Prof. Anuradha D. Thakare, Mrs. Shruti M. Chaudhari / International Journal of Engineering Research and Applications (IJERA) ISSN: 2248-9622 www.ijera.com
Vol. 2, Issue 6, November- December 2012, pp.1455-1459

Introducing a Hybrid Swarm Intelligence Based Technique for Document Clustering

Prof. Anuradha D. Thakare*, Mrs. Shruti M. Chaudhari**
*(Department of Computer Engineering, Pune University, Pune-44)
** (Department of Computer Engineering, Pune University, Pune-44)

ABSTRACT

Swarm intelligence (SI) is widely used in many complex optimization problems. It is a collective behavior of social systems such as honey bees (bee algorithm, BA) and birds (particle swarm optimization, PSO). This paper presents a detailed overview of Particle Swarm Optimization (PSO), its variants and hybridization of PSO with Bee Algorithm (BA). This paper also surveys various SI techniques presented by the researchers. The objective is to utilize the capability of this technique for document clustering which will be utilized to solve the issues of clustering by applying modifications to the Bee Algorithm and Particle Swarm Optimization.

Keywords- Bee Algorithm (BA), Clustering, Genetic Algorithm (GA), Particle Swarm Optimization (PSO), Swarm Intelligence (SI).

I. INTRODUCTION

Clustering is an important unsupervised classification technique. In clustering, a set of patterns, usually vectors in a multi dimensional space, are grouped into clusters in such a way that patterns in the same cluster are similar in some sense and patterns in different clusters are dissimilar in the same sense. For this it is necessary to first define a measure of similarity which will establish a rule for assigning patterns to the domain of a particular cluster centre. One such measure of similarity may be the Euclidean distance D between two patterns x and z defined by D=(x-z) [1]. Smaller the distance between x and z, greater is the similarity between the two and vice versa. Several clustering techniques are available in the literature. Some like the widely used K means algorithm, optimize of the distance criterion either by minimizing the within cluster spread (as implemented in this article), or by maximizing the inter-cluster separation. Other techniques like the graph theoretical approach, hierarchical approach, etc., are also available which perform clustering based on other criteria. The concept of clustering has been around for a long time. It has several applications, particularly in the context of information retrieval and in organizing web resources. The main purpose of clustering is to locate information and in the present day context, to locate most relevant electronic resources. The research in clustering eventually led to automatic indexing to index as well as to retrieve electronic records. Clustering is a method in which we make cluster of objects that are somehow similar in characteristics. The ultimate aim of the clustering is to provide a grouping of similar records. Clustering is often confused with classification, but there is some difference between the two. In classification the objects are assigned to pre defined classes, whereas in clustering the classes are formed. The term "class" is in fact frequently used as synonym to the term "cluster" [1]. Extensive studies dealing with comparative analysis of different clustering methods suggest that there is no general strategy which works equally well in different problem domains. However, it has been found that it is usually beneficial to run schemes that are simpler, and execute them several times, rather than using schemes that are very complex but need to be run only once. An intuitively simple criterion is the within cluster spread, which, as in the K-means algorithm, needs to be minimized for good clustering. However, unlike the K-means algorithm which may get stuck at values which are not optimal, the proposed technique should be able to provide good results irrespective of the starting configuration [2]. Document clustering is the process of grouping document into a number of clusters. The goal of document clustering is to make the data in the same cluster share a high degree of similarity while being very dissimilar to document from other clusters [3].

Swarm Intelligence and GA for Clustering:

Swarm Intelligence (SI) is a computational intelligence technique to solve complex real world problems. It involves the study of collective behavior of individuals in population. The individual interact locally with one another and with their environment in a decentralized control system. The term SI has come to represent the idea that it is possible to control and manage complex systems of interacting entities even though the interactions between and among the entities being controlled is, in some sense, minimal [4]. This notion therefore lends itself to forms of distributed control that may be much more efficient, scalable and effective for

large, complex systems. The underlying features of SI are based on observations of social insects. Ant colonies and beehives, for example, have the interesting property that large numbers of them seem to conduct their affairs in a very organized way with seemingly purposeful behavior that enhances their collective survival. Swarm intelligence, as demonstrated by natural biological swarms, has numerous powerful properties desirable in many engineering systems, such as network routing [5]. Swarm Intelligence (SI) is the property of a system whereby the collective behaviors of (unsophisticated) agents interacting locally with their environment cause coherent functional global patterns to emerge. SI provides a basis with which it is possible to explore collective (or distributed) problem solving without centralized control or the provision of a global model.

