International Journal of Advanced Technology and Engineering Exploration, Vol 7(64)
ISSN (Print): 2394-5443 ISSN (Online): 2394-7454
http://dx.doi.org/10.19101/IJATEE.2020.762022

Research Article

An efficient ICKM approach for similarity measurement and distance estimation based on k-means

Isha Kumari[1*] and Vivek Sharma[2]
M.Tech Student, Department of Computer Science, TIT, Bhopal, MP, India[1]
Professor, Department of Computer Science, TIT, Bhopal, MP, India[2]

Received: 20-December-2019; Revised: 21-March-2020; Accepted: 24-March-2020

Abstract
An iterative centroid initialization k-means (ICKM) based clustering has been proposed in this paper. In this approach first the dataset selection has been performed along with the option of choosing and selection as per the data use or the user can access partial data also based on the iterative centroid. Then the data preprocessing steps are followed for the data arrangement and analysis. There are four different distance algorithms have been considered with the k-means. These algorithms provide the complete variability for the distance estimation and production. The proposed method found to be useful along with different distance estimation and measures.

Keywords
K-means, Euclidean, ICKM, Similarity measurement, Centroid distances.

1.Introduction
The need of clustering and classification techniques has been increased due to the arising need in several areas [1, 2].

Depend on the need and demand in genera two or three groupings has been performed [3]. The major use of clustering techniques includes different domains like health, information processing and pattern discovery [4, 5]. The usability and the aspects are clear from the approach that can be considered in different domain and the aspectual views can be considered for the different scenario.

The main aim of any data mining approaches is to process and analyses the data in the way to scale the data in the algometric way for data clustering and data arrangement to find the refined clusters [6, 7]. The data is arranged according to the content or the attributes values [8]. The arrangement in such a way that the data normalization has been performed to utilize the data in a meaningful and computational process [9].

This computational process is then applied to the next process and then the clustering process is started. This step is capable in providing the data in such manner that it can be processed and the algometric calculations have been performed systematically and easily [10, 11].

The main aspect of knowledge discovery is to analyse the knowledge in such a way to summarize the approaches in the similar way based on the data attributes, similarity behaviour and attribute properties [12].

The other computational aspects also cover the behaviour, recognition and attribute filtering for the accurate similarity measurement [13−17]. So it is needed to initialize and check the object distance with different variability. For this different distance measures have been considered in our paper. It also provides the analytical view for future research in the direction of distance measures.

Table 1 show the analysis based on current trends.

*Author for correspondence

Table 1 Current trends analysis

S.No	Reference	Method	Approach	Results achieved
1	[18]	Parallel heuristic for a k-medoids clustering	They have proposed a hybrid heuristic algorithm for the modified k-medoids problem. It is based on the shared memory parallel implementation. They have suggested that the dual bound for the objective value is the main advantage of their algorithm.	They have presented computational results based on large-scale problem. There are several decision variables.
2	[19]	Computer vision approach to spatial identification	They have proposed an efficent technique based on computer vision and machine learning for the k-clusters identification. They have applied their approach on the unsupervised data clustering.	Their results shows correct identification of k clusters. It also shows no loss of information. It also eliminates data overlap. They have used Silhouette and Precision matrix for the methodology testing.
3	[20]	Electricity consumption analysis through k-means	They have applied k-means algorithm for the analysis of the electricity consumption at home. It is based on the electricity data points. For the optimal number findings they have used davis boulden index and Silhouette_score in the k-means algorithm.	The results supports the approach.
4	[21]	Evolutionary algorithm for graph clustering problem	Authors approach has been inspired by algorithms like krill herd (KH) and genetic algorithm (GA). They have proposed a new graph clustering algorithm. They have suggested that KH is an effective algorithm. It is capable in solving continuous space optimization problems.	The proposed algorithms initial results shows that the results are of higher quality compared to other related algorithms.
5	[22]	Label propagation algorithms	They have discussed about label propagation algorithm (LPA). They have suggested that it is one of the classical community detection algorithms. They have suggested the disadvantages like poor stability. They have proposed an improved approach that is adjustable parameter for the stability of label propagation algorithm.	Their results shows the capability in reducing the randomness of the label propagation algorithm.
6	[23]	Density-grid based clustering method	They have discussed and analyzed for the algorithm for the purpose of scaling as well as suitable with big data sets. They hav proposed a fast density-grid clustering algorithm. The working mechnism is based on dividing the information space into a lattice structure and afterward doling out a thickness estimation to every framework cell.	Their experimental results shows better in comparison to DBSCAN in terms of accuracy and found lower run time.
7	[24]	Hierarchical clustering for categorical data	They have proposed a hierarchical clustering framework. It has been used for the purpose of clustering ategorical data based on Multinomial and Bernoulli mixture models.	Their approach main benifit is it can cluster image as well as text with different extensive experiments using the bag of visual words model.
8	[25]	Association rules algorithm	They have proposed a new data mining algorithm based on association rule algorithm.For clustering they have used k-means algorithm.	Their results shows the accuracy and efficiency of the approach.
9	[26]	Block-diagonal subspace clustering	They have proposed a directly pursuing block-diagonal affinity matrix method. They called it block-diagonal subspace clustering with Laplacian rank constraint (BDLRC). It has been used for the subspace clustering.	The results supports the approach.

