IJCSN International Journal of Computer Science and Network, Volume 6, Issue 1, February 2017
ISSN (Online) : 2277-5420
www.IJCSN.org
Impact Factor: 1.5

A Review on Clustering Analysis based on Optimization Algorithm for Datamining

[1] Rashmi P. Dagde; [2] Snehlata Dongre

[1] Research Scholar, M. Tech Computer Science Engineering,
G. H. Raisoni College of Engineering,
Nagpur, India

[2] Assistance Professor, Computer Science Engineering,
G. H. Raisoni College of Engineering,
Nagpur, India

Abstract – Clustering analysis is one of the important concept of data mining. Many researchers are focus on the clustering problem it is one of the research based criteria. The clustering is belongs to the unsupervised learning in which teacher is absent. This paper shows to analysis the clustering problem. clustering is the data mining concept in which grouping are done with the help of the algorithm. For the clustering in this paper the Bisecting K-mean algorithm is used. It will find the clustering means it will arrange the data into group wise manner. In this paper the data set is collected from the UCI Repository. The Bisecting K-mean algorithm has some drawback like it will not find the centroid for these the clustering not found proper manner and to remove this drawback used the PSO algorithm. The particle swarm optimization algorithm is remove the drawback of the clustering. PSO algorithms find the optimal path. This integrated hybrid model increase the accuracy of the clustering.

Keywords - *UCI Repository data, Bisecting K-mean Algorithm, Particle swarm optimization technique*

1. Introduction

Data mining is the mining of the predictive information from database and it is new technology to help companies focus on the very important information in their data bases. Data mining means applying few mining technique on data set for do it useful. It is process of analyzing hidden predictive information from the big data set. It is used to examine the old data to find the information. In this paper clustering is used it is one of the popular technique of data mining. It is task of dividing a data into the number of similar clusters. Means it is task of grouping a set of object in a same group are similar to each other in the other group. Data clustering technology is to finding the similar hidden pattern from the given data set. It is the method to obtaining the cluster of the item without the class label related to the approximation of the item in one cluster. Clustering is the very big amount of the data set that contain the large number of records with high dimensions. And now a days it used for the identifying useful information from the historical data. In data clustering it has disadvantages such as dependency of the ability of this technique to initialize the cluster centre. Means it is hard to perform to initialize the centre of the cluster for that purpose to increase the efficiency of this method used the optimization technique. In this project optimization technique is use for the overcome the drawback of clustering. The optimization is used to find the global optimization solution. Now a days in real word the optimization problem are dynamic. It will not find the global optimal solution but also find the trajectory of changing optima over dynamic nature. The optimization technique will give the optimal or good solution from the complex optimization problem.

2. Literature Survey

The clustering analysis is very big problem specially in large data set. The clustering analysis means it will divide the given data into certain classes according to the main attribute of the dataset, thus the item of class to have the same or similar meaning. They have find attention due to their importance in data mining research and application. It is the unsupervised method that applied for prediction of structure of object by dividing them into

different set. The Clustering algorithm has some drawback for example K-mean have drawback in searching for optimal solution. It is very difficult to find the cluster center. Consider the drawbacks and limitation of single clustering technique, research has developed optimization algorithm to overcome the drawbacks of clustering technique.

In this paper the data mining technique is used for clustering. The clustering is the one of the problem in data mining that always affected many researchers. Clustering technique is one of the important unsupervised method. In this project the k-mean algorithm is used for the clustering. It will find the cluster. The clustering number must be defined initially. Each center of K-mean identify by mean location of data vector which find the cluster. And the practical swarms optimization algorithm will used for finding the optimal path of clustering. It is one of the important algorithm inspired by birds. Due to the limitation of the k-mean algorithm the PSO will used for overcome the k-mean drawback. The drawback of K-mean is to initialization of cluster center. This is one of the disadvantages of the k-mean. To overcome this advantages the PSO algorithm was used , using PSO it will find the cluster center will generated initial piratical are optimized to maintain the final cluster center.[1]

