ISSN (Online) 2278-1021
ISSN (Print) 2319-5940

IJARCCE

International Journal of Advanced Research in Computer and Communication Engineering

Vol. 7, Issue 8, August 2018

A Review and Analysis of Centroid Estimation in k-means Algorithm

Mrs. Mini Jain[1], Prof. Chetan Gupta[2]

M. Tech., Research Scholar, Department of CSE SIRTS, Bhopal[1]

Assistant Professor, Department of CSE SIRTS, Bhopal[2]

Abstract: This paper aim is to explore the centroid estimation analysis and distance measure variations from the previous methods of clustering and data mining techniques. This paper discusses the study based on these literatures so that methodological exploration may be possible. It is helpful in finding the advantages and disadvantages. Based on the gap identification new insights for the future development have been highlighted. This computation analysis also provides us the parametric exploration of the k-means clustering algorithm for the betterment in the efficiency of clustering.

Keywords: Software Metrics, Object Oriented Programming, Parameters, Quality Estimation

1. INTRODUCTION

Clustering algorithms have been used widely in different areas of research including health, business, student database etc. K-means algorithm is widely used clustering algorithm and simple in use [1, 2]. Centroid initialization and estimation is important in clustering [2]. There are mainly two categories of clustering algorithms based on the use [3]. These are partitioning algorithms and hierarchical algorithms [3]. In partitioning algorithms a limited number of sets. In case of hierarchical clustering and smaller sets in hierarchical way [4]. Clustering quality depends on the quality of centroids, its initialization, distance means and iterations [1, 2]. In the current circumstance in regular daily existence the database is ending up faster. So that to constraining information probability for prune is the best option in data mining [5].

Clustering is also an important part of data mining. The Data Mining (DM) and Knowledge Discovery in Databases (KDD) improvements have been established in transit that the authentic regard isn't in securing the data, yet rather in our ability to remove supportive reports and to find fascinating examples and associations [6-9]. The course of action of DM shapes used to independent and affirm plans in data is the focal point of the learning divulgence process. These techniques incorporate data decision, data preprocessing, data change, DM, and interpretation and evaluation of cases. Diverse experts have made proposals that zone data should lead the DM technique [10-12]. High-utility information mining is a noticeable task in the field on learning disclosure.

Customary information and the explorations that sweeps for social occasion of perpetual happened things has been connected by using the approach presented on [13]. Despite the support of unending case mining, it acknowledges that everything has equivalent essentialness and has single occasion in each trade [13-17]. High-utility illustration mining settles this obstruction by considering that everything may have a weight that will incorporate some accommodating information in examining for those things. A couple of utilizations can get accommodating information by mining the high utility item sets in esteem based databases, for instance, exhibit receptacle examination, click stream examination, and common applications. Correlated association also improves the clustering performance (Fig. 1).

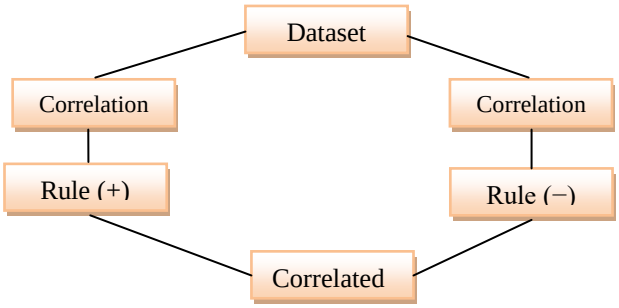

Figure 1. Correlation Analysis

DOI 10.17148/IJARCCE.2018.789

ISSN (Online) 2278-1021
ISSN (Print) 2319-5940

IJARCCE

International Journal of Advanced Research in Computer and Communication Engineering

Vol. 7, Issue 8, August 2018

The main objectives of this paper are as follows:

1. Review and analysis of k-means algorithms and the way of using it in different areas.
2. Study the factors which influence the results through k-means.
3. Impact and influence of centroid initialization.
4. Checking the impact through different distance measures.

