IJCSE International Journal of Computer Sciences and Engineering Open Access
Survey Paper Vol.-7, Special Issue-10, May 2019 E-ISSN: 2347-2693

GA-PSO Based Clustering Algorithm For Multi View Data: A Survey

Amitosh Patel[1*], Shuchita Mudgil[2]

[1]Department of Computer Science & Engg, Shri Ram Institute of Science and Technology, Jabalpur, India
[2]Department of Information Technology, Kalaniketan Polytechnic College, Jabalpur, India

Corresponding Author: amitoshpatel27@yahoo.com

DOI: https://doi.org/10.26438/ijcse/v7si10.101106 | Available online at: www.ijcseonline.org

***Abstract*—** Data mining an non-trivial extraction of novel, implicit, and actionable knowledge from large data sets is an evolving technology which is a direct result of the increasing use of computer databases in order to store and retrieve information effectively. This paper gives an idea of optimization algorithm by which the efficient result can be fetched. Optimization is a dire need for a huge amount of data processing. So that optimization is a challenging issue in data mining. It seems to be that there are many different approaches has been proposed by authors in order to optimize the results. Partial swam optimization and genetic algorithms are some sort of approach which can be used for optimization.

***Keywords*—** Data Mining, Clustering, Optimized Algorithm, PSO, GA

I. INTRODUCTION

Data mining (DM) is the extraction of useful and non-trivial information from the large amount of data that is possible to collect in many and diverse fields of science, business and engineering. DM is part of a bigger framework, referred to as knowledge discovery in databases (KDD) that covers a complex process from data preparation to knowledge modeling. Within this process, DM techniques and algorithms are the actual tools that analysts have at their disposal to find unknown patterns and correlation in the data. Typical DM tasks are classification (assign each record of a database to one of a predefined set of classes), clustering (find groups of records that are close according to some user defined metrics) and association rules (determine implication rules for a subset of record attributes). A considerable number of algorithms have been developed to perform these and others tasks, from many fields of science, from machine learning to statistics through various computing technologies like neural and fuzzy computing. What was a hand tailored set of case specific recipes, about ten years ago, is now recognized as a proper science. It is sufficient to consider the remarkable wide spectrum of applications where DM techniques are currently being applied to understand the ever growing interest from the research community in this domain. Data mining can be viewed as an essential step in the process of knowledge discovery. Data are normally preprocessed through data cleaning, data integration, data selection, and data transformation and prepared for the mining task. Started as little more than a dry extension of DM techniques, DM is now bringing important contributions in crucial fields of investigations and in the traditional sciences like astronomy, high energy physics, biology and medicine that have always provided a rich source of applications to data miners. An important field of application for data mining techniques is also the World Wide Web. The Web provides the ability to access one of the largest data repositories, which in most cases still remains to be analyzed and understood. Recently, data mining techniques are also being applied to social sciences, home land security and counter terrorism. A DM system is therefore composed of a software environment that provides all the functionalities to compose DM applications, and a hardware back-end onto which the DM applications are executed.

This work discusses about one of the application areas of partition based clustering algorithms k-Means and Fuzzy C-Means by means of an experimental approach choosing a real time telecommunication data. Many applications have been proposed by using different algorithms. Now, it is necessary to discuss some of the applications of related areas. This will be helpful to understand the related breakthrough in computations and engineering applications. They discuss that both theoretical and practical efforts in band images often neglect the characteristics having interactions and mutual influence among attributes or criteria, even in the stages of different brand life cycles. This study aims to create a hierarchical framework for brand image management. The analytical network process and fuzzy sets theory have

been applied to both mindshare in brand images and inherent interaction/interdependencies among diverse information resources. A real empirical application is demonstrated in the department store. Both the theoretical and practical background of this work have shown the fuzzy analytical network process can capture expert's knowledge existing in the form of incomplete and vague information for the mutual inspiration on attribute and criteria of brand image management.