In this paper the data mining technique is used for analysis of clustering. The clustering is one of the problem in data mining that always affected many researchers. Clustering is one of the important unsupervised classification method. In this project or research paper the k-mean algorithm is used for the clustering. It will find the number of cluster. The cluster number are must be defined firstly. Each center of K-mean denoted by mean location of data vector which represent to cluster. And the Mussels Wandering Optimization algorithm will used for finding the optimal path of clustering. It is one of the important algorithm related to the Swarm intelligence. It is one of the new effective global optimization algorithms. It aims to find an optimal solution by using or modifying Mussels leisurely locomotion behavior. It is very simple and easy to used algorithm. Due to the drawback of the k-mean clustering algorithm the MWO will used for overcome the k-mean drawback. The one of limitation of K-mean has drawback is to initialization of cluster center. This is one of the disadvantages of the k-mean. To overcome this disadvantages the MWO algorithm was used , using MWO it will find the cluster center belonging with other generated initial piratical are optimized to maintain the final cluster center.[2]

In this paper the data mining technique is used for analysis of clustering. The clustering is one of the problem in data mining that always affected many researchers. Clustering technique is unsupervised method. In this paper K-Harmonic mean algorithm is use for clustering. It has drawback like sensitivity to initial starting point and convergence problem to the local optimum. And Simplified swarm Optimization was used for the finding the optimal path. It is inspired by the Practical swarms optimization algorithm. It was originally propose to overcome the limitation of PSO for discrete type optimization. It was used for overcome the drawback of K-Harmonic mean algorithm..[3]

In this paper the data mining technique is used for analysis of clustering. The clustering is one of the problem in data mining that always focus on many researchers. Clustering technique is one of the important unsupervised classification method. In this paper C-mean algorithm was used for the clustering. It is most important popular fuzzy clustering algorithm. It will also used in real word application. It has drawback like sensitive to initialize and it is easily struck at local minima. And the practical swarms optimization algorithm will used for finding the optimal path of clustering. It is one of the important algorithms inspired by birds. Due to the drawback of the C-mean clustering algorithm the PSO will used for overcome the C-mean drawback. The drawback of C-mean is not find the to initialization of cluster center. This is one of the disadvantage of the C-mean. To overcome this advantages the PSO algorithm was used , using PSO it will find the cluster center along with other generated initial particle are optimized to maintain the final cluster center. [4]

In this paper the consensus clustering is used in which each discrete feature is viewed as simple clustering of the data. The plus points of this algorithm are it generate better clustering and it is less sensitive to noise. It has drawback like sensitivity to initial starting point and convergence problem to the local optimum. And the Piratical swarm optimization algorithm is use for finding the optimize path for improving the consensus clustering algorithm. [5]

In this research paper the swarm intelligence brain storm Algorithm is used. The optimization problem will be simply splits into uni modal problem and multi modal problem. The uni modal problem has one optimum solution and the multi modal problem has much optimum solution these concepts are use in this paper. It is one of the nature-inspired searching techniques. In this optimization technique the clustering information is to

reuse the landscape of problem and to give the information to every individual. Every individual in the (BSO) that is brain storm optimization algorithm is not a solution to optimize the problem. It is one of the algorithms which is important algorithm in swarm intelligence is related to the behavior of human being. In this algorithm there are three strategies that are result of clustering in which new individual's generation and selection. The solution are splits into several cluster and the good solution of every cluster are kept into the next generation. New individual are generated Related to one or more two individual in cluster it is one of the kind of search reduction algorithm.[6]

In this project the partition clustering is used for the clustering analysis. It will concern by the optimal cluster and the point of cluster center in hyper dimension space. In this research project two stage dynamic clustering method is used. In the state of the optimum set of solution is produced by multi-objective particle swarm intelligence algorithm and then use the decision making method. The decision tree method is the technique is order of important by similar to the solution and the division is done related to the optimal solution to choose the good one solution.[7]

In this paper the concept of e-learning is used. In this concept the learning analysis will perform with the help the clustering. The clustering is one of the data mining techniques which can use to analysis the learning efficiency of the student. The PSO is one of the clustering optimize technique that can be adapted and modified to handle the big data in learning analysis. The particle swarm optimize algorithm has been used to cluster student and measure their learning efficiency the parallel particle swarm optimization based clustering algorithm will reduce the processing time as a function of the available number of processor. The PSO used as student and it will measure their efficiency.[8]