2.Proposed work

This paper shows the mechanism of k-means with iterative centroid selection and random initialization of centroid that is ICKM. For exploring our approach in detail we have divided it into five different parts:
1. Dataset discussion
2. Pre-processing and data arrangement
3. ICKM
4. Similarity score calculation
5. Distance estimation

For the experimentation diabetes database have been considered. Data pre-processing has been performed and analysed. Then k-means approach has been applied along with the distance measures Euclidean (ED), Pearson Coefficient (PC), Chebyshev (Ch) and Canberra (Ca).

The flowchart is shown in *Figure 1*. It depicts and explores the procedure along with the clustering procedures. It shows the complete procedure along with the constraint satisfaction mechanism with the minimization and maximization procedure. The procedure starts with the input set selection.

Algorithm: K-means algorithm
Step 1: Input set has been selected from the preprocessed set.
Step 2: Initialized the cluster centers iteratively.
Step 3: Randomly determine the weight in each iterations.
Step 4: similarity score calculation for the minimum and maximum differences in each clusters
$X(c)=\sum_{j=1}^{k}\sum_{i=1}^{n}||d_i^{(j)}-c_j||^2$
$d_i - c_j$ is the Euclidean distance.
k is the number of cluster.
n shows the cases numbers.
Step 3: The cluster center is iteratively calculated until the similar conditions not match with the subsequent cases in finding minimum variance
$c_i = (\frac{1}{n_i})\sum_{j=1}^{n_i} d_i$
Step 4: Minimum variance based splitting has been assigned for each cluster.
Step 6: data point assignments have been accumulated.