In the two-level variable weighting method, the variable weights V are used to identify the important variables in each view, and the view weights W are used to identify compact cluster structures within these views. If the view contains compact cluster structures, a large view weight is assigned so as to enhance the effect of such view; on the contrary, if the view contains loose cluster structures, a small view weight is assigned to eliminate the effect of such view. Compared with the traditional variable weighting method, the new method can take both individual variables and multiple views into consideration and capture the differences among different views and variables.

The traditional variable weighting methods suffer from unbalanced phenomenon: the view with more variables will play more important role than the view with less variables. In the two-level variable weighting method, the view weights will be only determined in the view level, while the variable weights will be only determined in a view. Therefore, the two levels of variable weights will eliminate the unbalanced phenomenon and compute more objective weights.

II. PARTICLE SWARM OPTIMIZATION

TY Swarm Intelligence (SI) is an innovative distributed intelligent paradigm for solving optimization problems that originally took its inspiration from the biological examples by swarming, flocking and herding phenomena in vertebrates.

Particle Swarm Optimization (PSO) incorporates swarming behaviors observed in flocks of birds, schools of fish, or swarms of bees, and even human social behavior, from which the idea is emerged. PSO is a population-based optimization tool, which could be implemented and applied easily to solve various function optimization problems, or the problems that can be transformed to function optimization problems. As an algorithm, the main strength of PSO is its fast convergence, which compares favorably with many global optimization algorithms like Genetic Algorithms (GA), Simulated Annealing (SA) and other global optimization algorithms. For applying PSO successfully, one of the key issues is finding how to map the problem solution into the PSO particle, which directly affects its feasibility and performance.

A swarm is a large number of homogenous, simple agents interacting locally among themselves, and their environment, with no central control to allow a global interesting behavior to emerge. Swarm-based algorithms have recently emerged as a family of nature-inspired, population-based algorithms that are capable of producing low cost, fast, and robust solutions to several complex problems [1][2]. Swarm Intelligence (SI) can therefore be defined as a relatively new branch of Artificial Intelligence that is used to model the collective behavior of social swarms in nature, such as ant colonies, honey bees, and bird flocks. Although these agents (insects or swarm individuals) are relatively unsophisticated with limited capabilities on their own, they are interacting together with certain behavioural patterns to cooperatively achieve tasks necessary for their survival. The social interactions among swarm individuals can be either direct or indirect [3]. Examples of direct interaction are through visual or audio contact, such as the waggle dance of honey bees. Indirect interaction occurs when one individual changes the environment and the other individuals respond to the new environment, such as the pheromone trails of ants that they deposit on their way to search for food sources. This indirect type of interaction is referred to as stigmergy, which essentially means communication through the environment [4]. The area of research presented in this depth paper focuses on Swarm Intelligence. More specifically, this paper discusses two of the most popular models of swarm intelligence inspired by ant's stigmergic behavior and birds'flocking behavior.

Swarm intelligence models are referred to as computational models inspired by natural swarm systems. To date, several swarm intelligence models based on different natural swarm systems have been proposed in the literature, and successfully applied in many real-life applications. Examples of swarm intelligence models are: Ant Colony Optimization, Particle Swarm Optimization, Artificial Bee Colony, Bacterial Foraging, Cat Swarm Optimization, Artificial Immune System, and Glowworm Swarm Optimization. Here the primarily focus on two of the most popular swarm intelligences models, namely, Ant Colony Optimization and Particle Swarm Optimization.

Swarm Intelligence Applications SI techniques are population-based stochastic methods used in combinatorial optimization problems in which the collective behavior of relatively simple individuals arises from their local interactions with their environment to produce functional global patterns. There is no best optimization technique for all the problems. Each method has its advantages, and the set of parameters define the quality of the solution. Engineers are increasingly interested in swarm behavior since the resulting swarm intelligence can be applied in optimization (e.g. in telecommunication systems), robotics, traffi patterns in transportation systems, military applications, etc. More

and more new applications arise from the research in SI. Every problem, application, that in its base has some kind of optimization can be tackled with SI techniques. Swarm robotics is a rapidly developing fiel that gets the inspiration from swarm intelligence. The animal societies are good examples of what future robotic swarms might achieve, but they are not at all limited by biological plausibility. The efficien y, fl xibility, robustness, and cost are possible criteria that should be used in development of such systems.

