Particle Swarm Optimization with I-DBSCAN algorithm for Density-based Clustering

Neha, Prince Verma
Dept. of Computer Science Engineering, CT Institute of Engineering Management &Technology, Jalandhar, India

Abstract - Over the years, several data mining techniques have been designed. There are few traditional statistical methods applied to provide more reliable and scalable tools for making improvements. Since data mining is applied in many applications, improvements are being made with time. Clustering is known as the method using which the data is classified into groups based on its similarity patterns. Improving the performance of incremental DBSCAN is the objective of this research. The PSO algorithm is used in this research to improve the performance of incremental I-DBSCAN algorithm. Here, the Euclidian distance is calculated in dynamic manner and due to which the execution time of clustering is reduced thus, increasing its accuracy. Each point and their value will be taken by the PSO as input and at every clustering point the error will be calculated. The specific point at which accuracy of clustering is the highest is considered as the best point since at this point, the error is the least. The accurate point for clustering is defined by the efficient calculation of Euclidian distance. The similarity among data points for clustering is defined by distance.

Keywords: *I-DBSCAN, PSO, Density-based Clustering*

I. INTRODUCTION

Nowadays, the information is the perfect source to power and success in several applications. It is possible to collect the information from different sources [1]. The people use large amount of information generated from various sources for their own profits. The information is stored such that it can be used in future as per the needs [2]. In today's technology based world several devices like the huge digital storage devices and computers have been designed to store the required information [3]. There are several devices available for storing the variety of data being generated [4]. A structured database is designed for avoiding chaos. To achieve the objective, a Database Management System (DBMS) has been designed through which the data can be efficiently arranged [5]. By using DBMS, it is possible to efficiently retrieve the data as per the requirements of users [6]. It is possible to collect large amount of data due to the proliferation feature of DBMS. The data from varieties of fields can be handled here [7]. For efficient decision making it is not enough to use only the information retrieval method. To manage the data more efficiently several new methods have been proposed. For taking care of activities through which the data can be summarized, important data can be extracted and patterns can be discovered from raw data, these networks have been designed [8]. It is very important to analyze and interpret the largely stored databases and files. To provide important related information the data mining method is necessary such that effective decision making can be provided [9]. An improved end-user business model is applied for navigation of data warehouse through OLAP (On-Line Analytical Processing) server [10]. Based on the manner by which a user wants to see the business layout the depictions are designed with the help of multidimensional structures. This method helps in providing a proper product line and region. For embedding the ROL-focused business analysis with data mining, the data warehouse and OLAP are combined with data mining server [11]. Data mining is facing several issues. To advance the centric metadata template of applications, this research provides promotion optimization and prospecting [12]. It is possible to perform direct implementation and operation decisions tracking by combining the data warehouses. It is possible to mine and then apply the best services based on the organization's future decisions. It is possible to evolve the warehouse based on new decisions and results [13]. The method with which the data is categorized among similar object groups is known as clustering. During the involvement of less numbers of clusters, simplification level is achieved. Few finer details are lost however, due to the presence of less numbers of clusters [14]. Data modeling is performed using the clusters. An unsupervised mechanism is required for searching the hidden patterns of clusters as per the machine learning view. The data concept is defined by the system generated as output [15]. There are several steps that are performed in clustering. To perform data clustering, several new techniques have been designed apart from hierarchical and partition-based clustering algorithms. Based on the different data sets, the clustering techniques can vary [16]. Depending on the density objective functions, it is possible to group the objects by applying density-based clustering. The total number of objects existing in the neighbor of data object defines the density of a specific object [17]. The presence of spatial data is the major highlight of Grid-Based Clustering algorithms. There are geometric

structure of objects in the space, the properties and operations and relationships available in the spatial data. For solving the clustering related issues, one of the measures to be taken is the k-means clustering algorithm [18]. It is the easiest way to apply k-means clustering algorithm in comparison to the other unsupervised algorithms. For the classification of provided dataset with the help of certain number of clusters, a fixed apriori is provided. Introducing k-centers is the major objective for every cluster [19]. It is important to place these centers very carefully. The results are achieved based on the variations in location. The identified cluster size is increased until the density of neighbors is higher than certain threshold value. DBSCAN (Density Based Spatial Clustering of Applications with Noise) is an efficient clustering based algorithm that is based on density. This algorithm helps in separating the noise from large spatial databases by applying arbitrary shaped clusters.

