IJCSE International Journal of Computer Sciences and Engineering Open Access
Review Paper Vol.-6, Issue-9, Sept. 2018 E-ISSN: 2347-2693

A Review on Optimizing Clustering Technique for Data Stream using Genetic Algorithm

Neha Sharma[1], Pawan Makhija[2]

[1,2]Shri Govindram Seksaria Institute of Technology and Science, Indore (M. P.) , India

[*]*Corresponding Author: [1]Ne3haa@gmail.com, Tel.: +91-8871548408*

Available online at: www.ijcseonline.org

Accepted: 20/Sept/2018, Published: 30/Sept/2018

***Abstract*—** In the current world , various sources like sensors, social media, web logs, network monitoring devices, traffic monitoring devices are generating lots of data. This huge data is arriving continuously, with high speed and changing its nature with time. Extracting useful information from the data stream demands enhancement in existing technologies of Data Mining. Clustering is an important part of data mining in which similar data points are merge into one group. Use of genetic algorithm in clustering data stream is an emerging technology. In this paper, we are discussing clustering techniques for data stream using Genetic Algorithm.

***Keywords*—** Data Stream, Genetic Algorithms, Clustering

I. INTRODUCTION

Data Streams are data which are continuous, huge in volume, and changing in nature. Applying mining techniques on such data requires new techniques as existing techniques are used only for stored data. Data stream needs to be mined in an on-line manner with one scan of the data. Clustering is a useful technique of data mining. Genetic algorithm (GA) is a heuristic approach by which optimized solutions are produced. Evolution based Darwin's theory is a milestone in the field of data mining. GA work in five steps: initialize the population, evaluate the fitness function, Selection, Crossover and Mutation. Crossover and Mutation are genetic operator. In Crossover, the single or paired gene is exchanged in selected populations. The information generated by Crossover is altered in Mutation operator. Genetic Algorithms are used for optimizing clustering technique to apply on data streams. Figure 1 shows the procedure of Genetic Algorithm. The paper format is as in Section I contains the introduction of Genetic algorithms and data streams. Section II to IV contains the clustering techniques for data stream using various Genetic algorithms and Section V concludes research work of various researchers .

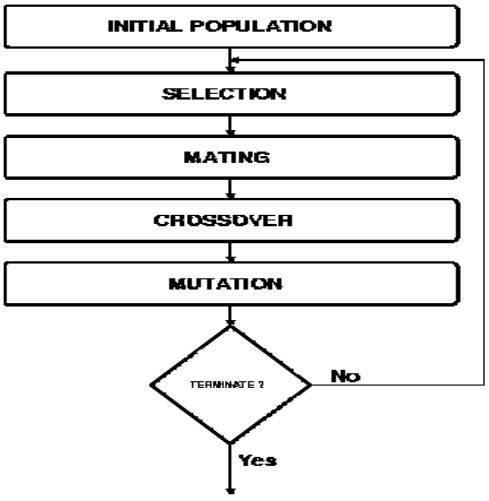

Figure 1: Procedure of Genetic Algorithm

II. CLUSTERING DATA STREAM USING GAUSSIAN MIXTURE MODEL GENETIC ALGORITHM (GMMGA) [3]

In clustering data stream, two major challenges are - to separate overlapping cluster and to identify a number of clusters. The solution is provided by extending Gaussian Mixture Model using Genetic Algorithm. Gaussian Mixture Model (GMM) works on the probability distribution of recently arrived data. In GMM, splitting is performed using the distance between two Gaussian component. This is measured by Euclidean distance. Merging of Gaussian component is performed by similarity measures between two Gaussian component. In GMMGA, the EM algorithm [4] is optimized using a Genetic Algorithm. Mainly the split and merge phases are modified. Let initially k Gaussian component be there. The samples are classified into Gaussian classes. Center of Gaussian component is selected randomly. These components are split into other classes using Euclidean distance. In each Gaussian class, N random time selection is made. This process is same as genetic operations. From the results the component with the largest fitness is chosen. Then the value of k is increased by one. In merging Euclidean norm vector is used. GMMGA's efficiency is measured by comparing with STREMA and Clustream. GMMGA is better to classify overlapping clusters.

