International Research Journal of Engineering and Technology (IRJET) e-ISSN: 2395 -0056
Volume: 03 Issue: 03 | Mar-2016 www.irjet.net p-ISSN: 2395-0072

A Survey over Various Techniques Used To Cluster Large Scale Data

Astendra Singh Chauhan[1], Dr. Mahesh Pawar[2]
[1]*School of Information Technology, RGPV, India*
[2]*Asst. Professor, Department of IT, UIT, RGPV, India*

Abstract - *In current scenario a huge amount of data is flooded over the worldwide web. That data comes from various sources and presents in unstructured format. Thus techniques are required to organize that data. Clustering techniques like the fuzzy C-means algorithms is put forward and applied to deal with very large data set or big data those data cannot load in your computer working memory. But there are drawbacks like limited convergence speed and there is no global optimal solution is provided, especially for big data scenario the overall speed is slow and hard to perform the clustering algorithm. A survey over various techniques which are used for clustering of large scale data is presented. To solve problems in existing techniques, a modified Hybrid Particle Swarm Optimization (FCM-FPSO) is presented. But that technique not able provide desired results thus a new hybrid technique which uses chaotic map based fitness function to improve the performance of clustering. Accelerated chaotic particle swarm optimization (ACPSO) based K-mean clustering technique is used to find optimal clusters.*

Key Words: Big Data, Clustering, Fuzzy C-Means, Particle Swarm Optimization,, Chaos Theory and Chaos Particle Swarm Optimization, PSO.

1. Introduction

In [15] recent years, there has been an unprecedented increase in the quantity and variety of data generated worldwide. According to the IDC's Digital Universe study, the world's information is doubling every two years and is predicted to reach 40ZB by2020. This increase in data, often referred to as a "data tsunami", is driven by the proliferation of social media along with an increase in mobile and networked devices (the Internet of Things),finance and online retail as well as advances in the physical and life sciences sectors.

As evidence of this, the online micro blogging service Twitter, processes approximately 12 TB of data per day, while Facebook receives more than five hundred million likes per day. In addition, the Cisco Internet Business Solutions Group (IBSG) predicts that there will be 25 billion devices connected to the Internet by 2015 and 50 billion by 2020. Such vast datasets are commonly referred to as "Big Data".

Big Data is characterized not only by its volume, but by a rich mix of data types and formats (variety) and it's time sensitive nature which marks a deviation from traditional batch processing (velocity). These characteristics are commonly referred to as the 3 V's.

Clustering [10] is a mathematical tool that attempts to discover structures or certain patterns in a data set, where the objects inside each cluster show a certain degree of similarity. Clustering is useful with database in Data Storage and Retrieval Process. When a query is made for the address of a Person the archived data is clustered according to the various criteria, e.g. - by similar street names, within the same zip code or by similar last name.

There [5] [8] have been many researches for cluster analysis. Fuzzy clustering is an extension of cluster analysis. For finding the similarity in the data and grouping the data many fuzzy clustering algorithms are defined in the literature, fuzzy C-Means algorithm one of them. Fuzzy clustering (also referred to as soft clustering) can allocate objects to clusters in a fuzzy way, where data objects can belong to more than one clusters and associated with different membership levels. Therefore, fuzzy clustering can reflect the real world in a more objective perspective. As one of the most popular fuzzy clustering algorithms, Fuzzy C-Means (FCM) clustering combines the fuzzy theory and K-Means clustering algorithm.

However, there are some problems with FCM clustering. First, FCM is very sensitive to the initialization condition, such as the determination of initial clusters. Second, the speed of convergence is limited, and the global optimal solution is hard to be guaranteed, especially for big data scenario. Third, for big data, the overall speed is slow, and it is hard to perform the clustering algorithm on all the original dataset. To solve above challenges, in this paper we propose a modified FCM algorithm especially for big data solution. First, to deal with the initialization sensitivity, we propose to improve FCM using Particle Swarm Optimization (PSO) by optimizing.

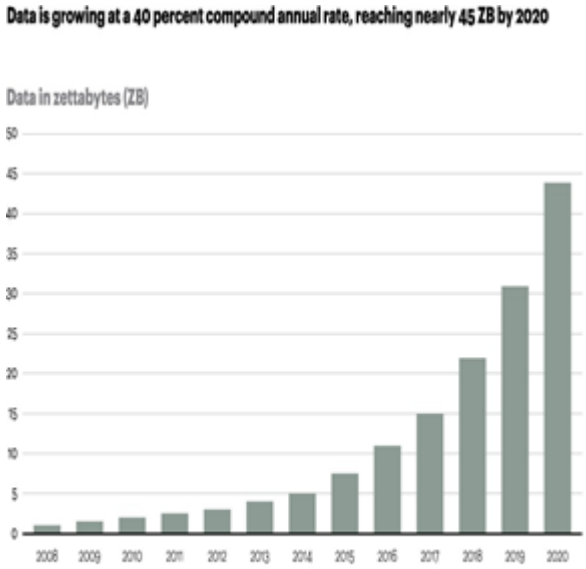

Figure 1 Data growth by 2020

Techniques to describe large scale of data know improve to add chaotic map particle (CPSO) so find optimal value but use in accelerated chaotic particle swarm optimization (ACPSO) to find cluster center and find global optimal value. And solve problem robust and clustering.

