International Journal of Combined Research & Development (IJCRD)
eISSN:2321-225X;pISSN:2321-2241 Volume: 7; Issue: 3; March -2018

Designing an Efficient Clustering Using Particle Swarm Optimization

Seema Patil[1], Dr. R J Anandhi [2]
[1]Research Scholar,
[2]Prof & Head,
[1,2] Dept of CSE ,The Oxford College of Engineering ,Bangalore ,India

***Abstract*—The efficiency and accuracy of clustering can be enhanced in data mining and knowledge discovery with the help of optimization. Clustering separates objects which are similar in one group and other objects into another group. In large datasets, the cluster analysis identifies groups of same data items. Data clustering have been emerging in the application of optimization based methods. To check the optimized solution of clustering many different optimization techniques can be applied. One of the such optimization technique is seen in swarm intelligence. The algorithms used in swarm optimization have shown successfully the best solutions for different data clustering techniques. This paper has proposed particle swarm optimization technique for data clustering with its literature review. The performance of particle swarm optimization is demonstrated with authenticated data set and artificial data set. Also, performance parameters of clustering is represented.**

Keywords—cluster,distance,particle swarm optimization

I. INTRODUCTION

Extraction of information like similar data into clusters, classification of data, associated data with rules, patterns which are sequential, models which are predictive from distinct types of data as text data.

Clustering of data is one of important method in data mining which categorize unlabeled data into diverse groups which is based on similar and dissimilar data. A process of clustering contains cluster of selected feature of data, similarity measurement of data and assessment of outcome of data

The clustering which divides the data into separate groups are highly dependent on similarity and dissimilar data which is classified as partition based clustering. Hierarchy based clustering. In each of clustering approach, calculation of similarity measurement differs. To improve the quality of cluster distance of data can be calculated based on its position of data and centroid position. Each time centroid position is updated as iteration progresses. The assignment of data to cluster centroid is continued until centroid remains unchanged.

Updating of centroid position improves the quality of cluster based on intercluster and intracluster calculation. Quality of cluster is relative to an objective function. Minimization of intracluster and maximization of intercluster improve the quality of clusters. In K-means clustering a data element belongs to only one cluster at a time. Under partition based clustering kmeans is considered as the best clustering approach. The efficiency of partitional data clustering is considered best as per selection of domain. At distinct level, similar and dissimilar data can be represented into nested tree structure which defines hierarchical data clustering. The splitting of one large cluster into sub cluster is classified as agglomerative hierarchical approach [1].

Many approaches [2] have proposed to show the performance of clustering process. To perform clustering independently or add clustering benefits to existing techniques, some optimization technique can be applied. One of the best optimization technique is swarm intelligence.

The intelligence of swarm has been inspired by behavior of birds which are biological. The intelligence in swarm has considered as innovative optimization technique. The various example of intelligence in swarm can be swarms of bees, fish schools, insect colonies. These swarm communicate with each other in the search of food. Also, these swarm socialize with each other in the search of their destination.

The intelligence in swarm behavior is based on self-organization and communication and cooperation among each other and with individual within the team. The individual swarm interaction is simple but results into difficult global behavior.

Many variants have been developed based on the intelligence of swarm. The example can be insect colony optimization which is based on the behavior of ants. Another example in intelligence of swarm is based on particle swarm which is considered as population based. The particle based swarm algorithm belong to the class of meta heuristics These particle swarm was inspired by the social behavior of bird,

where each particle in the swarm acts like unintelligent agent. These agents interact in the environment by communicating with each other to give the best optimized solution. Also, these agents cooperate with each other to give the best optimized solution.

The optimization based on particle swarm has main three parts. The first part is called as velocity and the second part is called as cognitive component and the third part is called as social component.

Swarm based particle optimization is one of the most popular and powerful algorithm to solve many optimization problems. One of the key points behind the swarm based particle optimization is its simplicity.

The paper contains following sections. The first section describes introduction part, the second section shows the review on PSO. The third section gives the optimized methods used in designing the clustering and the fourth section gives results.

II.Literature Review

This section describes the literature review on Particle swarm optimization. The author Shahira et al [4] has proposed version of the PSO for clustering which determines local clustering swarm based particle, it uses different neighborhood to search for the centroids for clusters which results into optimal solution. The proposed representation assures the diversity of the swarm by making no repetition of redundant particles.

