Fuzzy Clustering with Particle Swarm Intelligence for Large Dataset Classification

Ashit Kumar Dutta

Department of computer science, Shaqra University, Saudi Arabia

Abstract **– The most challenging problem in data mining is deriving knowledge from large dataset. Existing methods have better performance in medium – scale dataset but the level of performance degrades in large datasets. Swarm intelligence (SI) is a computational method to solve complex problems inspired from biological phenomena like flock of birds, shoal of fish, herd of sheep and swarm of bees. The ant colony optimization is the multi-agent system solving of problems through cooperation like ants. Particle swarm optimization (PSO) is one of the methods successfully implemented with fuzzy concepts and solved complex problems. The objective of the research is to classify the large dataset using fuzzy clustering with PSO. The experiment results proved that the proposed method is more effective and produces optimum accuracy.**

Keywords **– Swarm intelligence, Particle swarm optimization, Fuzzy clustering, Classification, Clustering.**

1. Introduction

Swarm Intelligence (SI) is a concept of artificial intelligence. It consists of simple agents interacting with one another and with their environment [1][2]. A swarm is designed in self – propelled particles by a set of particles with constant speed and each time increments the average direction of motion of the other particles in their local environment [3][4]. Evolutionary algorithms, particle swarm optimization (PSO), and ant colony optimization and their attributes are ruling the field of nature – inspired meta heuristics [5]. PSO is an optimization algorithm to deal with problems and generate the best solution in the representation of point or surface in an n-dimensional space [6][7]. The main advantage of this approach is simulated annealing that makes up the particle swarm making the technique resilient to the problem of local minima. Figure 1.1. specifies the fuzzy classification with PSO. The combination used for the purpose of the classification of the dataset. Section 2 depicts review of the literature. Section 3 clarifies the methodology of the research. Section 4 gives the experimental results of the proposed method.

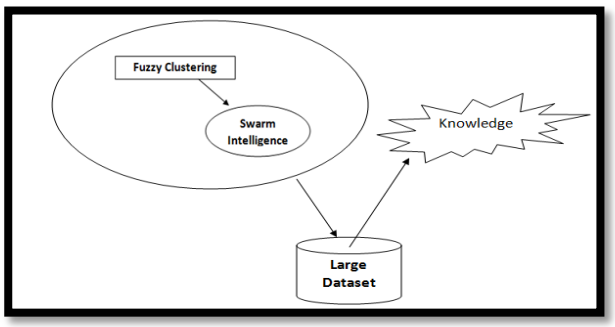

Figure 1. Fuzzy clustering with swarm Intelligence Technique

2. Review of Literature

Hassan M. Elragal [8] proposed a research on classifiers using both fuzzy and SI. Genetic algorithms (GA) are in practice to set right fuzzy rules. SI is the technique useful for classification related problems. PSO is the SI method to reduce the complexity in feature selection from large dataset. The positions and velocities will be in a vector

DOI: 10.18421/TEM74-06
https://dx.doi.org/10.18421/TEM74-06

Corresponding author: Ashit Kumar Dutta,
Department of computer science, Shaqra University, Saudi Arabia
Email: amela.jusic@untz.ba

Received: 06 April 2018.
Accepted: 24 September 2018.
Published: 26 November 2018.

format for the nth particle in the duration n+1 using fitness function. The research used both Takagi and Mamdani based inference systems to optimize parameters of input membership function. The performance of the proposed classifiers in the research showed higher classification accuracy with little number of fuzzy rules.

Rocio Perez Prado et al. [9]., have proposed a fuzzy rule based meta scheduling method with the support of SI for Grid computing. The quality of the knowledge produced by the system is the success of the method. The part of genetic fuzzy learning strategies used to provide effective population evolved during classification of dataset. Generalized knowledge acquisition with a SI method allows generalization to be utilised with fuzzy methods to minimize the number of stages required by the canonical strategy. The study reveals particles within the search space to achieve the optimum location for these objects. The updating process indicates the quality of each object in each iteration.

Y. Wang et al. [10] proposed a memory-based multiagent coevolution algorithm for robust tracking the moving objects. Each agent can remember, retrieve, or forget the appearance of the object through its own memory system by its own experience. Experimental results show that their proposed method can deal with large appearance changes and heavy occlusions when tracking a moving object.

The paper by Q. Ni and J. Deng [11] did a survey on the performance of PSO with the proposed random topologies and explores the relationship between population topology and the performance of PSO from the perspective of graph theory characteristics in population topologies. Further, in a relatively new PSO variant which named logistic dynamic particle optimization, an extensive simulation study is presented to discuss the effectiveness of the random topology and the design strategies of population topology.

