Svedberg Open
DISSEMINATION OF KNOWLEDGE

International Journal of Artificial Intelligence and Machine Learning

Publisher's Home Page: https://www.svedbergopen.com/



Research Paper Open Access

A Comprehensive Survey of K-Means Clustering Algorithms: Techniques, Advantages, Limitations, and Accuracy Improvement Strategies

M S Grace[1], B Prajna[2]

[1] Research Scholar, Department of Computer Science and Systems Engineering, Andhra University, Visakhapatnam, India. Email: msgrace.rs@gmail.com
[2]Professor, Department of Computer Science and Systems Engineering, Andhra University, Visakhapatnam, India. Email: prajna.mail@gmail.com

Abstract
Clustering is the most important stage in the analysis of data, which consists of categorizing the collection of objects into groups referred to as clusters. The most commonly used and essential method of clustering is K-means clustering. It's likewise referred to as "nearest neighbor searching." Several efforts have been taken to boost the accuracy of the K-means clustering method. The elements or investigation points are to be grouped into distinct groups using the k-means approach. The K-means clustering algorithms primary centroids are randomly selected, which poses an important concern because the algorithm's effectiveness largely depends on the initial centroids and may contribute to local modifications. Increasing the clustering outputs of the K-means clustering procedure was the primary aim of the current study. Relevant outcomes from multiple studies have shown the standard clustering technique and the prospects for accuracy improvements in clustering. The present publication provides an overview of the research conducted by several researchers employing K-means clustering. We have additionally expressed regarding the K-means clustering algorithm's benefits and limitations. The article delivers an extensive review of the numerous k-means clustering algorithm techniques that have been developed over time. This study may have significant effects on text-based clustering and topic depicted data.

Keywords: Cluster analysis, K-means clustering, closest neighbor searching, centroid, K-means clustering, unsupervised classification, clusters, and data mining.

1. INTRODUCTION

As a consequence of the growing information technology company operations and the increasing availability of hardware and software for computers, an enormous amount of data has been obtained and protected in data bases. The amount of understanding in the world is expected to double every 20 months, based on researchers. However, direct consumption of raw data is not permitted. By taking out information that is helpful for making decisions, its real worth can be measured. Data analysis was frequently carried out by hand in the vast majority of regions. People look for technological solutions to automate procedures when the quantity of data analysis and examination outweighs human capacity (Khadem et al.).

Data exploration is an emerging field of study in the area of computing that includes finding patterns or knowledge from enormous quantities of data that are fascinating implicit, previously unidentified, and potentially beneficial. Data mining is used to derive some valuable information from huge quantities of data. Researchers in data analysis have produced a variety of instruments and approaches for extracting similarities from data. Classification, clustering, rules for associations, regression analysis, outlier analysis, and other procedures are all useful to mine different patterns. (Vaidya et al. and Patil)

The clustering of data is a hot study topic with applications in a number of disciplines, including marketing, psychiatry, computer science, statistics, pattern recognition, data mining, and processing images. [(Anderberg), (Everitt et al.), Brohee and Helden]. However it can be utilized with univariate and bivariate data, clustering is usually thought of as a multivariate technique. As stated by Johnson and Wichern, clustering or grouping is one of the best multivariate analytical approaches and an established approach for statistical data analysis. It can be carried out based on similarities or distances.

In 1938, Zubin reintroduced clustering—also termed cluster analysis—to psychology. Initially invented in anthropology by (Driver and Kroeber), (Cattell) presented mathematical techniques for categorizing items based upon perceived shared characteristics. Prior to the publication of Principles of Numerical Taxonomy by Sokal and Sneath, strategies for clustering were not extensively employed in the world of science. As a result, methodologies for clustering are being researched all over the world, and a wide range of books have been

issued by researchers such as Anderson, Tryon and Bailey, Jardine and Sibson, Fisher, Hartigan, H. Spath, Aldenderfer and Blashfield, Romesburg, Fukunagaga, Kaufman and Rousseeuw, Berkhin, and Mikhin, among others. K-means, the clustering formulation that is most commonly used, aims to optimize the observed closeness between the input data components and the centroids of the cluster with which they are associated (Slonim et al.). The goal of this paper is to provide a comprehensive overview of the many k-means clustering approaches that were successfully developed in recent years.

2. CLUSTER ANALYSIS

An unsupervised method of classification referred to as cluster analysis, or clustering, divides a collection of information—generally multidimensional—classified in groups or clusters by considering the similarity of their members in accordance to a predefined criterion [(Hartigan), (Jain and Dubes), (Estivill)]. Based on the manner in which their output is organized, cluster analysis can be classified into two primary groupings: partitioning and hierarchical nonhierarchical cluster strategies. Equivalent objects are clustered together utilizing a concept named hierarchical clustering, or hierarchical clustering analysis. The similarity value is used in order to gradually merge the clusters (agglomerative methods) or break them (divisive approaches). The findings of a hierarchical clustering method indicate that a dendrogram, or tree diagram, can be utilised to visually represent agglomerative and divisive techniques. The dendrogram displays each stage of the hierarchical process, indicating the points at which clusters join based on differences or similarity. Partitioning processes have been employed for splitting the data object set into clusters, assuring that each combination of item collections is entirely distinct from the other or contains certain objects. Unless a locally optimal fraction is determined, cluster segmentation begins with an initial cluster division and improves progressively (Hartigan).

3. K-MEANS CLUSTERING

N objects are broken down into K distinct clusters using a method of repetition called K-means. K-means is probably the most recognized and often implemented partitioning-based method for clustering that uses center points for cluster visualization (Estivill). The quality of k-means clustering is assessed using the within-cluster squared error parameter [(J. MacQueen), (Yuan and Yang), (Hatie et al.)].