Particle swarm optimization was used by Van der Merwe and Engelbrecht[5] for data clustering which represent kmeans technique as the initial technique to give the initial swarm. The author proposes two algorithms named gbest PSO and a hybrid approach which were resulted in good cluster distance as required for data clustering.

The author Paulus et al [6] have proposed a work of alumni data with PSO. University management uses alumni data for taking decisions for developing communications. The proposed work has converted Abandoned and Reborn(AR) techniques into clustering problem with aim of achieving desired clusters. The generated clusters were evaluated with three parameters: closeness, separation and purity.

The author [7] have proposed the objective function which results into efficient cluster distance among clusters with fuzzy C means technique. The results were compared with existing kmeans approach.

A recent similar work [8] reports the results of k-means added with PSO and multiclass merging to perform data clustering. The author Shen[9] et al have proposed new learning strategy of PSO with the local search ability . The local search ability of a particle is very poor when the global best position is considered. With the improvement of local search ability, the exploitation of global best component is enhanced.

The author [10] have proposed clustering algorithm combining PSO with kmeans. The kmeans technique converges fast to global optima and also local optimal cluster also suffer in kmeans beacause of centroid initialization problem,so the author has combined both the approaches by utilizing global search capability of PSO and local search capability of kmeans.

Another similar work [11] reported where PSO is used for data clustering independently without being hybridized with any other clustering technique. Jay Prakash et al [12] has modified Binary Particle swarm optimization to improve information among particles to avoid local optima by using genetic crossover among particles to generate relevant set of features. Rene et al[13] has proposed PSO algorithm with modified version of hard k medoids clustering algorithm for relational data. The author et al [14] has proposed PSO with clustering multivariate time series data with hybridization between PCA and Mahalanobis distance.

III. Methodology

Swarm based particle optimization contains a population of candidate solution called as swarm. Every particle in swarm results into optimized problem. Every particle in the swarm has a position and velocity in the search space and the best solution among these solutions is considered.

Every particle is denoted by ith index and X and V are the two vectors of position and velocity and the search space is set of numbers. The position of particle is denoted by Xi(t) and velocity is denoted as Vi(t) and for each particle new velocity and new position should be calculated.

The velocity describes the movement of particle i in the sense of direction and time step which is located in their position. In addition to position and velocity every particle has a memory of its own best position which is best experience described as personal best and denoted by Pi(t) . In addition to this personal best, a comman best experience among the members of swarm is described by global best and denoted by g(t). The global best is not denoted with ith index because it belong to whole swarm that is whole population and not to a specific particle experience and is described as global best and global comman experience among the members of swarm. So there is personal list for every particle and global best which is best experience of all particles in the swarm.

The new position and velocity is calculated in following way. Let us define a vector from current position to personal best that is vector from Xi(t) to Pi(t) and another vector which connects current position to the global best that is vector from g(t) to Xi(t). A particle moves toward new position but uses all components of a particle. A particle can move somewhat parallel to vector of velocity or can move somewhat parallel to

personal best or can move parallel to global best. The new position is denoted by Xi(t+1) and new velocity is addition of these three vectors that is addition of previous velocity, global best and personal best. This is the simple model of PSO. PSO uses previous decision about the movement. The new position is created according to the previous velocity to personal best and to the global best, so this can be probably a better location because swarm uses previous decision about the movement of this particle and it uses previous experience of particle itself and it uses previous experience of whole swarm. So, this is probably better location based on this simple model and considering the rules of particles that each particle will cooperate to find the best location in the search space which can be best solution possible for optimization problem.

The swarm based particle optimization follows a mathematical model can be described as follows

1.One of the equation which is simpler for updating the position of particleXi(t+1) is

Xi(t+1) =Xi(t)+Vi(t+1)

2. The second equation completely describes the model of PSO and it updates the velocityVi(t+1) of particle

Vi(t+1) =w*Vi(t)+c1* (Pi(t)-Xi(t))+c2*(g(t)-Xi(t))

Where c1 and c2 are acceleration coefficient also a real valued coefficient, w is inertia weight also a real valued coefficient. To improve the optimization PSO, a random number r1 and r2 which are uniformly distributed between [0,1] can be applied to cognitive component and to social component.

