Improving the Performance of PSO based Clustering by using Surrogates

Priyanka Shrivastava*, Prof. Mangesh Khandelwal** and Prof. Vinod Kumar***
*M.Tech Scholar, Dept. of CSE., Bansal Institute of Research and Technology,Bhopal,Madhya Pradesh,India
**Assistant Professor, Dept. of CSE, Bansal Institute of Research and Technology, Bhopal,Madhya Pradesh,India
***HOD, Dept. of CSE, Bansal Institute of Research and Technology, Bhopal,Madhya Pradesh,India

Abstract: Clustering is the grouping of a specific set of objects based on their characteristics, aggregating them according to their similarities. Regarding to data mining, this methodology partitions the data implementing a specific combine algorithm, most appropriate for the desired information analysis. K-means is among the simplest unsupervised learning algorithms that resolve the well-known clustering problem. The procedure follows a simple and effortless way to classify a given data set through a certain number of clusters. However the method can be computationally costly in that a high number of function calls have to progress the swarm at each optimization iteration. In order to increase the efficiency of this algorithm an notion of Surrogate purpose is incorporated which functions as an stand in for pricey objective function. The work also aims to provide better evaluation of the proposed hybrid approach on the basis of acquired numerical results.

Keywords: Clustering, Data Mining, PSO Algorithm, K-Means Algorithm, Surrogate functions.

Introduction

Data mining can be extensively characterized as the disclosure and investigation of helpful data from a large data source specially web based. This portrays the programmed hunt of data assets accessible online and offline, There are around three information disclosure spaces that relate to data mining over distributed network which are Web Content Mining, Web Structure Mining, and Web Usage Mining.

- **Web content mining**: The methodology of separating learning from the substance of records or their portrayals. Web report content mining, asset revelation taking into account ideas indexing or specialists based innovation might likewise fall in this classification.
- **Web structure mining**: The procedure of inducing learning from the World Wide Web association and connections in the middle of references and referents in the Web.
- **Web usage mining**: Also known as Web Log Mining is the procedure of separating fascinating examples in web access.

Clustering

Clustering is process of grouping objects. Objects are clustered according to their own characteristics, aggregating them according to their similarities. Regarding to data mining, this methodology partitions the information implementing a specific combine algorithm, most appropriate for the desired data analysis. This clustering analysis allows an object not to be part of a bunch, or just belong to this, calling this sort of grouping tough partitioning. In the other hand, soft partitioning states that each and every object is a member of a cluster in a predetermined degree. More specific divisions can be potential to create like objects belonging to multiple clusters, to induce an item to take part in just one cluster or even assemble real trees on group associations. There are numerous different ways to implement this partitioning, based on different models. Distinct algorithms are applied to each Model, differentiating its properties and results.

Clustering Algorithms in Data Mining

Based on the newly described cluster Versions, there are Lot of clustering techniques that could be applied to a data set in order to partition the information which are as given below:

- **Centroid-based:** In this type of grouping method, every cluster is referenced by a vector of values. Each item is part of the audience whose value gap is minimum, comparing to other clusters. The Amount of clusters should be pre-defined, and this really is the biggest problem of this sort of algorithms. This methodology is the most near the classification subject and is enormously used for optimization problems.
- **Distributed-based:** Related to pre-defined statistical models, the dispersed methodology combines objects whose value belongs to the same distribution. Due to its random nature of value generation, this Process requires a well-

defined and complex model to interact in a better method with real data. However these processes can achieve an optimal answer And compute correlations and dependencies.

- **Connectivity-based:** On this type of algorithm, every Thing is linked to its neighbours, depending on the level of that Connection on the distance between them. Based on this assumption, Clusters are created with nearby objects, and can be described as a Maximum distance limitation. With this connection between members, these Clusters have hierarchical representations. The distance function varies On the focus of the analysis.
- **Density-based:** As stated by the high density of participants of a data collection, at a determined location. It aggregates some space notion to a density standard level To group members in clusters. Such processes may have less Performance on discovering the limit regions of the group.

