An Efficient K-Means Method Based on Centroid Handling for the Similarity Estimation

Girdhar Gopal Ladha, Ravi Kumar Singh Pippal

***Abstract:** The main aim of this paper is to handle centroid calculation in k-means efficiently. So that the distance estimation will be more accurate and prominent results will be fetched in terms of clustering. For this PIMA database has been considered. Data preprocessing has been performed for the unwanted data removal in terms of missing values. Then centroid initialization has been performed based on centroid tuning and randomization. For distance estimation Euclidean, Pearson Coefficient, Chebyshev and Canberra algorithms has been used. In this paper the evaluation has been performed based on the computational time analysis. The time calculation has been performed on different random sets. It is found to be prominent in all the cases considering the variations in all aspects of distance and population.*

***Keywords:** K-means, Centroid Handling, Distance measures, Similarity estimation.*

I. INTRODUCTION

The grouping the clustering algorithms mostly relies on the cases where similarity index can be generated based on the index [1]. It shows the iterative and approximation way or the unsupervised mechanism of finding elements groups. It can be labeled differently as per the requirement [2].

The impact of clustering algorithms is mostly depending on the accurate distance estimation [3]. It depends on the centroid estimation. The major precaution major should be done in terms of centroid initialization. In current era clustering algorithms have been used widely in different areas. The areas of applicability are engineering, e-commerce; health etc. K-means, fuzzy c-means and hierarchical clustering are mostly used in different areas [3, 4]. These algorithms efficiency is depends on the way the data cluster based on the inter centroid calculation [4].

In the present condition in normal day by day the data is growing. So, data pruning, estimation along with the data extraction is very important aspects in pattern finding and matching [5]. It is a part of data mining algorithms and it is helpful in efficient pattern identification. The extraction and arrangement process are the part of data mining. The knowledge discovery is the important in terms of data acquisition and data exploration in terms of information discovery [6-9]. The design and information of the structural goal is to be capable in finding the latest trends and technological aspects in terms of different aspects of experimentation and analysis of all the empirical and calculative way. It should be commenced and explored in terms of data grouping, knowledge representation and classification [10-12]. Standard data mining techniques have been used widely in the distance estimation tasks [13]. The concrete structure of any data mining exploration relies on the appropriate data representation in terms of the data scenic approach for the collaborative approach of the data centric approach in terms of data exploration and data recognition [14-16]. It also explores the mechanism for the adaptation of the correct methodology in terms of data mining methodology in terms of suitability and use. It should in such manner that it can be calculative and representatively fit to the acquiring way [17, 18]. It is also impactful in terms that it can be used in the same manner for data productivity and allocation variability for the resource generation and allocation. The major challenge is the selection of the appropriate method in the applicability and processing. The main objective of this paper is to estimate the similarity index along with the time estimation.

II. LITERATURE SURVEY

In 2019, Reddy et al. [19] discussed about the huge data handling and cost-effective mechanism for handling the huge data. They have devised a solution in handling data belongs to multiple clusters. The data belongs to health care big data. The algorithm used is fuzzy c-means algorithm. They have suggested the working mechanism of fuzzy c-means based on midpoint for its data point. According to the authors this type of fuzzy implementation may help in reducing data loss.

In 2019, Vanitha et al. [20] discussed about the modern technology and their prospect helpful in agriculture field for the farmers. They have suggested that the python can be used as a front end for analyzing the agricultural data set. They have suggested Jupyter notebook for the crop production prediction as the data mining tool. They have considered the precipitation, temperature, reference crop, evapotranspiration, area, production and yield as the parameters. The year considered here are from 2000 to 2018. They have used k-means, KNN, SVM, and Bayesian network algorithms. In 2019, Dai and Sheng [21] discussed about the evolutionary algorithms. They have also discussed the role of these algorithms in clustering. They have suggested that advance requirement of number of clustering in normal clustering algorithms.

Revised Manuscript Received on December 30, 2019.