A. Evaluation methods of clustering

Quality of clustering is measured according to the two criteria: F-measure and purity. F-measure and purity values are used to evaluate the accuracy of the clustering algorithms. The F-measure is a harmonic combination of the precision and recall values used in information retrieval. In general, the larger the F-measure is, the better the clustering results. Purity of a cluster represents the fraction of cluster corresponding to the largest class of data assigned to that cluster. Also the purity value should be larger for the clustering approach.Quality of clustering is measured according to the two criteria F-measure and purity. To evaluate the accuracy of clustering F-measure and purity can be calculated.

(IV) Experiment data set

There are two data sets taken among them one is synthetic data sets where as other is real data sets which is Iris data set[15] taken from UCI repository with 4 inputs and 3 clusters. There are 150 data vectors. The second data set is Synthetic data set which is two-dimensional data set having three clusters and in each cluster there are 10 data vectors. Each cluster contains value through following method:

$$\text{Class}=\begin{Bmatrix} 1 & then\ 1{:}10 \in [5 \quad 10], \\ 2 & then\ 11{:}20 \in [15\ 20], \\ 3 & then\ 21{:}30 \in [10\ 20], \end{Bmatrix}$$

Minimization of intra cluster distance by PSO has shown in figure 1. The purpose is to develop the compact clusters so that similar data have to appear in same cluster. At the beginning, there were random allocation of centroids, hence more distance appeared in a cluster but as iteration progress, there are improvement observed and it became stable after 50th iteration which indicate there is no progress further in betterment of cluster.

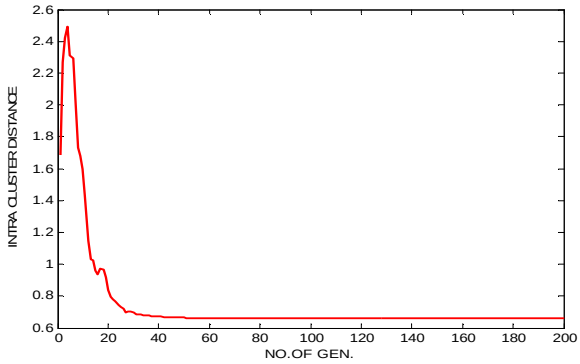

Figure 1: Plot of Number of Generations Vs Intracluster Distance for Iris data Population formation

Randomly data points selected as the starting value of centroids for all the three classes. There are 3 classes of IRIS which are Iris Setosa, Iris Versicolour, Iris Virginica and 4 parameters which are sepal length, sepal width, petal length, petal width, hence there is total 3*4 =12 dimension available which have to be searched by PSO. In each class there are total 50 data points available with number as [1-50],[51-100],[101-150]. With PSO cluster were formed efficiently by valid data points. Few data points were represented as error.

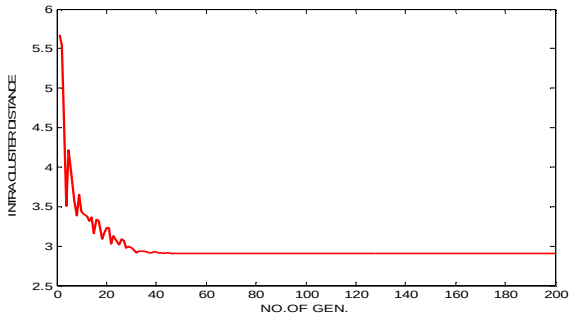

Figure 2: Plot of Number of Generations Vs Intracluster Distance for Synthetic data

The intra cluster distance within cluster must be less. As the number of iterations increases the intra cluster is also minimized in PSO with synthetic data which is shown in figure 2. Clusters were formed with valid data points but few resulted in error also.

The performance parameters of PSO on IRIS and Synthetic data is shown in Table 3. The table compares F-Measure and Purity for the PSO. It has observed that measurement parameters of PSO with IRIS data set that is F- Measure and Purity is high compare to Synthetic data.

