ISSN(Online) : 2319-8753
ISSN (Print) : 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(An ISO 3297: 2007 Certified Organization)

Website: www.ijirset.com

Vol. 6, Issue 7, July 2017

Analysing the Performance of PSO Clustering by Using Surrogates

Priyanka Shrivastava[1], Mangesh Khandelwal[2], Vinod Kumar[3]

M.Tech Scholar, Dept. of CSE., Bansal Institute of Research and Technology, Bhopal, Madhya Pradesh, India[1]

Assistant Professor, Dept. of CSE, Bansal Institute of Research and Technology, Bhopal, Madhya Pradesh, India [2]

HOD, Dept. of CSE, Bansal Institute of Research and Technology, Bhopal, Madhya Pradesh, India [3]

ABSTRACT: Data mining is the combination of data assembled by customary information mining philosophies and procedures with data accumulated over large no of data sources. Data mining is utilized to retrieve some useful information from a set of accumulated data-K-Means is one of the common and simplest unsupervised algorithm for clustering- It classifies data on the basis of Euclidian distance. The PSO algorithm is an optimization technique based on swarm intelligent which is further extended to data clustering. It is used to refine clusters produced by K-Means. The proposed work is aimed to find a solution for generating different .However the method can be computationally expensive in that a large number of function calls is required to advance the swarm at each optimization iteration. In order to increase the efficiency of the algorithm an concept of Surrogate function is incorporated which serves as an stand in for expensive objective function.The work also aims to provide better analysis of the proposed hybrid approach on the basis of obtained numerical results.

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

I. INTRODUCTION

Mobile Data mining, the extraction of concealed prescient data from expansive databases, is a compelling new innovation with incredible potential to help organizations concentrate on the most essential data in their information stockrooms. Most organizations effectively gather and refine enormous amounts of information. Information mining strategies can be actualized quickly on existing programming and equipment stages to improve the benefit of existing data assets, and can be incorporated with new items and frameworks as they are brought on the web. With the dangerous development of data sources accessible on the World Wide Web, it has gotten to be progressively vital for clients to use mechanized instruments in discover the coveted data assets, and to track and dissect their use designs. These factors give rise to the necessity of creating server side and client side intelligent systems that can effectively mine for knowledge.

1.1 Data Mining:

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.

ISSN(Online) : 2319-8753
ISSN (Print) : 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(An ISO 3297: 2007 Certified Organization)

Website: www.ijirset.com

Vol. 6, Issue 7, July 2017

- **Web usage mining**: Also known as Web Log Mining is the procedure of separating fascinating examples in web access.

II. RELATED WORK

Particle Swarm Optimization (PSO) is an evolutionary computation technique. Separate adjustment to inertia weight and learning factors in PSO undermines the integrity and intelligent characteristic in the evolutionary process of particle swarm to some extent, thus it is not suitable for solving most complicated optimization problems [1]. Data mining mainly relies on 3 V's namely, Volume, Varity and Velocity of processing data. Volume refers to the huge amount of data it collects, Velocity refers to the speed at which it process the data and Variety defines that multi-dimensional data which can be numbers, dates, strings, geospatial data, 3D data, audio files, video files, social files, etc. These data which is stored in big data will be from different source at different rate and of different type; hence it will not be synchronized. This is one of the biggest challenges in working with big data. Second challenge is related to mining the valuable and relevant information from such data adhering to 3rd V i.e. Velocity. Speed is highly important as it is associated with cost of processing [2].During the process of classification a lot of irrelevant features are covered by input data [3].Existence of these unrelated features will generate a dimension calamity. In many classification problems, its difficult to learn good classifiers prior to eliminate these undesirable features due to huge amount of data. Reducing the amount of unrelated or redundant features can reduce the running time of classification algorithms. An improved PSO based technique is employed to extract relevant features using dependent criteria for dimension reduction. Clustering algorithms have emerged and rapidly developed as an alternative powerful meta-learning tool to undertake a broad range of applications because it is particularly useful for segmenting large multidimensional data into distinguishable representative clusters. However, Data clustering is one of the most complicated engineering problems and it is considered as an NP-hard problem [4]. Fuzzy clustering is a popular unsupervised learning method used in cluster analysis which allows a point in large data sets belongs to two or more clusters. Prior work suggests that Particle Swarm Optimization based approach could be a powerful tool for solving clustering problems.

III. PROPOSED ALGORITHM

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. However, 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.