Mahout [5] has the potential of being a useful clustering tool. The algorithms have the advantage of using Hardtop Map Reduce and HDFS. It turns out that Mahout can be an inexpensive solution for solving high scalable clustering problems. However, the documentation is still obscure. There are no scientific publications describing how the algorithms are implemented.

On the other side, the open source nature makes the code accessible to everyone. It is not an easy or an intuitive tool to use. Often we need to check directly in the source code to find out how to use it. There are serious doubts that the typical cluster analyst or scientist will have the advanced IT skills necessary to use it with success.

The tools available for dimensionality reduction are not efficient. Often we saw different terms for the same concept, verbs, nouns, and even html tags in the feature vectors. Specifically in the field of document clustering where we have to deal with huge dimensions it is important to have efficient preprocessing tools. We recommend more effort on Mahout's analyzer tool. It will be extremely useful if Mahout could output a quality measure as the total within-cluster sum of squares.

2. Literature Review

In [1] PSO (Particle Swarm Optimization) and chaotic theory based clustering technique is presented. In that technique a fitness function is used and chaotic map used to modify the initialization of the PSO, then a new mutation function called chaotic mutation function and inertia weight are used to improve the results of the proposed technique. But this suffers problems like convergence of the parameters. To improve the performance of the technique, an enhanced PSO and FCM mechanism can be used to improve the performance of the technique.

In [2] a hybrid technique which uses C-mean clustering and PSO (particle swarm optimization) to provide better clustering solution for the large dataset is presented. C-Mean algorithm is not able to provide global optimal solution for the clustering problems on the other hand particle swarm optimization provides better optimal solutions. Thus fuzzy particle swarm optimization FPSO is used. In that results of the C-mean technique are provided as an initial parameter function for the PSO, then optimal results are generated by the PSO. But PSO also suffers defects like parameter dependency and some others thus an enhanced technique is required to improve the performance of the clustering process.

In [3] a fuzzy c-mean algorithm which poses the properties of the fuzzy theory and K-mean clustering technique is presented. In that technique c- mean with PSO algorithm is used to provide better and optimal solution for many problems. Because c-mean suffers defects like sensitive to initialization condition, speed of the convergence is limited, and there is guarantee of optimal solutions. Thus a PSO based technique is used with C-mean clustering to improve the performance of the whole clustering process.

In [4] a hybrid fuzzy C-mean based fuzzy PSO (Particle Swarm optimization) clustering technique is presented. In existing techniques C-mean clustering technique is used, but it suffers defects like sensitive to initialization, and no guaranteed global solution for clustering is provided. Thus a PSO technique is used which provides optimal solution. In that C-mean alters initialization of the PSO and then provides better solution for such problems.

In [9] a Fuzzy C-mean clustering technique to provide clustering solution for large scale data is presented. In that technique fuzzy means not clearly defined, and C stands for clustering. It is an enhanced version of K-mean technique, in K-mean only distance is calculated on the other hand in C-mean inverse distance weights are calculated to form

clusters. Thus FCM provides better solution for clustering problems.

In [14] fuzzy C-mean clustering technique for fuzzy database is presented. In existing techniques a SQL query or FSQL query based techniques are used to retrieve data from database. But when the size of data increases then it generate ambiguity, vagueness, uncertainty, in the data thus a simple SQL query based technique is not enough to provide optimal solution. Thus a fuzzy C-mean clustering technique is presented to provide optimal solution for clustering problems.

In [11] a hadoop framework based technique called Mahout is used to provide clustering solution for large scale data problems, is presented. In these techniques Hadoop framework with Mahout is used. In existing technique various techniques with Mahout are used but machine learning techniques with hadoop framework provide enhanced performance and having application in fields like pattern recognition, classification and many more. Thus a hadoop and Mahout based technique is presented. Which provides better solution for such problems?

In [17] a data clustering is a powerful clustering technique to describe large scale of data knows improve to add chaotic map particle (CPSO) so find optimal value but use in accelerated chaotic particle swarm optimization (ACPSO) to find cluster center and find global optimal value. And solve problem robust and clustering.