Related work

Particle Separate modification to inertia weight and learning factors in PSO undermines the integrity and intelligent characteristic from the evolutionary process of chemical swarm to some extent, thus it's not suitable for solving many complex optimization problems [1]. Volume refers to the massive amount of data it gathers, Velocity denotes the speed where it process the data and Variety defines that multi-dimensional data which could be figures, dates, strings, geospatial information, 3D data, sound files, video files, social files, etc.. These data that's stored in large data are going to be from various sources at different rate and of different type; hence it won't be synchronized. This is one of the biggest challenges in working with big data. Second challenge is related to mining the precious and relevant data from such information sticking to 3rd V i.e. Velocity. Speed is highly significant as it's connected with expense of processing [2]. During the procedure for classification a great deal of irrelevant attributes are covered by enter information [3]. Existence of these unrelated features will bring in a dimension calamity. In many classification problems, its hard to learn fantastic classifiers prior to eliminate these undesirable features as a result of huge amount of data. Reducing the total amount of redundant or unrelated features can decrease the working time of classification algorithms. A better PSO based technique is used to extract important features using dependent criteria for dimension reduction. Clustering algorithms have emerged and quickly developed instead powerful meta-learning tool to undertake a wide selection of applications because it is very helpful for segmenting large multidimensional data to distinguishable representative clusters. Fuzzy clustering is a popular unsupervised learning method used in audience analysis that allows a point in huge data sets belongs to 2 or more clusters. Prior work indicates that Particle Swarm Optimization based strategy could be a highly effective instrument for solving clustering issues.

Problem Identification

Some of the major problems identified in previous works are:

- In PSO based clustering K-Means combined with the heuristic approach which can be extremely slow.
- Particle swarm optimization (PSO) is a population-based, heuristic minimization technique that is based on social behaviour. The method has been shown to perform well on a variety of problems including those with nonconvex, nonsmooth objective functions with multiple local minima.
- The method can be computationally expensive in that a large number of function calls is required to advance the swarm at each optimization iteration. This is a significant drawback when function evaluations depend on output from an off-the-shelf simulation program, which is often the case in engineering applications.

METHODOLOGY

Proposed Algorithm

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.

Pre-processing

1. For each particle
   Initialize particle
END
2. Do
   For each particle
      Calculate fitness value
     Store the filtness value in a GlobalSurrogateDatabase
3. Calculate particle velocity according equation (a)

Assign the velocity to GlobalSurrogateVelocityDb
Update particle position according equation (b)
Assign the velocity to GlobalSurrogatePositionDb
End

Improved Solution

1.For each particle
Initialize particle
END
2.Do
For each particle
If the fitness value is better than the best fitness value (pBest) in GlobalSurrogateDb
Set current value as the new pBest
End
3.Choose the particle with the best fitness value of all the particles as the gBest
For each particle
4.Compare the position and velocity with GlobalSurrogatePositionDb and GlobalSurrogateVelocityDb
End
5. Assign the filtered particles to K-Means algorithm to form clusters.

K-Means Algorithm

The K-Means algorithm is a simple yet effective statistical clustering technique. The main idea is to define k centroids, one for each cluster. These centroids should be placed in a cunning way because of different location causes different result. So, the better choice is to place them as much as possible far away from each other. The next step is to take each point belonging to a given data set and associate it to the nearest centroid. When no point is pending, the first step is completed and an early group age is done. At this point we need to re-calculate k new centroids as barycenters of the clusters resulting from the previous step. After we have these k new centroids, a new binding has to be done between the same data set points and the nearest new centroid. A loop has been generated. As a result of this loop we may notice that the k centroids change their location step by step until no more changes are done.

Basic algorithm steps are:

1. Choose a value for K, for determining no of clusters.
2. Choose K data points) from dataset at random. These are the initial cluster centres.
3. Use simple Euclidean distance to assign the remaining instances to their closest cluster centre.
4. Use the instances in each cluster to calculate a new mean for each cluster.
5. If the new mean values are identical to the mean values of the previous iteration the process terminates. Otherwise, use the new means as cluster centres and repeat steps 3-5.

PSO (Particle Swarm Optimization) Algorithm

PSO is initialized Using a Group of random particles (solutions) and then searches for optima by upgrading generations. In every iteration, each particle is updated by subsequent two "best" values. The first one is the ideal solution (fitness) it's achieved up to now. (The fitness value can be stored.) This value is called pbest. Another "best" value that is monitored by the particle swarm optimizer is the very best value, acquired so far by any particle in the populace. This best value is a global best and known as gbest. When a particle chooses part of the population as its topological acquaintances, the best value is a neighbourhood best and is called lbest.After finding the two best values, the particle updates its velocity and positions with subsequent equation (a) and (b).