In this paper they improve the kernels rough fuzzy c-mean clustering algorithm by using optimize particle swarm optimize algorithm. The fuzzy set enables efficient handling the overlapping partition it will deals with uncertainty in class definition where as the kernel trick ensure linear separatibility of the complex cluster which is not linearly separable. The PSO find the near optimum values of the different parameters. In c-mean clustering the cluster center have used to form the cluster and then it will extended to fuzzy c-mean algorithm that is related to the concept of the fuzziness in order to handling the overlapping class in the real data set and then it will extended to the rough fuzzy c-mean algorithm. It will solved the uncertainty problem of the data set. The performance of this method are very careful to the choice of the parameters, optimal values are determine by using the PSO algorithm. By using this method each cluster are treated as an interval or rough set. [9]

The clustering algorithm that first collected the statistical data samples and then analyse their features and then finally classifies them to different clusters. The distributed clustering algorithm focus on the multidimensional is concerned. The distributed k-mean algorithm is a simple and efficient clustering algorithm it can process massive data efficiently. In this paper the distributed k-mean clustering algorithm based on map reduce computing is divided into the steps like randomly select the initial center and initial parameter, put all data object into function map and by comparing the distance from data object to the center, select the point of minimum distance as a center and output are form as a map fragment. To improve the distributed k-mean clustering algorithm the classic Euclidian algorithm will used for calculating the distance.[10]

In research paper the (CRO) coral reefs optimization algorithm used to clustering problem. It will give clustering partition for a data warehouse. The coral reefs optimization algorithm dealing with the real data. It is a bioinspired meta heuristic for problem. Thai is one of the optimization problem. To the one of good information, that optimization will not be used to clustering analysis for this purpose the analysis of the coral reefs optimization algorithm in clustering problem and improve the three modification of coral reefs optimization algorithm for improving the performance.[11]

3. Flowchart of System

In this project the data sets are taken from the UCI repository. It stands for the University of California Irvine. It is one of the machine learning repository having the collection of database and it is machine learning university. Then applying the clustering technique that is bisecting k-mean algorithm. It will find the k number of cluster of the apply data set. Then applying the optimization algorithm it will find the optimize path of the clustering and increase the accuracy of the integrated hybrid algorithm.

Collect Data Sets(UCI Repository Collect Data Sets(UCI Repository

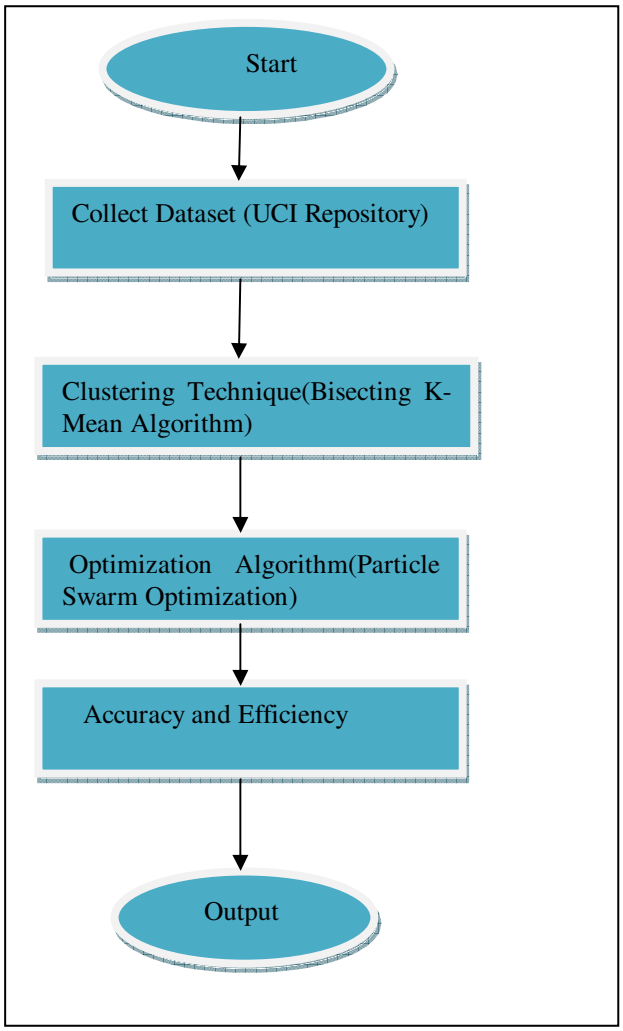

Fig 1:- System Architecture

4. Methodology

Clustering is the one of the popular technique of data mining. It is task of dividing a data into a number of similar cluster. Means it is task of grouping a set of objects in such way that object in same group are similar to each other in the other group.