* Correspondence Author
Girdhar Gopal Ladha*, Ph.D. Scholar, Department of Computer Science, RKDF University, Bhopal (MP), India.
Ravi Kumar Singh Pippal, Professor, Department of Computer Science RKDF University, Bhopal (MP), India.

They have also difficulty in the diversity preserving according to the authors. For solving these types of issues, they have proposed a multi-objective algorithm. It is a clustering ensemble algorithm. The name of the algorithm is modified version of dual-similarity clustering ensemble (MDSCE). The drawback suggested has been removed in this approach. Then they have proposed a k-means based process. For testing they have considered the real datasets. Their results show high-quality clustering solutions.

In 2019, Qin et al. [22] discussed the weighted k-nearest neighbor (WKNN). They have suggested that this algorithm is used for indoor positioning frequently. They have suggested the main drawback is the selection of k points dependency. It is mainly based on the received signal strength (RSS). They have suggested new algorithm for the drawback removal considering RSS physical distance and achieve better performance in different positioning experimentation.

In 2019, Cui and Gao [23] discussed about the cluster analysis technique. They have suggested that the mixed attribute data is the mostly used for clustering in different area. They have suggested that the conversion of the class attributes into numerical attributes is tough. They have suggested the need of advance clustering algorithm which is capable in processing mixed attribute data at the large scale.

In 2019, Chen and Lu [24] proposed a minimum spanning tree (MST) clustering algorithm. It is based on density. The background considered is based on the no effective support on the dataset with noises. Their approach employed split and merge stages. For the split stage density estimation method have been designed. For the merge stage connected sub graph is employed. They have performed their experimentation on the synthetic and real datasets. Their results show that their approach has the capability of detecting clusters with noises.

In 2019, Wu et al. [25] discussed data clustering as the Np-hard problem. They have suggested that fuzzy c-means (FCM) algorithm is one of the important fuzzy clustering approaches. They have suggested that FCM can trap in local optima easily and this is the major drawback. So, they have suggested the combination of genetic algorithm (GA) and particle swarm optimization (PSO) for solving this problem with FCM. They have also suggesting the importance of whale optimization algorithm (WOA). They have proposed a fuzzy clustering algorithm based on memetic fuzzy whale optimization (MFWO) algorithm. Their results support their approach.

In 2019, Hanyang et al. [26] discussed about the automatic identification system (AIS). K-means based and analytical approved space-based AIS(S-AIS) data. They have used elbow rule for the optimal number of clusters determination. Then they calculated the normalized standard deviation of course over ground (COG) and speed over ground (SOG) of vessels in South Africa area.

III. PROPOSED WORK

An efficient k-means method based on centroid handling for the similarity estimation has been proposed in this paper. It is categorized in the following parts:

1. Dataset selection
2. Preprocessing
3. Centroid estimation and initialization
4. Mapping and distance acquisition
5. Time computation

Pima Indians diabetes database have been considered for the data analysis and experimentation. In the first phase preprocessing and data arrangement have been performed. The data is arranged according to the algometric scale for the data clustering and data arrangement to find the refined clusters. The data is arranged according to the content or the patient attributes values. The arrangement in such a way that the data normalization has been performed to utilize the data in a meaningful and computational process. For clustering purpose and centroid calculation, k-means algorithm has been used. The centroid estimation and calculation have been done in the bi-directional iterative process which can efficiently handle the data in the same way in terms to generate unbiased centroid in each cycle with different calculative way of finding it. K-means is applied on the data preprocessed value in our case. The data preprocessed values are the values obtained by the dataset. Euclidean (E), Pearson Coefficient (P), Chebyshev (Ch) and Canberra (Ca) are the algorithms which are used here as these are common and very famous. Based on these algorithms the calculation process is computed.

Figure 1 show the similarity mapping based on unbiased centroid initialization. It is clear from figure 1 that in the first phase the data has been selected and preprocessing has been performed. Then based on the generated unbiased centroid similarity mapping has been performed. Then distance calculation has been performed and then time computation has been done.