v[] = v[] + c1 * rand() * (pbest[] - present[]) + c2 * rand() * (gbest[] - present[]) ……(1)

present[] = present[] + v[]…… (b)

v[] is the particle velocity, present[] is the current particle (solution). pbest[] and gbest[] are defined as stated before. rand () is a random number between (0,1). c1, c2 are learning factors. usually c1 = c2 = 2. The pseudo code of the procedure is as follows:

1.For each particle
Initialize particle

END
2.Do
For each particle
Calculate fitness value
If the fitness value is better than the best fitness value (pBest) in history
set current value as the new pBest
End
3.Choose the particle with the best fitness value of all the particles as the gBest
For each particle
4.Calculate particle velocity according equation (a)
Update particle position according equation (b)
End
While maximum iterations or minimal error criteria is not attained .Clamped to a maximum speed Vmax. If the amount of accelerations would cause the Velocity on such dimension to transcend Vmax, which is a parameter given by the user. Then the velocity on that dimension is limited to Vmax.

Experimental Results

The experiment was conducted on Visual Studio 2010 platform with C# as language. The operating system used was Windows 10.The sample dataset was created and used for the purpose of demonstration. The results of the experiments cleanly shows that the improved algorithm has better space and time complexities as compared to simple PSO clustering algorithm.

Table 1: Space Analysis

	PSO Clustering	Improved PSO Clustering
Space (In Bytes)	1706448	1364748

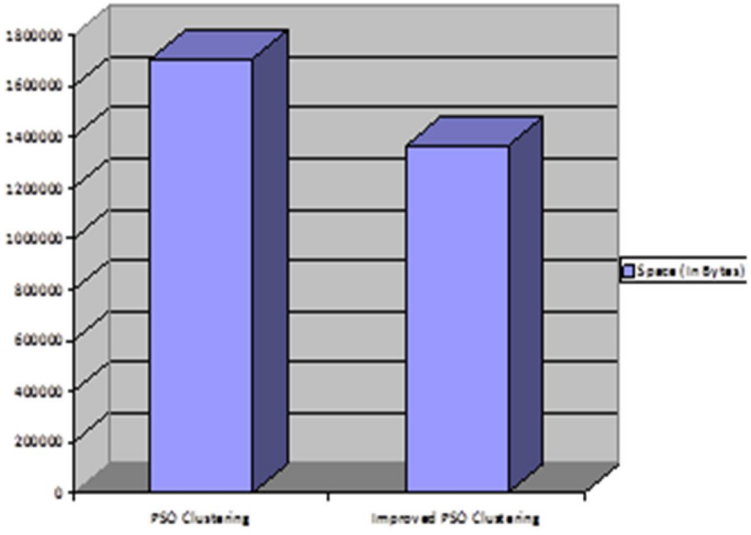

Fig 1: Space Analysis

Table 2: Time Analysis

	PSO Clustering	Improved PSO Clustering
Time (In Millis)	1910	1425

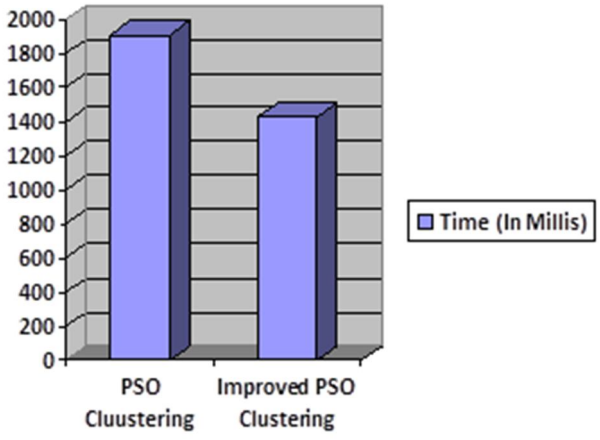

Fig 2: Time Analysis

Conclusion and Future Work

Standard Particle Swarm Optimization has advantages and disadvantages, to overcome the dearth of PSO. There are several standard version of PSO. The basic variants as stated previously have supported controlling the velocity and the stable convergence. In the other hands, modified variant PSO assist the PSO to process other conditions that Can't be solved with the basic PSO. The proposed work aims to increase efficiency of data mining algorithms using clustering techniques. Future work aims to increase the efficiency by using better parameters of other algorithms also. It is not possible to develop a system that makes all the requirements of the user. User requirements keep changing as the system is being used. Some of the future enhancements that can be done to this system are:

- As the technology emerges, it is possible to upgrade the system and can be adaptable to desired environment.
- Because it is based on object-oriented design, any further changes can be easily adaptable.
- The efficiency of algorithm can be further increased by applying more efficient data mining algorithms in near future. More work is possible on security of data in cloud servers.
- Security can be increased by applying efficient encryption/decryption algorithms.