The section organization in this paper is as follows. Literature review and the analysis of the literature have been presented in section 2. Analysis based on the methods has been discussed in section 3. Gap analysis has been presented in section 4. Section 5 discusses the conclusion based on the study and the future work.

2. LITERATURE SURVEY

In 2014, Jacob and Nazeer [18] suggested that the data mining tools are helpful for retrieving useful information from huge biological databases. For data grouping clustering is used extensively and efficient. For removing the drawback of k-means algorithm, they have combined k-means clustering algorithm and Improved Clustering Process Ant Colony Algorithm (lCPACA). The joined calculation is fit for deciding the ideal number of bunches and their comparing centroids. It likewise wipes out the issues because of nearby ideal arrangements and reliance on starting centroids.

In 2016, Wang et al. [19] suggested that the clustering results heavily depend on the initial centroids. A versatile technique for scatter centroids is proposed to enhance the soundness and exactness of the grouping result. The adaptively disperse centroids k-means algorithm (ADC-k-means) have been proposed using MapReduce model on hadoop platform. It is then compared with the k-means algorithm. The exploratory outcome demonstrates that proposed calculation is viable.

In 2017, Olukanmi and Twala [20] suggested that the traditional k-means algorithm is easily misled by outliers. For this, they have updated the centroid update step. With the aim that problem is maintained a strategic distance from when new centroids are figured. They have proposed k-means-sharp (k-means) to detect exceptions naturally by methods for a worldwide limit got from the circulation of point-to-centroid separations. The approach requires neither client mediation nor earlier learning of the quantity of exceptions. Since it saves k-means' structure, k-means# acquires the previous' simplicity of usage and if wanted, it can profit by other existing k-implies changes.

In 2017, Kumar and Vashistha [21] applied k-means algorithm on medical dataset. They have performed their experimentation using real and artificial datasets on MATLAB. They have shown the improvement in accuracy in case of diabetes dataset. The results were looked at by utilizing conventional separation work versus proposed separate capacity for customary k-Means calculation. The outcomes demonstrates an enhanced k-implies characterization calculation by applying proposed technique for discovering least separation between centroid. At the point when thought about centroid separate by applying customary Euclidean, Canberra in k-Means calculation, the proposed adjusted calculation demonstrates least separation. A reproduction demonstrates Canberra separate capacity perform better as contrasted with Euclidean and proposed model.

In 2017, Prem kumar and Ganesh [22] suggested that the clustering is useful in the area like medical, business and education. They have suggested that the k-means is widely used clustering algorithm. They suggested that it has been suffer from the selection of random initial centroids. It may degrades the performance of the
Clustering results. So they have suggested median based initial centroids. It has been applied on the experimental set for the performance validation. Their results shown that the accuracy of clustering with reduced number of iterations has been improved.

In 2017, Trivedi, Kanungo [23] suggested that the microarray data play an important role in monitoring the articulation profile of substantial number of genes. The k-implies bunching calculation is picking up ubiquity in the information disclosure area for adequately investigating this information by looking at quality articulation profiles or test articulation profiles. Notwithstanding, the procedure utilized in this calculation is computationally costly as respect to time intricacy and choice of introductory centroids. It chooses introductory centroids arbitrarily that influence the nature of coming about bunches. With a specific end goal to effectively handle this issue, they have proposed a variation technique for discovering beginning centroids by utilizing entropy based most distant neighbor approach. Our exploratory outcomes show the exactness of quality groups with less number of cycles rather than the conventional k-implies bunching.

In 2017, Rahim and Ahmad [24] proposed a new method based on radial and angular coordinates for the selection of coordinates. To check the attainability of the proposed strategy, they have contrasted their technique and the standard K-implies calculation. For the examination, we utilize manufactured informational indexes with various size of examples and number of bunches. The analysis demonstrates that in the vast majority of the cases the proposed

ISSN (Online) 2278-1021
ISSN (Print) 2319-5940

IJARCCE

International Journal of Advanced Research in Computer and Communication Engineering

Vol. 7, Issue 8, August 2018

technique unmistakably commands over the standard k-implies calculation as far as execution time and required number of cycles.