IV. RESULT AND DISCUSSION

Figure 2-Figure 5 shows the computational analysis based on time from the set 1, set 2, set 3, set 4 and set 5 from the complete dataset. Figure 2 shows the time analysis based on the computational parameters and similarity mapping for set 1. Figure 3 shows the time analysis based on the computational parameters and similarity mapping for set 2. Figure 4 shows the time analysis based on the computational parameters and similarity mapping for set 3. Figure 5 shows the time analysis based on the computational parameters and similarity mapping for set 4. It clearly shows that the time computation in case of Pearson is more. In case of Chebyshev and Canberra the performance is good. Overall Euclidean is better.

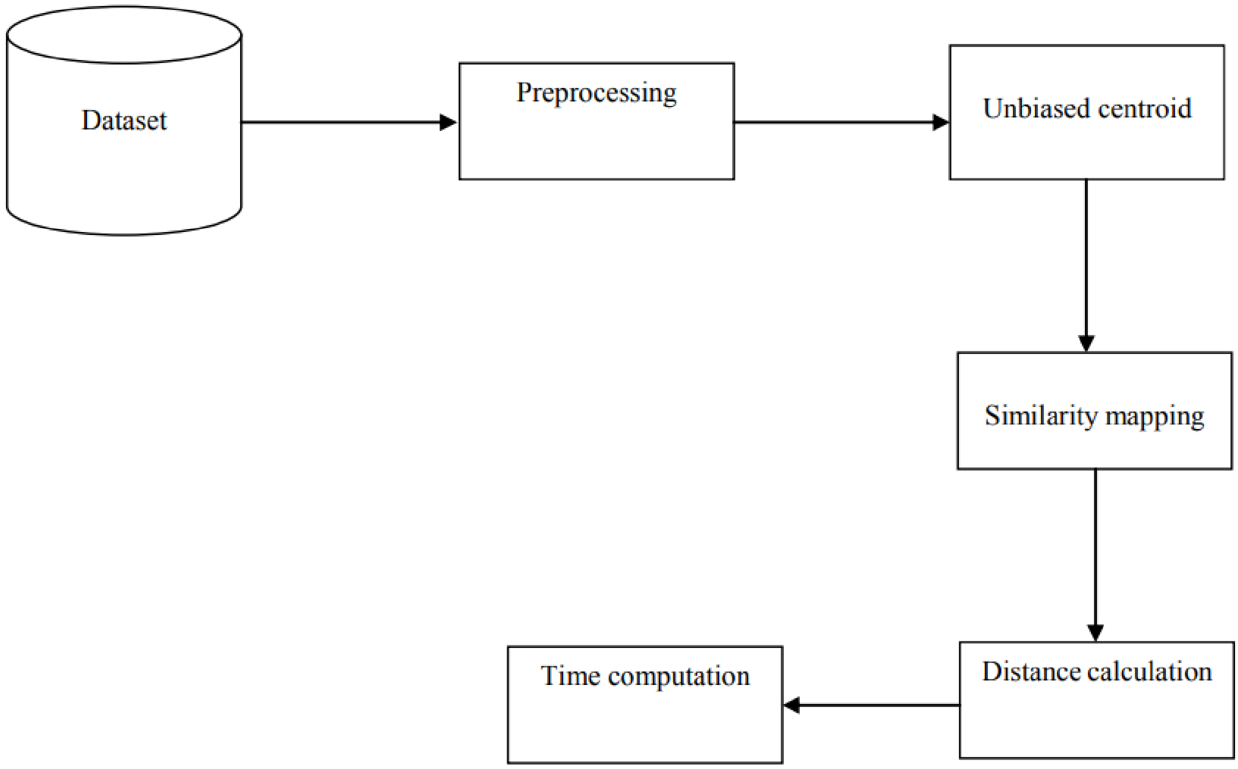

Figure 1 Similarity mapping based on unbiased centroid initialization

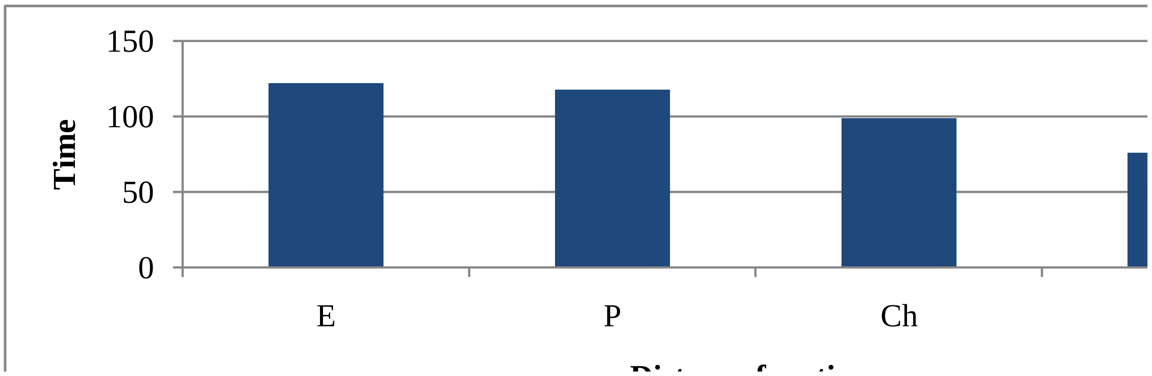

Figure 2: Time analysis based on the computational parameters and similarity mapping for set 1

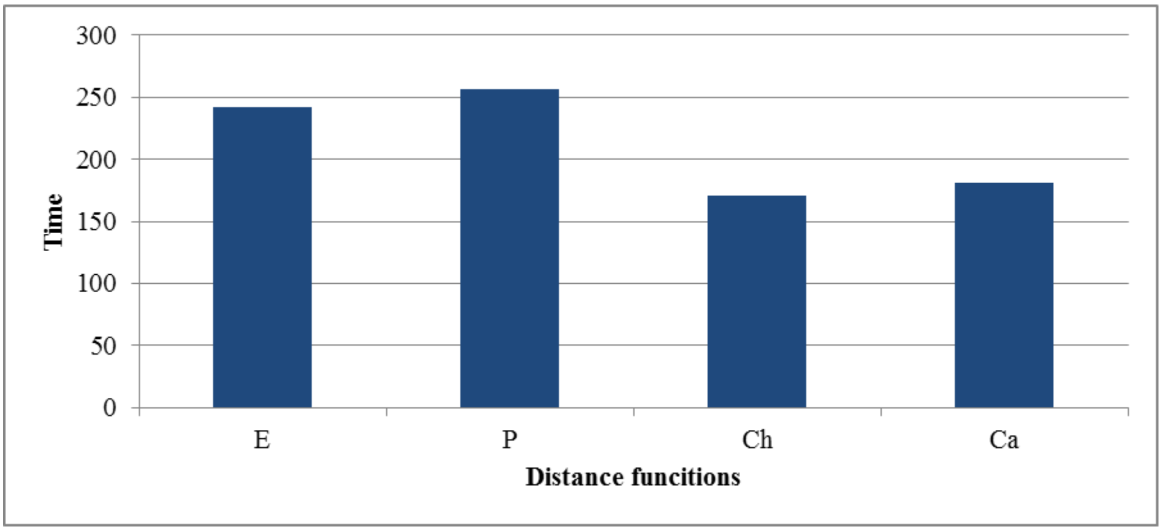

Figure 3: Time analysis based on the computational parameters and similarity mapping for set 2

