International Journal of Advanced Technology and Engineering Exploration, Vol 7(73)
ISSN (Print): 2394-5443 ISSN (Online): 2394-7454
http://dx.doi.org/10.19101/IJATEE.2020.762109

Research Article

An efficient distance estimation and centroid selection based on k-means clustering for small and large dataset

Girdhar Gopal Ladha[1*] and Ravi Kumar Singh Pippal[2]
Ph.D. Scholar, Department of Computer Science, RKDF University, Bhopal (MP), India[1]
Professor, Department of Computer Science RKDF University, Bhopal (MP), India[2]

Received: 10-October-2020; Revised: 25-December-2020; Accepted: 28-December-2020

Abstract
In this paper an efficient distance estimation and centroid selection based on k-means clustering for small and large dataset. Data pre-processing was performed first on the dataset. For the complete study and analysis PIMA Indian diabetes dataset was considered. After pre-processing distance and centroid estimation was performed. It includes initial selection based on randomization and then centroids updations were performed till the iterations or epochs determined. Distance measures used here are Euclidean distance (Ed), Pearson Coefficient distance (PCd), Chebyshev distance (Csd) and Canberra distance (Cad). The results indicate that all the distance algorithms performed approximately well in case of clustering but in terms of time Cad outperforms in comparison to other algorithms.

Keywords
K-means, Distance estimation, Centroid selection, Distance methods.

1.Introduction
In terms of clustering algorithms, the major task is the centroid selection, distance estimation and appropriate grouping finalization [1─4]. Clustering provides an unsupervised way of clustering. It groups the cluster based on the similarity mapping and the approximation based on different estimation [1, 2]. There are different clustering algorithms in which k-means and fuzzy-c means algorithms are widely used. Clustering algorithms can be applied in the following areas [5─10]:

- Healthcare
- Business analytics and E-commerce
- Decision support system
- Industry data exploration and grouping

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].

In this paper k-means clustering algorithm was used. If we specifically discussed the applicability of k-means, then it can be widely used in the development of disease diagnosis system, search engine, performance grouping, etc. [8–10]. In case of k-means algorithm there is not any labelled data [11, 12]. It is capable in the division [13–16]. This division is in the form of objects. Based on this division different segments were created. This segment provides the groups similar and non-similar objects based on the division [17, 18].

In this paper k-means clustering algorithm was properly utilized in the following scenario:
1. Distance estimation
2. Centroid selection
3. In terms of small and large dataset
4. Computational analysis

So, the main objective of this paper is to explore k-means clustering algorithm for the efficient distance estimation and centroid selection for small and large dataset.

*Author for correspondence

2.Literature survey

In 2017, Mahajan et al. [19] discussed regarding the classification of diabetes. They presented genetically optimized neural network classifier for the diabetes diagnosis. Their proposed method is mainly based on NN, principal component analysis (PCA) and genetic algorithm (GA). NN was used as the classifier and PCA was used for the dimensionality reduction. In 2017, Jasim et al. [20] discussed about the classification process. They have applied k-nearest neighbor (KNN) and artificial neural network (ANN) for the classification and evaluation purpose. They have considered Pima-Indian-Diabetes dataset. Then they applied T-test. The result shows the prominence of ANN over KNN. In 2017, Kalyankar et al. [21] discussed about diabetic mellitus (DM). They have implemented machine learning algorithm. It was implemented in the Hadoop MapReduce environment. It has been considered for the Pima Indian diabetes dataset. It is used to fetch out missing values. It is capable in the prediction of types of diabetes. In 2017, Kaur and Batra [22] discussed regarding the diabetes diagnosis. They used boosting and bagging to improve the classifier performance. They compare their approach that is Hierarchical and Progressive Combination of Classifiers with boosting and bagging. Experimentation was performed on PIMA Indian diabetes dataset. In 2018, Kaur et al. [23] discussed about smart healthcare technology. They have proposed a Cloud IoT based framework. It has been proposed for the diabetes prediction. They considered sensors in smart wearable devices for the monitoring and collections of blood glucose. They used ensemble model for the diabetes prediction in patients. Their result indicates that the highest accuracy obtained was 94.5% through decision tree (DT) and neural network (NN). In 2018, Huang and Lu [24] suggested the key for reducing the mortality is the early detection. They have combined information gain and deep neural network (DNN) for the early detection. The information gain has been used for decreasing the attributes. They achieved 90.26% of classification accuracy. In 2018, Kohli and Arora [25] discussed machine learning in terms of medical diagnosis. They have applied different classification algorithms. They considered three different datasets for the experimentation. These are Heart, Breast cancer and Diabetes dataset from the UCI repository. They applied p-value test for the feature selection by backward modeling. In 2018, Rani and Kautish [26] discussed regarding the large amount of health-related data in terms of health system. They suggested that the data processing and extraction may be difficult in case of size of the data. They suggested that the association clustering and time series-based data mining may be helpful in developing warning system. In 2018, Li and Ye [27] explored 152 type-2 diabetes. It belongs to Traditional Chinese Medicine records. They applied K-Medoids method for the clustering of TCM records. They applied FP-Growth algorithm for the extraction of hidden relationship. They achieved 71% accuracy. In Guttikonda et al. [28] discussed data mining techniques in terms of diabetes prediction. They considered hue for checking the nature of disease persistency. The Pima Indian database was considered for the experimentation. They have also used support vector machine (SVM). This combination is effective in the classification process. In 2019, Kim et al. [29] discussed about type 2 diabetes (T2D). They applied topological data analysis (TDA) for the clustering analysis on the T2D data. In 2019, Karthikeyan et al. [30] discussed about rule-based system. They proposed rule-based classification technique. They suggested that this technique is effective in large kind of problems. They compared different classifiers in case of diabetes dataset. In 2020, Devasena et al. [31] discussed about the data analytics. They have applied predictive diabetes diagnosis. It has been used for the prediction of type 2 diabetes. In their proposed system k-means clustering and random forest algorithms were used. Their results were found to be effective.

