



IGNITED MINDS
Journals

*International Journal of
Information Technology
and Management*

*Vol. VII, Issue No. IX,
August-2014, ISSN 2249-
4510*

**“COMPARATIVE ANALYSIS OF CLUSTERING
USING OPTIMIZATION ALGORITHMS”**

AN
INTERNATIONALLY
INDEXED PEER
REVIEWED &
REFEREED JOURNAL

“Comparative Analysis of Clustering Using Optimization Algorithms”

Altekar

Research Scholar, Singhania University, Jhunjhunu, Rajasthan

Abstract – In this paper we present Comparative Analysis of Clustering using Optimization Algorithms. The data clustering is recognized data analysis technique in data mining this paper is also based on comparative study of GA; ACO & PSO based Data Clustering techniques. To compare the results we use different metrics such as weighted arithmetic mean, standard deviation, Normalized absolute error & Precision value that measured the performance to compare and analyze the results.

Keywords: Clustering, Clustering techniques, Compare, Precision.

INTRODUCTION

Clustering techniques have become a revolution step in microarray data analysis because they can identify groups of genes or samples displaying a similar expression profile. Clustering is an unsupervised classification (grouping). attach label to each data points in a set, so that object in each set can share some common trait. i.e. maintaining students (name, Roll-id, Branch, collage name). Microarray technology has been recently introduced and provides solutions to a wide range of problems in medicine, health and environment, drug development, etc. Microarrays, widely recognized as the next revolution in molecular biology, enable scientists to analyze genes, proteins and other biological molecules on a genomic scale [M.Schena]. A microarray is a collection of spots containing DNA deposited on the solid surface of glass slide. Each of the spot contains multiple copies of single DNA sequence [Wei-Bang Chen]. Microarray expression technology helps in the monitoring of gene expression for tens and thousands of genes in parallel.

Clustering is a technique that attempts to organize unlabeled data objects into clusters or groups of similar objects. A cluster is a collection of data objects that are similar to one another with in the same cluster and are dissimilar to objects in other clusters. Clustering techniques have been used in a variety of fields like machine learning, artificial intelligence, web mining, image segmentation, life science and medicine, earth science, social science and economics. A comprehensive review of the state-of-the-art clustering techniques can be found in [Xu and Wunsch]. In recent years, due to the increasing computational speed of computers, heuristics are used to solve clustering problems. Various heuristic algorithms have already been proposed in the

literature such as Genetic Algorithm (GA), Ant Colony Optimization (ACO), Differential Evolution (DE) and Particle Swarm Optimization (PSO). Clustering techniques based on Evolutionary Computing and Swarm Intelligence algorithms have outperformed many classical techniques of clustering.

PSO was first introduced to optimize various continuous nonlinear functions by [Kennedy and Eberhart]. PSO algorithms have shown to successfully optimize a wide range of continuous functions. Many variants of PSO algorithms were developed over the years and applied to solve the various optimization problems. Literature review reveals that only few attempts has been made to solve the clustering problem using PSO algorithms and also there is no cross comparison among many PSO variants derived over the years for solving clustering problems. The performance of the well-known PSO algorithms are studied with the consideration of three clustering metrics such as TRace Within criteria (TRW), Variance Ratio Criteria (VRC) and Marriott Criteria (MC) using real world data sets. The results are compared with the published results of the basic PSO, GA and DE for all the clustering metrics. A detailed performance analysis of the PSO algorithms has been carried out based on Run Length Distribution (RLD).

PSO ALGORITHM IN CLUSTERING

PSO based clustering algorithm was first proposed by [Merwe] [Xiao] a hybrid approach to cluster the gene data. Self-Organizing Maps (SOM) trains the weights of the nodes in the first stage and weights were optimized using PSO approach. Chen and Ye [15] employed a PSO representation in which each particle corresponds to the centroids of the clusters.

