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

Isha Kumari¹ and Vivek Sharma²
M.Tech Student, Department of Computer Science, TIT, Bhopal, MP, India¹
Professor, Department of Computer Science, TIT, Bhopal, MP, India²

Received: 20-December-2019; Revised: 21-March-2020; Accepted: 24-March-2020
©2020 Isha Kumari and Vivek Sharma. This is an open access article distributed under the Creative Commons Attribution (CC BY) License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

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

<table>
<thead>
<tr>
<th>S.No</th>
<th>Reference</th>
<th>Method</th>
<th>Approach</th>
<th>Results achieved</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>[18]</td>
<td>Parallel heuristic for a k-medoids clustering</td>
<td>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.</td>
<td>They have presented computational results based on large-scale problem. There are several decision variables.</td>
</tr>
<tr>
<td>2</td>
<td>[19]</td>
<td>Computer vision approach to spatial identification</td>
<td>They have proposed an efficient technique based on computer vision and machine learning for the k-clusters identification. They have applied their approach on the unsupervised data clustering.</td>
<td>Their results show 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.</td>
</tr>
<tr>
<td>3</td>
<td>[20]</td>
<td>Electricity consumption analysis through k-means</td>
<td>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.</td>
<td>The results supports the approach.</td>
</tr>
<tr>
<td>4</td>
<td>[21]</td>
<td>Evolutionary algorithm for graph clustering problem</td>
<td>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.</td>
<td>The proposed algorithms initial results shows that the results are of higher quality compared to other related algorithms.</td>
</tr>
<tr>
<td>5</td>
<td>[22]</td>
<td>Label propagation algorithms</td>
<td>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.</td>
<td>Their results shows the capability in reducing the randomness of the label propagation algorithm.</td>
</tr>
<tr>
<td>6</td>
<td>[23]</td>
<td>Density-grid based clustering method</td>
<td>They have discussed and analyzed for the algorithm for the purpose of scaling as well as suitable with big data sets. They have proposed a fast density-grid clustering algorithm. The working mechanism is based on dividing the information space into a lattice structure and afterward doing out a thickness estimation to every framework cell.</td>
<td>Their experimental results shows better in comparison to DBSCAN in terms of accuracy and found lower run time.</td>
</tr>
<tr>
<td>7</td>
<td>[24]</td>
<td>Hierarchical clustering for categorical data</td>
<td>They have proposed a hierarchical clustering framework. It has been used for the purpose of clustering categorical data based on Multinomial and Bernoulli mixture models.</td>
<td>Their approach main benefit is it can cluster image as well as text with different extensive experiments using the bag of visual words model. Their results shows the accuracy and efficiency of the approach.</td>
</tr>
<tr>
<td>8</td>
<td>[25]</td>
<td>Association rules algorithm</td>
<td>They have proposed a new data mining algorithm based on association rule algorithm. For clustering they have used k-means algorithm.</td>
<td>Their results shows the accuracy and efficiency of the approach.</td>
</tr>
<tr>
<td>9</td>
<td>[26]</td>
<td>Block-diagonal subspace clustering</td>
<td>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.</td>
<td>The results supports the approach.</td>
</tr>
</tbody>
</table>
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

\[ X(c) = \sum_{j=1}^{k} \sum_{i=1}^{n} ||d_{ij}^{(j)} - c_{j}||^2 \]

\(d_{ij}\) - 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

\[ c_{i} = \left( \frac{1}{n_{i}} \right) \sum_{j=1}^{n_{i}} d_{ij} \]

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 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


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.


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.

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.


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.


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

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.

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.


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.

Alalayan 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.


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