Genetic Algorithms are a family of computational models inspired by evolution. These algorithms encode a potential solution to a specific problem on a simple chromosome-like data structure and apply recombination operators to these structures as to preserve critical information. Genetic algorithms are often viewed as function optimizer, although the ranges of problems to which genetic algorithms have been applied are quite broad. An implementation of genetic algorithm begins with a population of (typically random) chromosomes. One then evaluates these structures and allocated reproductive opportunities in such a way that these chromosomes which represent a better solution to the target problem are given more chances to `reproduce' than those chromosomes which are poorer solutions. The 'goodness' of a solution is typically defined with respect to the current population [6].

Genetic algorithms (GAs) are randomized search and optimization techniques guided by the principles of evolution and natural genetics. It is having a large amount of implicit parallelism. GAs perform search in complex, large and multimodal landscapes, and provide near-optimal solutions for objective or fitness function of an optimization problem. In GAs, the parameters of the search space are encoded in the form of strings (called chromosomes). A collection of such strings is called a population. Initially, a random population is created, which represents different points in the search space. An objective and fitness function is associated with each string that represents the degree of goodness of the string. Based on the principle of survival of the fittest, a few of the strings are selected and each is assigned a number of copies that go into the mating pool. Biologically inspired operators like cross- over and mutation are applied on these strings to yield a new generation of strings. The process of selection, crossover and mutation continues for a fixed number of generations or till a termination condition is satisfied. Recently, an application of GAs has been reported in the area of (supervised) pattern classification in RN for designing a GA-classifier. It attempts to approximate the class boundaries of a given data set with a fixed number (say H) of hyper-planes in such a manner that the associated misclassification of data points is minimized during training. When the only data available are unlabeled, the classification problems are sometimes referred to as unsupervised classification [7].

Particle Swarm Optimization (PSO) is very popular SI algorithm for global optimization over continuous search spaces. Since its advent in 1995, PSO has attracted the attention of several researchers all over the world resulting into a huge number of variants of the basic algorithm as well as many parameter automation strategies [7]. Solving an optimization problem is one of the common scenarios that occur in most engineering applications. It is a stochastic-based search technique that has its roots in artificial life and social psychology, as well as in engineering and computer science. The concept of Particle Swarms, although initially introduced for simulating human social behaviors, has become very popular these days as an efficient search and optimization technique. The Particle Swarm Optimization (PSO), as it is called now, does not require any gradient information of the function to be optimized uses only primitive mathematical operators and is conceptually very simple [8].

The original PSO method is a population-based optimization technique, where a population is called a swarm. Every particle in swarm is analogous to an individual “fish” in a school, and it can be seemed a swarm consists of *N* particles moving around a *D*-dimensional search space. Every particle makes use of its own memory and knowledge gained by the swarm as a whole to find the best solution. The *pBestPosition(p,i)* is introduced as the best previously visited position of the *i*th particle. The *gBestPosition(i)* is the global best position of the all individual *pBestPosition(p,i)* values. The position of the *i*th particle is represented by *R (p,i) = xMin + (xMax-xMin) * rand;* and its velocity is represented as *V(p,i) = vMin + (vMax-vMin) * rand.* The position and velocity of the *i*th particle are updated by *pBestValue (p)* and *gBestValue* in each generation. Then, the movement of each particle naturally evolves to an optimal or near-optimal solution. The update equations can be formulated as:

$$R(p,i) = R(p,i) + V(p,i) \quad (1)$$

V(p,i)= V(p,i) + w*(rand * C1 * (pBestPosition(p,i)-R(p,i)) + rand * C2 * (gBestPosition(i) - R(p,i))) (2)

Where *rand* is random number between (0, 1); *C1, C2* are learning factors. V (p,i) denote the velocity of the particle, R(p,i) is a updated particle position.