Y. Zhou and H. Zheng [12] proposed a novel complex valued cuckoo search algorithm. They use complex-valued encoding to expand the information of nest individuals and denote the gene of individuals by plurality. The value of independent variables for objective function is determined by modules, and a sign of them is determined by angles. The position of nest is divided into real part gene and imaginary gene. Six typical functions are tested, and the usefulness of the proposed algorithm is verified.

The paper by R. Alwee et al.[13] proposed a hybrid model that combines support vector regression (SVR) and autoregressive integrated moving average (ARIMA) to be applied in crime rates forecasting. Particle swarm optimization is used to estimate the parameters of the SVR and ARIMA models. The experimental results show that their proposed hybrid model is able to produce more accurate forecasting results as compared to the individual models.

K. S. Lim et al. [14] described an improved Vector Evaluated Particle Swarm Optimization algorithm by incorporating the non dominated solutions as guidance for a swarm rather than using the best solution from another swarm. The results suggest that the improved Vector Evaluated Particle Swarm Optimization algorithm has impressive performance compared with the conventional Vector Evaluated Particle Swarm Optimization algorithm.

Ha Zhenya et al. [15] proposed a four layer fuzzy neural network to realize knowledge from samples. The structure of the network provides flexible rules for easy training of the system. The combination of fuzzy and neural networks showed effective knowledge acquisition. The research produced knowledge from large dataset using an adaptive fuzzy neural network.

Caichang Ding et al. [16] did a survey on SI and its applications. SI is based on the methods underlying the performance of systems having multiple agents and highly distributed control. SI follows bottom up approach consisting of simple set of rules. SI is a new way to control multiple agent systems. The structure of SI does not need any supervision for the effective results. SI provides an example of a highly decentralized architecture to improve simplicity. It is a massively parallel system, emphasizing self – organization capabilities. It consists of finite set of similar agents with limited capabilities. Stigmergy communication is a communication that takes place between individuals through environment. Pheromone based stigmergy is the capability of SI to find the shortest path from a food source to the nest.

Hisao Ishibuchi et al. [17] proposed a fuzzy rule based classifier to visualize classification results for the users. The advantage of fuzzy rule based classifiers is interpretability. It is easy to interpret linguistics using fuzzy rules in a human understandable manner. It is necessary to use multiple fuzzy partitions with different granularities for fuzzy rule generation. GA is used to choose only a small number of candidate rules to construct a compact fuzzy rule based classifier. Fuzzy rules should be used with two antecedent conditions to understand the search space.

3. Research Methodology

The proposed method uses FC and PSO for the classification of dataset, PSO is used to calibrate the attributes of input membership functions, output membership functions and structure of the fuzzy rules [18] [19]. The number of fuzzy rules is gradually changed from 3 to 12.

The procedure involved in the classification of dataset using proposed method is as follows:

Step 1: Divide the database into training and testing data set.

Step 2: Create N number of iteration for the system to learn the environment. Each object consists of:

i. Parameters of input and output membership functions
ii. Random rules generated according to the assigned number of rules.

Step 3: For each object, construct fuzzy classifier with PSO, use training data set and calculate time, F1 score and accuracy.

Step 4: Implement PSO with fuzzy classifier to produce results by adjusting fitness function and classify the dataset.

Step 5: Repeat steps 2 to 4 until a solution results in perfect classification of all training samples.

Step 6: Select the best performing fuzzy classifier based on the training data set results recorded in step 5. The best performing classifier is the one with maximum classification accuracy and minimum number of rules.

Step 7: Apply testing data set on the selected fuzzy classifier and record classification accuracy. To identify the size of each particle, assume there are N inputs and one output.

The total number of variables for input and output membership functions will be (Ni*3+Nc)*2. Assume there are N rules and each rule has (Ni+l) variables. Then, the total number of variables that need to be optimized (particle size) will be (Ni*3+Nc)*2 + (Ni+l)* Nr.

4. Results and Experiment

The research has used 4 real world datasets IRIS, Phoneme, Diabetes and Heart. Support vector Machines (SVM) and Random Forest (RF) are the other methods compared with the proposed Fuzzy classifier with PSO (FCPSO). The IRIS datasets were manually classified into 3 classes. The Phoneme, Diabetes and Heart datasets were classified into 2 classes [20][21]. The datasets were divided into Training and Testing samples. The training phase gives an opportunity to the method to learn the dataset and produce the effective results.