Table 3: Measurement parameters for PSO on IRIS and Synthetic data

Measurement Parameters	IRIS	Synthetic data
F-MEASURE	0.8562	0.7766
PURITY MEASURE	0.8533	0.6667

Also, the performance parameters of PSO on IRIS and kmeans on IRIS data is shown in Table 4. The table compares F-Measure and Purity for the PSO and K-means. It has observed that measurement parameters of PSO with IRIS data set that is F- Measure and Purity is high compare to K means.

Table 4: Measurement parameters for PSO and kmeans on IRIS dataset

Measurement Parameters	PSO	K-means
F-MEASURE	0.8562	0.84
PURITY MEASURE	0.8533	0.84

Conclusion

In this paper, the evolution of clustering techniques based on Particle Swarm optimization is discussed. Selection of different performance measurement such as intra-cluster distance which describes compactness of clusters and accuracy which discuss correctness in clustering has been represented. Intra-cluster distances are indicators of how good the clusters are in terms of the position of each data element within its corresponding cluster as well as against the elements of different clusters. Low intra-cluster distance is better than high intra-cluster. The performance measurement of PSO and k-means are measured successfully. Hence PSO can be used as efficient data clustering approach.

References

[1]U.M.Fayyad,G.Piatetsky-Shapiro,P.Smyth, R. Uthurusamy, “Advances in Knowledge Discovery and Data Mining”, The MIT Press, CA, USA, 1996.

[2]A.Abraham, H. Guo, H. Liu “Swarm intelligence: foundations, perspectives and applications”, Swarm Intelligent System, Springer, 2006, pp. 3–25.

[3] J. Kennedy, R. Eberhart, “Particle swarm optimization” Proceedings of the IEEE International Conference on Neural Networks, vol. 4, IEEE, 1995, pp. 1942–1948.

[4] Shahira, Mohamed Farouk Abdel, Hesham Ahmed Hefty, “Local Best Particle Swarm Optimization for Partitioning Data Clustering”,Evolutionarycomputation,2016, vol 1, IEEE,2016, pp 41-47

[5] D. Van der Merwe, A. Engelbrecht, “ Data clustering using particle swarm optimization” 2003 Congress on Evolutionary Computation, 2003. CEC'03., vol. 1, IEEE, 2003, pp. 215–220.

[6] Paulus Mudjihartono, Thitipong Tanprasert, Rachsuda Jiamthapthaksin, “Clustering Analysis on Alumni Data Using Abandoned and Reborn Particle Swarm Optimization”,8th International conference on knowledge and Smart technology 2016,IEEE,pp 22-26.

[7] C.-Y. Chen, F. Ye, “Particle swarm optimization algorithm and its application to clustering analysis”,2004 IEEE International Conference on Networking, Sensing and Control, vol. 2, IEEE, 2004, pp. 789–794.

[8] Y. Lin, N. Tong, M. Shi, K. Fan, D. Yuan, L. Qu, Q. Fu, “K-means optimization clustering algorithm based on particle swarm optimization and multiclass merging”,Advances in Computer Science and Information Engineering, Springer, pp. 569–578,2012

[9] Yuanxia Shen,Guuyin Wang, “Study on the local search Ability of Particle Swarm Optimization”, Springer, ICSI 2010,Part I,LNCS 6145 pp11-18,2010

[10] Chunqin,Qian, “Clustering Algorithm Combining CPSO and Kmeans”, International Conference on Advances in Mechanical Engineering and industrial informatics, Atlantis Press,2015,pp 749-755.

[11] T. Cura, “A particle swarm optimization approach to clustering”, Expert Syst. Appl. 39 (1) (2012) 1582–1588.

[12] Jay Prakash, Pramod Kumar Singh, “Particle Swarm Optimization with K-means for Simultaneous Feature Selection and Data Clustering”, 2015 Second International Conference on Soft Computing and Machine Intelligence,2015, pp 74-77.

[13] Rene Pereira, Federal de Sergipe, “Particle Swarm Optimization applied to Relational Data Clustering”, 2016 IEEE International Conference on Systems, Man, and Cybernetics,2016,pp 001690-001695

[14] Abbas Ahmadi, Atefeh Mozafarinia, Azadeh Moheb, “Clustering of Multivariate Time Series Data Using Particle Swarm Optimization”,2015 international Symposium on Artificial intelligence and Signal Processing,2015, pp 176-181.