II. LITERATURE REVIEW

Ahmad M. Bakr, et.al (2015) proposed a modified adaptation of incremental DBSCAN algorithm. The main objective of this approach was to construct and upgrade the shaped clusters available in big datasets. The modification should be done in an incremented manner to modify this procedure [20]. This algorithm gave more accurate results by applying this approach on big datasets having big dimensions. In the future, this approach will be modified further. This approach should perform better in parallel manner for making enhancements in the nearby future. The incremental DBSCAN algorithm was implemented in parallel manner within every partition. This technique can be applied more easily after identifying the individual partitions.

Iyer Aurobind Venkatkumar, et.al (2016) stated that massive volume of data was generated daily. This data should be secured from the access of outsiders. The included had different types of patterns and associations. The hiding of all such information could be helpful in different situations [21]. The accurateness of prediction was not required all the times. Moreover, there was no guarantee for their correctness. The datasets were classified among particular predefined sets by using the application of some statistical models. These models were called classifiers. The hidden correlations could be computed amid the massive volume of available data in association technique. The data clustering algorithms were analyzed on the basis of merits and demerits of the respective algorithms. In this work, four main clustering algorithms were implemented. These algorithms included k-means, BIRCH, DBSCAN and STING. These algorithms were compared to understand their different features.

Qi Xianting, et.al (2016) stated that the density-based clustering algorithm was one of the most popular algorithms. This algorithm was used to remove noise within the applications. In this work, a modified DBSCAN algorithm was proposed to give solution different arising concerns. This approach was identified as feature selection based DBSCAN algorithm [22]. This algorithm was provided on several real world datasets. Different simulations were implemented on this algorithm. This phenomenon provided help to test the performance of this recently proposed algorithm. The tested results demonstrated that the proposed algorithm gave better performance than existing approaches. The proposed algorithm could also deal with large data efficiently.

Kuan-Teng Liao, et.al (2016) proposed two algorithms. These algorithms were known as centroid based clustering algorithms and UKmeans algorithm. The similarity of an application could be improved using earlier approach [23]. In this work, a modified approach was proposed to manage both time and efficiency of the system equally. This easy similarity technique was used to minimize time. This approach provided two additional factors called intersection and density of clusters are brought up. The efficiency of the clustering approach can be improved by using these two factors. In this technique, the range could be minimized using square root boundary technique. This helped to limit the upper bound of the locations achievable for the centroids. This helped to increment the efficiency of clustering technique. The simulation results depicted that the proposed approach gave better outcomes than existing approaches.

Wenbin Wu et.al (2016) presented a new scheme for improving prediction accuracy. This approach also managed the dynamics of training sample. The k-means clustering algorithm was implemented with neural network. This algorithm was applied in those situations that included small terms of WPF [24]. Amid different techniques using k-means clustering algorithms, several categories were made as per the similarities. This approach involved the information relevant to meteorological conditions and the other existing available data. The simulation results of this proposed approach were quire effectual in comparison with the baseline and previous short-term WPF techniques. In future, novel researches will be performed for the efficient meteorological prediction. These researches will help to improve the accurateness of prediction.

Vadlana Baby, et.al (2016) presented an effectual distributed threshold privacy-preserving k-means clustering algorithm. This approach employed code based threshold secret sharing as a privacy-preserving technique. Involving a code based scheme allowed the partition of data into different parts. These parts were further processed at several servers [25]. The proposed approach had fewer amounts of iterations than the existing approaches. No trust was needed from the end of servers. In this work, different methodologies were compared.

