I J C T A, 10(8), 2017, pp. 177-183 ISSN: 0974-5572

Data clustering using hybrid swarm intelligence method with multi objective functions

Anuradha Thakare*, Dipali Kharche and C.A. Dhote*****

ABSTRACT

Clustering groups the objects which have intrinsic similarity. The standardization of clustering algorithms is difficult though various clustering techniques are proven to be good. The research in clustering results in developing the methods which may give best outcome for specific datasets only. Therefore, there is a need to design and develop effective clustering method which can produce optimized results for all the datasets.

This work presents a new hybrid multi-objective clustering method based on Ant colony optimization and Particle swarm optimization method which partitions the data into an appropriate number of clusters. In order to get appropriate partitioning and to detect correct number of clusters, Ant Colony Optimization algorithm is used. The clustering objective functions are evaluated using Ant Colony Optimization clustering algorithm which results in finding random cluster centroids and associated data points. Three objective functions, one reflecting the total compactness of the partitioning based on the Euclidean distance, the other replicating the total symmetry of the clusters, and the last reflecting the cluster connectedness, are considered here. These are optimized simultaneously using hybrid ACPSO method and improvement in the accuracy of the result is observed.

Index Terms: Data Clustering; Ant Colony Optimization (ACO); Particle Swarm Optimization (PSO); Compactness; Connectedness; Symmetry

1. INTRODUCTION

Data mining is a technology, for extracting hidden and predictive information. There is a tremendous growth in Data mining field with the growth in real time data. The knowledge-driven results can be extracted using Data mining tools to predict future tendencies which can further be used for business intelligence applications. The process of information finding from databases requires automatic and fast clustering of large datasets with numerous attributes of different types. Though various tools are developed, there are certain challenges with clustering techniques. Swarm Intelligence (SI) is a technique from the family of nature inspired algorithms. Several researchers from the field of pattern recognition and clustering are working on SI for good clustering results. Clustering techniques based on the Swarm intelligence have reportedly outperformed many classical methods of partitioning a complex real world dataset.

2. RELATED WORK

Clustering is a collection of objects, which are similar in some pattern to each other and are dissimilar to the objects in other clusters. The clustering algorithms which gives better result for some data sets, may gives poor result for the other data sets. Therefore, determining number of accurate clusters is the main issue for clustering algorithms [1]. The cluster validity indices are introduced by the researchers for validating

* Pimpri Chinchwad College of Engineering Pune, *Email: anuradha.thakare@pccoepune.org*

** Pimpri Chinchwad College of Engineering Pune, *Email: Kharche.deepali@gmail.com*

*** PMITRA, Badnera, Amravati, *Email: vikasdhote@rediffmail.com*

the results of clustering. The objective of cluster validity is to impose an ordering of the clusters for their goodness. A graph theory with a clustering tendency index for cluster partitioning is proposed [2] in which the number of clusters and the partition that best fits the data set are choose according to the optimal cluster tendency index value. An analysis of design philosophy is perfectly used in essential validity indices for clustering. A variety of existing cluster validity indices are reviewed [3]. The techniques to overcome the limitations of the existing indices are proposed.

In clustering techniques, optimization algorithms are used to get the optimal results [1, 4, 5]. Genetic algorithm results in optimal partitions at the initial stage and later Swarm Intelligence algorithms [6, 7] are applied for further optimization. Researchers have developed hybrid clustering algorithms to improve the performance.

The paper is organized as follows: Section II presents the Related Work and section III presents proposed Model for optimal data clustering. Section IV presents the experimental results and final conclusion is given in section V.

3. PROPOSED HYBRID METHOD FOR DATA CLUSTERING

K-means clustering algorithm has tackled more challenges to achieve exact clustering with the issue of cluster initialization. The random initial centroids led algorithm to converge in local optimal solutions, and restrict the formation of accurate clusters. Hence, researchers have suggested many improvements in K-means algorithm with respect to selection of initial centroids. The feasibility behind the works have become compound because of wide data delivery with high dimension. In contrast, a thought on changing the cluster centroids at run-time has been presented. These improvements have reached good outcome, because they have intended to find the total centroids throughout the process, rather than focusing on initialization part. Also, finding to search the optimal cluster is tedious task if the element of the data is huge. Therefore, developing a method to handle both initialization problem and optimal finding procedure is a major challenge for cluster analysis.

