International Journal of Advanced Technology and Engineering Exploration, Vol 7(63)
ISSN (Print): 2394-5443 ISSN (Online): 2394-7454
http://dx.doi.org/10.19101/IJATEE.2020.762021

Review Article

A review for the efficient clustering based on distance and the calculation of centroid

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: 10-November-2019; Revised: 15-February-2020; Accepted: 25-February-2020

Abstract
Clustering is helpful in different areas of interdisciplinary engineering. It helps in finding the alike element in a single label. The clustering efficiency depends on the centroid calculation and the nearest distance estimation. This paper's main aim is to review and analysis the method in finding the better clustering mechanism to extract the higher efficiency. In this regard different methods from the previous approaches have been discussed and their advantages have been highlighted. Based on the identified gaps, future suggestions have been listed for the efficient clustering mechanism.

Keywords
Distance calculation, Centroid estimation, Clustering, Distance measures.

1.Introduction
There are several clustering algorithms. It has been extensively applied in different sectors like engineering, e-commerce; health etc. K-means, fuzzy c-means and hierarchical clustering are mostly used in different areas [1, 2]. These algorithms efficiency depends on the way the data cluster based on the inter centroid calculation [2].

There are mostly two classes of grouping calculations dependent on the utilization [3]. These are parceling calculations and progressive calculations [3]. In parceling calculations, a set number of sets. If there should arise an occurrence of various leveled bunching and littler sets in a progressive way [4]. Bunching quality relies upon the nature of centroids, its instatement, separate methods and emphases [1, 2].

In the present condition in normal day by day presence the database is winding up quicker. So that to the obliging data likelihood for prune is the best alternative in information mining [5].

It is a part of data mining algorithms and it is helpful inefficient pattern identification. The data mining (DM) and knowledge discovery in databases (KDD) are the important aspect in knowledge discovery [6−9]. The strategy of DM shapes used to autonomous and insist designs in information is the point of convergence of the learning exposure process. These strategies fuse information choice, information pre-processing, information change, DM, and elucidation and assessment of cases. Different specialists have made recommendations that zone information should lead the DM strategy [10−12]. High-utility data mining is a discernible errand in the field on learning revelation.

Standard data and the investigations that clear for the social event of interminable happened things have been associated with utilizing the methodology displayed on [13]. In spite of the help of unending case mining, it recognizes that everything has proportional vitality and has a single event in each exchange [13−17]. High-utility representation mining settles this deterrent, by thinking about that everything may have a load that will fuse some pleasing data in looking at for those things. A few usages can get obliging data by mining the high utility itemsets in regard-based databases, for

*Author for correspondence

example, display container examination, click stream examination, and basic applications.

The main objectives of this paper are as follows:

1. To analyze clustering algorithms in different aspects and estimation parameters.
2. To find out the factors that are helpful in achieving higher accuracy.
3. To find out the factors which influence the centroid estimation.
4. To analyze different distance factors for the centroid distance.

Section structure organization of this paper is as follows. Section 2 discusses the related work and analysis based on the previous work. Section 3 provides the detail analysis and exploration based on the literature and some new insights. Section 4 provides the detail analysis based on the problem identified. Section 5 discusses the conclusion and the future suggestions.

2.Literature survey

In 2010, Mahmuddin and Yusof [18] proposed a test-and-generate approach. This approach is used for the estimation of the total numbers. For this they have applied the hybrid bees algorithm and cluster validity index. Based on their modified approach they are capable of finding the near optimal cluster centers. They have experimented their approach with benchmark datasets and found their approach to be useful.

In 2017, Limungkura and Vateekul [19] suggested that the traditional algorithms of clustering works on single cluster assignment. They focused on the handling of multiple data points. For this they have proposed an overlapping clustering concept. One of best in class segment-based covering bunching procedure is "Nonexhaustive, Covering K-Means" or "NEO-K-Means" to put it plainly, which is an augmentation of K-Means bunching calculation. Despite the fact that NEO-K-Means works adequately for most genuine world multicategory information. Be that as it may, the way toward doling out bunch's centers just on the base separation from the information to the centroid and disregards other fundamental parameters, for example, separate between groups and the range of bunches. Also, the quantity of estimation of information focuses on covering region is additionally still not sufficiently exact. These are colossal disadvantages of NEO-K-Means that makes bunching exactness lower than it ought to be. They expect to defeat this impediment in NEO-K-Means by utilizing span of bunch and separation between groups for helping in assessing information and doling out bunches. The test results demonstrate that theirstrategy essentially outflanks NEO-K-Means on nine genuine multi-class informational indexes as far as F1.

In 2018, Aryuni et al. [20] discussed about the internet banking customers. They have suggested that the customer segmentation is relied on internet banking data. They have applied k-means and k-medoids for the scoring based on the customer internet banking transactions. Their result performance shows that the k-means is better based on intra cluster distance.