Many aspects of the collective activities of social behavior in nature are self-organized. Self organization (SO) is a set of dynamical mechanisms whereby structures appear at the global level of a system from interactions among its lower-level components. SO relies on four basic ingredients:

- Positive feedback (amplification) examples are recruitment and reinforcement. For instance, recruitment to a food source is a positive feedback that relies on trail-laying and trail-following in some ant species, or dances in bees.
- Negative feedback: counterbalances positive feedback and helps to stabilize the collective pattern in the form of saturation, exhaustion or competition.
- Amplificatio of fluctuations randomness is often crucial since it enables discovery of new solutions.
- Multiple interactions: a minimal density of mutually tolerant individuals is required to generate a self-organized structure.

The objective of SI is to model the simple behavior of the individuals, their local interactions with the environment and neighboring individuals, in order to obtain more complex behaviors that can be used to solve complex problems, mostly optimization problems. A critical number of individuals are required for ”intelligence” to arise. The two best known SI algorithms are: Particle Swarm Optimization (PSO) and Ant Colony Optimization (ACO).

The simplest mathematical models of animal swarms generally represent individual animals as following three rules:

1 Move in the same direction as your neighbor
2 Remain close to your neighbors
3 Avoid collisions with your neighbors

Particle Swarm Optimization (PSO) was originally inspired by the flockin behavior of birds. In terms of this bird flockin analogy, a particle swarm optimizer consists of a number of particles, or birds, that fl around and search space, or the sky, for the best location. The individuals communicate either directly or indirectly with one another search directions (gradients). Each of the particles in a swarm corresponds to a simple agent that moves through a multi-dimensional search space sampling an objective function at various positions. The best solution can be represented as a point or surface in the search space. Potential solutions are plotted in this space and seeded with an initial velocity. The motion of a given particle is dictated by its velocity which is continuously updated in order to pull it towards its own best position and the best positions experienced by the neighbors in the swarm. The performance of each particle is evaluated using a predefine fitnes function which encapsulates the characteristics of the optimization problem. Over time, particles accelerate towards those with better fitness values. PSO is a simple, but powerful search technique. It has few parameters to adjust and is easy to implement.

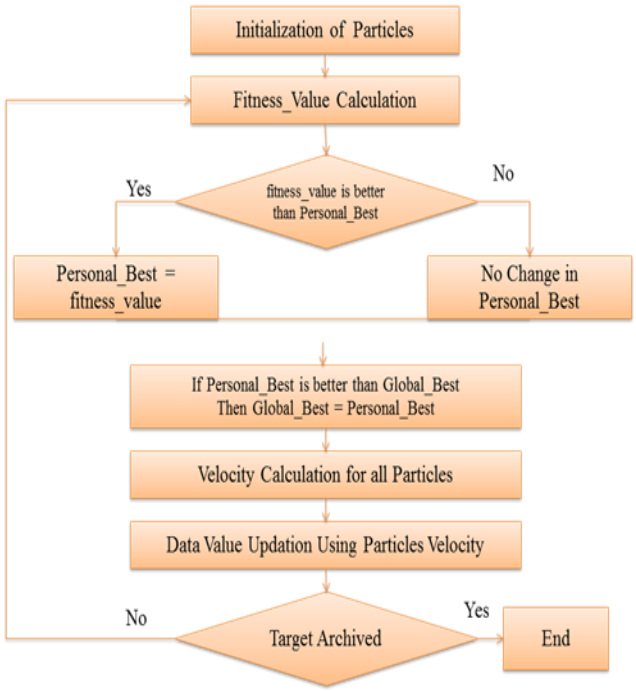

Figure 1: Procedures for particle swarm optimization

Many current models use variations on these rules, often implementing them by means of concentric "zones" around each animal. In the zone of repulsion, very close to the animal, the focal animal will seek to distance itself from its neighbors to avoid collision. Slightly further away, in the zone of alignment, the focal animal will seek to align its direction of motion with its neighbors. In the outermost zone of attraction

Particle Swarm Optimization Bird flocking and fish schooling are the inspirations from nature behind particle swarm optimization algorithms. It was first proposed by Eberhart and Kennedy. Mimicking physical quantities such as velocity and position in bird flocking, artificial particles are constructed to “fly” inside the search space of optimization problems.