References

[1] A Hybrid Clustering Technique Combining A PSO Algorithm with K-Means pp 487–499,Kripa Shankar Bopche & Anurag Jain *International Journal of Computer Applications (0975 – 8887) Volume 137 – No.1, March 2016*

[2] Accelerated PSO Swarm Search Feature Selection with SVM for Data Stream Mining Big Data ,Himani Patel , International Journal of Innovative Research in Computer and Communication Engineering , Vol. 4, Issue 9, September 2016 .

[3] Towards Better Classification Using Improved Particle Swarm Optimization Algorithm and Decision Tree for Dengue Datasets,B.R Devi ,K.N.Rao & S.P.Shetty,International Journal of Soft Computing 11(1) : 18-25 , 2016.

[4] A Novel Data Clustering Algorithm based on Modified Adaptive Particle Swarm Optimization, Ganglong Duan, Wenxiu Hu, Zhiguang Zhang, International Journal of Signal Processing, Image Processing and Pattern Recognition,2016.

[5] Improved Particle Swarm Optimization Method Directed By Indirect Surrogate Modeling , Y. Volkan Pehlivanoğlu, Serdar Ay & Faruk Gül , Journal Of Aeronautics And Space Technologies January 2015

[6] A two-layer surrogate-assisted particle swarm optimization algorithm, Chaoli Sun,Yaochu Jin,Jianchao Zeng & Yang Yu , April 2014.

[7] Chunming Yang and Dan Simon, “A New Particle Swarm Optimization Technique”, 2005.

[8] Hui Wang, Yong Liu, Sanyou Zeng , Hui Li and Changhe Li, “Opposition based Particle swarm Algorithm with Cauchy Mutation”, IEEE 2007.

[9] Marco A. Montes de Oca and Thomas Stutzle, “Convergence Behavior of the Fully Informed Particle Swarm Optimization Algorithm”, 2008.

[10] Marco A. Montes de Oca, Thomas Stützle, Mauro Birattari and Marco Dorigo, “Frankenstein’s PSO: A Composite Particle Swarm Optimization Algorithm”, IEEE TRANSACTIONS ON EVOLUTIONARY COMPUTATION, VOL. 13, NO. 5, OCTOBER 2009.

[11] George I. Evers and Mounir Ben Ghalia, “Regrouping Particle Swarm Optimization: A New Global Optimization Algorithm with Improved Performance Consistency Across Benchmarks”, 2009.

[12] George I. Evers, a thesis on “AN AUTOMATIC REGROUPING MECHANISM TO DEAL WITH STAGNATION IN PARTICLE SWARM OPTIMIZATION”, 2009.

[13] Praveen Kumar Tripathi, Sanghamitra Bandyopadhyay, Sankar Kumar Pal, “Multi-Objective Particle Swarm Optimization with time variant inertia and acceleration coefficients”, 2007.

[14] M. A. Montes de Oca, J. Pena, T. Stutzle, C. Pinciroli, and M. Dorigo, “Heterogeneous Particle Swarm Optimizers”, Jan 2009.

[15] Magnus Erik, Hvass Pedersen, Andrew John Chipperfield, “Simplifying Particle Swarm Optimization”, 2009.
[16] Gary G. Yen and Wen Fung Leong, “Dynamic Multiple Swarms in Multiobjective Particle Swarm Optimization”, IEEE TRANSACTIONS ON SYSTEMS, MAN, AND CYBERNETICS —PART A: SYSTEMS AND HUMANS, VOL. 39, NO. 4, JULY 2009.
[17] Qinghai Bai, “Analysis of Particle Swarm Optimization Algorithm”, CCSE, Vol 3, No. 1, February 2010.
[18] Magnus Erik, Hvass Pedersen, “Good Parameters for Particle Swarm Optimization”, Technical Report no. HL1001, 2010.
[19] Prithwish Chakraborty, Swagatam Das, Ajith Abraham, Václav Snasel and Gourab Ghosh Roy, “On Convergence of Multi-objective Particle Swarm Optimizers”, IEEE, 2010.
[20] Shailendra S. Aote et al. / International Journal of Computer Science Engineering (IJCSE)