The K-means challenge can be solved using the algorithm known as K-means. There are several modifications, which are covered in the section after this one. To take advantage of either of the K-means algorithm's variations, one needs to first identify the number of clusters detected in the data; this may require numerous runs or trials in order to find the best number of clusters. There is no functional k-means method since the capacity to arrive at a global optimum is dependent upon the features of the data collection, the amount it contains, and the number of elements implicated in the cases. Two iteration phases make up the k-means clustering methods: the central point upgrade stage, in which the center points of the cluster are altered depending on the partition obtained by the previous stage and he positioning or initialization section, which is a continual process where every information point is allocated to its nearest the central point using the Euclidean metric units. When either a certain number of iterations have been made or no data point moves clusters, the iterative procedure eventually comes to an end (Slonim et al.).

Knowledge extraction, pattern identification, image processing, and other disciplines have all availed utilization of K-means clustering, the most commonly utilized clustering algorithm. K-means clustering separates n points of data into k clusters that constitute collections of related data points. Using a procedure that is iterative, each point gets allocated to the cluster whose centroid is closest. The centroid of these groups is subsequently established once more by averaging them. This initial process demonstrates the essential K-means clustering technique (Kijsipongse and U-ruekolan).

1. K randomly chosen centroids are picked from the dataset in order to create an initial grouping.
2. Measure the distance that lies among every centroid and the information point, and declare the centroid that is closest to the point of data as an individual.
3. To reconstitute the newly added cluster centroids, use an average of all the information elements allotted to the clusters.
4. Continue with step 2 till convergence.

The K-Means Clustering Initial Algorithm [14]

A clear explanation of Algorithm 1's behavior can be obtained by referring to Figure 2, which offers an example. Figure 2 depicts a schematic description of the K-means approach in progress. Initially there are two distinct categories of items that are associated. The centroids of the two sets are

figured out after that. Another time, the centroid is used for constructing the clusters, which produces unique database groupings. This approach is carried out unless the best clustering can be identified. Countless easily accessible resources are available for data analysis. Matlab, Orange, Tanagra, Statistica, Sas or Sas Professional Mining, Weka or Pentaho, and Rapid Miner are a few of them.

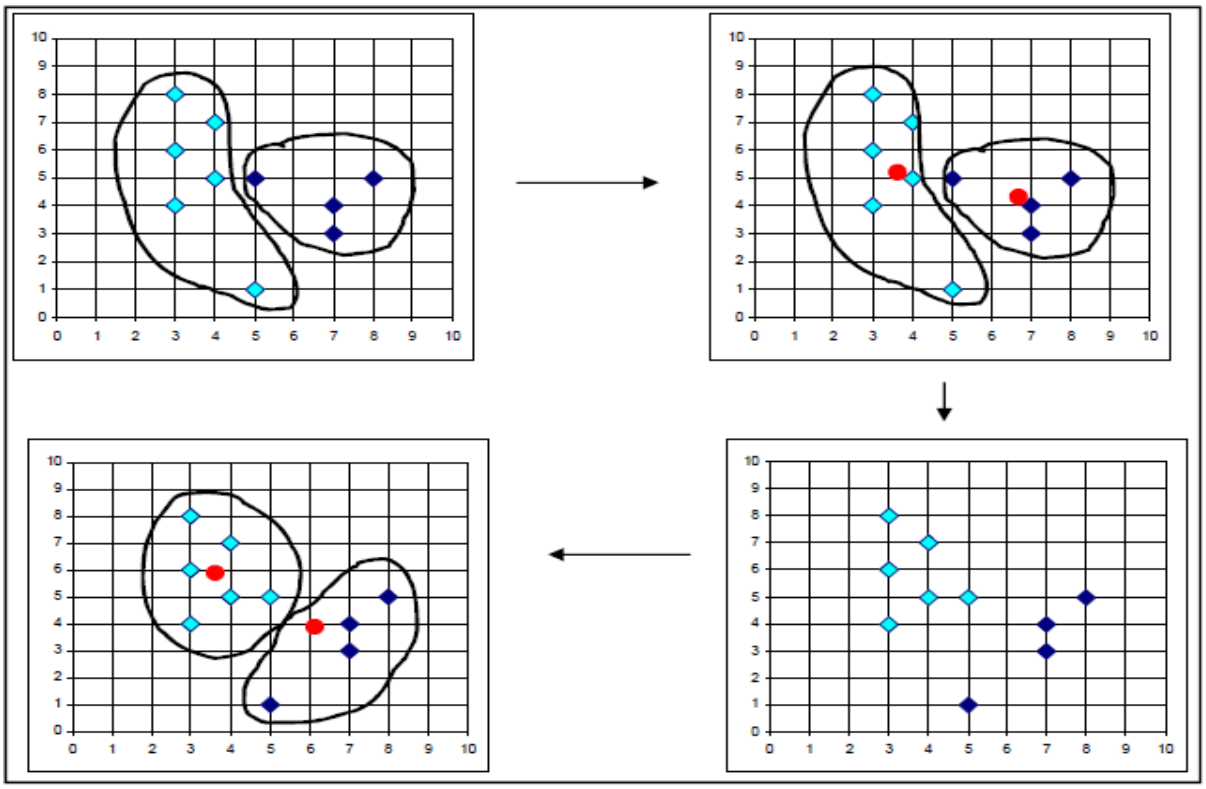


Figure 2 K-mean clustering process's functioning

4. EVOLUTION- LITERATURE SURVEY

K-means had been suggested by James MacQueen in 1967. This land has evidently underwent significant renovations. Each research project performed from 1967 to 1998 centered around implementing K-means into the statistical clustering business. The grouping using K-means followed by observing all of the alterations and advances. (Alsabti and others) They offered an efficient way to cluster by generating a structure in a k-d tree that makes it easier to identify the necessary layout. (Kamal and others) introduced a strategy that uses simultaneously labeled and unlabeled documents and is based on an estimation machine and classifiers. They came to an agreement that their algorithm produces better classification findings. Mike and companions displayed using the technique of clustering using K-means on an Annapolis wild star that exhibited a two-order-of-magnitude speeding.