3.Proposed work

In this paper an efficient distance estimation and centroid selection based on k-means clustering for small and large dataset.

Steps of k-means algorithm are as follows:
Step 1: Assign the centroids randomly.
Step 2: Then there is the need of determining the distance of the data points. The distance will be measured from both of the object considering initial randomization.
Step 3: Then centroids updation was performed till the iterations or epochs determined.
Step 4: Distance and repositioning calculation were performed till the final cluster.
Step 5: Final clustered data

The complete approach is divided into following parts:
1. PIMA Indian diabetes dataset selection
2. Data pre-processing
3. Distance and centroid estimation
4. Impact on small and large dataset
5. Computation analysis

For the complete study and analysis PIMA Indian diabetes dataset was considered. Data pre-processing was performed on the dataset. It has been performed for the pruning of unwanted data, null instances and other non-relevant data. After pre-processing distance and centroid estimation was performed. The distance will be measured from both of the object considering initial randomization. Then centroids updation were performed till the iterations or epochs determined. Distance measures used here are Euclidean distance (Ed), Pearson Coefficient distance (PCd), Chebyshev distance (Csd) and Canberra distance (Cad). Ed determines the distance between two points. It provides the distance between the two points. It shows the estimation in Euclidean space. Other calculative coefficients were used in case of PCd, Csd and Cad.

Then impact on small and large dataset was considered for the further data estimation and analysis. For this part data division has been performed randomly. Data has been divided in different segments. Some segments are small and some segments are large. By this segmentation impact analysis were performed for the large and small segments. The segments considered were on the basis of random point selection. Finally, computational analysis was performed on the same data.

Figure 1 shows the complete working process mechanism of data selection and computational analysis