In this work, the security analysis of this proposed approach was too provided. The k-means clustering algorithm was used along with code based threshold secret sharing approach in this work. The approaches were used to preserve privacy. The clustering process was performed mutually. In this process, the reliability of third party was avoided. The proposed approach offered ideal conservation of the client data.

III. RESEARCH METHODOLOGY

The data density is used to create clusters in density based clustering approach. The two values called EPS and Euclidian distance are used by the I-DBSCAN algorithm. The cluster radius is defined by the EPS value. The radius of the data is defined by EPS value in the earlier study in static manner. In this study, the PSO algorithm is used. This algorithm measures EPS value in dynamic manner. The objective function is defined dynamically in the PSO algorithm. The present iteration and earlier iterations are compared on the basis of swarm value. The objective function is identified using the swarm value having maximum iteration. The following expression describes the dynamic objective function. Execution of every iteration changes the value.

$$v_{i+1} = v_i + c * rand * (p_{best} - x_i) + c * rand * (g_{best} - x_i)$$

In the above equation, V_i represents the element velocity. The variable p_{best} represents the optimum value among accessible options. The variable "rand.x" represents random number. This is the value given to every feature of the website. The "c" variable defines this value. This procedure selects the optimum value recognized from overall population and demonstrates it as p_{best}. The best value selected after each iteration is represented by "g_{best}". The obtained value is added with the traverse value of every attribute for concluding the objective function. This phenomenon is given as:

$$x_{i+1} = x_i + v_{i+1}$$

The "$x_{(i+1)}$" denotes position vector. These multi-objective optimization issues are solved by using dynamic PSO algorithms regarding the best computed value. The PSO algorithm gives the data utilized for encryption as input. The key utilized for encoding provides support to generate enhanced value.