(Kiri et al.) Designed a thorough approach that extracts context-relevant data from the K means system for clustering using limitations. (Tapas et al.) Applied Lloyd's clustering with K-means technique and referred to it as a filtering algorithm. (Cheung) has offered an improved version of J. Macqueen's K-means clustering methodology. Even in circumstances when the beginning number of groups is unresolved this approach facilitates rapid generation of clusters. (Jimenez et al.) Employed the LSI and K-means clustering approaches for carrying out a basic study examining the effects of semantic in the IR.

(Modha and Spangler) have discovered a strategy for integrating various, distinct feature regions in the algorithm for clustering K-means. (Tian and others) Conducts an examination of the simultaneous K-means clustering approach and indicates modified launching centers technology that decreases the number of steps required for grouping. (Pham and others) established the K-means grouping coefficient. They reached different findings on the total number of units across different databases. Wu Xindong et al. performed a survey. Their survey lists the top ten methods for data mining, along with their shortcomings as well as potential improvements in the future.

(Xiong et al.) Recently released results about the influence of a skewed distribution of information on K-means classification. They presented a well-organized analysis of K-means and cluster verification parameters from the point of view of data allocation. Actually, their primary objective was finding the relationships between K-means clustering and data distribution, in along with the temperature measure and the f-measure. Fuzzy

choice of features was put forward by (Xiuyun Li et al.) as an approach for enhancing K-means segmentation. They relied on the feature significant variable to figure out the corresponding contribution of each element to the categorization. (Osama Abu Abbas) executed an analysis comparing several data clustering methodologies. Yan Zhu et al. released a novel approach that uses clustering exemplars generated through affinity propagation for clustering initialization. In addition, they have lowered the clusters' total squared error. (Taoying Li et al.) Provided a novel method for fuzzy K-means clustering and ended up arriving at a conclusion with increased precision, lessened processing time, and improved efficiency. (Viet-Vu vu et al.) Have put forward an effective strategy for active seed selection that optimizes the coverage of the entire data set and is based on a min max approach. By employing a genetic algorithm, (Zhu and Wang) have offered a more effective clustering technique.

(Oyelade et al.) Utilized an assortment of understood characteristics for assessing the academic achievement of students through the use of the K-means clustering method. (Wu and Yao) applied their improved technique on a clustering examination of transiting data with the identical site name but locations that were distinct. (Yufen Sun et al.) Recently developed the universal K-means clustering methodology for identifying naturally occurring clusters in databases. They additionally showed a high level of precision in the results they achieved. It has been proven by (Wang and Yin) that their technique is more precise than the initial K-means clustering. The approach to clustering has been modified by (Li Xinwu) to offer greater dependability and reliability. Napoleon and Lakshmi undertook an analysis of the time needed for execution of the original K-means method for clustering along with their recommended K-means direction.

(Hesam et al.) Have made ramifications to the guided K-means approach algorithm that enables handling of astrophysical data bases. Provide an easy technique for grouping data points into clusters (Shi Na et al.). Their improved method operates with good accuracy in O (nk) time. As stated by (Shamir and Tishby), K-means remains stable in the regime of huge sample. The most contentious part of clustering, or the ideal amount of clusters to choose, is addressed by (Mark and Boris).

An approach for PCA formed clustering of k-means of partial databases was recently proposed in (Honda et al.). They reached the conclusion that K-means clustering with PCA assistance is highly resistant than K-means clustering method along with PDS. (Sathya et al.) Have offered a method for rapidly finding the cluster search, whereas the comparison depends on the query's co-occurrence word and document similarity. A novel algorithm, kMkNN, has been proposed by (Xueyi Wang) to solve the closest neighbor seeking problem. He has thought of applying triangle inequality and K-means clustering. A K-means strategy of clustering based on coefficients of variability has been put forward (Ren and Fan). They exhibit how the approach they use outperforms the K-means algorithm of clustering process with respect to of results.

A pair of methods for centering initialization was evaluated by (Son and Anh). There are two kinds of them: the kd tree and the CF tree. (Thomas A. Runkler) focussed on partially controlled clustering while creating a partially guided k-harmonic means clustering technology. A decentralized K-means technique has been suggested by (Fatta et al.). They ultimately reached the conclusion that the suggested course of action was efficient as well as precise. The hybrid K-means clustering approach was initially introduced by Zhang and Murugesan. K-means and UPGMA can be separated using top-down and bottom-up approaches, respectively. (Sarma et al.)'s efficient K-means clustering performance confirmed useful for massive databases. They are now aware of how their approach improves up kernel K-means classification. (Wang and Su), who completed their test results on iris, wine, and abalone information sets, changed the clustering technique. They enhanced the algorithmic processes with regard to of both outcome accuracy and duration.

(Tripathy et al.) Put forward a technique for the conventional kernel-based K-means clustering procedure, which would afterward use the rough set thoughts to update the central point value. (Abhay et al.) Established a model for yes/no results estimating in K-means clustering using meteorological data. The maximal triangle rule was proposed by Feng et al. in order to enhance the K-means clustering algorithm. By employing KMTR to improve clusters, they have solved K-means's difficulty. Parallel K-means are being developed by (Ekasit et al.) on GPU clusters. They utilize the task pool methodology for dynamic load rebalancing to spread the workload equitably among the several GPUs connected to the clusters, improving the performance of the parallel K-Means at the inter-node basis.

(Jing et al.) Established a closure-based K-means clustering methodology that is easy to understand easily parallelized, and highly efficient. A clustering algorithm on self-adaptive weights has been proposed by (Zhang et al.). The final result of their experiment indicates that it is more consistent and accurate. (Mahmud et al.) Demonstrated how the approach could be strengthened by eliminating the primary plant seeds point restriction by employing an average with weights. They have additionally decreased the amount of repetitions in the clustering strategy. (Wang et al.) Employed the density concept to improve K-means clustering. They were capable to get higher requirements function E and clustering precision.