4.1 Bisecting K-Mean Technique:-

In this system Bisecting K-mean clustering technique is used. After applying K mean algorithm toe cluster will form now we apply Bisecting K-mean algorithm on the obtain clusters. Firstly select any cluster among the two cluster for doing these and then calculate centered for both the cluster separately that is cent1 and cent2 for cluster1 and cluster2 respectively. Later distance is calculated between the centered and data set.

Steps:-

(1) If dist1>dist2 then again divide the cluster1 into two more cluster.
(2) If dist2>dist1 then again divide the cluster2 into two more cluster.
(3) This process is continue till K clusters are obtain

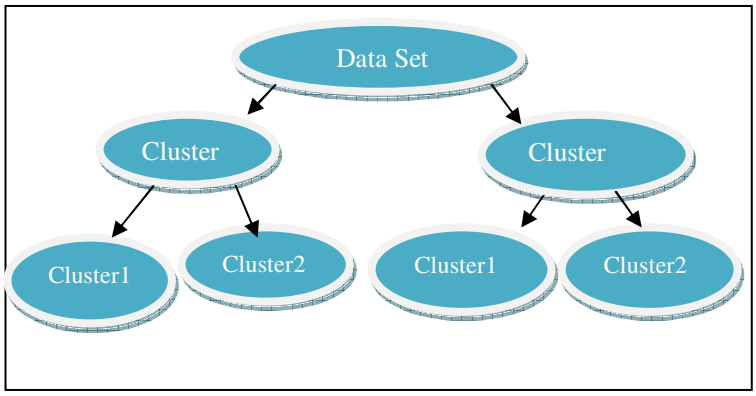

Fig 2:- Tree Diagram for k=4 cluster formation

4.2 Particle Swarm Optimization algorithm:-

The K-mean has drawback like the dependency of method to initialize the cluster center. And to overcome this limitation the particle swarm optimization algorithm is taken. The k mean used to main the cluster center and each cluster center is same as a particle of the particle swarm optimization algorithm.

Later on using PSO algorithm these clusters center along with generated initial particle are obtained to maintain the final cluster center.

Using particle to show a clustering result was introduced which has a better unity with the architecture of PSO algorithm because of the common form of every particle of PSO should be a accurate solution obtained. In this way a particle confident of the clusters midpoint**.**

5. Conclusion

The Bisecting K-mean algorithm is one of the clustering algorithm used in the large sets data. It will find out the

accuracy of the clustering. The particle swarm optimization will find the optimize path. This integrated clustering algorithm increases the accuracy.

References

[1] Habibollah Agh Atabay,Mohammad Javad Sheikhzadeh, Mehdi Torshizi, "A Clustering Algorithm Based on Integration of K-Means and PSO," Conference on Swarm Intelligence and Evolutionary Computation (CSIEC2016), Higher Education Complex of Bam, Iran, 2016, pp. 59-63.

[2] Peng Yan, ShiYao Liu, Qi Kang,Bing Yao Huang,MengChu Zhou "A Data Clustering Algorithm Based on Mussels Wandering Optimization," IEEE International Conference on Computer Science and Service System, pp. 713-718.

[3] Chia-Ling Huang,Wei-Chang Yeh,"A New K-Harmonic Means based Simpified Swarm Optimization for Data Mining," IEEE International Conference on Computer Science and Service System, 2014 pp. 1-10.

[4] O.A. Mohamed Jafar, R. Sivakumar, "A Study on Fuzzy and particle Swarm Optimization Algorithm and their Application to clustering Problem," IEEE Internatinal Conference on Advanced Communication Control and Computing Technologies(ICACCCT), pp. 462-466.

[5] Parul Agarwal,Shikha Mehta, "Comparative Analysis of Nature Inspired Algorithms on Data Clustering," IEEE International Conference on Research in Computational and Communication Network(ICRCICN), pp. 119-124.

[6] Shi Cheng, Yuhui Shi, Quande Qin, Shujing Gao, "Solution Clustering Analysis in Brain Storm Optimization Algorithm," IEEE Symposium on Swarm Intelligence (SIS), 2013, pp. 111-118.