A new ACPSO method is developed for optimal clustering process. The works mainly focus on hybridization of ACO and PSO Algorithm for Optimization in Data Clustering. The input data is given to ACO algorithm to obtain initial cluster centroids, these cluster centroids are optimized further using PSO algorithm for finding optimal clustering. Those two optimization algorithms make use of Multi Objective functions to evaluate the clustering results. The final optimal clustering output is evaluated using the external performance metrics like, F-measure and comparison is made with existing algorithms. The multiple objective functions are used to improve the quality of clusters. The objective functions, such as Cluster Symmetry, Cluster Connectedness, and Cluster Compactness [1] are used for cluster optimization.

The dataset is given as input to the ACO algorithm [8, 9, 10,11] to acquire the initial centroids for clustering method. The ants can made up with the scope for the number of ant positions which can be defined here as quantity of likely data group in the input data space. It differs between the minimum values to the maximum value of the given input datasets. So, every result is the position of ants placed with the range of data.

Initially, ants are placed in random way and cluster centroids are constructed. For this solution, cost is measured using the different cost calculation metrics as multiple objective functions given below. In the next iteration, the position of ants should be reformed to improve the solution whereas the pheromone update is critical [8, 9, 10, 11]. The pheromone updating is linked with good results.

$$T_{ij} \leftarrow (1-\rho).T_{ij} + \sum_{k=1}^{m} \Delta T_{ij}^{k}$$

Where, ρ is the evaporation amount, m is the number of ants, and ΔT_{ij}^{k} is the quantity of pheromone placed on edge (*i*, *j*) by ant *k*.

$$\Delta T_{ij}^{k} = \begin{cases} Q/L & \textit{if ant k used edge } (i, j) \textit{ in the solution,} \\ 0 & \textit{otherwise.} \end{cases}$$

Where Q is a constant, and L_k is the length of the outcome made by ant *k*.

In the building of outcome, ants select the following element to be continued through a mechanism. When ant *k* is in location *i* and has so far built the small results *p*, the chance of working to place *j* is given by [8, 9, 10, 11]:

$$p_{ij}^{k} = \begin{cases} \dfrac{T_{ij}^{\alpha} .\eta_{ij}^{\beta}}{\Sigma c_{il} \in N(s^{p}) T_{il}^{\alpha} . \ \eta_{ij}^{\beta}} & \textit{if } \ c_{ij} \in N(s^{p}) \\ 0 & \textit{otherwise,} \end{cases}$$

Where $N(s^p)$ is the set of likely mechanisms; that is, edges (*i*, *l*) where *l* is a location not yet stayed by the ant *k*. The inspirations α and β control the comparative meaning of the pheromone in contradiction of the empirical information η_{ij}.

Based on the calculation, every ants position are simplified for every repetition which will give a new result until it reaches n repetition. The final iteration gives the centroid set which is given for the PSO algorithm.

3.1. PSO Algorithm for Optimal Clustering

The initial group of centroids obtained from ACO algorithm is then passes to the particle solution in PSO algorithm [5, 7]. The ‘p’ solution which is good in terms of cost estimation is given as a first population of particles for PSO clustering. Here, each solution (particles) is a centroid set of the input data. So, the population scope of the PSO clustering is p X (k*d) matrix and the velocity cost of each particle is initialized to zero.

For each outcome, fitness is computed based on multi objective functions. Based on objective of minimization or maximization for correct clustering, p_{best} and g_{best} are found out. Here, the particle which is having the minimum fitness is set as p_{best} for the present iteration and the particle having minimum fitness in all the iterations performed and update it as g_{best}.

After finding p_{best} and g_{best}, the particles (centroids) velocities are originate using the following equation.

$$v_{t+1} = v_t + \phi_1 * rnd() * (p_{best} - x_t) + \phi_2 * rnd() * (g_{best} - x_t)$$

Where, ϕ1 and ϕ2 are set as two usually. v_t is the old velocity of the particle and *rnd*() is a random number between (0, 1). x_t is the current particle taken for finding new velocity.