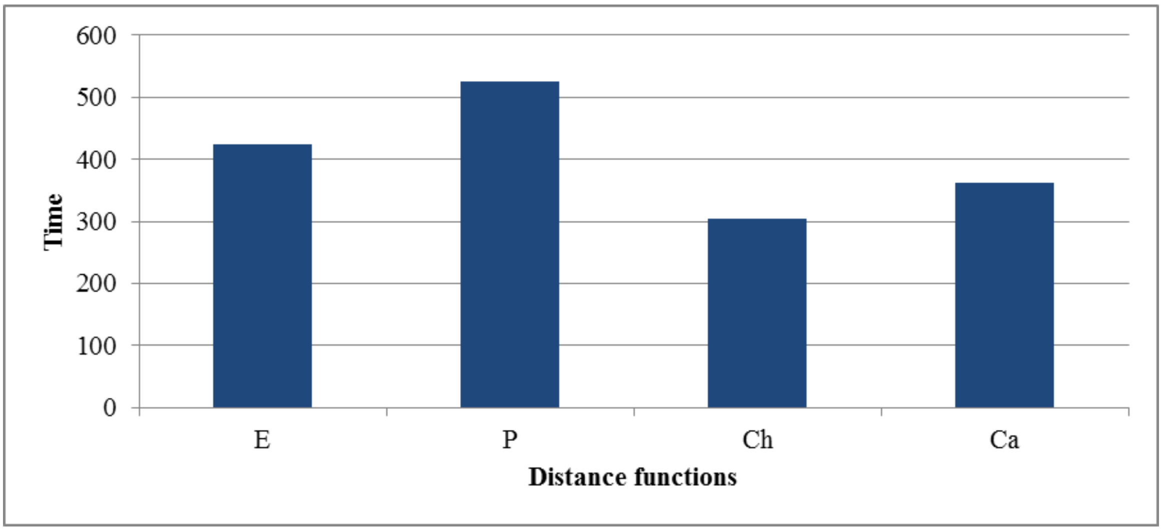

Figure 4: Time analysis based on the computational parameters and similarity mapping for set 3

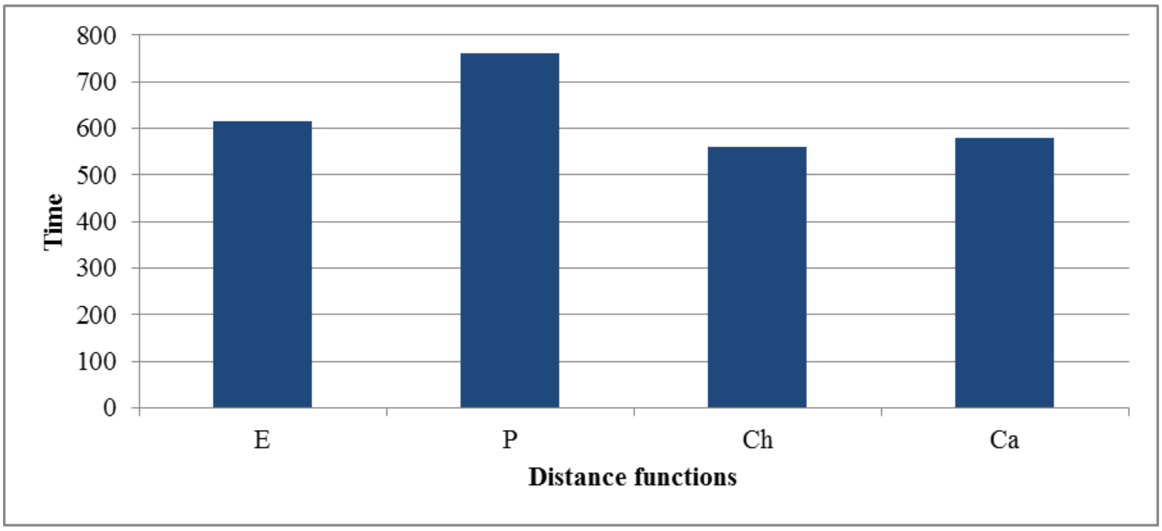

Figure 5: Time analysis based on the computational parameters and similarity mapping for set 4

V. CONCLUSION

This paper explores the centroid handling mechanism for the purpose of better similarity mapping in terms of distance estimation. This paper provides an empirical way with different parameters considering unbiased centroid initialization by the help of k-means clustering algorithm for the support of computation variability. The algorithms used for similarity ranking and distance measurements are E, P, Ch and Ca. These distance algorithms have been used for the centroid calculation. The comparative analysis shows the comparative analysis for different distance algorithms.