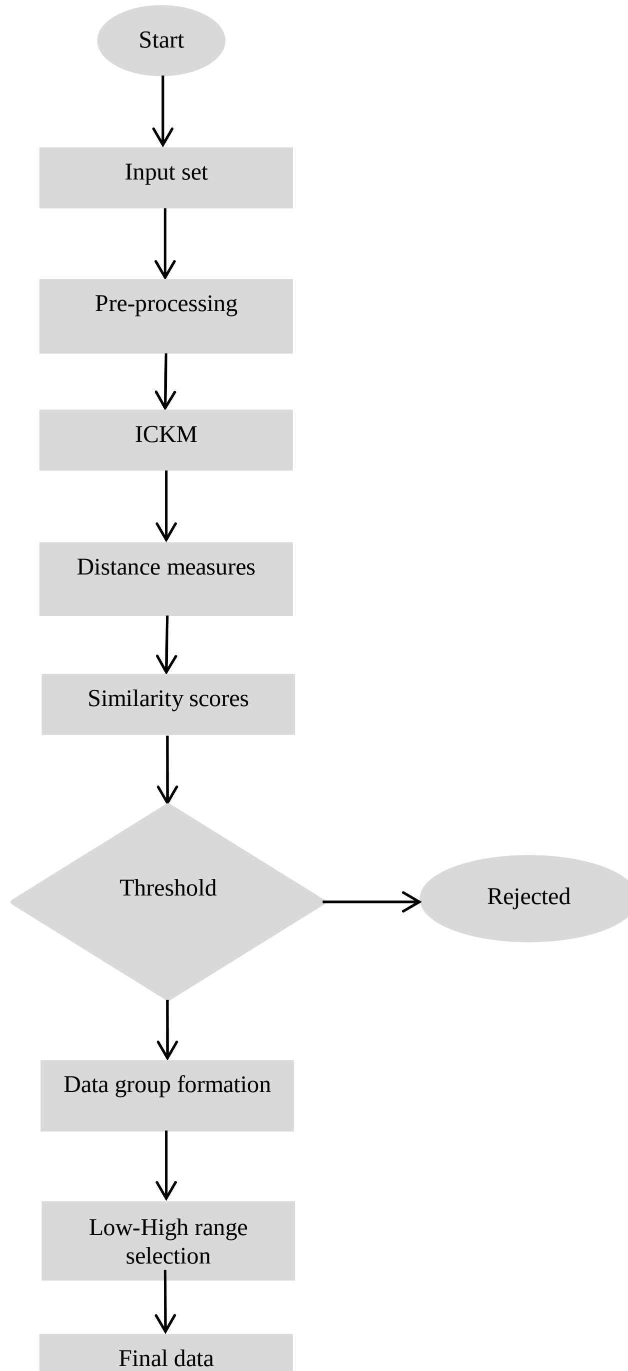

Figure 1 Flowchart of the proposed approach

3.Results and discussion

Three random sets have been selected from the complete data set for the time comparison.

Figure 2 shows the result analysis based on computational time in case of partial dataset 1. *Figure 3* shows the result analysis based on computational time in case of partial dataset 2. *Figure 4* shows the result analysis based on computational time in case of partial dataset 3. *Figure 5* shows the result analysis based on computational time in case of partial dataset 3.

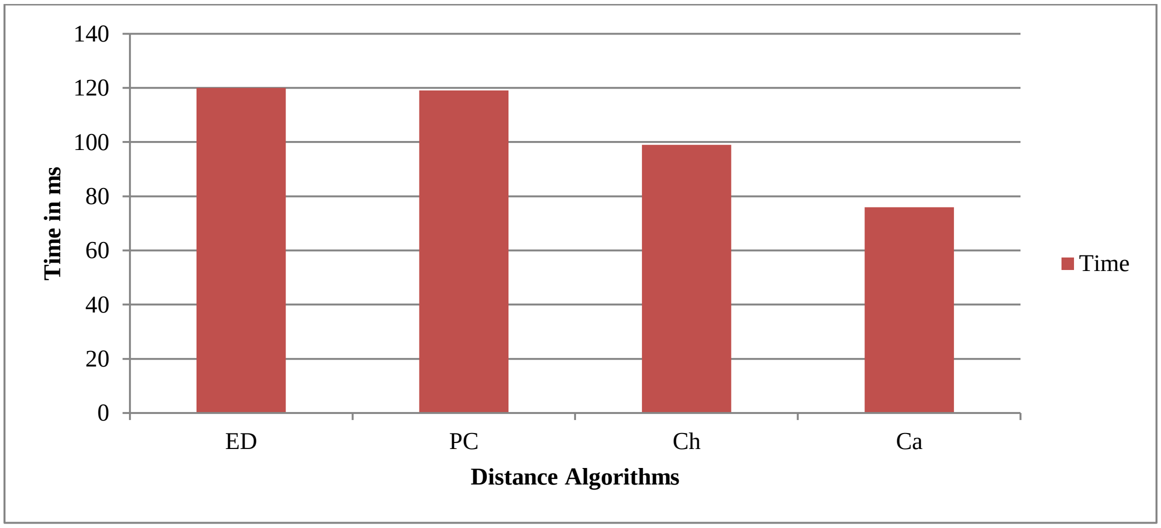

Figure 2 Result analysis based on computational time in case of partial dataset 1