Bee algorithm (BA) is an optimization algorithm inspired by the natural foraging behavior of honeybees (Eberhart, Shi, & Kennedy, 2001). BA requires the setting of a number of parameters, including number of scout bees (n), number of elite sites selected from n visited sites (e), number of best sites out of n visited sites (b), number of bees recruited for elite e sites (n1), number of bees recruited for best b sites (n2), number of bees recruited for other visited sites (r), and neighborhood (ngh) of bees dance search and stopping criterion [8][9].

The bee algorithm is a population-based search algorithm first developed in 2005. It is based on the food foraging behavior of swarms of honey bees. In its basic version, the algorithm performs a kind of neighborhood search combined with random search. It can be used for both combinatorial optimization and functional optimization. A colony of honey bees can extend itself over long distances (up to 14 km) and in multiple directions simultaneously to exploit a large number of food sources. A colony prospers by deploying its foragers to good fields. In principle, flower patches with plentiful amounts of nectar or pollen that can be collected with less effort should be visited by more bees, whereas patches with less nectar or pollen should receive fewer bees. The foraging process begins in a colony by scout bees being sent to search for promising flower patches [11]. Scout bees move randomly from one patch to another. During the harvesting season, a colony continues its exploration, keeping a percentage of the population as scout bees. When they return to the hive, those scout bees that found a patch which is rated above a certain quality threshold (measured as a combination of some constituents, such as sugar content) deposit their nectar or pollen and go to the "dance floor" to perform a dance known as the waggle dance. This dance is essential for colony communication, and contains three pieces of information regarding a flower patch: the direction in which it will be found, its distance from the hive and its quality rating (or fitness). This information helps the colony to send its bees to flower patches precisely, without using guides or maps. Each individual's knowledge of the outside environment is gleaned solely from the waggle dance. This dance enables the colony to evaluate the relative merit of different patches according to both the quality of the food they provide and the amount of energy needed to harvest it. After waggle dancing inside the hive, the dancer (i.e. the scout bee) goes back to the flower patch with follower bees that were waiting inside the hive. More follower bees are sent to more promising patches. This allows the colony to gather food quickly and efficiently. While harvesting from a patch, the bees monitor its food level. This is necessary to decide upon the next waggle dance when they return to the hive. If the patch is still good enough as a food source, then it will be advertised in the waggle dance and more bees will be recruited to that source.

The clustering behavior can be classified into two stages: the global searching stage and the local refining stage. The global searching stage guarantees each particle searches widely enough to cover the whole problem space. The refining stage makes all particles converge to the optima when a particle reaches the vicinity of the optimal solution [3]. The hybrid PSO algorithm combines the ability of globalized searching of the PSO algorithm and the fast convergence of the K-means algorithm and avoids the drawback of both algorithms.

A new way of integrating GA with PSO explored in the paper has shown better clustering performance not only on data sets but also on different documents [5]. PSO is one such method where global optimization is done. But the disadvantage of the PSO algorithm is, in PSO, when we select an initial population, say 10; we are constrained to the 10 particles.

In order to integrate BA global search ability with the local search advantages of PSO [8], the study proposes a new optimization hybrid swarm algorithm – the Particle-Bee Algorithm (PBA). It integrates the intelligent swarming behavior of honeybees and birds. The study compares the performance of PBA with that of genetic algorithm (GA), differential evolution (DE), bee algorithm (BA) and particle swarm optimization (PSO) for multi-dimensional benchmark numerical problems.

The algorithm named as Genetic Bee Tabu K-means Clustering Algorithm (*GBTKC)* is a novel hybrid algorithm. In this algorithm, the benefits of K-means algorithm are used in order to improve its efficiency [13]. *GBTKC* is based on basic Honey Bee Algorithm (HBA) and a mixture of Genetic Algorithm (GA), Tabu Search (TS) and K-Means Method is used to design it. It does not stuck on locally optimal solutions. The quality of findings and answers provided by this algorithm is much better than those of the previous studied algorithms in subject literature.