1. For all particles
2. {
    a. Initialization of particle
3. }
4. Do for maximum iterations
5. {
    a. For all particles

b. {
    i. Calculation of fitness_value
    ii. If the fitness_value is better than Personal_Best
c. {
d. Set Personal_Best = fitness_value

6. }
7. If Personal_Best is better than Global_Best
8. {
    a. Set Global_Best = Personal_Best
9. }
10. }
11. For all particles
12. {
    a. Calculate particle Velocity
    b. Use Global_Best & Velocity to update particle Data
13. }

III. RELATED WORK

This section gives an extensive literature survey on the multiple relational classification using genetic algorithms. We have studied various research paper and journal and know about data classification. All methodology and process are not described here. But some related work in the field of association classification discuss by the name of authors and their respective title.

Xiaojun Chen [1] proposes TW-k-means, an automated two-level variable weighting clustering algorithm for multi-view data, which can simultaneously compute weights for views and individual variables. In this algorithm, a view weight is assigned to each view to identify the compactness of the view and a variable weight is also assigned to each variable in the view to identify the importance of the variable. Both view weights and variable weights are used in the distance function to determine the clusters of objects. In the new algorithm, two additional steps are added to the iterative k-means clustering process to automatically compute the view weights and the variable weights. They used two real-life data sets to investigate the properties of two types of weights in TW-k-means and investigated the difference between the weights of TW-k-means and the weights of the individual variable weighting method. The experiments have revealed the convergence property of the view weights in TW-k-means. They compared TW-k-means with five clustering algorithms on three real-life data sets and the results have shown that the TW-k-means algorithm significantly outperformed the other five clustering algorithms in four evaluation indices.

Many real-world applications expose the nonlinear manifold structure of the lower dimension rather than its high-dimensional input space. This greatly challenges most existing clustering and representative selection algorithms which do not take the manifold characteristics into consideration. The performance of the corresponding learning algorithms can be greatly improved if the manifold structure is considered. Enmei Tu [2] propose a graph-based k-means algorithm, GKM, which bears the simplicity of classic k-means while incorporating global information of data geometric distribution. GKM fully exploits the intrinsic manifold structure for appropriate data clustering and representative selection. GKM is evaluated on both synthetic and real-life data sets and achieves very impressive results compared to the state-of-the-art approaches, including classic k-means, kernel k-means, spectral clustering, and clustering through ranking and for representative selection. Given the widespread appearance of manifold structures in real world problems, GKM shows promising potential for partitioning manifold-distributed data.

Data mining is the process of discovering meaningful new correlation, patterns and trends by sifting through large amounts of data, using pattern recognition technologies as well as statistical and mathematical techniques. Cluster analysis is often used as one of the major data analysis technique widely applied for many practical applications in emerging areas of data mining. Two of the most delegated, partition based clustering algorithms namely k-Means and Fuzzy C-Means are analyzed in this research work. These algorithms are implemented by means of practical approach to analyze its performance, based on their computational time. The telecommunication data is the source data for this analysis. The connection oriented broad band data is used to find the performance of the chosen algorithms. T. Velmurugan [3] evaluate distance (Euclidian distance) between the server locations and their connections are rearranged after processing the data. The computational complexity (execution time) of each algorithm is analyzed and the results are compared with one another. By comparing the result of this practical approach, it was found that the results obtained are more accurate, easy to understand and above all the time taken to process the data was substantially high in Fuzzy C-Means algorithm than the k-Means.