REFERENCES

1. Guttikonda G, Katamaneni M, Pandala M. Diabetes Data Prediction Using Spark and Analysis in Hue Over Big Data. In2019 3rd International Conference on Computing Methodologies and Communication (ICCMC) 2019 Mar 27 (pp. 1112-1117). IEEE.
2. Tavse P, Khandelwal A. An Efficient K-means Clustering approach in Wireless Network for data sharing. International Journal of Advanced Technology and Engineering Exploration. 2015;2(2):9.
3. Dubey AK, Gupta U, Jain S. Analysis of k-means clustering approach on the breast cancer Wisconsin dataset. International journal of computer assisted radiology and surgery. 2016; 11(11):2033-47.
4. Pan Q, Xiang L, Jin Y. Rare Association Rules Mining of Diabetic Complications Based on Improved Rarity Algorithm. In2019 IEEE 7th International Conference on Bioinformatics and Computational Biology (ICBCB) 2019 Mar 21 (pp. 115-119). IEEE.
5. Cios KJ, Moore GW. Uniqueness of medical data mining. Artificial intelligence in medicine. 2002 Sep 1;26(1-2):1-24.
6. Kavakiotis I, Tsave O, Salifoglou A, Maglaveras N, Vlahavas I, Chouvarda I. Machine learning and data mining methods in diabetes research. Computational and structural biotechnology journal. 2017 Jan 1; 15:104-16.
7. Aljumah AA, Ahamad MG, Siddiqui MK. Application of data mining: Diabetes health care in young and old patients. Journal of King Saud University-Computer and Information Sciences. 2013 Jul 1;25(2):127-36.
8. Mishra A, Mohapatro M. An IoT framework for Bio-medical sensor data acquisition and machine learning for early detection. International Journal of Advanced Technology and Engineering Exploration. 2019; 6 (54): 112-125.
9. Yousefi L, Swift S, Arzoky M, Sacchi L, Chiovato L, Tucker A. Opening the Black Box: Exploring Temporal Pattern of Type 2 Diabetes Complications in Patient Clustering Using Association Rules and Hidden Variable Discovery. In2019 IEEE 32nd International Symposium on Computer-Based Medical Systems (CBMS) 2019 Jun 5 (pp. 198-203). IEEE.
10. Pebesma J, Martinez-Millana A, Sacchi L, Fernandez-Llatas C, De Cata P, Chiovato L, Bellazzi R, Traver V. Clustering Cardiovascular Risk Trajectories of Patients with Type 2 Diabetes Using Process Mining. In2019 41st Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC) 2019 Jul 23 (pp. 341-344). IEEE.