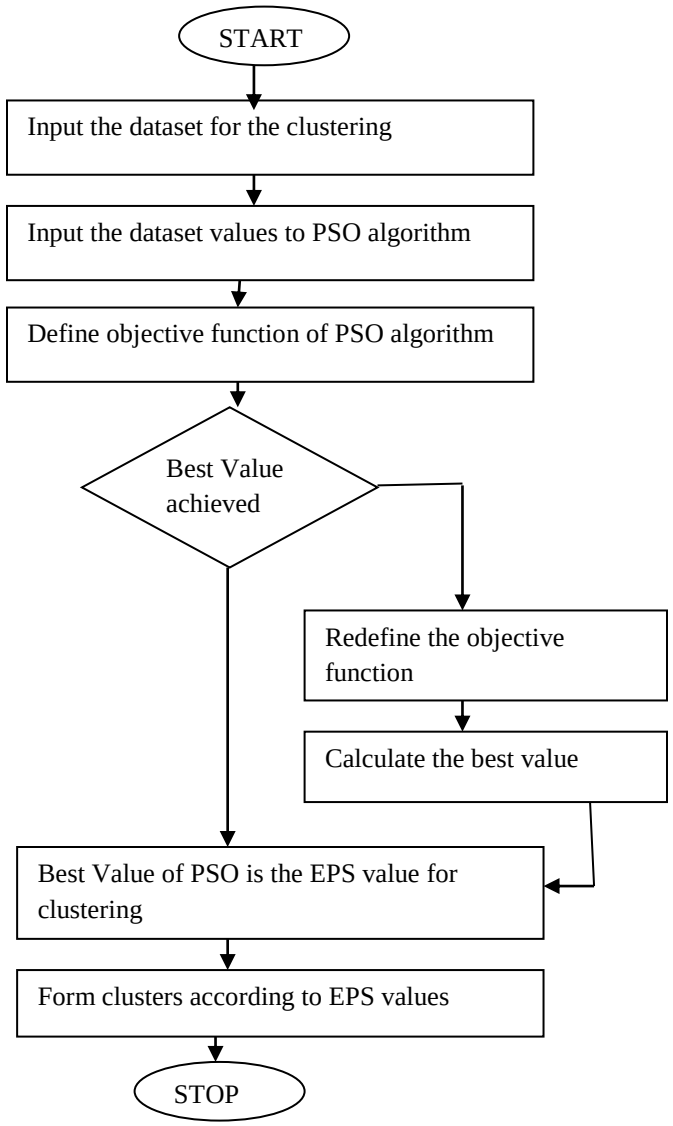

Figure 1: Proposed Flowchart

IV. EXPERIMENTAL RESULTS

The MATLAB is the tool which is used for the implementation of I-DBSCAN algorithm and proposed I-DBSCAN algorithm. The dataset is collected from different internet sources. The performance of the I-DBSCAN algorithm and proposed I-DBSCAN algorithm is analyzed in terms of accuracy and execution time on different sizes of dataset like 2000 instances,4000 instances, 6000 instances and 8000 instances

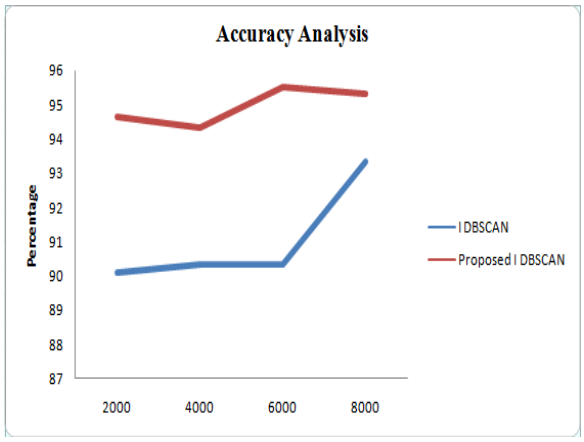

Fig 2: Accuracy Analysis

As shown in figure 2, the accuracy of the I-DBSCAN algorithm proposed I-DBSCAN is compared for the performance analysis. The performance of the algorithm is compared on different set of data values. It is analyzed proposed DBSCAN algorithm has high accuracy as compared to I-DBSCAN algorithm

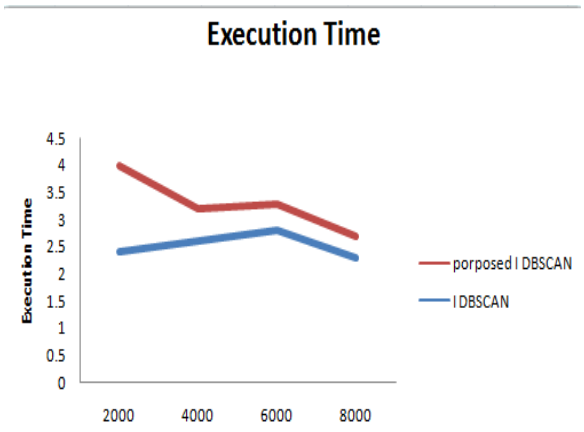

Fig 3: Execution Time Analysis

As shown in figure 3, the execution time of the I-DBSCAN algorithm proposed I-DBSCAN is compared for the performance analysis. The performance of the algorithm is compared on different set of data values. It is analyzed proposed DBSCAN algorithm has low execution time as compared to I-DBSCAN algorithm

V. CONCLUSION

The technique used to cluster similar and dissimilar type of data for analyzing complex data is called clustering or cluster analysis. In this work, the density based clustering algorithm is implemented to cluster similar and dissimilar data on the basis of density of data in the input dataset. The density based clustering algorithm calculates most dense area. The similarity method is used to compute similar and dissimilar data from this region. The EPS value is computed by implementing DBSCAN algorithm. The EPS value will be the center of dataset. In order to obtain maximal accuracy, the EPS value is computed in dynamic manner. The similarity amid the data points is computed by applying Euclidian distance method. In future, PSO algorithm will be implemented for increasing clustering accuracy. This algorithm will compute Euclidian distance dynamically.