In 2018, Hafezi et al. [21] suggested the challenges in the direction-of-arrival (DOA) estimation in case of multiple active acoustic sources. They have proposed a density-based clustering for the source counting in case of DOA estimation. Multiple density-based spatial clustering of applications with noise (DBSCAN) has been applied to obtain weighted centroids. A self-sufficient DBSCAN is at long last kept running on the weighted centroids to extricate the last DOA gauges. The outcomes utilizing created and assessed DOAs demonstrate that the proposed strategy altogether beats the traditional histogram top picking too as the first DBSCAN and varieties of Kmeans with $\leq 4°$ DOA estimation exactness and enhances the source checking.

In 2018, Malik et al. [22] discussed about clustering techniques. Bunching is generally used to distinguish concealed example in multidimensional complex information and, these shrouded examples give bases to making choices. They have suggested the drawback of k-means algorithm that it is not suitable for high dimension data. So, they have applied k-means with particle swarm optimization (PSO) and principle component analysis (PCA) for achieving the good results.

In 2018, Divya and Devi [23] discussed on efficient clustering mechanism for the high dimensional dataset. They have used k-means clustering algorithm. But they have noted that the results of this clustering algorithm are influenced by initial centroid points. They have used silhouette with PCA algorithms. It is used for the empty cluster reduction. It is also helpful in the new initial centroid estimation. They have used iris, wine, thyroid, yeast and solar datasets for the experimentation. Their results proved the effectiveness.

In 2018, Huan et al. [24] discussed the k-means algorithm in reference to the time and precision at the time of clustering. They have proposed classical vector space model for the textual representation. In light of the most extreme separation strategy, K information focuses with substantial dispersion contrast are chosen as the beginning bunch focuses, and the likeness between the group focuses and the example information is gotten through KL uniqueness. And after that put what share the similitude in a group, shape the group count equation and the separation measure capacity of the iterative focus, and figure the cycle until the example informational index is unfilled. The trial demonstrates that the enhanced content grouping calculation proposed in this paper not just diminishes the absolute utilization time of grouping, yet additionally enhances the precision of grouping in the meantime.

In 2018, Sing et al. [25] proposed the nearest instance centroid estimation (NICE) latent dirichlet allocation (LDA) algorithm. For experimentation they have considered different datasets from UCI repository. They have suggested that their approach is robust. They likewise build up a productive total technique in view of example based discovering that executes this course blend of classifiers in an a lot less complex way computationally. They exhibited that our technique for information bunching what's more, LDA usage, while presenting just a single free parameter, prompts results that are comparative and regularly superior to those accomplished by the best in class part RBF SVMs. Their approach shows the effectiveness based on other related approaches based on the approach combination.

3.Analysis

The following *Table 1* show the analysis based on the problems identified in terms of the method used in the previous literature. It also highlights the method used and how it is utilized with the parametric and computational contributions.

Table 1 Comparison based on related methods

S. NO	Sources	Algorithm used	Problems identified
1	[26]	K-means clustering algorithm using uniform distribution data points	Distance mapping is missing.
2	[27]	Cluster size constraints using a modified k-means algorithm	Centroid initialization and calculation can be random.
3	[28]	Clustering algorithm with genetic algorithm	Error identification is missing.
4	[29]	Improvement in k-means algorithm	They have recommended the fundamental downside of k-implies is to give suitable number of groups. Course of action of number of packs before applying the count is outstandingly unfeasible and requires significant learning of bunching field. They have proposed an enhancement in the introduction of the centroids. Separation estimates utilized are Manhattan remove, dice separation and cosine separate. The limit esteems can be connected in the introduction of the centroid. It very well may be stretched out to the halting condition moreover.
5	[30]	Rapid centroid estimation	They have suggested for the entropy minimization objective.

4.Problem statements

Based on the literature exploration and analysis the problems identified are as follows:

1. Minimum and maximum centroid threshold calculation mechanism is missing.
2. The different distance algorithm can be used and tested for the best accuracy.
3. Starting and stopping initialization is missing.
4. Automatic parameter selection and computation calculation are missing.
5. Cluster matching factor based on the combination of attributes is missing.

Figure 1 shows the concept of information management.