(Lee and Lin) have developed a K-means algorithm for selection and elimination. The authors employed an excessive amount of clusters to improve efficiency. Patil and Vaidya undertook a review of a number of clustering methods. The cluchunk system was put forward by (Cheng et al.). The unprocessed web data that

includes chunklet information has been organized by this system. (Bikram and others) Employing DI and DBI parameters for cluster validation, we modified the standard K-means clustering approach. The execution of a quantum-based K-means clustering by (Ellis et al.) exhibits the gains in accuracy and precision. (Nuno and collaborators) made use of the interconnected community frameworks of a tag chain to boost the K-means clustering of text methods.

(Anoop and Satyam) conducted an assessment on the most commonly utilized methods for clustering in data extraction in the past few decades. (Vijayalakshmi and Renuka) highlighted about the different methods and aspects connected with distinct algorithms used for clustering. They even talked about challenges with various clustering algorithms that have been applied to big datasets. (Kurt et al.) Presented necessary extensions and spherical K-means clustering. Furthermore, the R extension package skmeams has been released. (Deepti et al.) Conducted research on various clustering methods, covering an overview, applications, disadvantages and requirements. (Maryam et al.) Investigated every kind of clustering in order to identify which is the most effective in discovering duplicate items. An upgraded K-means clustering method to handle two-dimensional data has been offered by (Rupali and Suresh).

An approach that evaluates the security of clustering algorithms in aggressive circumstances has been presented by (Biggio et al.). (Khadem and others) provide a survey on methods and instruments for data mining. A technique for categorizing visitors to websites using the ART1 neural network-based clustering approach algorithm was laid out by Yogish and Raju. This approach's performance will be analyzed along with the K-means and SOM methods of clustering. (Silva et al.) Highlighted the findings from the channel clustering analysis conducted just recently. Ichikawa and Morishita have developed a novel and user-friendly technology that decreases information technology time and integrates a heuristic component. (Pattabiraman and others) Forum threads were clustered applying three different methods for clustering, and the advancement of accuracy was investigated.

A new word, "MFCC," or Multitype Feature Co-selected for Clustering, was recently introduced by (Parimala and Palanisamy). The technique makes use of numerous function class types to cluster web pages. In addition, the creators addressed a few search engine challenges. AFS global K-means algorithm was published by (Wang et al.). In this approach, the initial cluster center is established utilizing the distance based on the neighborhood of the AFS topology. The K-means technique clustering algorithm's execution time was minimized by (Lin and Lee).

A hybrid approach based on a prototype was previously proposed by (Sarma et al.) in order to speed up the traditional K-means clustering method. To boost cluster matching, (Lam et al.) have created a PSO basis method of clustering for gene extraction. (Sharma and Fotedar) examined the various data mining methods employed for software development projects estimation. (Krey et al.) Have set up order-constrained substances in K-means as a more consistent tool for clustering audio components. (Su and Huwang) applied statistical evaluation of data for changing the standard K-means algorithm. To deal with clustering difficulties (Xue and Liu) recommended an original approach that combines membrane technology together with the K-means algorithmic design.

5. LIMITATIONS

There are numerous issues with K-means clustering that need to be overcome. Several investigators experienced certain limitations when using the K-means algorithm for their research. We talk over a few of the regular constraints below.

- **Outliers**

Several investigators have reported that when there are outliers in the data, there will be differences in the outcomes, which is indicating that there are no uniform outcomes based on different executions on the same data. These kinds of items are known as outliers; they exist in the set of data but have no effect on clusters to form. In addition, within clusters, outliers could increase the total of squared mistakes. Thus, it is very essential to separate irregularities from the information collected. By employing preprocessing techniques on the original dataset, outliers can be reduced (Kumar et al.).

- **Number of Clusters**

With the use of the K-means clustering technique, calculating out how many clusters one should begin with wasn't ever easy. Selecting the right amount of clusters at the outset is essential. It has been noticed that the number of classes contained within the dataset is sometimes utilized for figuring out the number of clusters. The matter of how many groups should be granted remained controversial (A. Abbas, 2008).

- **Vacant clusters**

The empty clusters occur whenever a cluster obtains no points during the assignment step. A previous version of traditional clustering (k-means) methodology exists a similar problem (Kumar et al.).

- **Non-globular dimensions and patterns**

The K-means clustering algorithm delivers results that aren't ideal if the clusters have shapes that are irregular,

different densities, and a variety of sizes. The convex forms of the clusters that occur consistently pose an obstacle (Patil and Vaidya).

6. APPLICATIONS

Clustering techniques include a wide range of possible uses in the scientific, financial, medical treatment, communications, and Web domains. These applications are addressed in some detail below.

- **Applying a Clustering Algorithm to recognize Cancerous Data**

A dataset's cancerous data record can be identified through the clustering algorithm. Several users explored with this application, evaluating known dataset samples as aggressive or non-cancerous. Using a random mixture of the data samples, different methods to cluster were used. In order to identify which samples were correctly clustered, the clustering result was reviewed. Since sample labels were initially recognized, clustering accuracy can be determined very easily (Girolami and He), (Wang and Garibaldi).

- **Search Engine Mechanism for The Clustering Method**

A crucial aspect of search engine operations is the clustering approach. It will thereby act as a framework for search engines. Similar objects have been put together in one cluster by search engines, while dissimilar ones are placed together in group. The success or failure of the clustering processes influences how well Search Engines function. A more effective clustering technique improves the possibility of achieving the outcomes you want on the top page (Liu et al.).

- **Clustering Algorithm in Academics**

Monitoring the academic advancement of learners has been an essential concern for the higher education community as a whole. The process of clustering makes it simple to handle the issue at hand. The pupils' outcomes are used to divide them into various clusters, each of which reflects a distinct performance level. The average performance of a class as a whole can be calculated by collecting the number of learners in each cluster. (Oyelade and others).