Cost estimation: The ant cost can be estimated using a set of inner evaluation metrics given as: Let us consider D be dataset having N points characterized as, $D = \{d_1, d_2, ..., d_N\}$. Here, all data point d_i have d-dimensional feature assessment, $d_i = \{f_1, f_2, ..., f_d\}$. The objective is to catch the k-centroids, $C = \{c_1, c_2, ..., c_k\}$ by reducing different objective function. The objective functions, based on the Euclidean distance-one reflecting the total compactness of the clustering, the other reflecting the total symmetry of the clusters, and the third reflecting the cluster connectedness, are considered for optimization.

Objective Function 1: Compactness [1] - The total compactness of the partitioning based on the Euclidean distance.

$$I(K) = (\frac{1}{K} \times \frac{\varepsilon_1}{\varepsilon_k} \times D_K)^{\rho}$$

$$\varepsilon_k = \sum\nolimits_{k=1}^{k} \sum\nolimits_{j=1}^{nk} de(\overline{c_k, x_j})$$

$$D_k = \max\nolimits_{i,j=1}^{k} de(\overline{c_i}, \overline{c_j})$$

Where K is the number of clusters, The index I is a composition of three factors namely, 1/K, $\varepsilon_1/\varepsilon_k$, and DK.

Objective Function 2: Connectedness [1] - Connectedness present in a partitioning will be measured using the relative neighborhood graph concept.

$$dps\left(\overline{x}, \overline{c}\right) = \frac{\sum_{i=1}^{k_{near}} de}{k_{near}} * de\left(\overline{x}, \overline{c}\right)$$

$$d_{short}(X, Y) = \min\nolimits_{i=1}^{p} \max\nolimits_{j=1}^{nedge_j} w\left(ed_j^i\right)$$

Where, p is the number of paths between X and Y nodes.

Objective Function 3: Symmetry [1] - The symmetry present in a partitioning will be measured using a newly developed point symmetry based distance.

$$dps\left(\overline{x}, \overline{c}\right) = dsym\left(\overline{x}, \overline{c}\right) * de\left(\overline{x}, \overline{c}\right)$$

Where de(x, c) is the Euclidean distance between the point x and c, and dsym(x, c) is a symmetry measure of x with respect to c. where de(x, c) is the Euclidean distance between the point x and c, and dsym(x, c) is a symmetry measure of x with respect to c.

Thus the proposed system is able to detect the appropriate number of clusters and the appropriate partitioning from data sets having either well-separated clusters of any shape or symmetrical clusters with or without overlaps [14, 15, 16, 17].

3.2. Cluster Performance using external Metrics

After performing the multi objective function the performance evaluated in terms of F-Measure.

F-Measure: A combination of both precision and recall that measures the extent to which a cluster contains only objects of a particular class and all objects of that class. The F-measure of cluster i with respect to class j is [1]

$$F(i, j) = \frac{(2 \times precision(i, j) \times recall(i, j))}{(precision(i, j) + recall(i, j))}$$

The overall F-measure of the whole partitioning is calculated as:

$$F = \sum_{j} \frac{m_j}{m} \max F(i, j)$$

Where the maximum is taken over all clusters i at all levels, mj is the number of objects in class j, and m is the total number of objects. F-measure (FM) [1] is a measure of the quality of a solution given the true clustering. For F-measure, the optimum score is 1.

The new locations for all the particles are calculated based on the new velocity and earlier positions. The formulae used for calculating new position are given as follows:

$$x_{t+1} = x_t + v_{t+1}$$

The similar process is repeated until ending criteria's are fulfilled. The cluster centroid stored in final g_{best} is taken as the concluding centroid which can group the data points based on the minimum distance [12, 13].

4. EXPERIMENTAL RESULTS AND ANALYSIS

The proposed ACPSO clustering is implemented with a system of having i5 processor and main memory of 2 GB RAM using MATLAB (R2012b).

The experimentation is carried out on six real life data sets from a UCI machine learning repository [1]. These data sets are described in terms of the number of points present, dimensions, and the number of clusters in Table 1.

Table 1
Datasets

Sr. No	*Dataset*	*No. of Instances*	*No. of Attribute*	*No. of Classes*
1	Iris	150	4	3
2	Wine	178	13	3
3	Sonar	208	60	2
4	Pima-Indians-diabetes	768	8	2
5	Indian Liver Patient Dataset (ILPD)	583	10	2
6	Glass	214	9	6
7	Hepatitis	155	19	2

In order to detect the proper cluster centroids and the number of clusters, ACPSO algorithm with multiple objective functions is implemented. Three objective functions are used in order to get compact clusters, symmetric clusters and connected clusters.