3.1 PSO (Particle Swarm Optimization) Algorithm

PSO is initialized with a group of random particles (solutions) and then searches for optima by updating generations. In every iteration, each particle is updated by following two "best" values. The first one is the best solution (fitness) it has achieved so far. (The fitness value is also stored.) This value is called pbest. Another "best" value that is tracked by the particle swarm optimizer is the best value, obtained so far by any particle in the population. This best value is a global best and called gbest. When a particle takes part of the population as its topological neighbours, the best value is a local best and is called lbest.After finding the two best values, the particle updates its velocity and positions with following equation (a) and (b).

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

present[] = present[] + v[]…….. (b)

ISSN(Online) : 2319-8753
ISSN (Print) : 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(An ISO 3297: 2007 Certified Organization)

Website: www.ijirset.com

Vol. 6, Issue 7, July 2017

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 minimum error criteria is not attained .Particles' velocities on each dimension are clamped to a maximum velocity Vmax. If the sum of accelerations would cause the velocity on that dimension to exceed Vmax, which is a parameter specified by the user. Then the velocity on that dimension is limited to Vmax.

K-Means Algorithm

The K-Means algorithm is a simple yet effective statistical clustering technique. Basic 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.*

4.3 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.

PreProcessing:

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)

ISSN(Online) : 2319-8753
ISSN (Print) : 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(An ISO 3297: 2007 Certified Organization)

Website: www.ijirset.com

Vol. 6, Issue 7, July 2017

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
4.Assign the filterered particles to K-Means algorithm to form clusters.

IV. EXPERIMENTAL RESULTS

The clustered version of Apriori algorithm clearly performed in much improved as compared to simple version. As per the application of proposed work on a list of transactions carried out by local customers the analysis clearly showed that the clustered version was much better as compared to other algorithms.

Table 1: Space Analysis

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

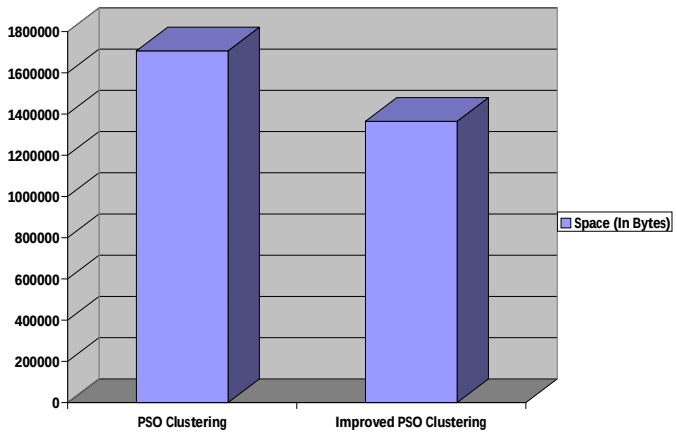

Fig 1: Space Analysis

ISSN(Online) : 2319-8753
ISSN (Print) : 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(An ISO 3297: 2007 Certified Organization)

Website: www.ijirset.com

Vol. 6, Issue 7, July 2017

Table 2: Time Analysis

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

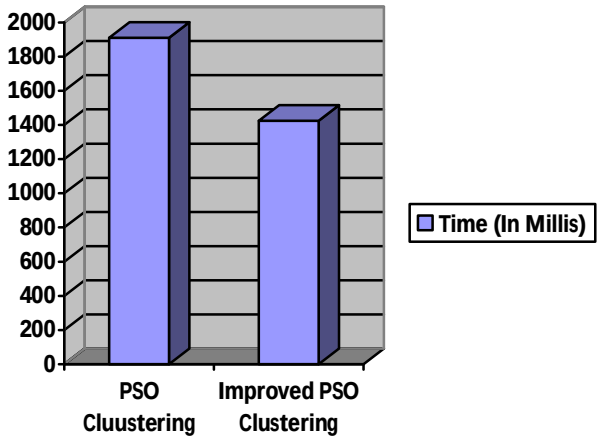

Fig 2: Time Analysis

The analysis cleanly shows that the improved algorithm has better space and time complexities as compared to simple PSO clustering algorithm. Also the clustered algorithm is more useful as the much better analysis can be performed for customer purchasing behavior as the transactions are clustered as seasonal sale trends.

V. CONCLUSION AND FUTURE WORK

Data mining is emerging technologies dealing with major efficiency issues. 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.

ISSN(Online) : 2319-8753
ISSN (Print) : 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(An ISO 3297: 2007 Certified Organization)

Website: www.ijirset.com

Vol. 6, Issue 7, July 2017

[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