VI. REFERENCES

[1] Zhe Zhang, Junxi Zhang, Huifeng Xue," Improved K-means clustering algorithm," 2008, Congress on Image and Signal Processing CISP, vol. 5, May pp. 169–172
[2] L. Li, J. You, G. Han, H. Chen, Double partition around medoids based cluster ensemble, 2012, International Conference on Machine Learning and Cybernetics, vol. 4, pp. 1390– 1394
[3] H. Du, Y. Li," An improved BIRCH clustering algorithm and application in thermal power," 2010, International Conference on Web Information Systems and Mining (WISM), vol. 1, Oct pp. 53–56
[4] R.T. Ng, J. Han," CLARANS: a method for clustering objects for spatial data mining," 2002, IEEE Trans. Knowl. Data Eng. 14 (5) 1003–1016
[5] S. Guha, R. Rastogi, K. Shim," CURE: an efficient clustering algorithm for large databases," 1998, Proceedings of the ACMSIGMOD International Conference Management of Data (SIGMOD'98), pp. 73–84
[6] G. Karypis, H. Eui-Hong, V. Kuma," Chameleon: hierarchical clustering using dynamic modeling", 1999, Computer 32 (8) 68–75
[7] M. Ester, H. Kriegel, J. Sander, X. Xu," A density-based algorithm for discovering clusters in large spatial databases with noise," 1996, Proc. 2nd International Conference on Knowledge Discovery and Data Mining, pp. 226–231
[8] Carla Agurto, Victor Murray, Eduardo Barriga, Sergio Murillo, Marios Pattichis, Herbert Davis, Stephen Russell, Michael Abramoff, Peter Soliz, "Multiscale AM-FM Methods for Diabetic Retinopathy Lesion Detection", IEEE Transactions on Medical Imaging, 2010, Volume: 29, Issue: 2, Pages: 502 - 512
[7] Lei Zhang, Qin Li, Jane You, David Zhang, "A Modified Matched Filter With Double-Sided Thresholding for Screening Proliferative Diabetic Retinopathy", IEEE Transactions on Information Technology in Biomedicine, 2009, Volume: 13, Issue: 4, Pages: 528 – 534
[8] Sameti M., Ward R.K., Parkes J.M., Palcic B., "Image Feature Extraction in the Last Screening Mammograms Prior to Detection of Breast Cancer", IEEE journal of selected topics in signal processing, VOL. 3, NO. 1, February 2009
[9] Gandhi T., Trivedi M.M., "Pedestrian Protection Systems: Issues, Survey, and Challenges", IEEE transactions on intelligent transportation systems, VOL. 8, NO. 3, September 2007
[10] V. Kumar, T. Lal, P. Dhuliya, and Diwaker Pant, "A study and comparison of different image segmentation algorithms", In Advances in Computing, Communication, & Automation (ICACCA)(Fall), International Conference on, IEEE 2016, pp. 1-6