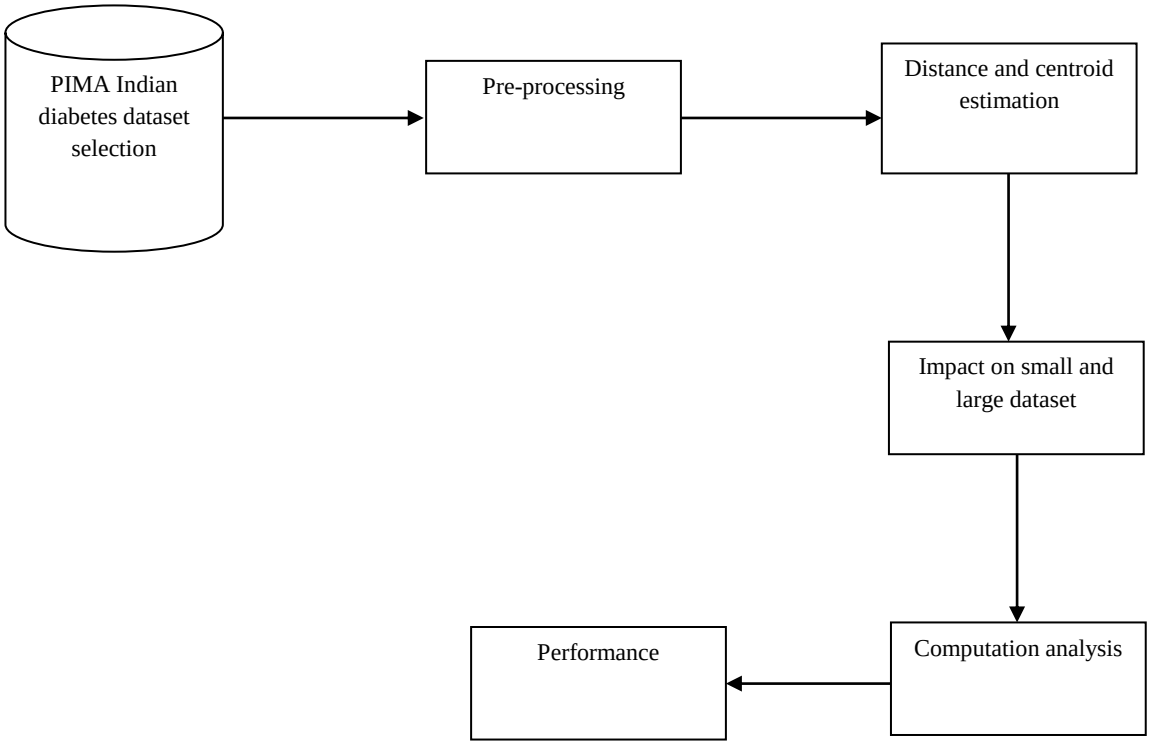

Figure 1 Complete working process mechanism of data selection and computational analysis

4.Results and discussion

The PIMA Indian diabetes dataset was considered for the iteration computation. Distance measures considered here are Ed, PCd, Csd and Cad. Ed determines the distance between two points. *Figure 2* shows the total elapsed iteration for small random dataset segment 1. *Figure 3* shows the total elapsed iteration for small random dataset segment 2. *Figure 4* shows the total elapsed iteration for large random dataset segment 1. *Figure 5* shows the total elapsed iteration for large random dataset segment 2.

The analysis indicates the following:

1. In case of Ed more iteration are needed.
2. In case of Cad a smaller number of iterations are needed.
3. Different trials suggest the same mechanism in case of large and small dataset.

4. Overall, all algorithms are found to be prominent in case of clustering data.
5. Cad needed less time in comparison to all.