11. Iyer A, Jeyalatha S, Sumbaly R. Diagnosis of diabetes using classification mining techniques. arXiv preprint arXiv:1502.03774. 2015 Feb 12.
12. Hao J, Zheng Y, Xu C, Yan Z, Li H. Feature Assessment and Classification of Diabetes Employing Concept Lattice. In2019 IEEE 23rd International Conference on Computer Supported Cooperative Work in Design (CSCWD) 2019 May 6 (pp. 333-338). IEEE.
13. Yaacob H, Omar H, Handayani D, Hassan R. Emotional profiling through supervised machine learning of interrupted EEG interpolation. International Journal of Advanced Computer Research. 2019 Jul 1;9(43):242-51.
14. Syafitri N, Labellapansa A, Kadir EA, Saian R, Zahari NN, Anwar NH, Shaharuddin NE. Early detection of fire hazard using fuzzy logic approach. International Journal of Advanced Computer Research. 2019 Jul 1;9(43):252-9.
15. Abood LH, Karam EH, Issa AH. Design of adaptive neuro sliding mode controller for anesthesia drug delivery based on biogeography based optimization. International Journal of Advanced Computer Research. 2019 May 1;9(42):146-55.
16. Karthikeyan R, Geetha P, Ramaraj E. Rule Based System for Better Prediction of Diabetes. In2019 3rd International Conference on Computing and Communications Technologies (ICCCT) 2019 Feb 21 (pp. 195-203). IEEE.
17. Wu H, Yang S, Huang Z, He J, Wang X. Type 2 diabetes mellitus prediction model based on data mining. Informatics in Medicine Unlocked. 2018 Jan 1; 10:100-7.
18. Dubey AK. An Efficient Variable Distance Measure K-Means [VDMKM] Algorithm for Cluster Head Selection in WSN. International Journal of Innovative Technology and Exploring Engineering. 2019; 9(1): 87-92.
19. Reddy BR, Kumar YV, Prabhakar M. Clustering large amounts of healthcare datasets using fuzzy c-means algorithm. In2019 5th International Conference on Advanced Computing & Communication Systems (ICACCS) 2019 Mar 15 (pp. 93-97). IEEE.
20. Vanitha CN, Archana N, Sowmiya R. Agriculture Analysis Using Data Mining and Machine Learning Techniques. In2019 5th International Conference on Advanced Computing & Communication Systems (ICACCS) 2019 Mar 15 (pp. 984-990). IEEE.
21. Dai H, Sheng W. A Multi-objective Clustering Ensemble Algorithm with Automatic k-Determination. In2019 IEEE 4th International Conference on Cloud Computing and Big Data Analysis (ICCCBDA) 2019 Apr 12 (pp. 333-337). IEEE.
22. Qin H, Shi S, Tong X. A new weighted indoor positioning algorithm based on the physical distance and clustering. In2019 15th International Wireless Communications & Mobile Computing Conference (IWCMC) 2019 Jun 24 (pp. 237-242). IEEE.
23. Cui G, Gao H. Rough Set Processing Outliers in Cluster Analysis. In2019 IEEE 4th International Conference on Cloud Computing and Big Data Analysis (ICCCBDA) 2019 Apr 12 (pp. 111-115). IEEE.
24. Chen J, Lu J. A Clustering Algorithm Based on Minimum Spanning Tree and Density. In2019 IEEE 4th International Conference on Big Data Analytics (ICBDA) 2019 Mar 15 (pp. 1-4). IEEE.
25. Wu ZX, Huang KW, Chen JL, Yang CS. A Memetic Fuzzy Whale Optimization Algorithm for Data Clustering. In2019 IEEE Congress on Evolutionary Computation (CEC) 2019 Jun 10 (pp. 1446-1452). IEEE.
26. Hanyang Z, Xin S, Zhenguo Y. Vessel Sailing Patterns Analysis from S-AIS Data Dased on K-means Clustering Algorithm. In2019 IEEE 4th International Conference on Big Data Analytics (ICBDA) 2019 Mar 15 (pp. 10-13). IEEE.

AUTHORS PROFILE

Girdhar Gopal Ladha has completed MCA in 1999 from MACT Bhopal. I have completed my M.TECH (IT) from BUIT Bhopal. Currently I am pursuing my PhD in Computer Science.

Dr. Ravi Kumar Singh Pippal is presently working with R.K.D.F. University, Bhopal as Associate Professor. He has received Ph.D. from ABV-Indian Institute of Information Technology and Management, Gwalior, M.Tech. from S.A.T.I., Vidisha and B.E. from J.E.C., Jabalpur. His research work has been widely recognized and appreciated by various international bodies. He has 27 research publications in international journals and conferences. He is a reviewer for international journals of great repute. Currently, he is guiding one Ph.D. scholar. His research areas of interest include Cryptology (Cryptography, Cryptanalysis), Steganography, Key Management, Image Encryption and many more.