- **The Algorithm for Clustering in Wireless Sensing Network Software Applications**

Applications that are based on wireless sensor networks are capable of employing the clustering algorithm. It's suitable for the detection of landmines. One of the main elements of the clustering process involves recognizing the members of a cluster who assemble all the data in that specific cluster. (Akkaya and others), (Naik and Saurabh)

7. CONCLUSION

In the present article, we have carried out a study on various scholars' work using the K-means clustering strategy. We examined current k-means clustering approaches. This research project demonstrates the variety of k-means clustering algorithms that have been invented since the 1960s to address various k-means algorithmic shortcomings.

We also discussed about the K-means clustering approach past events, obstacles, as well as applicability. Furthermore, we highlighted that the K-means method algorithm's performance has been substantially enhanced over the years that have passed. The most emphasis is put into developing the clusters more precise and productive. There's constantly space for improvement in this field. It's rarely simple deciding on the right variety of clusters to start with. Ultimately, it comes to light that even with the significant amount of work carried out on the K-means clustering approach, there is still an opportunity for improvement. We'll analyze the computational time complexity of various k-means clustering algorithm modifications in the future, as well as a study regarding their relative efficiency and its accuracy.

8. REFERENCES

1) K. Jain and R. Dubes, Algorithms for Clustering Data. Englewood Cliffs, NJ: Prentice-Hall, (1988).
2) A. K. Jain, prof. S. Maheshwari, “survey of recent clustering techniques in data mining”, international journal of computer science and management research, vol 1 issue 1 Aug 2012.
3) A. Kumar, R. Sinha, V. Bhattacherjee, D. S. Verma, S. Singh, “modeling using K-means clustering algorithm”, IEEE 2012, 1st international conference on recent advances in information technology(RAIT).
4) A. Saurabh, A. Naik, "Wireless sensor network based adaptive landmine detection algorithm, " 2011 3rd International Conference on Electronics Computer Technology (ICECT), vol.1, no., pp.220, 224, 8-10 April 2011.
5) B. Biggio, I. Pillai, S. R. Bulò, D. Ariu, M. Pelillo, F. Roli, “is data clustering in adversial settings secure?”, Proceedings of the 2013 ACM workshop on Artificial intelligence and security, Pages 87-98.
6) B. Everitt, S. Landau, M Leese. and D. Stahl, Cluster Analysis, 5th ed., John Wiley and Sons, (2011).
7) B. K Tripathy, A. Ghosh, G.K. Panda, "Kernel based K-means clustering using rough set, " 2012 International Conference on Computer Communication and Informatics (ICCCI), vol., no., pp.1, 5, 10-12 Jan. 2012.

8) B. K. Mishra, N. R. Nayak, A. Rath, S. Swain, "far efficient K-means clustering algorithm", Proceedings of the 2012 ACM's International Conference on Advances in Computing, Communications and Informatics, Pages 106-110.
9) B. Mirkin, Clustering: A Data Recovery Approach, 2nd ed., Chapman and Hall/CRC, (2013).
10) C. Romesburg, Cluster Analysis for Researchers, London, Wadesworth, (1984).
11) C. Yuan and H. Yang, "Research on k-value selection method of k-means clustering algorithm",. Multidisciplinary Scientific Journal, 2019, 2(2), 226-235.
12) D. Jimenez, E. Ferretti, V. Vidal, P. Rosso, and C. F. Enguix, "The influence of semantics in IR using LSI and K-means clustering techniques", ACM Proceedings of the 1st international symposium on Information and communication technologies, Pages 279-284.
13) D. Napoleon, P.G. Lakshmi, "An efficient K-Means clustering algorithm for reducing time complexity using uniform distribution data points, " Trendz in Information Sciences & Computing (TISC), IEEE 2010, vol., no., pp.42, 45, 17-19 Dec. 2010.
14) D. S. Modha, W. S. Spangler, "Feature Weighting in k-Means Clustering", ACM Journal of Machine Learning, Volume 52 Issue 3, September 2003, Pages 217 – 237.
15) D. Sisodia, L. Singh, S. Sisodia, K. Saxena, "Clustering Techniques: A Brief Survey of Different Clustering Algorithms", International Journal of Latest Trends in Engineering and Technology (IJLTET), Vol. 1 Issue 3 September 2012.
16) D. T. Pham, S. S. Dimov, and C. D. Nguyen, "Selection of K in K-means clustering", Proc. IMechE Vol. 219 Part C: J. Mechanical Engineering Science, IMechE 2005.
17) E. A. Khadem, E. F. Nezhad, M. Sharifi, "Data Mining: Methods & Utilities", Researcher2013; 5(12):47-59. (ISSN: 1553-9865).
18) E. Casper, C. C. Hung, E. Jung, and M. Yang, "A quantum-modeled K means clustering algorithm for multi-band image segmentation", Proceedings of the 2012 ACM Research in Applied Computation, Pages 158-163.
19) E. Kijsipongse, S. U-ruekolan, "Dynamic load balancing on GPU clusters for large-scale K-Means clustering, " 2012 IEEE International Joint Conference on Computer Science and Software Engineering (JCSSE), vol., no., pp.346, 350, May 30 2012-June 1 2012.
20) G. D. Fatta, F. Blasa, S. Cafiero, G. Fortino, "Epidemic K-Means Clustering, "2011 IEEE 11th International Conference on Data Mining Workshops (ICDMW), vol., no., pp.151, 158, 11-11 Dec. 2011.
21) H. E. Driver and A. L. Kroeber , "Quantitative Expression of Cultural Relationships," University of California Publications of American Archaeology and Ethnology, 1932, vol. 4, pp.211-256.
22) H. Sp´th, Cluster Analysis Algorithms. West Sussex, UK: Ellis Horwood Limited, (1980).
23) H. Sp´th, Cluster Dissection and Analysis: theory, FORTRAN programs, examples. (Translator: Johannes Goldschmidt.) Ellis Horwood Ltd Wiley, Chichester, (1985).
24) H. Xiong; J. Wu; J. Chen, "K-Means Clustering Versus Validation Measures: A Data-Distribution Perspective, " IEEE Transactions on Systems, Man, and Cybernetics, Part B: Cybernetics, vol.39, no.2, pp.318, 331, April 2009.
25) H. Xiuchang, SU Wei, "An Improved K-means Clustering Algorithm", JOURNAL OF NETWORKS, VOL. 9, NO. 1, JANUARY 2014.
26) H.K. Yogish, G.T. Raju, "Clustering of Preprocessed Web Usage Data Using ART1 Neural Network and Comparative Analysis of ART1, K-Means and SOM Clustering Techniques," 2013 5th IEEE International Conference on Computational Intelligence and Communication Networks (CICN), vol., no., pp.322, 326, 27-29 Sept. 2013.
27) H.T Dashti, T Simas, R.A Ribeiro, A Assadi, A Moitinho, "MK-means- Modified K-means clustering algorithm, " The 2010 IEEE International Joint Conference on Neural Networks (IJCNN), vol., no., pp.1, 6, 18-23 July 2010.
28) J. A. Hartigan ,Clustering Algorithms, John Wiley & Sons. Inc., New York, (1075).
29) J. A. Silva, E. R. Faria, R. C. Barros, E. R. Hruschka, A. C. P. L. F. d. Carvalho, J. Gama, "Data stream clustering: A survey", ACM Computing Surveys (CSUR), Volume 46 Issue 1, October 2013, Article No. 13.
30) J. Feng; Z. Lu; P. Yang; X. Xu, "A K-means clustering algorithm based on the maximum triangle rule, " 2012 IEEE International Conference on Mechatronics and Automation (ICMA), vol., no., pp.1456, 1461, 5-8 Aug.
31) J. MacQueen, "Some methods for classification and analysis of multivariate observations." Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability, Volume 1: Statistics, 281--297, University of California Press, Berkeley, Calif., 1967.
32) J. MacQueen, "Some methods for classification and analysis of multivariate observations". In: Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability, 1967, vol 1, 281-297.
33) J. Wang; J. Wang; Q. Ke; G. Zeng; S. Li, "Fast approximate k-means via cluster closures, " 2012 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), vol., no., pp.3037, 3044, 16-21 June 2012.
34) J. Wang; X. Su, "An improved K-Means clustering algorithm," 2011 IEEE 3rd International Conference on