[11] R. Radha, and S. Jeyalakshmi, "An effective algorithm for edges and veins detection in leaf images", In Computing and Communication Technologies (WCCCT), 2014 World Congress on, IEEE 2014, pp. 128-131
[12] Sohini Roychowdhury, Dara D. Koozekanani, Keshab K. Parhi, "DREAM: Diabetic Retinopathy Analysis Using Machine Learning", IEEE Journal of Biomedical and Health Informatics, 2014, Volume: 18, Issue: 5, Pages: 1717 - 1728
[13] Gary G. Yen, Wen-Fung Leong, "A Sorting System for Hierarchical Grading of Diabetic Fundus Images: A Preliminary Study", IEEE Transactions on Information Technology in Biomedicine, 2008, Volume: 12, Issue: 1, Pages: 118 - 130
[14] Lama Seoud, Thomas Hurtut, Jihed Chelbi, Farida Cheriet, J. M. Pierre Langlois, "Red Lesion Detection Using Dynamic Shape Features for Diabetic Retinopathy Screening", IEEE Transactions on Medical Imaging, 2016, Volume: 35, Issue: 4, Pages: 1116 – 1126
[15] M.M. Fraza, S.A. Barmana, P. Remagninoa, A. Hoppea, A. Basitb, B. Uyyanonvarac, A.R. Rudnickad, C.G. Owend, "An Approach To Localize The Retinal Blood Vessels Using BitPlanes And Centerline Detection", Comput. Methods Programs Biomed, 2011
[16] Lassada Sukkaew, Bunyarit Uyyanonvara, Sarah Barman, "Automatic Extraction of the Structure of the Retinal Blood Vessel Network of Premature Infants", J Med Assoc Thai Vol. 90 No. 9 2007
[17] A. Devbrat, and J. Jha. "A Review on Content Based Image Retrieval Using Feature Extraction" International Journal of Advanced Research in Computer Science and Software Engineering Volume3, March 2016.
[18] Harini R, Sheela N Rao, "Feature Extraction and Classification of Retinal Images for Automated Detection of Diabetic Retinopathy", 2016 Second International Conference on Cognitive Computing and Information Processing (CCIP)
[19] Sharad Kumar Yadav, Shailesh Kumar, Basant Kumar, Rajiv Gupta, "Comparative Analysis of Fundus Image Enhancement in Detection of Diabetic Retinopathy", 2016 IEEE Region 10 Humanitarian Technology Conference (R10-HTC)
[20] Ahmad M. Bakr, Nagia M. Ghanem, Mohamed A. Ismail," Efficient incremental density-based algorithm for clustering large datasets", 2015, Elsevier B.V.
[21] Iyer Aurobind Venkatkumar, Sanatkumar Jayantibhai, Kondhol Shardaben," Comparative study of Data Mining Clustering algorithms", 2016, IEEE
[22] Qi Xianting, Wang Pan," A density-based clustering algorithm for high-dimensional data with feature selection", 2016, IEEE
[23] Kuan-Teng Liao, Chuan-Ming Liu," An Effective Clustering Mechanism for Uncertain Data Mining Using Centroid Boundary in UKmeans", 2016, IEEE
[24] Wenbin Wu and Mugen Peng," A Data Mining Approach Combining K-Means Clustering with Bagging Neural Network for Short-term Wind Power Forecasting", 2016, IEEE
[25] Vadlana Baby, Dr. N. Subhash Chandra," Distributed threshold k-means clustering for privacy preserving data mining", 2016, IEEE
[26] Neha , Prince Verma, I-DBSCAN Algorithm with PSO for Density Based Clustering ", 2019,
E-ISSN: 2347-2693

Miss. Neha pursed Bachelor of technology in informatiom technology from DAVIET & jalandhar, punjab, India in 2017. She is currently pursuing Master of Technology in computer Science Engineering from CTIEMT, Jalandhar,Punjab, India. Her email contact as Neha47025@gmaill.com. Her main research work focuses on data mining, big data analytics, Machine learning algorithms, and Particle Swarm Optimization with I-DBSCAN algorithm for Density-based Clustering.

Mr. Prince verma pursed B.tech degree in Computer science and Engineering From MIMIT, Malout, Punjab, India in 2008 and M.tech in computer science and engineering in 2013 from DAVIET, Jalandhar , Punjab, India. He is currently purusing his PHDin the areas of big data analytics from IKG Punjab technical University, Kapurthala, Punjab, India . He is currently the head and assistant Professor in the department of computer science and engineering at CTIEMT, Lalandhar, Punjab, India . His research interest lies in data mining algorthim optimization techniques, big data Analytics. He has more than 35 research publicatons in reputed international journals.