The research objective is to present an effective classifier with optimum accuracy in limited time interval and reduced number of fuzzy rules. Tables 4.1, 4.2, 4.3 and 4.4 show the F1 measure of the methods for the datasets. The proposed method has reached an average F1 score of 88 % and other methods have acquired an average of 85% during training phase [22][23]. The training phase results may differ in the testing phase as the samples are unique and there is not relationship between training and testing dataset.

Table 1. IRIS Dataset – Training phase

Metrics / Algorithm	Precision	Recall	F1 Measure
SVM	81.2	83.1	82.14
RF	84.3	85.2	84.75
FCPSO	87.3	88.3	87.8

Table 2. Phoneme Dataset – Training phase

Metrics / Algorithm	Precision	Recall	F1 Measure
SVM	80.6	79.6	80.1
RF	82.3	80.7	81.49
FCPSO	83.5	84.5	84

Table 3. Diabetes Dataset – Training phase

Metrics / Algorithm	Precision	Recall	F1 Measure
SVM	82.6	84.3	83.44
RF	83.4	84.9	84.14
FCPSO	85.6	85.2	85.4

Table 4. Heart Dataset – Training phase

Metrics / Algorithm	Precision	Recall	F1 Measure
SVM	80.4	86.7	83.43
RF	81.4	84.9	83.11
FCPSO	83.4	86.1	84.73

The testing phase reveals the actual image of the system used for the classification. Precision, recall and F1 measures are the evaluation method to indicate the retrieval capacity of the system. Tables 4.5, 4.6, 4.7 and 4.8 are the results of the testing phase. F1 measure of FCPSO is more effective than other methods. FCPSO has achieved an average of 87% of F1 score for all dataset.

Table 5. IRIS Dataset – Testing phase

Metrics / Algorithm	Precision	Recall	F1 Measure
SVM	86.4	84.6	85.49
RF	85.8	85.3	85.55
FCPSO	88.4	88.9	88.65

Table 6. Phoneme Dataset – Testing phase

Metrics / Algorithm	Precision	Recall	F1 Measure
SVM	84.3	85.3	84.8
RF	85.3	86.2	85.75
FCPSO	87.6	86.8	87.2

Table 7. Diabetes Dataset – Testing phase

Metrics / Algorithm	Precision	Recall	F1 Measure
SVM	84.3	84.9	84.6
RF	85.1	86.2	85.65
FCPSO	88.6	89.3	88.95

Table 8. Heart Dataset – Testing phase

Metrics / Algorithm	Precision	Recall	F1 Measure
SVM	83.5	84.3	83.9
RF	84.6	85.1	84.85
FCPSO	87.6	88.3	87.95

Accuracy differs from F1 measure. Dataset used for the work were classified manually and calculation of accuracy carried out accordingly. Table 4.9 shows the accuracy of methods used in the research. SVM and RF have reached similar accuracy. The proposed method has reached an average of 92% for IRIS dataset, 89% for Phoneme dataset, 92% for diabetes dataset and 91% for heart dataset. Number of rules are similar for all dataset. SVM and RF got less number of rules than the proposed method but accuracy is much less than the proposed method. Figure 4.1 shows the accuracy details of the methods used in the research.

Table 9. Comparison of accuracy produced by classifiers

Dataset		SVM	RF	FCPSO
IRIS	Class 1	86.3	87.2	89.6
	Class 2	87.6	85.6	91.2
	Class 3	84.9	83.9	94.6
	No. of Rules	11	9	9
Phoneme	Class 1	84.6	87.6	89.7
	Class 2	85.3	86.9	90.8
	No. of Rules	8	9	8
Diabetes	Class 1	88.7	87.9	91.3
	Class 2	84.6	85.6	94.6
	No. of Rules	9	10	9
Heart	Class 1	84.9	85.6	92.4
	Class 2	87.8	88.9	91.6
	No. of Rules	8	7	9