Two dimensional and three-dimensional data were used for evaluation. [Orman] proposed a dynamic clustering system based binary PSO and K-means algorithm. The algorithm automatically identifies the number of clusters and employs a validity index to evaluate the clusters. [Cohen] proposed a Particle Swarm Clustering (PSC) algorithm where each particle represents a centroid in the input data space. The whole population is needed to present the final clustering solution. Sandra and [Krink] compared the performance of Differential Evolution (DE), Random Search (RS), PSO and GA for partitioned clustering problems. The empirical results show that PSO and DE perform better compared to GA and K means algorithms. Recently, [Swagatham] proposed an automatic clustering technique using an improved differential evolution algorithm.

GENETIC ALGORITHM (GA)

Genetic algorithms can be considered as a search technique whose algorithm is based on the mechanics of natural selection and genetics. It has been used in realms as diverse as search optimization and machine learning problems since they are not restricted by problem specific assumptions such as continuity or unimodality. In rough terms a genetic algorithm creates a collection of possible solutions to a specific problem. Initially the solutions are typically randomly generated solutions so their initial performance is normally poor. No matters how bad, there will be small segments of our collection of solutions that will be nearby our desired solution that is partially correct answers. Genetic Algorithms exploit this characteristic by recombination and progressively creating better solutions so that by the end of the run one have achieved one solution that is at least nearly optimal.

The flow steps of genetic algorithm for finding a solution of a given problem may be summarized as follows.

- Step 1: Initialize population for possible solution
- Step 2: Generate chromosomes of a population with 0's and 1's randomly
- Step 3: If the solution is satisfied then terminate else jump to next step
- Step 4: Compute population fitness value
- Step 5: Initialize number of generation
- Step 6: While number of generation * 2 ≤ termination condition; do
- Step 7: Select all the genetic solutions which can propagate to next generation
- Step 8: Increment number of generation
- Step 9: Identify each bit in genetic solution

Step 10: Perform crossover operation up to until 50% of bits are crossed

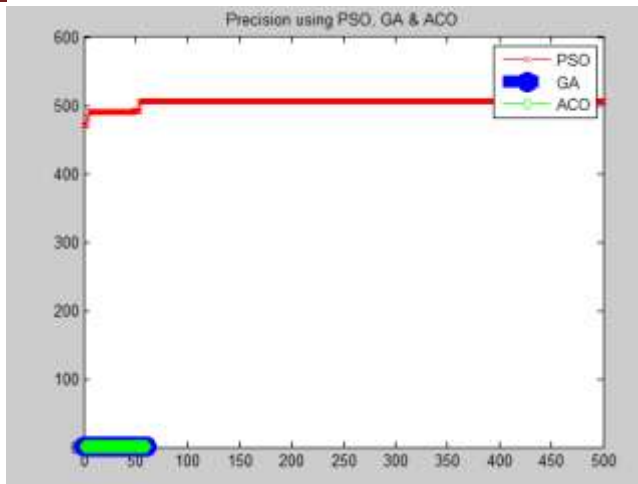
Step 11: end while

ANT COLONY OPTIMIZATION (ACO)

ACO algorithm was inspired by the behavior of ant colonies. Ants are social insects being interested mainly in the colony survival rather than individual survival. The interest of ants is the ability to find the shortest path from their nest to food. This idea was the source of the proposed algorithms inspired from ants' behavior. When searching for food ants initially explore the area surrounding their nest in a random manner and then while ants move it may leave a chemical pheromone trail on the ground. These ants are guided by pheromone smell. It tend to choose the paths marked by the strongest pheromone concentration .When an ant finds a food source then it evaluates the quantity and the quality of the food and carries some of it back to the nest. While ants returns the quantity of pheromone that an ant leaves on the ground may depend on the quantity and quality of the food. This pheromone trails will guide other ants to the food source. This type of indirect communication between the ants via pheromone trails enables them to find shortest paths between their nest and food sources. Figure 4 shows the cluster forming by using Ant Colony Optimization Algorithm.