In 2018, Wang et al. [25] proposed a split-merge-evolve algorithm for clustering data into k number of clusters. The calculation haphazardly separates information into k bunches at first, at that point more than once parts terrible bunches and unions nearest groups to advance the last bunching outcome. A key metric amid the bunching procedure of the Split-consolidate develop calculation is a client picked or characterized bunching quality metric or then again inward assessment. The calculation develops the grouping result towards the client expected fantastic outcome, despite the fact that there is no ground truth or named information required amid the grouping procedure. The calculation configuration is adaptable in its execution, with different normal strategies, for example, centroid and availability based measures that can be utilized as a part of its usage. The calculation is anything but difficult to execute and viable. With 4 datasets, including 2 genuine datasets in our tests, the Split-blend advance calculation performs better than both most regularly utilized k-implies and agglomerative various leveled calculations.

In 2018, Chouhan and Purohit [26] suggested that the WWW is largest source of shared information. To viably sort out, abridge and explore through the data on the web in a quick and top notch way, report grouping calculations are required. Different bunching calculations are proposed by the scientists in whom the k-means is broadly utilized apportioning grouping calculation which is simple for execution, has quick merging property in neighborhood, what's more, sets aside less time for execution. Be that as it may, significant downside of this technique is its arbitrary decision of starting group centroids. To conquer this issue, an approach for record grouping utilizing particle swarm optimization (PSO) strategy is proposed. PSO strategy is connected before K-means for finding the ideal focuses in the inquiry space and these focuses are utilized as introductory bunch centroids for k-means calculation to discover last groups of archives. Consequences of bunching calculations are tried on four distinctive archive datasets. The result demonstrates that the most productive grouping comes about are created than conventional k-means.

3. ANALYSIS

Based on the review analysis, the related methods have been shown with the method advantages and the gaps identified. The comparison shown in Table 1 show the sources, method highlights and the gaps identified.

Table 1. Comparison Based on Related Methods

S. No	Sources	Method	Gap identification
1	[27]	K-means clustering algorithm using uniform distribution data points	They have presented an efficient k-means algorithm. It is based on uniform distribution data points. Distance mapping can be added in this method for uniformity in mapping also.
2	[28]	Cluster size constraints using a modified k-means algorithm	The altered k-implies calculation can be utilized to get bunches in favored sizes. A potential application would acquire bunches with measure up to group estimate. In addition, the adjusted calculation makes utilization of earlier information of the given informational index for specifically introducing the group centroids which helps getting away from neighborhood minima. The outcomes on multidimensional information exhibit that the k-implies calculation with the proposed alterations can satisfy group measure limitations and prompt more exact and powerful outcomes. It can be extended with the random size selection.
3	[29]	Clustering algorithm with genetic algorithm	Error can be calculated with the margin values from each iteration to improve the accuracy.
4	[30]	Improvement in k-means algorithm	They have suggested the main drawback of k-means is to provide appropriate number of clusters. Arrangement of number of bunches before applying the calculation is exceptionally unfeasible and requires profound learning of clustering field. They have proposed an improvement in the initialization of the centroids. Distance measures used are Manhattan distance, dice distance and cosine distance. The boundary values can be

ISSN (Online) 2278-1021
ISSN (Print) 2319-5940

IJARCCE

International Journal of Advanced Research in Computer and Communication Engineering