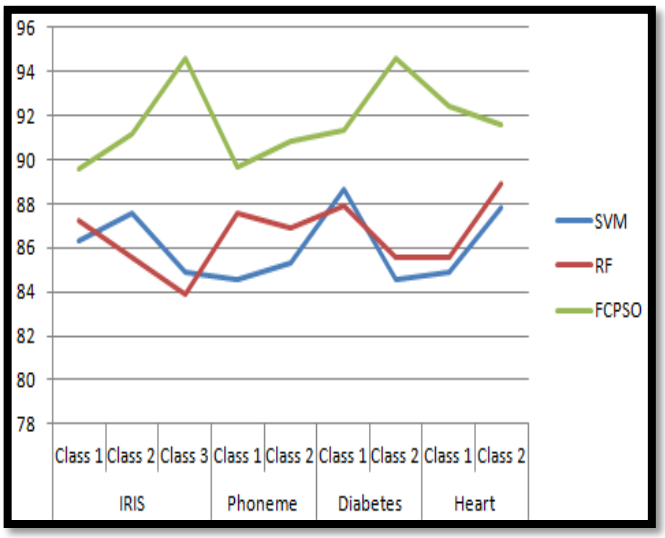

Figure 2. Comparison of accuracy of methods

Table 4.10 shows the time taken by the methods during training phase. Figure 4.2 depicts the training time denoted in Table 4.9.

Table 10. Training Time (Seconds) of Methods

Dataset		SVM	RF	FCPSO
IRIS	Class 1	0.89	0.81	0.79
	Class 2	0.73	0.83	0.76
	Class 3	0.84	0.84	0.81
Phoneme	Class 1	0.87	0.79	0.86
	Class 2	0.89	0.81	0.82
Diabetes	Class 1	0.85	0.86	0.84
	Class 2	0.78	0.84	0.86
Heart	Class 1	0.69	0.83	0.87
	Class 2	0.77	0.89	0.82

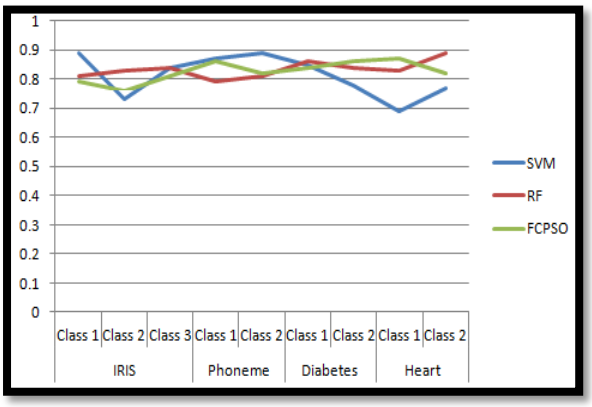

Figure 3. Comparison of Training time

Table 4.11 shows the testing time of the methods. SVM took a minimum duration of 0.58 seconds for IRIS dataset. RF has taken a minimum duration of 0.54 seconds for IRIS dataset. FCPSO has taken 0.49 seconds for Diabetes dataset. The proposed FCPSO took least time to classify the dataset with few fuzzy rules. The part of PSO in the research is the key for the performance of the work. Figure 4.3 shows the testing time of the proposed research.

Table 11. Testing Time(Seconds) of Methods

Dataset		SVM	RF	FCPSO
IRIS	Class 1	0.71	0.79	0.64
	Class 2	0.58	0.54	0.51
	Class 3	0.64	0.56	0.53
Phoneme	Class 1	0.72	0.59	0.56
	Class 2	0.76	0.71	0.67
Diabetes	Class 1	0.69	0.68	0.64
	Class 2	0.71	0.66	0.49
Heart	Class 1	0.81	0.62	0.52
	Class 2	0.86	0.69	0.51

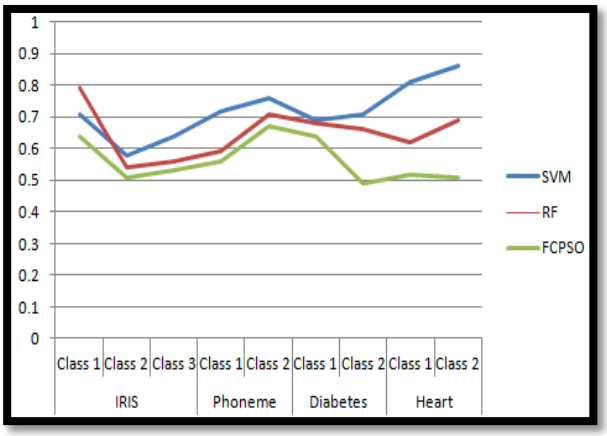

Figure 4. Comparison of Testing time

5. Conclusion

The proposed research classifies the large dataset into desired classified results. PSO is the new technique of SI used to tune the machine learning methods to produce effective results. The objective of the research is to classify the dataset with limited fuzzy rules. The output of the research is satisfied and the accuracy, F1 measure and time clearly specifies the success of the work. The work has achieved an average F1 score of 88% and 89% of accuracy. The real time dataset were used and the result will be useful for the research based on SI.