Applying k-Means to minimize the sum of the intra-cluster variances is the most popular clustering approach. However, after a bad initialization, poor local optima can be easily obtained. To tackle the initialization problem of k-Means, Grigorios Tzortzis [4] propose the MinMax k-Means algorithm, a method that assigns weights to the clusters relative to their variance and optimizes a weighted version of the k-Means objective.Weights are learned together with the cluster assignments, through an iterative procedure. The proposed weighting scheme limits the emergence of large

variance clusters and allows high quality solutions to be systematically uncovered, irrespective of the initialization. Experiments verify the effectiveness of our approach and its robustness over bad initializations, as it compares favorably to both k-Means and other methods from the literature that consider the k-Means initialization problem.

Cheng-Huang [5] propose a new searching scheme, candidate group search, for solving the K-harmonic means clustering problem. The candidate view is based on the distance between the centroid to all entities. For every centroid, they screening the possible entities which distance is near the any pth quartile. By taking each center as a core and defining possible candidate group, CGS can reduce the computational time since the searching size is smaller comparing with VNS. At the same time, CGS also maintain good converge because this candidate group possibility to escape from local optimum. This explains why CGS has better chance to find the optimum solution and reduces the computational time significantly comparing with VNS which searches in the whole set by a random approach. Computational results showed CGS gets considerably improvement than two heuristic methods, VNS and tabu search. Moreover, computational results in much larger test instances are also presented. These results convinced CGS is better than tabu search and VNS. Future research may include testing CGS in larger instances. Converge properties of CGS could be an interesting problem for future research.
Fuzzy c-means clustering algorithm (FCM) is a method that is frequently used in pattern recognition. It has the advantage of giving good modeling results in many cases, although, it is not capable of specifying the number of clusters by itself. Yi Ding [6] aimed at the problems existed in the FCM clustering algorithm, a kernel-based fuzzy c-means (KFCM) is clustering algorithm is proposed to optimize fuzzy c-means clustering, based on the Genetic Algorithm (GA) optimization which is combined of the improved genetic algorithm and the kernel technique (GAKFCM). In this algorithm, the improved adaptive genetic algorithm is used to optimize the initial clustering center firstly, and then the KFCM algorithm is availed to guide the categorization, so as to improve the clustering performance of the FCM algorithm. In the paper, Matlab is used to realize the simulation, and the performance of FCM algorithm, KFCM algorithm and GAKFCM algorithm is testified by test datasets. The results proved that the GAKFCM algorithm proposed overcomes FCM′s defects efficiently and improves the clustering performance greatly.

Mukhopadhyay,[12] suggest a multi objective genetic algorithm-based approach for fuzzy clustering of categorical data is proposed that encodes the cluster modes and simultaneously optimizes fuzzy compactness and fuzzy separation of the clusters. Moreover, a novel method for obtaining the final clustering solution from the set of resultant Pareto-optimal solutions in proposed. This is based on majority voting among Pareto front solutions followed by k-nn classification. The performance of the proposed fuzzy categorical data-clustering techniques has been compared with that of some other widely used algorithms, both quantitatively and qualitatively. For this purpose, various synthetic and real-life categorical datasets have been considered. Also, a statistical significance test has been conducted to establish the significant superiority of the proposed multi objective approach.

J. Zhang [13] presents the use of fuzzy logic to adaptively adjust the values of px and pm in GA. By applying the K-means algorithm, distribution of the population in the search space is clustered in each generation. A fuzzy system is used to adjust the values of px and pm. It is based on considering the relative size of the cluster containing the best chromosome and the one containing the worst chromosome. The proposed method has been applied to optimize a buck regulator that requires satisfying several static and dynamic operational requirements. The optimized circuit component values, the regulator's performance, and the convergence rate in the training are favorably compared with the GA using fixed values of px and pm. The effectiveness of the fuzzy-controlled crossover and mutation probabilities is also demonstrated by optimizing eight multidimensional mathematical functions

In Gen Cluster two-level variable weighting method [6], high complexity of the genetic algorithms including Gen Cluster can cause problems for clustering data sets with huge number of records. A possible solution to this problem can be as follows: (1) take a random sample of a manageable number of records (as many as possible) into a sample data set, (2) apply Gen Cluster on the sample data set and get the best chromosome, (3) use the genes of the best chromosome as the initial seeds of the K-Means algorithm which is applied on the whole (not just the sample) data set. Due to the low complexity of K-Means its application on the whole data set should not be a problem.