Vol. 7, Issue 8, August 2018

			applied in the initialization of the centroid. It can be extended to the stopping condition also.
5	[31]	Modified k-means algorithm	They have proposed a modified k-means algorithm. It is based on the sensitivity of initial center. This calculation partitions the entire space into various fragments also, ascertains the recurrence of information point in each portion. The quantity of centroid (k) will be given by the client in a similar way like the customary K-mean calculation and the quantity of division will be k*k ('k' vertically and in addition 'k' evenly). In the event that the most astounding recurrence of information point is same in various portions and the upper bound of Portion crosses the limit 'k' at that point converging of various portions end up compulsory and afterward take the most astounding k portion for figuring the underlying centroid (seed point) of clusters. Time synchronization can be considered.
6	[32]	K-means text clustering algorithm	They have proposed an improved k-means text clustering algorithm. In this they have optimized the initial cluster centers. The calculation initially ascertains the thickness of every datum question in the informational collection, and after that judge which information protest is a segregated point. In the wake of evacuating all of separated focuses, an arrangement of information objects with high thickness is gotten. A while later, picks k high information protests as the underlying group focuses, where the remove between the information objects is the biggest. The trial comes about demonstrate that the enhanced k-means calculation can improve the stability and accuracy of content clusters. It can be applied to different domains.

4. PROBLEM STATEMENTS

The following gaps have been identified for the betterment in the previous approaches and from the literature suggested.

1. There is the need of automatic selection of distance algorithm according to the global rank selection mechanism.
2. Distance algorithm like Manhattan and Pearson coefficients are missing in the previous literature.
3. Variations in stopping conditions are missing in the previous literature.
4. Random centroid selection and initialization is missing in the previous research.
5. Control in the initial centroid variations and the mechanism for finding the solution based on the cluster matching factor is missing in the previous research work.

5. CONCLUSION AND FUTURE WORK

This paper explores the k-means algorithm in different aspects including the parametric, centroid initialization, distance algorithm and attribute exploration. This study explores the pros and cons along with the analytical view of analyzing the approach in detail. This study focus on the enhancements and the attributes responsible for the enhancement.

Based on the study and analysis the future suggestions are as follows:

1) Distance metric selection can be validated with the ranking mechanism.
2) Different combinations of distance measures may be used.
3) Controlling the boundary values for the stopping condition may be a future challenge.
4) Optimizing the attribute with the random selection for the best value mechanism is the need of future enhancement.

ISSN (Online) 2278-1021
ISSN (Print) 2319-5940

IJARCCE

International Journal of Advanced Research in Computer and Communication Engineering