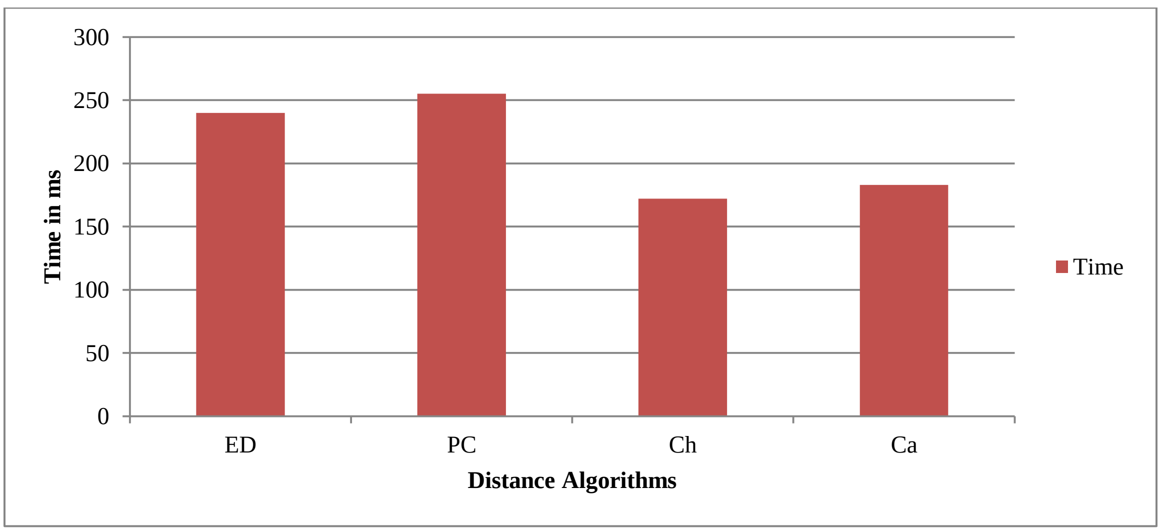

Figure 3 Result analysis based on computational time in case of partial dataset 2

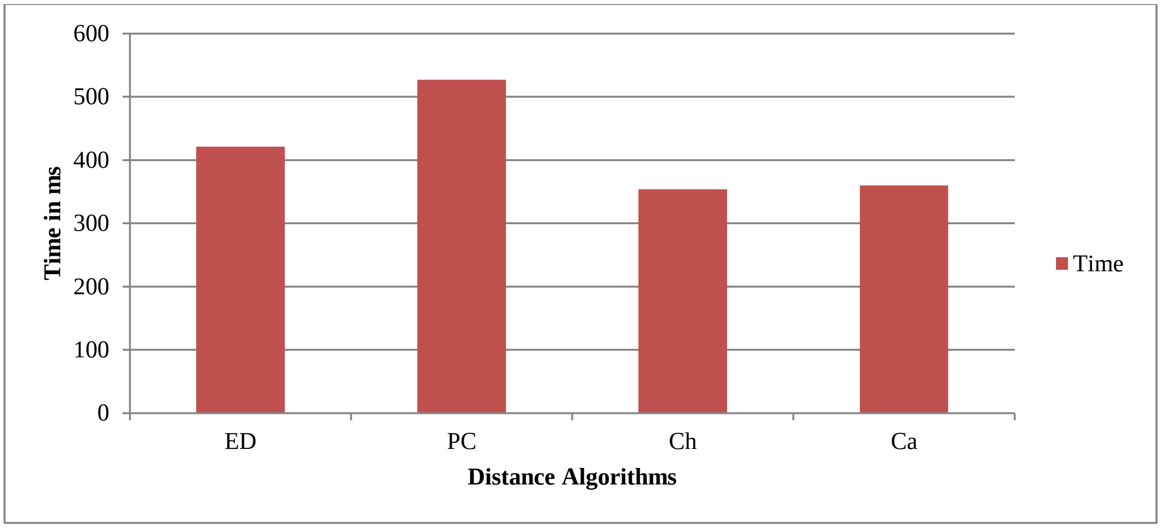

Figure 4 Result analysis based on computational time in case of partial dataset 3