Gauss chaotic map adopts a random sequence with a random starting point as a parameter. It relies on the Gauss Chaotic map parameter to update the positions and velocities of the particles [14]. It provides the significant chaos distribution to balance the exploration and

exploitation capability for search process. This easy and fast function generates a random seed processes, and further improve the performance of PSO due to their unpredictability.

II. PROPOSED SYSTEM

From the study of research done in SI, the proposed system is to hybridize PSO and BA [5]. There are two techniques for hybridization, namely, transitional technique and parallel technique. Transitional technique can be used for hybridization of PSO and BA. The idea is to integrate PSO and BA. The proposed algorithm can runs PSO for some time and then makes transitions to BA and runs BA for some time and transits back to PSO.
There are two techniques for integrating P.S.O and B.A.

Transitional Technique:
This technique is used to integrate P.S.O and B.A. The algorithm runs P.S.O for some time and then makes a transition to B.A and it runs in B.A for some time and transits back to P.S.O. The steps are as follows:

1. Select a population of size n randomly and initialize the population.
2. Start with number of iterations equal to zero.
3. Perform either B.A operations or P.S.O operations.
4. Generate output of the algorithms.
5. Evaluate fitness for each individual.
6. Perform these transitions until termination condition is reached.
7.

Parallel Technique:
In this parallel method, the population is divided into two parts and it is evolved with the two techniques respectively. The algorithm executes the two techniques simultaneously and selects user specified number of best individuals from each system exchanging after a user defined number of iterations. The individual with larger fitness value is more often selected. The steps are as follows:

1. Select a population of size n randomly and Initialize the population.
2. Evaluate fitness for each individual.
3. If a termination criterion is not met, split the population to do selective reproduction and velocity updating. Depending on the algorithm employed.
4. If algorithm used is B.A perform operations. Else perform personal best calculation.
5. Repeat this process until final solution is reached.

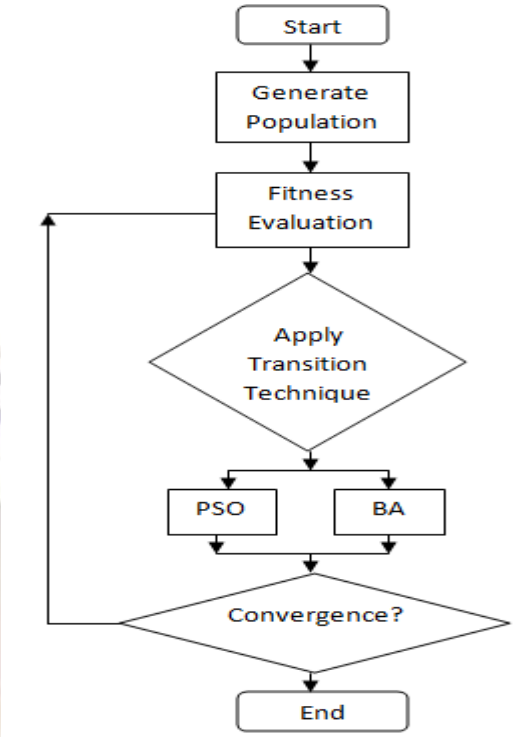

Figure 1: Flowchart for proposed Hybrid System

The objective of the proposed system is to get the more optimum results for clustering using the hybridization of Particle Swarm Optimization and Bee Algorithm. The hybridization can overcome the drawbacks of both algorithms. In future, it can be applied on different datasets.

III. CONCLUSION

In this paper, a particle swarm optimization (PSO) algorithm, which using characterizes bird flocking or fish schooling behavior and Bee Algorithm (BA), which is based on food foraging behavior of swarms of honey bees, are studied to solve the clustering problem. Differing from many of the previously-proposed approaches, the PSO algorithm can be applied both when the number of clusters is known as well as when this number is unknown. Every particle makes use of its own memory and knowledge gained by the swarm as a whole to find the best solution. The proposed algorithm which will be utilized to solve the issues of clustering is the modified version of the basic Bee Algorithm and Particle Swarm Optimization. In future, it can be applied to solve other optimization problems.