Communication Software and Networks (ICCSN), vol., no., pp.44, 46, 27-29 May 2011.
35) J. Xue, X. Liu, "A K-nearest Based Clustering Algorithm by P Systems with Active Membranes", Journal of Software, Vol 9, No 3 (2014), 716-725, Mar 2014.
36) J. Zhu; H. Wang, "An improved K-means clustering algorithm, " 2010 The 2nd IEEE International Conference on Information Management and Engineering (ICIME), vol., no., pp.190, 192, 16-18 April 2010.
37) J. Zubin, "A technique for measuring like-mindedness", The Journal of Abnormal and Social Psychology. 1938, vol. 4. pp508-516.
38) K. Akkaya, F. Senel, B. McLaughlan, "Clustering of Wireless Sensor and Actor Networks based on Sensor Distribution and Inter-actor Connectivity" ACM Journal of Parallel and Distributed Computing, volume 69, Issue 6, June, 2009, Pages 573-587.
39) K. Alsabti, S. Ranka, V. Singh, "An efficient k-means clustering algorithm", 1998 Proceedings of IPPS/SPDP Workshop on High Performance Data Mining.(date location).
40) K. Fukunaga, Introduction to Statistical Pattern Recognition, 2nd ed., Academic Press, (1990).
41) K. Honda, R. Nonoguchi, A. Notsu, H. Ichihashi, "PCA-guided k-Means clustering with incomplete data, " 2011 IEEE International Conference on Fuzzy Systems (FUZZ), vol., no., pp.1710, 1714, 27-30 June 2011.
42) K. Hornik, I. Feinerer, M. Kober, C. Buchta, "Spherical k-Means Clustering", Journal of Statistical Software, September 2012, Volume 50, Issue 10.
43) K. Ichikawa, S. Morishita, "A simple but powerful heuristic method for accelerating k-means clustering of large-scale data in life science," 2013 IEEE Transactions on Computational Biology and Bioinformatics, vol.PP, no.99, pp.1, 1.
44) K. Murugesan, J. Zhang, "Hybrid Bisect K-Means Clustering Algorithm, " 2011 International Conference on Business Computing and Global Informatization (BCGIN), vol., no., pp.216, 219, 29-31 July 2011.
45) K. NIGAM, A. K. MCCALLUM, S. THRUN, T. MITCHELL, "Text Classification from Labeled and Unlabeled Documents using EM", ACM journal of Machine Learning Special issue on information retrieval 1999.
46) K. Parimala, Dr. V. Palanisamy, " Enhanced Performance of Search Engine with Multitype Feature Co-Selection of K-Means Clustering Algorithm ", IJASCSE, Vol 2, Issue 1, 2013.
47) K. Pattabiraman, P. Sondhi, C. Zhai, "exploiting forum thread structures to improve clustering", ACM Proceedings of the 2013 Conference on the Theory of Information Retrieval, Pages 15.
48) K. Wagsta, C. Cardie, S. Rogers, S. Schroedl, "Constrained K-means Clustering with Background Knowledge", ACM Proceedings of the Eighteenth International Conference on Machine Learning, 2001, p. 577-584.
49) L. Kaufman and P. J. Rousseeuw, Finding Groups in Data: An Introduction to Cluster Analysis, John Wiley & Sons, Inc. New York, (1990).
50) L. Wang, X. Liu, Y. Mu, "The Global k-Means Clustering Analysis Based on Multi-Granulations Nearness Neighborhood", Mathematics in Computer Science, March 2013, Volume 7, Issue 1, pp 113-124.
51) L. Xinwu, "Research on text clustering algorithm based on improved Kmeans, " 2010 International Conference on Computer Design and Applications (ICCDA), vol.4, no., pp.V4-573, V4-576, 25-27 June 2010.
52) M. Bakhshi, M.R.F. Derakhshi, E. Zafarani, "Review and Comparison between Clustering Algorithms with Duplicate Entities Detection Purpose", International Journal of Computer Science Emerging Technology, Vol-3, No 3, June, 2012.
53) M. Estlick, M. Leeser, J. Theiler, J.J. Szymanski, "algorithmic transformations in the implementation of K-means clustering on reconfigurable hardware", Proceedings of the 2001 ACM/SIGDA ninth international symposium on Field programmable gate arrays, Pages 103 110.
54) M. Girolami, C. He, "Probability density estimation from optimally condensed data samples, " IEEE Transactions on Pattern Analysis and Machine Intelligence, vol.25, no.10, pp.1253, 1264, Oct. 2003.
55) M. R. Anderberg, "Cluster Analysis for Applications", New York: Academic Press, (1973).
56) M. S. Aldenderfer, and, R. K.. Blashfield, Cluster Analysis. Beverly Hills, CA: Sage Publications, (1984).
57) M. Sathya, J. Jayanthi, N. Basker, "Link based K-Means clustering algorithm for information retrieval, " 2011 IEEE International Conference on Recent Trends in Information Technology (ICRTIT), vol., no., pp.1111, 1115, 3-5 June 2011.
58) M. Sharma, N. Fotedar, "Software Effort Estimation with Data Mining Techniques-A Review", International journal of engineering sciences and research technology, 3(3): March, 2014, ISSN: 2277-9655.
59) M. Vijayalakshmi, M.R. Devi, " A Survey of Different Issue of Different clustering Algorithms Used in Large Data sets", International Journal of Advanced Research in Computer Science and Software Engineering, Volume 2, Issue 3, March 2012 ISSN: 2277 128X.
60) M.M.T. Chiang, B. Mirkin, "Intelligent Choice of the Number of Clusters in K-Means Clustering: An Experimental Study with Different Cluster Spreads", Journal of Classification, March 2010, Volume 27, Issue 1, pp 3-40.