III. CLUSTERING DATA STREAM BASED ON EVOLUTION INCENTIVE [5]

On Proposed approach, the algorithm is working in two parts. Initially high density partition based algorithm (DBSCAN) is used to generate clusters in an on-line manner. These clusters are called micro-clusters. On these micro clusters, the improved quantum genetic algorithm and evolution incentive function is used for optimizing the cluster's center. Next the adaptive mutation operator is used for optimal variation operation on population. The micro clusters are stored in real storage spaces. For each micro cluster, binary chromosomes are generated based on micro cluster radius. The fitness value of chromosomes is calculated. Particle swarm algorithm is used to optimize quantum rotation angle and offspring's are generated. Variation of individual is finding out by using a mutation operator. Based on the individual progress macro clusters are defined.

IV. OPTIMIZED K-MEANS CLUSTERING ALGORITHM ALONG WITH GENETIC ALGORITHM

K-means is a very popular algorithm for clustering stored data. K-means can be improved by using Genetic Algorithm to make it suitable for stream data. K-means algorithm requires initial centroids to proceed. In paper [6], a genetic based method is proposed to choose best initial centroids. Fitness of data points from the population k are calculated by Mean Square Error. After that Selection,Crossover and Mutation performed to generate new population. The output is used as an initial centroid for K-means. The experiments show that the proposed method is better than other existing techniques.

V. CONCLUSION AND FUTURE SCOPE

The Genetic algorithm is an emerging technique capable in optimization. Various clustering techniques are optimized using Genetic algorithm for stored data. Clustering data that are arriving continuously, with a massive speed and in a huge volume is not possible using traditional clustering techniques. Clustering technique can be improved using Genetic Algorithm. Our paper reviews some optimization of clustering technique using Genetic Algorithm to cluster data streams.

REFERENCES

[1] Gaber, Mohamed Medhat, Arkady Zaslavsky, and Shonali Krishnaswamy. "Mining data streams: a review." ACM Sigmod Record 34.2 (2005): 18-26.
[2] Mahdiraji, Alireza Rezaei. "Clustering data stream: A survey of algorithms." International Journal of Knowledge-based and Intelligent Engineering Systems 13.2 (2009): 39-44.
[3] Gao, Ming-ming, Chang Tai-hua, and Xiang-xiang Gao. "Application of Gaussian mixture model genetic algorithm in data stream clustering analysis."Intelligent Computing and Intelligent Systems (ICIS), 2010 IEEE International Conference on. Vol. 3. IEEE, 2010.
[4] Zhou, Aoying, et al. "Distributed data stream clustering: A fast EM-based approach." Data Engineering, 2007. ICDE 2007. IEEE 23rd International Conference on. IEEE, 2007.
[5] Heng, Liang. "Fast Clustering Optimization Method of Large-Scale Online Data Flow Based on Evolution Incentive."2014 Fifth International Conference on Intelligent Systems Design and Engineering Applications (ISDEA). IEEE, 2014.
[6] Alsayat, Ahmed, and Hoda El-Sayed. "Social media analysis using optimized K-Means clustering." Software Engineering Research, Management and Applications (SERA), 2016 IEEE 14th International Conference on. IEEE, 2016.

Authors Profile

Mrs. Neha Sharma received Bachelor of Engineering from RGPV University, Bhopal, in 2011 and a Master of Engineering from RGPV University in the year 2014. She is currently pursuing PhD. and currently working as Assistant Professor in Information Technology Department, SGSITS, Indore since 2014. She has published more than 06 research papers in reputed international journals . Her main research work focuses on Data Mining and Data Stream Mining. She has 5 years of teaching experience.

Mr. Pawan Makhija was born in Indore, India, in 1987. He received the B.E. degree in Information Technology from Rajiv gandhi proudyogiki vishwavidyalaya, Indore, India, in 2009, and the M.E. in Information Technology from the Institute of engineering and technology (IET) DAVV, Indore(M.P), India, in 2014, respectively. He is pursuing his PhD. In Computer engineering from Institute of engineering and technology (IET) DAVV, Indore (M.P). In 2009, he joined the Department of Information technology, Acropolis Institute of technology and research, Indore(M.P) as a Lecturer. Since July ,2013, he has been with the Department of Information Technology, Shri G. S. Institute of technology and science, Indore (M.P) where he is working as an Assistant Professor. He was a visiting faculty in Institute of engineering and technology (IET) DAVV, Indore (M.P). His current research interests include Data Mining, Stream Mining, Word Sense Disambiguation, Text Mining. He is also an associate faculty of the Indian Institute of technology , IIT Bombay. He was one of the top performer among 5000 teachers awarded with the SAP award of excellence by IIT Bombay under T10KT project .