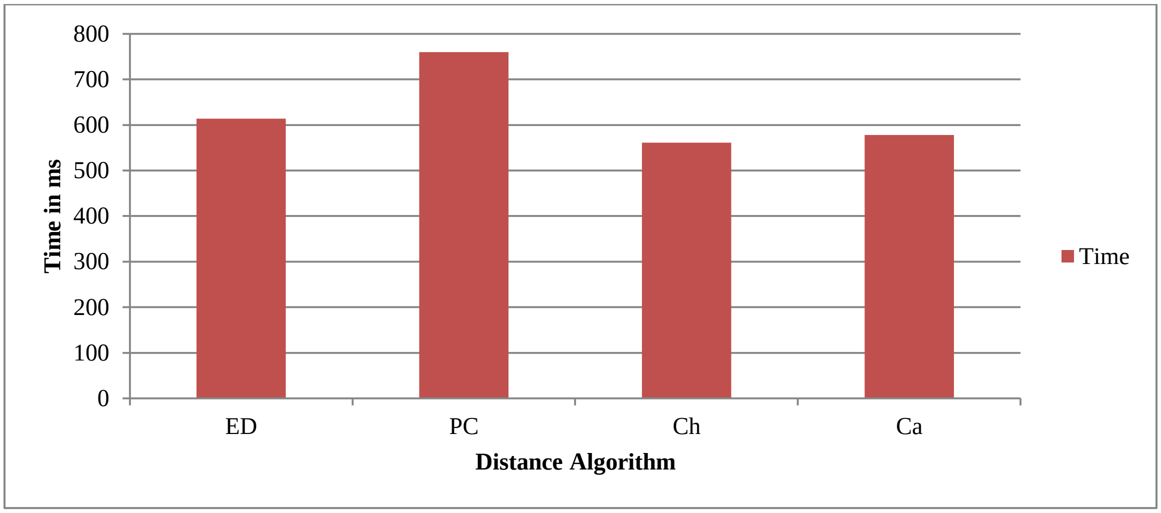

Figure 5 Result analysis based on computational time in case of partial dataset 3

4.Conclusion

In this paper different distance algorithm with the k-means algorithms for finding the centroid distance for improving the similarity matrix. An efficient iterative centroid k-means (ICKM) have been proposed with the experimental analysis. Finally different aspects have been considered for the analysis of the centroid distance impact and there initialization for the data clustering.

Acknowledgment

None.

Conflicts of interest

The authors have no conflicts of interest to declare.

References

[1] 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.

[2] 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.

[3] Mahmud MS, Rahman MM, Akhtar MN. Improvement of K-means clustering algorithm with better initial centroids based on weighted average. In international conference on electrical and computer engineering 2012 (pp. 647-50). IEEE.

[4] Margaret H. Data mining-“introductory and advanced concepts”. Pearson.

[5] Khandelwal A, Jain YK. An efficient k-means algorithm for the cluster head selection based on SAW and WPM. International Journal of Advanced Computer Research. 2018; 8(37):191-202.

[6] Pei J, Han J, Lu H, Nishio S, Tang S, Yang D. H-mine: hyper-structure mining of frequent patterns in large databases. In proceedings of international conference on data mining 2001 (pp. 441-8). IEEE.

[7] Dubey AK, Dubey AK, Agarwal V, Khandagre Y. Knowledge discovery with a subset-superset approach for mining heterogeneous data with dynamic support. In CSI sixth international conference on software engineering 2012 (pp. 1-6). IEEE.