The initial cluster centroids and partitions are obtained using ACO algorithm. These partitions are further optimized with PSO algorithm to achieve various objectives. Three objective functions are applied

Table 2
Clustering Results Intermsof F-measure Values

	Clustering Objective Functions			
Dataset	*Objective 1-Compactness*	*Objective 2-Connectedness*	*Objective 3-Symmetry*	*Multi objective (Objective1+ Objective2+ Objective3)*
Iris	0.92	0.96	0.97	0.97
Wine	0.9	0.91	0.91	0.92
Sonar	0.62	0.57	0.57	0.62
Pima Indians diabetes	0.94	0.88	0.52	0.95
ILPD	0.9	0.88	0.88	0.92
Glass	0.96	0.88	0.88	0.97
Hepatitis	0.9	0.95	0.69	0.96

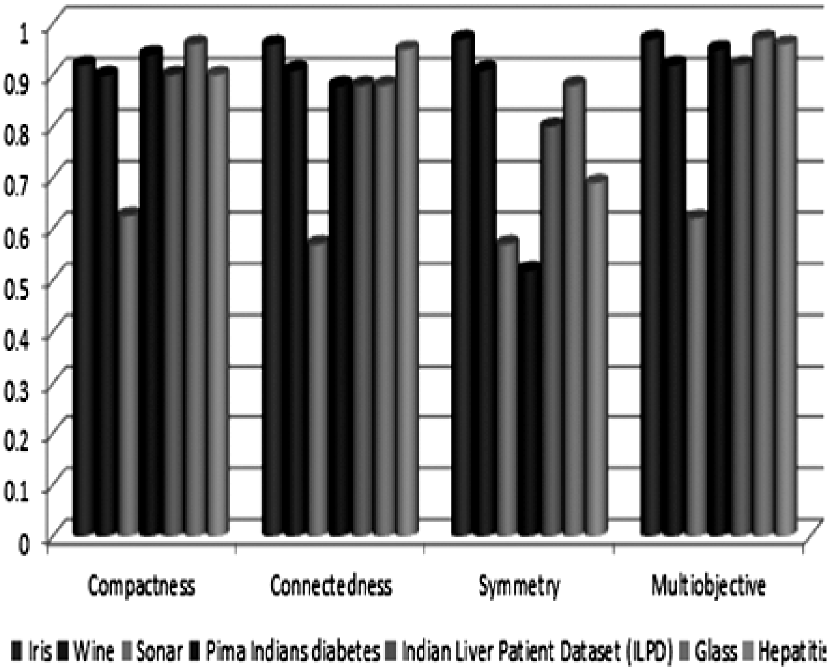

Figure 1: Analysis of single objective and multiple objectives of clustering In terms of F-Measure

individually first to compare the effect of clustering. The compactness of the clusters is calculated by using the Euclidean distance formula then by using the point symmetry based distance symmetry is calculated and lastly, the connected clusters are obtained by using the relative neighborhood graph concept [18,19,20]. The performance is checked by using a single objective function and multiple objective functions on the same dataset. The F-measure values are as shown in Table 2.and Fig.1 represents the Analysis of single objective and multiple objectives of cluster in terms of F-Measure [21].

5. CONCLUSION

The new hybrid Algorithm ACPSO for Multiple Function Optimization is introduced to obtain Symmetry of clusters, Compactness of cluster and Connectedness of clusters. Two optimization algorithms like ACO and PSO algorithm are effectively combined for data clustering. ACO algorithm is proven to be best for cluster initialization with shorter iterations. The hybridization of PSO with ACO improved the performance of PSO for clustering. The performance of proposed method is tested using F-measure and real life datasets. The proposed Hybrid clustering Algorithm performs better in terms of cluster quality for almost all the datasets.