3. Proposed Work

In existing techniques, PSO, CPSO, FPSO are used to provide solution for the various clustering problems but these techniques also suffers some inherent defects to resolve these defects accelerated chaotic particle swarm optimization (ACPSO) algorithms can be used to find the optimal value to form a cluster. This technique provides an enhanced performance to form clusters. In this technique a chaotic map and Particle swarm optimization technique used with Fuzzy K-mean clustering to provide efficient and accurate mechanism to form clusters

4. Conclusions:

To cluster large scale data is a difficult task to do, because data is presented in unstructured and heterogeneous format. Fuzzy C-mean technique is used to cluster large scale data but fuzzy c-means algorithm is sensitive to initialization and is easily trapped in local optima. On the other hand, the particle swarm algorithm is a global stochastic tool which provides optimal solution for many classification and clustering problems. A survey over various techniques used to provide clustering solution for many problems is presented in section II. For proposed work a hybrid technique which uses FCM-FPSO and Chaotic map based fitness function to provide improve solution to cluster large scale data.

5. References:

[1] SamanPoursiahNavi, EhsanToreini, Maryam MehrnejadAnd S.KazemShekofteh "Analysis Of The Usage Of Chaotic Theory In Data Clustering Using Particle Swarm Optimization" Indian J.Sci.Res, 2014.
[2] Dr. M.Seetha, G. Malini Devi, Dr. K.V.N.Sunitha"An Efficient Hybrid Particle Swarm Optimization For Data Clustering" International Journal of Data Mining & Knowledge Management Process (IJDKP) Vol.2, No.6, November 2012.
[3] YandXianfend And Liu Pengfei " Tailoring Fuzzy C Means Clustering Algorithm For Big Data Using Random Sampling And Particle Swarm Optimization " International Journal Of Database Theory And Application Vol.8.No 3 (2015).
[4] Hesam Izakian and Ajith Abraham'' Fuzzy C-Means and Fuzzy Swarm for Fuzzy Clustering Problem'' Expert Systems with Applications 38 (2011) 1835–1838
[5] RuiMáximoEstevesAndChunmingRong "Using Mahout for Clustering Wikipedia's Latest Articles .A Comparison between K-Means and Fuzzy C-Means in the Cloud" 2011 Third IEEE International Conference on Coud Computing Technology and Science.
[6] V.Kumutha, And S. Palaniammal "Improved Fuzzy Clustering Method Based On Intuitionistic Fuzzy Particle Swarm Optimization" Journal of Theoretical and Applied Information Technology 10 The April 2014. Vol. 62 No.1.
[7] Yinghua Lu and, Tinghuai Ma" Implementation Of The Fuzzy C-Means Clustering Algorithm In Meteorological Data" International Journal Of Database Theory And Applicationvol.6, No.6 (2013).
[8] Sara Nasser And, Rawan Alkhaldi" A Modified Fuzzy K-Means Clustering Using Expectation Maximization" University Of Nevada Reno, Reno Nv 89557, Usa.
[9] Tejwant Singh and, Mr. Manish Mahajan "Performance Comparison Of Fuzzy C Means With Respect To Other Clustering Algorithm" Volume 4, Issue 5, May 2014, Issn: 2277 128x.
[10] QuangHieu And Xingjian Huang "An Improved Fuzzy C-Means Clustering Algorithm Based On Pso" Journal Of Software, Vol. 6, No. 5, May 2011.
[11] G.Jeeva, And E.K.Subramanian "Improving Search Performance For Big data Processing Using Machine Learning Algorithm" International Journal Of Scientific & Engineering Research, Volume 5, Issue 4, April-2014 451issn 2229-5518.
[12] Arantxa Duque Barrachina And AislingO'driscoll "A Big Data Methodology For Categorising Technical Support Requests Using Hadoop And Mahout" Duque Barrachina And O'driscolljournal Of Big Data2014, 1:1.
[13] SamanPoursiahNavi And, EhsanToreini" Analysis Of The Usage Of Chaotic Theory In Data Clustering Using

Particle Swarm Optimization" Indian J.Sci.Res. 4 (3): 335-353, 2014 ,Issn: 0976-2876 (Print).

[14] Neha Jain And, Seema Shukla "Fuzzy Databases Using Extended Fuzzy C-Means Clustering" International Journal Of Engineering Research And Applications (Ijera) ISSN: 2248-9622 Vol. 2, Issue 3, May-Jun 2012, Pp.1444-1451.

[15] Timothy C.Havens and James C.Bezdek "Fuzzy C-Means Algorithms For Very Large Data" IEEE Transactions On Fuzzy Systems Vol. 20. No. 6, December 2012.

[16] Cheng-Hong Yang and Li-yeh Chuang "Accelerated chaotic particle swarm optimization for data clustering" International Conference on Machine Learning And Computing, IPCSIT vol.3 2011.

[18] Min Chen and Simone "Fuzzy clustering using automatic particle swarm optimization".