IV. CONCLUSION

Ac PSO based variable weighted clustering algorithm for multi view data will optimized the clustering technique in effective and efficient way. The rules discovered are generally with high accuracy, generalization and comprehensibility. Particle swarm optimization (PSO) is a population based stochastic optimization technique developed by Dr. Eberhart and Dr. Kennedy in 1995, inspired by social behavior of bird flocking or fish schooling. PSO shares many similarities with evolutionary computation techniques such as Genetic Algorithms (GA). The system is initialized with a population of random solutions and searches for optima by updating generations. However, unlike GA, PSO has no evolution

operators such as crossover and mutation. In PSO, the potential solutions, called particles, fly through the problem space by following the current optimum particles.

REFERENCES

[1] Satyasai Jagannath Nanda, Ganapati Panda, Automatic clustering algorithm based on multi-objective Immunized PSO to classify actions of 3D human models, Engineering Applications of Artificial Intelligence, Volume 26, Issues 5–6, May–June 2013, Pages 1429-1441

[2] Tuğrul Çavdar, PSO tuned ANFIS equalizer based on fuzzy C-means clustering algorithm, AEU - International Journal of Electronics and Communications, Volume 70, Issue 6, June 2016, Pages 799-807,

[3] Amin Khatami, Saeed Mirghasemi, Abbas Khosravi, Chee Peng Lim, Saeid Nahavandi, A new PSO-based approach to fire flame detection using K-Medoids clustering, Expert Systems with Applications, Volume 68, February 2017, Pages 69-80

[4] Hui-Liang Ling, Jian-Sheng Wu, Yi Zhou, Wei-Shi Zheng, How many clusters? A robust PSO-based local density model, Neurocomputing, Volume 207, 26 September 2016, Pages 264-275

[5] Chen Jinyin, Lin Xiang, Zheng Haibing, Bao Xintong, A novel cluster center fast determination clustering algorithm, Applied Soft Computing, Volume 57, August 2017, Pages 539-55

[6] Xiaojun Chen, Xiaofei Xu, Joshua Zhexue Huang, and Yunming Ye "TW-k-Means: Automated Two-Level Variable Weighting Clustering Algorithm for Multi-view Data" in IEEE Transactions On Knowledge And Data Engineering, Vol. 25, No. 4, April 2013

[7] Enmei Tu , Longbing Cao , Jie Yang , Nicola Kasabov "A novel graph-based k-means for nonlinear manifold clustering and representative selection" in Elsevier transaction of Neuro computing 143 (2014) 109–122

[8] T. Velmurugan "Performance based analysis between k-Means and Fuzzy C-Means clustering algorithms for connection oriented telecommunication data" in Elsevier transaction of Applied Soft Computing 19 (2014) 134–146

[9] Grigorios Tzortzis "The Min Max k-Means clustering algorithm" in Elsevier transaction of Pattern Recognition 47 (2014) 2505–2516

[10] Cheng-Huang Hung , Hua-Min Chiou b, Wei-Ning Yang "Candidate groups search for K-harmonic means data clustering" in Elsevier transaction of Applied Mathematical Modelling 37 (2013) 10123–10128

[11] Yi Ding, Xian Fu, Kernel-based fuzzy c-means clustering algorithm based on genetic algorithm, Neuro Computing, Volume 188, 5 May 2016, Pages 233-238

[12] A. Mukhopadhyay, U. Maulik and S. Bandyopadhyay, "Multiobjective Genetic Algorithm-Based Fuzzy Clustering of Categorical Attributes," in *IEEE Transactions on Evolutionary Computation*, vol. 13, no. 5, pp. 991-1005, Oct. 2009.

[13] J. Zhang, H. S. H. Chung and W. L. Lo, "Clustering-Based Adaptive Crossover and Mutation Probabilities for Genetic Algorithms," in *IEEE Transactions on Evolutionary Computation*, vol. 11, no. 3, pp. 326-335, June 2007.