61) M.S. Mahmud, M.M. Rahman, M.N. Akhtar, "Improvement of K-means clustering algorithm with better initial centroids based on weighted average, " 2012 7th IEEE International Conference on Electrical & Computer Engineering (ICECE), vol., no., pp.647, 650, 20-22 Dec. 2012
62) N. Cravino, J. Devezas, Á. Figueira, "Using the overlapping community structure of a network of tags to improve text clustering", Proceedings of the 23rd ACM conference on Hypertext and social media, Pages 239-244.
63) N. Jardine and R Sibson,. Mathematical Taxonomy, John Wiley and Sons, Ltd, Chichester, (1871)
64) N. Slonim, E. Aharoni and K. Crammer, "Hartigan's K-Means Versus Lloyd's K-Means-Is It Time for a Change? Proceedings of the Twenty-Third International Joint Conference on Artificial Intelligence, 2013, pp. 1677-1684.
65) N. T. Son; D. T. Anh, "Two Different Methods for Initialization the I-k-Means Clustering of Time Series Data", 2011 Third International Conference on Knowledge and Systems Engineering (KSE), vol., no., pp.3, 10, 14-17 Oct. 2011.
66) O. A. Abbas, "comparisons between data clustering algorithms", the international Arab journal of information technology, vol. 5, no. 3, July 2008.
67) O. J Oyelade, O. O Oladipupo, I. C Obagbuwa, "Application of k-Means Clustering algorithm for prediction of Students' Academic Performance", (IJCSIS) International Journal of Computer Science and Information Security, Vol. 7, _o. 1, 2010.
68) O. Shamir, N. Tishby, "Stability and model selection in k-means clustering", Machine Learning, September 2010, Volume 80, Issue 2-3, pp 213-243.
69) P. Berkhin, A Survey of Clustering Data Mining Techniques. In: Kogan J., Nicholas C., Teboulle M. (eds) Grouping Multidimensional Data, Springer, Berlin, Heidelberg, (2006).
70) R. A. Johnson and D. W. Wichern, Applied Multivariate Statistical Analysis, 5th ed.,, Eaglewood Cliffs, NJ: Prentice- Hall, (2002).
71) R. C. Tryon and D. E. Bailey, Cluster Analysis,. McGraw Hill, New York, (1970)
72) R. Cattell, " and other coefficients of pattern similarity", Psychometrika, 1949, vol. 4. pp.279-298.
73) R. R. Sokal and P. H. A.. Sneath, Principles of Numerical Taxonomy. San Francisco: California, (1963).
74) R. Vij, S. Kumar, "Improved k-means clustering algorithm for two dimensional data", ACM Proceedings of the Second International Conference on Computational Science, Engineering and Information Technology, Pages 665-670.
75) S. Broh´e and J. V. Helden, "Evaluation of clustering algorithms for protein-protein interaction networks", BMC Bioinformatics, (2006), 7(1): 488.
76) S. Krey, U. Ligges, F. Leisch, "Music and timbre segmentation by recursive constrained K-means clustering", Computational Statistics, February 2014, Volume 29, Issue 1-2, pp 37-50.
77) S. Na; L. Xumin; G. Yong, "Research on k-means Clustering Algorithm: An Improved k-means Clustering Algorithm, " 2010 Third IEEE International Symposium on Intelligent Information Technology and Security Informatics (IITSI), vol., no., pp.63, 67, 2-4 April 2010.
78) S. Ren; A. Fan, "K-means clustering algorithm based on coefficient of variation, " 2011 4th International Congress on Image and Signal Processing (CISP), vol.4, no., pp.2076, 2079, 15-17 Oct. 2011.
79) S.S. Lee, J.C. Lin, "An accelerated K-means clustering algorithm using selection and erasure rules", Journal of Zhejiang University SCIENCE C, October 2012, Volume 13, Issue 10, pp 761-768.
80) S.S. Lee, J.C. Lin, "Fast K-means clustering using deletion by center displacement and norms product (CDNP)", Journal of Pattern Recognition and Image Analysis, Volume 23 Issue 2, April 2013, Pages 199-206.
81) T. H. Sarma, P. Viswanath, B. E. Reddy, "A hybrid approach to speed-up the k-means clustering method", International Journal of Machine Learning and Cybernetics, April 2013, Volume 4, Issue 2, pp 107-117.
82) T. Hastie, R. Tiiranibsh and J. Friedman., The Elements of Statistical Learning: Data Mining, Inference and Prediction, 2nd ed., Springer-Verlag, (2009).
83) T. Jinlan, Z. Lin, Z. Suqin, L. Lu, "Improvement and Parallelism of k-Means Clustering Algorithm", Tsinghua Science & Technology, Volume 10, Issue 3, June 2005, Pages 277–281.
84) T. Kanungo, D. M Mount, N.S. Netanyahu, C.D. Piatko, R. Silverman, A. Y. Wu, "An efficient k-means clustering algorithm: analysis and implementation, " IEEE Transactions on Pattern Analysis and Machine Intelligence, vol.24, no.7, pp.881, 892, Jul 2002.
85) T. Li; Y. Chen; X. Mu; M. Yang, "An improved fuzzy k-means clustering with k-center initialization, " 2010 Third IEEE International Workshop on Advanced Computational Intelligence (IWACI), vol., no., pp.157, 161, 25-27 Aug. 2010.
86) T. Liu, C. Rosenberg, H.A. Rowley, "Clustering Billions of Images with Large Scale Nearest Neighbor Search, " Applications of Computer Vision, 2007. IEEE Workshop on WACV '07, vol., no., pp.28, 28, Feb. 2007.
87) T.A Runkler, "Partially supervised k-harmonic means clustering, " 2011 IEEE Symposium on Computational

