K-Mean Clustering and PSO: A Review

Gursharan Saini, Harpreet Kaur

Abstract- Clustering is a method which divides data objects into groups based on the information found in data that describes the objects and relationships among them. There are a variety of algorithms have been developed in recent years for solving problems of data clustering. Data clustering algorithms can be either hierarchical or partitioned. Most promising among them are K-means algorithm which is partitioned clustering algorithm .Moreover k-mean Algorithm is an efficient Clustering Algorithm but it can generate a local optimal solution. On the other hand, Particle Swarm Optimization is used for global optimization. Thus K-means algorithm shows improved results when used with the combination of PSO (Particle Swarm Optimization).

Index Terms- Data clustering, Data Mining, K-Mean PSO

I. INTRODUCTION

Data mining has attracted a great deal of attention in the information industry because of the wide availability of large amounts of data and the imminent need for turning such data into useful information and knowledge. The importance of data mining is increasing exponentially since last decade and in recent time where there is very tough competition in the market where the quality of information and information on time play a very crucial role in decision making. There is very large amount of data availability in real world and it is very difficult to access the useful information from this huge database and provide the information to which it is needed within time limit and in required pattern. Clustering is a method which divides data objects into groups based on the information found in data that describes the objects and relationships among them, their feature values which can be used everywhere in many applications, such as knowledge discovery, vector quantization, pattern recognition, data mining and data dredging etc. There are mainly two methods for clustering: hierarchical clustering and partitioned clustering [2]. Different algorithms have been proposed which take into account the nature of the data and the input parameters in order to cluster the data. The k-means algorithm is most widely used well known unsupervised partitioning method for data clustering. K-means clustering method grouped data based on their closeness to each other according to Euclidean distance. In this clustering approach user decide that how many clusters should be and on the basis of closeness of data vector to the centroid, which is mean of the data vector of cluster is assigned to that cluster which shows minimum distance. The result depends on the number of cluster (k value) and the initial centroid chosen by the K-Mean algorithm.

Manuscript Received on June 2014.

Gursharan Saini, M.Tech in Computer Science and Engineering from Sant Baba Bhag Singh Institute of Engineering & Technology, Padhiana, India.

Harpreet Kaur, is M.Tech ,pursuing Ph.D and Working as an A.P.(Senior Scale) at Sant Baba Bhag Singh Institute of Engineering & Technology, Padhiana, India.

The K-means algorithm is responsible for generating a local optimal solution and in contrast to that Particle Swarm Optimization (PSO) technique offers a globalized search methodology that can be used in K- Means algorithm to find the global optimal solution. Using the advantages of both the algorithms a Novel approach can be made that overcomes drawbacks of both the algorithms and generate a combined algorithm for the solution.

II. K-MEAN

The k-means algorithm is most widely used well known unsupervised partitioning method for data clustering. K-means clustering method grouped data based on their closeness to each other according to Euclidean distance. In this clustering approach user decide that how many clusters should be and on the basis of closeness of data vector to the centroid, which is mean of the data vector of cluster is assigned to that cluster which shows minimum distance. The result depends on the number of cluster (k value) and the initial centroid chosen by the K-Mean algorithm.

A. K-Mean Algorithm

Step 1:
Initialize the cluster's centroid vectors randomly.
Step 2:
For each data vector calculate the distance between data vector and each cluster centroid which will minimum data vector will assign with that cluster and distance calculate using equation (1). Where d is the dimension.

$$d(Z_p, M_j) = \sqrt{\sum_{k=1}^{d}(Z_p, k - M_j, k)} \quad \text{..... (1)}$$

Where Zp is the pth data point, MJ is centroid of jth cluster.
Step 3:
Recalculate the centroid of cluster using equation (2)

$$M_j = \frac{1}{n}\left(\sum_{\forall Z_p \in C_j} Z_p\right) \quad \text{..... (2)}$$

Where nJ is the number of data point in cluster j.
Step 4:
Repeat step 3&4 until stopping criteria satisfied.
The satisfying criteria can be either number of iteration or change of position of centroid in consecutive iterations.

III. PARTICLE SWARM OPTIMIZATION (PSO)

PSO is a population-based search algorithm which is initialized with a population of random solutions, called particles. As against the other evolutionary computation techniques, each particle in this algorithm, called PSO is also associated with a velocity [4]. In PSO, swarm is composed of a set of particles $P = \{ p1 , p2 , p3 ,..... pn \}$. The position of particle corresponds to a candidate solution of the optimization problem. In PSO a number of simple entities "the particles" are placed in the search space of some problem or function, and each one of these evaluates the objective function at its current location. Thereafter, each particle then determines its movement through the search space by combining some aspect of the history of its

Published By:
Blue Eyes Intelligence Engineering
& Sciences Publication Pvt. Ltd.

own current and best (best-fitness) locations with those of one or more members of the swarm, with some random perturbations. The next iteration takes place after all particles have moved. Eventually the swarm as a whole, like a flock of birds collectively foraging for food, is likely to move close to an optimum of the fitness function. The basic PSO algorithm consists of three steps, namely, generation of particles and their information, movements and new information vector. This can be considered as generating particle's positions and velocities, velocity update and finally, position update [6]. The main advantage of Particle Swarm Optimization (PSO) technique is that it offers a globalized search methodology that can be used in K- Mean algorithm to find the optimal solution. Particle swarm optimization is an evolutionary computation technique which finds optimum solution in many applications. Using the PSO optimized clustering results in the components, in order to get a more precise clustering efficiency [2]. Clustering with swarm-based algorithms (PSO) is emerging as an alternative to more conventional clustering techniques. PSO is a population-based stochastic search algorithm that mimics the capability of swarm (cognitive and social behavior). Data clustering with PSO algorithms have recently been shown to produce good results in a wide variety of real-world data [4].