References

[1]. Al-Sultana, K. S., & Khan, M. M. (1996). Computational experience on four algorithms for the hard clustering problem. *Pattern recognition letters*, *17*(3), 295-308.

[2]. Anderberg, M. R. (1973). *Cluster analysis for applications* (No. OAS-TR-73-9). Office of the Assistant for Study Support Kirtland AFB N MEX.

[3]. Berkhin, P. (2006). A survey of clustering data mining techniques. In *Grouping multidimensional data* (pp. 25-71). Springer, Berlin, Heidelberg.

[4]. Carlisle, A., & Dozier, G. (2001, April). An off-the-shelf PSO. In *Proceedings of the workshop on particle swarm optimization*(Vol. 1, pp. 1-6).

[5]. Cios K., Pedrycs W., Swiniarski R., (1998). Data Mining - *Methods for Knowledge Discovery*, *Kluwer Academic Publishers*.

[6]. Cui, X., Potok, T. E., & Palathingal, P. (2005, June). Document clustering using particle swarm optimization. In *Swarm Intelligence Symposium, 2005. SIS 2005. Proceedings 2005 IEEE* (pp. 185-191). IEEE.

[7]. Eberhart, R. C., & Shi, Y. (2000). Comparing inertia weights and constriction factors in particle swarm optimization. In *Evolutionary Computation, 2000. Proceedings of the 2000 Congress on* (Vol. 1, pp. 84-88). IEEE.

[8]. Everitt, B., Cluster Analysis.(1980). 2 Edition. *Halsted Press, New York*.

[9]. Jain, A. K., Murty, M. N., & Flynn, P. J. (1999). Data clustering: a review. *ACM computing surveys (CSUR)*, *31*(3), 264-323.

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

[11]. Kennedy J., Eberhart R. C. and Shi Y.,(2001). Swarm Intelligence, *Morgan Kaufmann, New York*.

[12]. Omran, M. G., Engelbrecht, A. P., & Salman, A. (2004). Image classification using particle swarm optimization. In *Recent Advances in Simulated Evolution and Learning* (pp. 347-365).

[13]. Eberhart, R. C., & Shi, Y. (1998, March). Comparison between genetic algorithms and particle swarm optimization. In *International conference on evolutionary programming* (pp. 611-616). Springer, Berlin, Heidelberg.

[14]. Kenney, J. (1995). Particle swarm optimization. In *Proc. IEEE International Conference on Neural Networks, 1995* (pp. 1942-1948).

[15]. H. Kim. (2006). Improvement of Genetic Algorithm Using PSO and Euclidean Data Distance, *International Journal of Information Technology*, 12(3), 142-148.

[16]. M. Settles, T. Soule.(2005). Breeding Swarms: A GA/PSO Hybrid, *Proc. of Genetic and Evolutionary Computation Conference* (GECCO'05), pp. 161-168.

[17]. J. Robinson, S. Sinton, Y. Rahmat-Samii.(2002). Particle Swarm Genetic Algorithm and their Hybrids: Optimization of a Profiled Corrugated Horn Antenna, *IEEE International Symposium on Antennas & Propagation*.

[18]. Wang, C., Liu, Y., Zhao, Y., Chen, Y.(2014). A Hybrid Topology Scale-free Gaussian-dynamic Particle Swarm Optimization Algorithm Applied to Real Power Loss Minimization. *Eng. Appl. Artif. Intel*. 32, 63–75.

[19]. Gong, Y.J., Zhang, J. (2013). Small-world Particle Swarm Optimization with Topology Adaptation. *In: 15th Annual Conference on Genetic and Evolutionary Computation Conference, pp. 25–32. ACM, New York.*

[20]. Zambrano-Bigiarini, M., Clerc, M., & Rojas, R. (2013, June). Standard particle swarm optimisation 2011 at cec-2013: A baseline for future pso improvements. In *Evolutionary Computation (CEC), 2013 IEEE Congress on* (pp. 2337-2344). IEEE.

[21]. Bonyadi, M. R., & Michalewicz, Z. (2014). A locally convergent rotationally invariant particle swarm optimization algorithm. *Swarm intelligence*, *8*(3), 159-198.

[22]. Du KL., Swamy M.N.S. (2016). Particle Swarm Optimization. In: Search and Optimization by *Metaheuristics. Birkhäuser, Cham.*

[23]. Li, F., & Guo, J. (2014, October). Topology optimization of particle swarm optimization. In *International Conference in Swarm Intelligence* (pp. 142-149). Springer, Cham.