Vol. 7, Issue 8, August 2018

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 Electrical & Computer Engineering (ICECE), 2012 7th International Conference on 2012 Dec 20 (pp. 647-650). IEEE.
[4]. Dunham MH. "Data mining-introductory and advanced concepts". Pearson Education; 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 Data Mining, 2001. ICDM 2001, Proceedings IEEE International Conference on 2001 (pp. 441-448). 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 Software Engineering (CONSEG), 2012 CSI Sixth International Conference on 2012 Sep 5 (pp. 1-6). IEEE.
[8]. Babu DB, Prasad RS, Umamaheswararao Y. "Efficient frequent pattern tree construction. International Journal of Advanced Computer Research". 2014 Mar 1;4(1):331.
[9]. Li K, Cui L. "A kernel fuzzy clustering algorithm with generalized entropy based on weighted sample". International Journal of Advanced Computer Research. 2014 Jun 1;4(2):596.
[10]. Horeis T, Sick B. Collaborative knowledge discovery & data mining: From knowledge to experience. In Computational Intelligence and Data Mining, 2007. CIDM 2007. IEEE Symposium on 2007 Mar 1 (pp. 421-428). IEEE.
[11]. Feng Y, Wu Z, Zhou Z. "Enhancing reliability throughout knowledge discovery process". In Data Mining Workshops, 2006. ICDM Workshops 2006. Sixth IEEE International Conference on 2006 Dec (pp. 754-758). 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 May 1;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 Sep 1;7(32):176.
[14]. Lan GC, Hong TP, Tseng VS. "An efficient projection-based indexing approach for mining high utility item sets". Knowledge and information systems. 2014 Jan 1;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 Apr 1;5(41):70-7.
[16]. Dubey AK, Shandilya SK. "Exploiting need of data mining services in mobile computing environments". In Computational Intelligence and Communication Networks (CICN), 2010 International Conference on 2010 Nov 26 (pp. 409-414). IEEE.
[17]. Li K, Cui L. "A kernel fuzzy clustering algorithm with generalized entropy based on weighted sample". International Journal of Advanced Computer Research. 2014 Jun 1;4(2):596.
[18]. Jacob C, Nazeer KA. "An improved ICPACA based K-means algorithm with self-determined centroids". In Data Science & Engineering (ICDSE), 2014 International Conference on 2014 Aug 26 (pp. 89-93). IEEE.
[19]. Wang B, Lv Z, Zhao J, Wang X, Zhang T. "An Adaptively Disperse Centroids K-Means Algorithm Based on Map Reduce Model". In Computational Intelligence and Security (CIS), 2016 12th International Conference on 2016 Dec 16 (pp. 142-146). IEEE.
[20]. Olukanmi PO, Twala B. "K-means-sharp: modified centroid update for outlier-robust k-means clustering". In Pattern Recognition Association of South Africa and Robotics and Mechatronics (PRASA-RobMech), 2017 Nov 30 (pp. 14-19). IEEE.
[21]. Kumar J, Vashistha R. "Estimation of inter-centroid distance quality in data clustering problem using hybridized K-means algorithm". In Electrical, Computer and Communication Technologies (ICECCT), 2017 Second International Conference on 2017 Feb 22 (pp. 1-7). IEEE.
[22]. Premkumar MS, Ganesh SH. "A Median Based External Initial Centroid Selection Method for K-Means Clustering". In 2017 World Congress on Computing and Communication Technologies (WCCCT) 2017 Feb 1 (pp. 143-146). IEEE.
[23]. Trivedi N, Kanungo S. "Performance enhancement of K-means clustering algorithm for gene expression data using entropy-based centroid selection". In Computing, Communication and Automation (ICCCA), 2017 International Conference on 2017 May 5 (pp. 143-148). IEEE.
[24]. Rahim MS, Ahmed T. "An initial centroid selection method based on radial and angular coordinates for K-means algorithm". In Computer and Information Technology (ICCIT), 2017 20th International Conference of 2017 Dec 22 (pp. 1-6). IEEE.
[25]. Wang M, Huang V, Bosneag AM. "A novel Split-merge-evolve k clustering algorithm. International Conference on Big Data Computing Service and Applications" (pp. 229-36). IEEE.
[26]. Chouhan R, Purohit A. "An approach for document clustering using PSO and K-means algorithm". In 2018 2nd International Conference on Inventive Systems and Control (ICISC) 2018 Jan 19. IEEE.
[27]. Napoleon D, Lakshmi PG. "An efficient K-Means clustering algorithm for reducing time complexity using uniform distribution data points". In Trends in Information Sciences & Computing (TISC), 2010 Dec 17 (pp. 42-45). IEEE.
[28]. Ganganath N, Cheng CT, Tse CK. "Data clustering with cluster size constraints using a modified k-means algorithm". International conference on cyber-enabled distributed computing and knowledge discovery (pp. 158-61). IEEE.
[29]. Kapil S, Chawla M, Ansari MD. On "K-means data clustering algorithm with genetic algorithm". In Parallel, Distributed and Grid Computing (PDGC), 2016 Fourth International Conference on 2016 Dec 22 (pp. 202-206). IEEE.
[30]. Rajeswari K, Acharya O, Sharma M, Kopnar M, Karandikar K. "Improvement in K-means clustering algorithm using data clustering". In Computing Communication Control and Automation (ICCUBEA), 2015 International Conference on 2015 Feb 26 (pp. 367-369). IEEE.
[31]. Singh RV, Bhatia MS. "Data clustering with modified K-means algorithm". In Recent Trends in Information Technology (ICRTIT), 2011 International Conference on 2011 Jun 3 (pp. 717-721). IEEE.
[32]. Xiong C, Hua Z, Lv K, Li X. "An Improved K-means text clustering algorithm By Optimizing initial cluster centers". In Cloud Computing and Big Data (CCBD), 2016 7th International Conference on 2016 Nov 16 (pp. 265-268). IEEE.