[8] Babu DB, Prasad RS, Umamaheswararao Y. Efficient frequent pattern tree construction. International Journal of Advanced Computer Research. 2014; 4(14):331-6.

[9] Li K, Cui L. A kernel fuzzy clustering algorithm with generalized entropy based on weighted sample. International Journal of Advanced Computer Research. 2014; 4(2):596-600.

[10] Horeis T, Sick B. Collaborative knowledge discovery & data mining: from knowledge to experience. In symposium on computational intelligence and data mining 2007 (pp. 421-8). IEEE.

[11] Zhou Z, Wu Z, Feng Y. Enhancing reliability throughout knowledge discovery process. In sixth international conference on data mining-workshops 2006 (pp. 754-8). IEEE.

[12] Mansour AM. Decision tree-based expert system for adverse drug reaction detection using fuzzy logic and genetic algorithm. International Journal of Advanced Computer Research. 2018; 8(36):110-28.

[13] Jamil A, Salam A, Amin F. Performance evaluation of top-k sequential mining methods on synthetic and real datasets. International Journal of Advanced Computer Research. 2017; 7(32):176-84.

[14] Lan GC, Hong TP, Tseng VS. An efficient projection-based indexing approach for mining high utility itemsets. Knowledge and Information Systems. 2014; 38(1):85-107.

[15] Singh B, Dubey V, Sheetlani J. A review and analysis on knowledge discovery and data mining techniques. International Journal of Advanced Technology and Engineering Exploration. 2018; 5(41):70-7.

[16] Dubey AK, Shandilya SK. Exploiting need of data mining services in mobile computing environments. In international conference on computational intelligence and communication networks 2010 (pp. 409-14). IEEE.

[17] Kumar J, Vashistha R. Estimation of inter-centroid distance quality in data clustering problem using hybridized K-means algorithm. In second international conference on electrical, computer and communication technologies 2017 (pp. 1-7). IEEE.

[18] Ushakov AV, Vasilyev I. A parallel heuristic for a k-medoids clustering problem with unfixed number of clusters. In international convention on information and communication technology, electronics and microelectronics 2019 (pp. 1116-20). IEEE.

[19] Contreras GF, Delgado BM, Ibarra DG, De Castro CL, Jaimes BR. Cluster CV2: a computer vision approach to spatial identification of data clusters. In symposium on image, signal processing and artificial vision 2019 (pp. 1-5). IEEE.

[20] Choi HW, Qureshi NM, Shin DR. Analysis of electricity consumption at home using K-means clustering algorithm. In international conference on advanced communication technology 2019 (pp. 639-43). IEEE.

[21] Akbari M, Izadkhah H. GAKH: a new evolutionary algorithm for graph clustering problem. In international conference on pattern recognition and image analysis 2019 (pp. 159-62). IEEE.

[22] Wang M, Xu Y. Research on label propagation algorithms based on clustering coefficient. In 4th international conference on cloud computing and big data analysis 2019 (pp. 348-52). IEEE.

[23] Brown D, Japa A, Shi Y. A fast density-grid based clustering method. In 9th annual computing and communication workshop and conference 2019 (pp. 48-54). IEEE.

[24] Alalyan F, Zamzami N, Bouguila N. Model-based hierarchical clustering for categorical data. In IEEE 28th international symposium on industrial electronics 2019 (pp. 1424-9). IEEE.

[25] Zhang G, Liu C, Men T. Research on data mining technology based on association rules algorithm. In 8th joint international information technology and artificial intelligence conference 2019 (pp. 526-30). IEEE.

[26] Yang Y, Zhang X. Block-diagonal subspace clustering with laplacian rank constraint. In information technology, networking, electronic and automation control conference 2019 (pp. 1556-9). IEEE.

Isha Kumari is currently pursuing her M.Tech in Computer Technology and Application from Technocrats Group of Institutions, Bhopal, MP, India. Her interest areas are Data Mining, Machine Learning and Artificial Intelligence.
Email: nagaiach.isha@gmail.com