COMPARATIVE ANALYSIS

The Figure defines percentage of precision using genetic, Ant Colony and Particle Swarm Optimization Algorithms. In the graph shown below, x-axis represents the 'number of person' having database be selected and y-axis represents the percentage accuracy. At some point of time GA has a constant precision value while dataset value has been increased. This means GA doesn't give the good results with large datasets. At some point of time ACO has 100% accuracy but it has decreased while dataset value has been increased. This means ACO doesn't give the good results with large datasets. But PSO has high accuracy with large datasets. This means PSO give the good results with large datasets.



CONCLUSION:

In this paper we obtainable the comparative study of GA, ACO & PSO based on the Data classification technique. The technique uses different metrics such as weighted arithmetic mean, Normalized absolute error & Precision value that measured the performance to compare and analyze the results. To assess the comparisons of performance of classification techniques is the difficult task unless same performance measures are used. The aim of classification is to achieve better classification using data set for detection of findings.

REFERENCES:

- [1] M.Schena, D.Shalon, Ronald W.davis and Patrick O.Brown, "Quantitative Monitoring of gene expression patterns with a complementary DNA microarray", Science, 270,199,pp:467-470.
- [2] Wei-Bang Chen, Chengcui Zhang and Wen-Lin Liu, "An Automated Gridding and Segmentation technique for cDNA Microarray Image Analysis", 19th IEEE Symposium on Computer- Based Medical Systems.
- [3] Rui Xu and Donald Wunsch II, Survey of Clustering Algorithms, IEEE Transactions on Neural Network, Vol. 16, No 3, May 2005.
- [4] Kennedy, J., Eberhart, R. C., Particle Swarm Optimization, Proceedings of the 1995 IEEE International Conference on Neural Networks, IEEE Press Piscataway, NJ, Vol. 4, pp. 1942 – 1948, 1995.
- [5] Jain, A.K., Murty, M.N., and P.J. Flynn, Data Clustering: A Review, ACM computing Survey, vol. 31, no 3, 1999.
- [6] Sandra Paterlini, Krink, T., Differential evolution and particle swarm optimization in partitional clustering, computational statistics & data analysis 50(5), 1220-1247, 2006.
- [7] Shi, Y., Eberhart, R., A modified particle swarm optimizer, In Proceeding of the 1998 IEEE World Congress on Computational Intelligence, Page 69-73, Piscataway, NJ, IEEE Press, 1998.
- [8] Shi, Y., Eberhart, R., Empirical study of particle swarm optimization. In Proceeding of the 1999 IEEE World Congress on Evolutionary Computing, Page 1945-1950. Piscataway, NJ, IEEE Press, 1999.
- [9] Van Der Merwe, D.W., Engelbrecht, A.P., Data clustering using particle swarm optimization, Proceedings of IEEE Congress on Evolutionary Computing 2003, Canberra, Australia, 215-220, 2003.
- [10] Xiao X, Dow ER, Eberhart RC, Miled ZB and Oppelt RJ, Gene Clustering Using Self-Organizing Maps and Particle Swarm Optimization, Proc of the 17th International Symposium on Parallel and Distributed Processing (PDPS '03),IEEE Computer Society, Washington DC, 2003.
- [11] Chen, C.Y., Ye, F., Particle swarm optimization algorithm and its applications to clustering analysis, Proceedings of IEEE International conference on Networking, Sensing and Control, 789-794, 2004.
- [12] Orman, M.G.H., Salman, A., Engelbrecht, A.P., Dynamic clustering using Particle Swarm Optimization with application in image segmentation, Pattern Analysis and Application, Vol. 8, No. 4, 332 – 344, 2005.
- [13] Cohen, S.C.M., and De Castro, L.N., Data Clustering with Particle swarms, IEEE Congress on Evolutionary Computations, Vancouver, Canada, 2006.
- [14] Swagatam Das, Ajith Abraham and Amit Konar, Automatic Clustering Using an Improved Differential Evolution Algorithm, IEEE Transactions on Systems Man and Cybernetics – PART A, IEEE Press, USA, 2008.