REFERENCES

[1] Jiawei Han, Micheline Kamber," *Data Mining: Concepts and Techniques*".
[2] Xiaohui Cui, Thomas E. Potok, Paul Palathingal," Document Clustering using

Particle Swarm Optimization", *0-7803-8916-6/05/$20.00 ©2005 IEEE.*

[3] Wenping Zou,1, 2 Yunlong Zhu,1 Hanning Chen,1 and Xin Sui, " A Clustering Approach Using Cooperative Artificial Bee Colony Algorithm", Hindawi Publishing Corporation Discrete *Dynamics in Nature and Society Volume 2010, Article ID 459796, 16 pages doi:10.1155/2010/459796*

[4] Mark Fleischer," Foundations of Swarm Intelligence: From Principles to Practice", *SWARMING: NETWORK ENABLED C4ISR 2003 BY MARK FLEISCHER. ALL RIGHTS RESERVED.*

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

[6] Ujjwal Maulik, Sanghamitra Bandyopadhyay," Genetic algorithm-based clustering technique", *Pattern Recognition 33 (2000) 1455}1465, Received 24 June 1998; received in revised form 29 April 1999; accepted 29 April 1999.*

[7] Yamille Del Valle, Ganesh Kumar Venayagamoorthy, Salman Mohagheghi, Ronald G. Harley, "Particle Swarm Optimization: Basic Concepts, Variants and Applications in Power Systems", *IEEE Transaction on Evolutionary Transaction, Vol. 12, No. 2, April 2008.*

[8] Li-Chuan Lien, Min-Yuan Cheng, "A hybrid swarm intelligence based particle-bee algorithm for construction site layout optimization", *journal homepage: www.elsevier.com/locate/eswa(2012), Expert Systems with Applications 39 (2012) 9642–9650.*

[9] Changsheng Zhang, Dantong Ouyang, Jiaxu Ning," An artificial bee colony approach for clustering", *journal homepage: www.elsevier.com/locate/eswa, Expert Systems with Applications 37 (2010) 4761–4767.*

[10] K. Premalatha, Dr. A.M. Natarajan," Discrete PSO with GA Operators for Document Clustering", *International Journal of Recent Trends in Engineering, Vol 1, No. 1, May 2009.*

[11] M.A.Behrang, A.Ghanbarzadeh, E.Assareh, "Comparison of the Bees Algorithm (BA) and Particle Swarm Optimisation (PSO) abilities on natural gas demand forecasting in Iran's residential-commercial sector", *IPROMS-2009.*

[12] H. Gozde, M.C. Taplamacioglu, I. Kocaarslan," Application of Artificial Bees Colony Algorithm in an Automatic Voltage Regulator (AVR) System ", *International Journal on"Technical and Physical Problems of Engineering"(IJTPE) Published by International Organization on TPE (IOTPE), ISSN 2077-3528 IJTPE Journal www.iotpe.com ijtpe@iotpe.com, September 2010 Issue 4 Volume 2 Number 3 Pages 88-92.*

[13] Mohammad Ali Shafia, Mohammad Rahimi Moghaddam, Rozita Tavakolian, "A Hybrid Algorithm for Data Clustering Using Honey Bee Algorithm, Genetic Algorithm and K-Means Method" *Journal of Advanced Computer Science and Technology Research 1 (2011) 110-125.*

[14] Li-Yeh Chuang, Yu-Da Lin, and Cheng-Hong Yang," An Improved Particle Swarm Optimization for Data Clustering" *proceedings of the International MultiConference of Engineers & Computer Scientist 2012, vol I, IMECS 2012, March 14-16, 2012, Hong Kong, ISBN: 978-988-19251-1-4, ISSN: 2078-0958(print), ISSN: 2078-0966(online)*