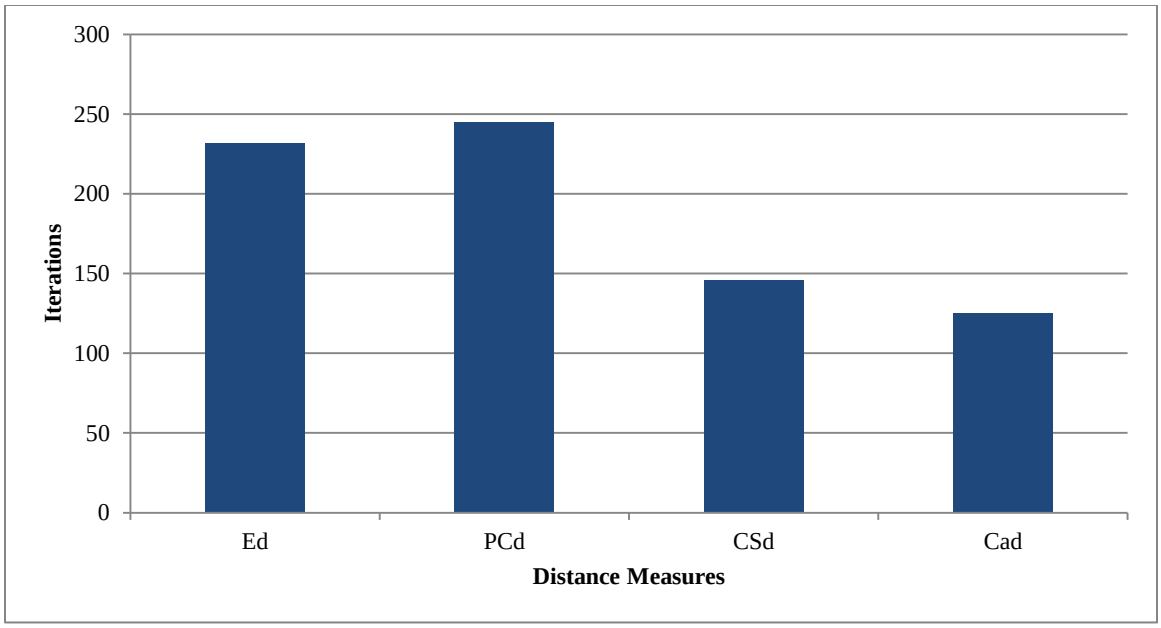

Figure 2 Total elapsed iteration for small random dataset segment 1

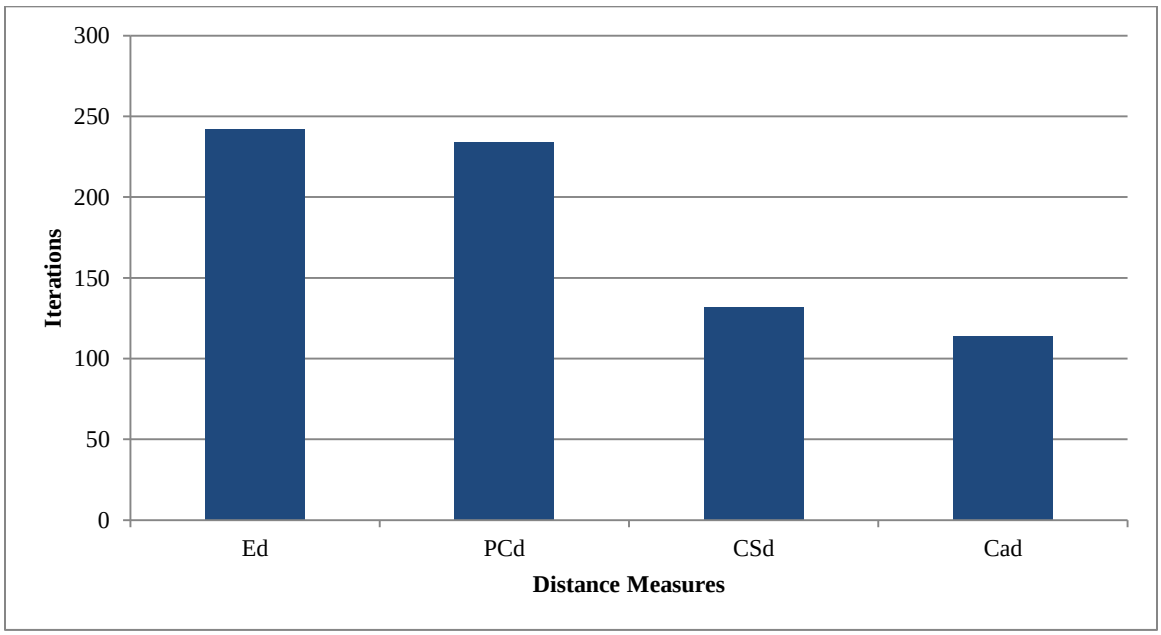

Figure 3 Total elapsed iteration for small random dataset segment 2

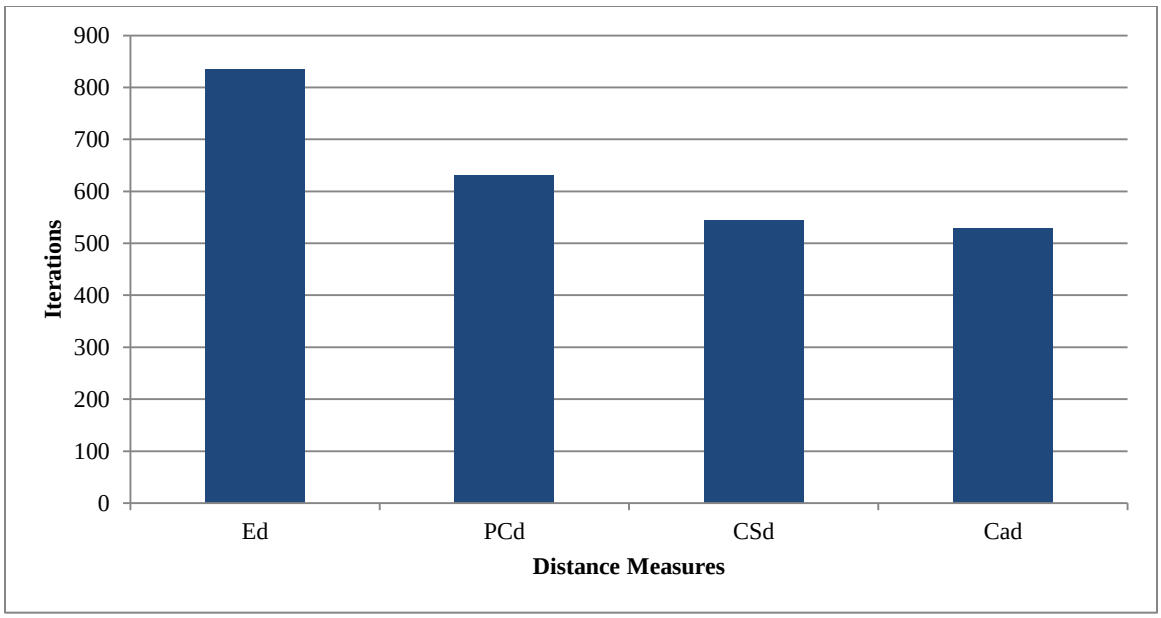

Figure 4 Total elapsed iteration for large random dataset segment 1

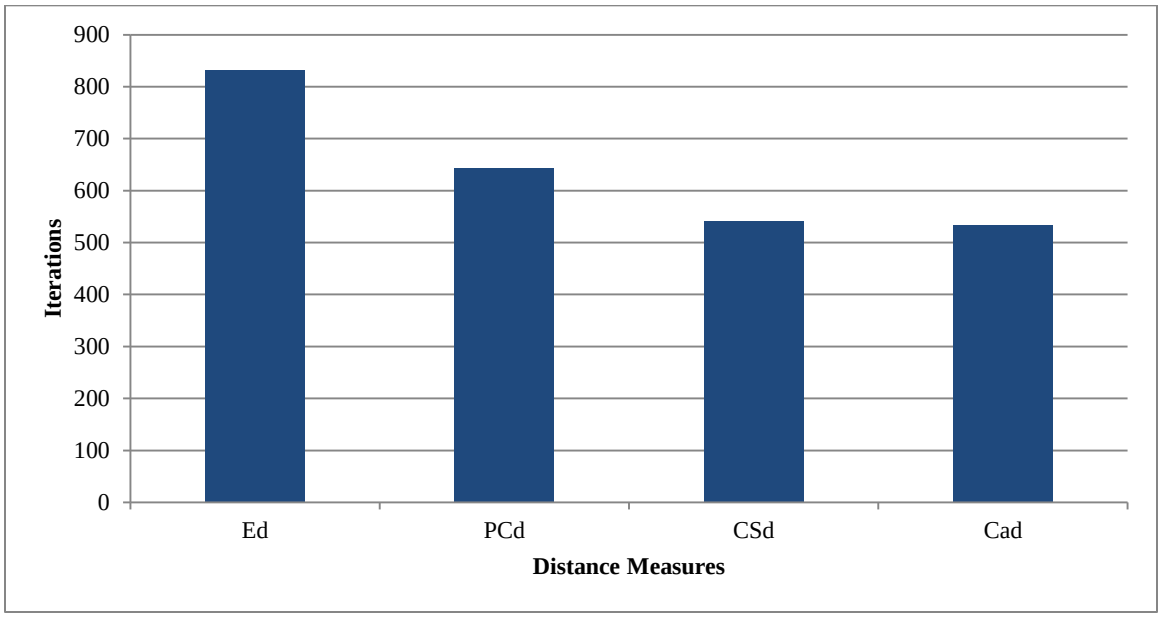

Figure 5 Total elapsed iteration for large random dataset segment 2

5. Conclusion

In this paper k-means clustering algorithm has been explored and analyzed. It has been explored in terms of four measures: distance estimation, centroid selection, small and large dataset and computational analysis. The results have been explored in terms of all the majors considered. In case of distance estimation four different distance algorithms were considered. Centroid selection was completely random. Different trials suggest the same mechanism in case of large and small dataset. Overall, all algorithms are found to be prominent in case of clustering data. Cad needed less time in comparison to all.

Acknowledgment

None.

Conflicts of interest

The authors have no conflicts of interest to declare.

References

[1] Fard MM, Thonet T, Gaussier E. Deep k-means: Jointly clustering with k-means and learning representations. Pattern Recognition Letters. 2020;138:185-92.

[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):9-16.

[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. In international conference on bioinformatics and computational biology 2019 (pp. 115-9). IEEE.

[5] Cios KJ, Moore GW. Uniqueness of medical data mining. Artificial Intelligence in Medicine. 2002; 26(1-2):1-24.