IV. LIMITATIONS OF K-MEAN AND PSO

Limitations of K-Mean

1. Initial selection of the number of cluster must be previously known and specified by the user.
2. Results directly depend on the initial centroid of cluster chosen by algorithm.
3. It can contain the dead unit problem.
4. Due to its sensitiveness to initial partition it can only generate a local optimal solution.

Limitations of PSO

1. The method cannot work on the problems of non-coordinate systems like the solution of energy field and the moving rules for the particles in the energy field.
2. When the search space is high its convergence speed becomes very slow.
3. Another PSO problem is its nature to a fast and premature convergence in mid optimum points.
4. PSO is a good clustering method, it does not perform well when the dataset is large or complex [6].

V. BENEFITS OF USING K-MEAN AND PSO IN COMBINATION

K-mean clustering is widely used to minimize squared distance between features values of two points reside in the same cluster. Particle swarm optimization is an evolutionary computation technique which finds optimum solution in many applications. Using the PSO optimized clustering results in the components, in order to get a more precise clustering efficiency [2].

PSO is a population-based stochastic search algorithm that mimics the capability of swarm (cognitive and social behavior). Data clustering with PSO algorithms have recently been shown to produce good results in a wide variety of real-world data [4]. PSO performs global search ability K-mean performs local search ability [5]. One of the big issue with K-mean clustering algorithm was to define the number of clusters at the start of the clustering process by the user. To overcome such a problem, particle swarm optimization (PSO) and fuzzy theorem which automatically determines the appropriate number of clusters and their centers [4]. PSO in sequence with K-Means algorithm for data clustering. when used overcome drawbacks of both algorithms, improves clustering and avoids being trapped in a local optimal solution. In this algorithm initial process starts by PSO due to its fast convergence and then the result of PSO algorithm is tuned by the K-Means near optimal solutions [4]

VI. CONCLUSION

In this paper, a review of various researches done in the areas of K-mean and PSO is presented and it is concluded that when K-mean is used in combination with PSO it produces efficient results in terms of efficiency and accuracy because both the algorithms overcome the drawbacks of their own.

VII. ACKNOWLEDGMENT

Author of this paper is thankful to Asst. Prof. Harpreet Kaur for providing her valuable ideas. Author would also like thank to their colleagues, friends, teachers and other guides for providing their support and helpful comments.

REFERENCES

[1] Pallavi Purohit and Ritesh Joshi March 2013 A New Efficient Approach towards k-means Clustering Algorithm ,International Journal of Computer Applications (0975-8887) Volume 65-No.11.

[2] Pritesh Vora, Bhavesh Oza February 2013 A Survey on K-mean Clustering and Particle Swarm Optimization, International Journal of Science and Modern Engineering, (IJISME) ISSN: 2319-6386, Volume-1, Issue-3.

[3] Manpreet Kaur and Usvir Kaur 2013A Survey on Clustering Principles with K-means Clustering Algorithm Using Different Methods in Detail , IJCSMC, Vol.2 Issue.5.

[4] Sunita Sarkar, Arindam Roy, Bipul Shyam Purkayastha 2013 Application of Particle Swarm Optimization in Data Clustering, International Journal of Computer Applications (0975 – 8887) Volume 65– No.25, 2013.

[5] Mehdi Neshat, Shima Farshchian Yazdi, Daneyal Yazdani and Mehdi Sargolzaei 2012 A New Cooperative Algorithm Based on PSO and K-Means for Data Clustering , Journal of Computer Science 8 (2): 188-194.

[6] Sandeep Rana, Sanjay Jasola and Rajesh Kumar 2010 A hybrid sequential approach for data clustering using K-Means and particle swarm optimization algorithm ,International Journal of Engineering, Science and Technology Vol. 2, No. 6.

[7] Rajeev Kumar, Rajeshwar Puran and Joydip Dhar 2010 Enhanced K-Means Clustering Algorithm Using Red Black Tree and Min-Heap , International Journal of Innovation, Management and Technology, Vol. 2, No. 1.

[8] Baolin Yi, HaiquanQiao, Fan Yang 2010 An Improved Initialization Center Algorithm for K-Means Clustering CSE International Conference.

[9] Napoleon and P. Ganga lakshmi 2010 An efficient K-Means clustering algorithm for reducing time complexity using uniform distribution data points, Trendz in Information Sciences & Computing (TISC).

[10] David Pettinger and Giuseppe Di Fatta Dec.2010 Space Partitioning for Scalable K-Means, Machine Learning and Applications (ICMLA), 2010 Ninth International Conference.

Published By:
Blue Eyes Intelligence Engineering
& Sciences Publication Pvt. Ltd.

Gursharan Saini obtained her B.Tech degree in computer Science and Engineering from PTU, University, in 2007 and doing M.Tech in Computer Science and Engineering from Sant Baba Bhag Singh Institute of Engineering & Technology, Padhiana. A Lifetime member of Indian Society for Technical Education.

Harpreet Kaur is M.Tech ,pursuing Ph.D and Working as an A.P.(Senior Scale) at Sant Baba Bhag Singh Institute of Engineering & Technology, Padhiana. She has published 15 papers in international Journals and conferences & 16 papers in national conferences.

Published By:
Blue Eyes Intelligence Engineering
& Sciences Publication Pvt. Ltd.