[7] Amin Alizadeh Naeini, Saeid Homayouni,"Improving the Dynamic Clustering of Hyperspectral Data Based on the Integration of Swarm Optimization and Decision Analysis",IEEE JOURNAL OF SELECTED TOPICS IN APPLIED EARTH OBSERVATIONS AND REMOTE SENSING, VOL. 7, NO. 6, JUNE 2014, pp.2161-2173

[8] Athabasca University,2Anna University,Kannan Govindarajan1, David Boulanger1, Jérémie Seanosky1, Jason Bell1, Colin Pinnell1, Vivekanandan Suresh Kumar1, Kinshuk1, Thamarai Selvi Somasundaram2, "Performance Analysis of Parallel Particle Swarm Optimization Based Clustering of Students," IEEE 15th International Conference on Advanced Learning Technologies DOI 10.1109, 2015, pp. 446-450.

[9] Imicio G. Medeiros and J6ao C. Xavier-Junior,Anne M. P. Canuto, "Applying the Coral Reefs Optimization Algorithm to Clustering Problems," IEEE 2015.

[10] Anindya Halder, "Kernel based Rough Fuzzy c-Means clustering optimized using Particle Swarm Optimization," International Symposium on Advanced Computing and Communication (lSACC), 2015.

[11] Song Ling, Qi Yunfeng,"Optimization of the Distributed K-means Clustering Algorithm Based on Pair Analysis", 8th International Congress on Image and Signal Processing (CISP 2015) 2015, pp. 1593-1598.

[12] Surjodoy Ghosh Dastider, Himanshu Kashyap,Shashwata Mandal, Abhinandan Ghosh,Saptarsi Goswami, "Feature Subset Selection for Clustering using Binary Particle Swarm Optimization," International Conference on Information Technology DOI 10.1109, 2015, pp. 159-164.

[13] Paulus Mudjihartono, Thitipong Tanprasert, Rachsuda Jiamthapthaksin, "Clustering Analysis on Alumni Data Using Abandoned and Reborn Particle Swarm Optimization,"IEEE, 2016, pp. 22-26.

[14] Ioan-Daniel Borlea, Radu-Emil Precup and Florin Dragan, "On the Architecture of a Clustering Platform for the Analysis of Big Volumes of Data," IEEE International Symposium on Applied Computational Intelligence and Informatics, 2016, pp. 145-150.

[15] Mohammad Reza Farmani and Giuliano Armano, "Clustering Analysis using Opposition-based API Algorithm",IEEE,2015.

[16] Min Chen and Simone A. Ludwig,"Fuzzy Clustering Using Automatic Particle Swarm Optimization",IEEE International Conference on Fuzzy Systems (FUZZ-IEEE),2015 pp. 1545-1552.

[17] ching-yi Chen, Fun Ye, "Particle Swarms Optimization Algorithm and its Application to Clustering Analysis," International Conference on Networking, Sensing & Control,2004,pp. 789-794.

[18] Simon Fong, Raymond Wong, and Athanasios V. Vasilakos, Senior Member, "IAccelerated PSO Swarm Search Feature Selection for Data Stream Mining Big Data,"IEEE Transactions On Services Computing, Vol. 9, No. 1, January/February 2016, Pp. 33-45.

[19] Pushpalatha K,Ananthanarayana V S "A New Glowworm Swarm Optimization Based Clustering Algorithm for Multimedia Documents," IEEE International Symposium on Multimedia, 2015 DOI 10.1109, pp. 262-265.

[20] Yong Zhang, Dun-wei Gong, "Multi-objective Particle Swarm Optimization Approach for Cost-based Feature Selection in Classification,"IEEE Transactions On Journal Name, Manuscript Id Doi 10.1109/Tcbb.2015, Pp. 1545-5963.

[21] Shafiq Alam, Gillian Dobbie, Yun Sing Koh, Patricia Riddle, "Web Bots Detection Using Particle Swarm Optimization Based Clustering," IEEE Transactions On Instrumentation And Measurement, Vol. 64, No. 12, December 2015,pp.3588-3600.

[22] Ruchika R. Patil,Amreen Khan, "Bisecting K-Means for Clustering Web Log data," International Journal of Computer Applications (0975 – 8887),2015, pp. 36-41.

[24] Zhang Ke1, Huang Lei, Chai Yi,"An Algorithm to Adaptive Determination of Density Threshold

for Density-based Clustering",Proceedings of the 35th Chinese Control Conference July 27-29, 2016, pp. 3929-3935.

[25] ChaitraH.V,Dr. Ravikumar G.K,"A secure and energy efficient cluster optimization by using hierarchal clustering technique",Third International Conference on Devices, Circuits and Systems (ICDCS'16) ,2016,pp.93-97