[6] Chahar R, Kaur D. A systematic review of the machine learning algorithms for the computational analysis in different domains. International Journal of Advanced Technology and Engineering Exploration. 2020; 7 (71): 147-64.

[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; 25(2):127-36.

[8] Kumari I, Sharma V. A review for the efficient clustering based on distance and the calculation of centroid. International Journal of Advanced Technology and Engineering Exploration. 2020; 7(63):48-52.

[9] Dubey AK, Gupta U, Jain S. Comparative study of K-means and fuzzy C-means algorithms on the breast cancer data. International Journal on Advanced Science, Engineering and Information Technology. 2018; 8(1):18-29.

[10] Pebesma J, Martinez-Millana A, Sacchi L, Fernandez-Llatas C, De Cata P, Chiovato L, et al. Clustering cardiovascular risk trajectories of patients with type 2 diabetes using process mining. In annual international conference of the engineering in medicine and biology society 2019 (pp. 341-4). IEEE.

[11] Iyer A, Jeyalatha S, Sumbaly R. Diagnosis of diabetes using classification mining techniques. arXiv preprint arXiv:1502.03774. 2015.

[12] Hao J, Zheng Y, Xu C, Yan Z, Li H. Feature assessment and classification of diabetes employing concept lattice. In 23rd international conference on computer supported cooperative work in design 2019 (pp. 333-8). 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; 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; 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; 9(42):146-55.

[16] Wang F, Wang Q, Nie F, Li Z, Yu W, Ren F. A linear multivariate binary decision tree classifier based on K-means splitting. Pattern Recognition. 2020; 107:107521.

[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; 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] Mahajan A, Kumar S, Bansal R. Diagnosis of diabetes mellitus using PCA and genetically optimized neural network. In international conference on computing, communication and automation 2017 (pp. 334-8). IEEE.

[20] Jasim IS, Duru AD, Shaker K, Abed BM, Saleh HM. Evaluation and measuring classifiers of diabetes diseases. In international conference on engineering and technology 2017 (pp. 1-4). IEEE.

[21] Kalyankar GD, Poojara SR, Dharwadkar NV. Predictive analysis of diabetic patient data using machine learning and Hadoop. In international conference on I-SMAC (IoT in social, mobile, analytics and cloud) (I-SMAC) 2017 (pp. 619-24). IEEE.

[22] Kaur H, Batra S. HPCC: An ensembled framework for the prediction of the onset of diabetes. In 4th international conference on signal processing, computing and control (ISPCC) 2017 (pp. 216-22). IEEE.

[23] Kaur P, Sharma N, Singh A, Gill B. CI-DPF: A cloud IoT based framework for diabetes prediction. In annual information technology, electronics and mobile communication conference 2018 (pp. 654-60). IEEE.

[24] Huang L, Lu C. Intelligent diagnosis of diabetes based on information gain and deep neural network. In international conference on cloud computing and intelligence systems 2018 (pp. 493-6). IEEE.

[25] Kohli PS, Arora S. Application of machine learning in disease prediction. In international conference on computing communication and automation 2018 (pp. 1-4). IEEE.

[26] Rani S, Kautish S. Association clustering and time series based data mining in continuous data for diabetes prediction. In second international conference on intelligent computing and control systems (ICICCS) 2018 (pp. 1209-14). IEEE.

[27] Li Y, Ye H. An analysis and research of type-2 diabetes TCM records based on text mining. In international conference on bioinformatics and biomedicine 2018 (pp. 1872-5). IEEE.
[28] Guttikonda G, Katamaneni M, Pandala M. Diabetes Data Prediction Using Spark and Analysis in Hue Over Big Data. In international conference on computing methodologies and communication 2019 (pp. 1112-17). IEEE.
[29] Kim HS, Yi C, Kim Y, Park U, Kook W, Oh B, Kim H, Park T. Topological data analysis can extract sub-groups with high incidence rates of Type 2 diabetes. International Journal of Data Mining and Bioinformatics. 2019; 22(1):44-60.
[30] Karthikeyan R, Geetha P, Ramaraj E. Rule Based System for Better Prediction of Diabetes. In 3rd international conference on computing and communications technologies 2019 (pp. 195-203). IEEE.
[31] Devasena MG, Grace RK, Gopu G. PDD: predictive diabetes diagnosis using datamining algorithms. In international conference on computer communication and informatics 2020 (pp. 1-4). IEEE.

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 from the Department of Computer Science, RKDF University, Bhopal (MP), India.

Email:ggladdha@gmail.com

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 and M.Tech. from S.A.T.I, Vidisha, M.P, India.