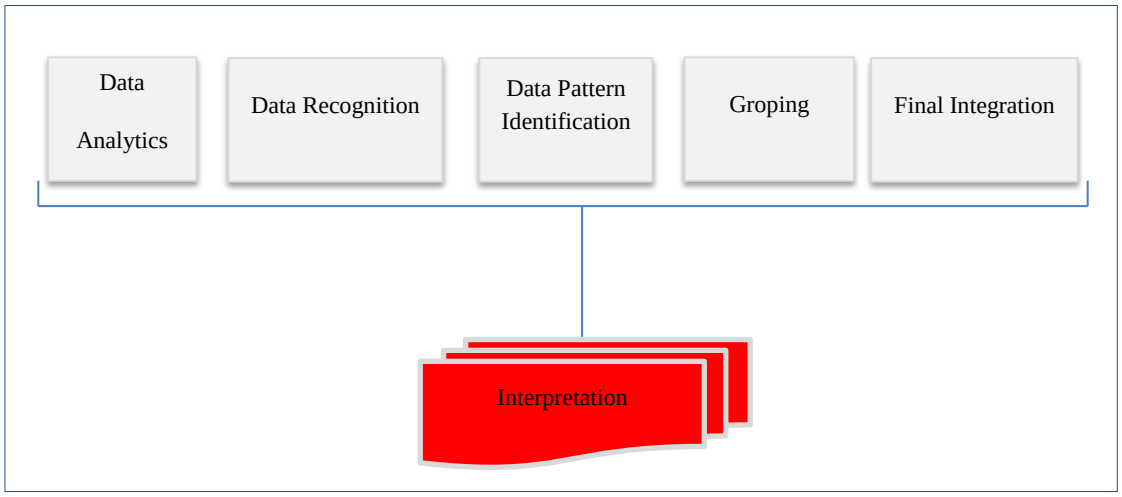

Figure 1 Information management

5.Conclusion and future work

In this paper parametric and computational analysis based on different parameters like centroid initialization, distance algorithm and threshold optimization have been discussed and analyzed. This study provides a detailed way to analyze the previous research work and find out the advantages and disadvantages and highlighted the related aspects based on different examples.

The future suggestions based on the above discussion and analyses are as follows:

1. Distance measurement parameters can be extended to validate the results.
2. Threshold based data control and centroid initialization can be helpful.
3. Data selection based on the centroid randomization can be used for the accurate estimation.

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".2006.

[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 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(15):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] Dubey AK, Gupta U, Jain S. Computational measure of cancer using data mining and optimization. In international conference on sustainable communication networks and application 2019 (pp. 626-32). Springer, Cham.
[18] Mahmuddin M, Yusof Y. Automatic estimation total number of cluster using a hybrid test-and-generate and K-means algorithm. In international conference on computer applications and industrial electronics 2010 (pp. 593-6). IEEE.
[19] Limungkura T, Vateekul P. Partition-based overlapping clustering using cluster's parameters and relations. In international conference on knowledge and smart technology 2017 (pp. 144-9). IEEE.
[20] Aryuni M, Madyatmadja ED, Miranda E. Customer segmentation in XYZ bank using K-means and K-medoids clustering. In international conference on information management and technology 2018 (pp. 412-6). IEEE.
[21] Hafezi S, Moore AH, Naylor PA. Robust source counting and acoustic DOA estimation using density-based clustering. In 10th sensor array and multichannel signal processing workshop 2018 (pp. 395-9). IEEE.
[22] Malik H, Sangrasi DM, Dayo ZA. Comparative analysis of hybrid clustering algorithm on different dataset. In 8th international conference on electronics information and emergency communication 2018 (pp. 25-30). IEEE.
[23] Divya V, Devi KN. An efficient approach to determine number of clusters using principal component analysis. In international conference on current trends towards converging technologies 2018 (pp. 1-6). IEEE.
[24] Huan Z, Pengzhou Z, Zeyang G. K-means text dynamic clustering algorithm based on KL divergence. In IEEE/ACIS 17th international conference on computer and information science 2018 (pp. 659-63). IEEE.
[25] Singh R, Li K, Principe JC. Nearest-instance-centroid-estimation linear discriminant analysis (Nice Lda). In international conference on acoustics, speech and signal processing 2018 (pp. 2846-50). IEEE.
[26] Napoleon D, Lakshmi PG. An efficient K-means clustering algorithm for reducing time complexity using uniform distribution data points. In trendz in information sciences & computing 2010 (pp. 42-5). IEEE.
[27] Ganganath N, Cheng CT, Chi KT. Data clustering with cluster size constraints using a modified k-means algorithm. In international conference on cyber-enabled distributed computing and knowledge discovery 2014 (pp. 158-61). IEEE.
[28] Kapil S, Chawla M, Ansari MD. On K-means data clustering algorithm with genetic algorithm. In fourth international conference on parallel, distributed and grid computing 2016 (pp. 202-6). IEEE.
[29] Rajeswari K, Acharya O, Sharma M, Kopnar M, Karandikar K. Improvement in K-means clustering algorithm using data clustering. In international conference on computing communication control and automation 2015 (pp. 367-9). IEEE.
[30] Yuwono M, Su SW, Moulton BD, Nguyen HT. Data clustering using variants of rapid centroid estimation. IEEE Transactions on Evolutionary Computation. 2013; 18(3):366-77.

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