D, Ozturk C (2011) A novel clustering approach: Artificial bee colony (abc) algorithm. Appl Soft Comput 11(1):652–657

[55]. Zhang C, Ouyang D, Ning J (2010) An artificial bee colony approach for clustering. Expert Syst Appl 37(7):4761–4767

[56]. Karaboga D, Ozturk C (2010) Fuzzy clustering with artificial bee colony algorithm. Sci Res Essay 5(14):1899–1902

[57]. Hsieh TJ,YehWC (2011) Knowledge discovery employing grid scheme least squares support vector machines based on orthogonal design bee colony algorithm. IEEE Trans Syst Man Cybern, Part B: Cybern 41(5):1198–1212

[58]. Zhang YF, Su ZG, Wang PH (2011f) A convenient version of t-s fuzzy model with enhanced performance. In: 2011 eighth international conference on fuzzy systems and knowledge discovery (FSKD), vol 2, pp 1074–1079

[59]. Li H, Li J, Kang F (2011c) Risk analysis of dam based on artificial bee colony algorithm with fuzzy c-means clustering. Can J Civ Eng 38(5):483–492

[60]. Zhao X, Zhang S (2011) An improved kfcm algorithm based on artificial bee colony. In: Deng H, Miao D, Wang FL, Lei J (eds) Emerging research in artificial intelligence and computational intelligence, Communications in computer and information science, vol 237. Springer, Berlin, pp 190–198

[61]. Shukran MAM, Chung YY, Yeh WC, Wahid N, Zaidi AMA (2011) Artificial bee colony based data mining algorithms for classification tasks. Mod Appl Sci 5(4):217–231

[62]. A. Santosa and M. K. Ningrum, Cat Swarm Optimization for Clustering, 2009 International Conference of Soft Computing and Pattern Recognition,IEEE, pp 54-59

[63]. Komarasamy, G., and Wahi, A., (2012). An optimized K-means clustering technique using bat algorithm, European J. Scientific Research, Vol. 84, No. 2, pp. 263-273.

[64]. Xiujuan Lei; Shuang Wu ; Liang Ge ; Aidong Zhang (2011). 'Clustering PPI Data Based on Bacteria Foraging Optimization Algorithm', 2011 IEEE International Conference on Bioinformatics and Biomedicine (BIBM), IEEE, pp. 96-99

[65]. Pitchaimanickam, B.; Radhakrishnan, S. (2013) . 'Bacteria Foraging Algorithm based clustering in Wireless Sensor Networks' , Fifth International Conference on Advanced Computing (ICoAC), IEEE, pp. 190-195