REFERENCES

[1] Sriparna Sahaa, Sanghamitra Bandyopadhyayb, "A generalized automatic clustering algorithm in a multiobjective framework", Department of Computer Science and Engineering, Indian Institute of Technology Patna, India, Applied Soft Computing 13 (2013) 89–108

[2] H.C. Chou, M.C. Su, E. Lai, A new cluster validity measure and its application to image compression, Pattern Analysis and Applications 7 (July) (2004) 205–220.

[3] W. Wang, Y. Zhang, On fuzzy cluster validity indices, Fuzzy Sets and Systems 158 (October (19)) (2007) 2095–2117.

[4] Crina GroŸan and D. Dumitrescu,"A comparison of multiobjective evolutionary algorithms" acta universitatis apulensis

[5] K. Premalatha and A.M. Natarajan, "A New Approach for Data Clustering Based on PSO with Local Search", Computer and Information Science, Vol. 1, No. 4, 2008.

[6] Miao Wan, Lixiang Li,Jinghua Xiao, Cong Wang and Yixian Yang,"Data clustering using bacterial foraging optimization", Journal of Intelligent Information Systems, Vol.38, No. 2, pp. 321-341, Apr.2012.

[7] R.J. Kuo a,Y.J. Syu, Zhen-Yao Chen and F.C. Tien, “Integration of particle swarm optimization and genetic algorithm for dynamic clustering”, Journal of Information Sciences, Vol. 19, pp. 124-140, July 2012.

[8] P.S. Shelokar, V.K. Jayaraman, B.D. Kulkarni, “An ant colony approach for clustering”, Analytica Chimica Acta, vol. 509, pp. 187195, 2004.

[9] Chen, Wei-neng; Zhang, Jun, “A novel set-based particle swarm optimization method for discrete optimization problem”, IEEE Transactions on Evolutionary Computation, vol. 14, no. 2, pp. 278300, 2010.

[10] Enrique Amig´o, Julio Gonzalo, Javier Artiles, Felisa Verdejo, “A comparison of Extrinsic Clustering Evaluation Metrics based on Formal Constraints”, Information Retrieval, Vol. 12, no. 4, pp 461-486, 2009.

[11] H.C. Chou, M.C. Su, E. Lai, A new cluster validity measure and its application to image compression, Pattern Analysis and Applications 7 (July) (2004) 205–220.

[12] U. Maulik, S. Bandyopadhyay, Performance evaluation of some clustering algorithms and validity indices, IEEE Transactions on Pattern Analysis and Machine Intelligence 24 (12) (2002) 1650–1654

[13] P.B. Helena Brás Silva, J.P. da Costa, A partitional clustering algorithm validated by a clustering tendency index based on graph theory, Pattern Recognition 39 (May (5)) (2006) 776–788.

[14] Sriparna Saha, Sanghamitra Bandyopadhyay, “A symmetry based multiobjective clustering technique for automatic evolution of clusters”, Pattern Recognition. vol. 43, pp. 7381-751, 2010.

[15] Sriparna Saha, Sanghamitra Bandyopadhyay, “Some connectivity based cluster validity indices”, Applied Soft Computing, vol.12, pp. 15551565,2012.

[16] Sriparna Saha, Sanghamitra Bandyopadhyay, *“A symmetry based multiobjective clustering technique for automatic evolution of clusters”*, Pattern Recognition. vol. 43, pp. 7381-751, 2010.

[17] S. Bandyopadhyay, U. Maulik, Nonparametric genetic clustering: comparison of validity indices, IEEE Transactions On Systems, Man and Cybernetics, Part C 31 (1) (2001) 120–125.

[18] G.T. Toussaint, The relative neighborhood graph of a finite planar set, PatternRecognition89912(1980)261–268

[19] CA Dhote, AD Thakare, SM Chaudhari, Data clustering using particle swarm optimization and bee algorithm, Computing, Communications and Networking Technologies (ICCCNT), 2013 Fourth International Conference on, IEEE (1-5)

[20] D Kharche, A Thakare, ACPSO: Hybridization of ant colony and particle swarm algorithm for optimization in data clustering using multiple objective functions, Communication Technologies (GCCT), 2015 Global Conference on, (854-859)

[21] ADThakare, CA Dhote,”Novel Multi-stage Genetic Clustering for Multiobjective Optimization in Data Clustering”, Computing Communication Control and Automation (ICCUBEA), 2015 International Conference on , IEEE(402-407).