Intelligence and Data Mining (CIDM), vol., no., pp.96, 103, 11-15 April 2011.
88) T.H Sarma, P. Viswanath, B.E. Reddy, "A fast approximate kernel k-means clustering method for large data sets, " Recent Advances in Intelligent Computational Systems (RAICS), 2011 IEEE, vol., no., pp.545, 550, 22-24 Sept. 2011.
89) V. Estivill-Castro Why so many clustering algorithms: a position paper. ACM SIGKDD Explorations Newsletter, 2002, 4(1), 65-75.
90) V.V. Vu, N. Labroche, B.B. Meunier, "Active Learning for Semi-Supervised K-Means Clustering, " 2010 22nd IEEE International Conference on Tools with Artificial Intelligence (ICTAI), vol.1, no., pp.12, 15, 27-29 Oct. 2010.
91) W. D. Fisher, Clustering and Aggregation in Economics, Johns Hopkins Press, Baltimore, Maryland., (1968)
92) W. Min; Y. Siqing, "Improved K-means clustering based on genetic algorithm, " 2010 IEEE International Conference on Computer Application and System Modeling (ICCASM), vol.6, no., pp.V6-636, V6-639, 22-24 Oct. 2010.
93) W. Yintong; L. Wanlong; G. Rujia, "An improved k-means clustering algorithm," World Automation Congress (WAC), 2012, vol., no., pp.1, 3, 24- 28 June 2012.
94) X. Li; J. Yang; Q. Wang; J. Fan; P. Liu, "Research and Application of Improved K-Means Algorithm Based on Fuzzy Feature Selection, " FSKD '08. Fifth International Conference on Fuzzy Systems and Knowledge Discovery. vol.1, no., pp.401, 405, 18-20 Oct. 2008.
95) X. Wang, "A fast exact k-nearest neighbors algorithm for high dimensional search using k-means clustering and triangle inequality", The 2011 International Joint Conference on Neural Networks (IJCNN),, vol., no., pp.1293, 1299, July 31 2011-Aug. 5 2011.
96) X. Wu, C. Yao, "Application of improved K-means clustering algorithm in transit data collection," 2010 3rd International Conference on Biomedical Engineering and Informatics (BMEI), vol.7, no., pp.3028, 3030, 16-18 Oct. 2010.
97) X. Wu, V. Kumar, J. R. Quinlan, J. Ghosh, Q. Yang, H. Motoda, G. J. McLachlan, A. Ng, B. Liu, P. S. Yu, Z.H. Zhou, M. Steinbach, D.J. Hand, D. Steinberg, "Top 10 algorithms in data mining", Knowledge and Information Systems, January 2008, Volume 14, Issue 1, pp 1-37.
98) X. Y. Wang, J. M. Garibaldi. "A comparison of fuzzy and non-fuzzy clustering techniques in cancer diagnosis", Proceedings of the 2nd International Conference on Computational Intelligence in Medicine and Healthcare, 2005.
99) Y. Cheng, Y. Xie, K. Zhang, A. Agrawal, A. Choudhary, "CluChunk: clustering large scale user-generated content incorporating chunklet information", ACM Proceedings of the 1st International Workshop on Big Data, Streams and Heterogeneous Source Mining: Algorithms, Systems, Programming Models and Applications, Pages 12-19. 2012.
100) Y. S. Patil, M.B. Vaidya, "A Technical Survey on cluster analysis in data mining", International Journal of Emerging Technology and Advanced Engineering, ISSN 2250-2459, Volume 2, Issue 9, September 2012.
101) Y. Sun; G. Liu; K. Xu, "A k-Means-Based Projected Clustering Algorithm, " 2010 Third International Joint Conference on Computational Science and Optimization (CSO), vol.1, no., pp.466, 470, 28-31 May 2010.
102) Y. Zhang, H. Shi, D. Zhang, "K-means clustering based on self-adaptive weight, " 2012 2nd International Conference on Computer Science and Network Technology (ICCSNT), vol., no., pp.1540, 1544, 29-31 Dec. 2012.
103) Y. Zhu, J. Yu, C. Jia, "Initializing K-means Clustering Using Affinity Propagation, " Ninth International Conference on Hybrid Intelligent Systems, 2009. HIS '09. vol.1, no., pp.338, 343, 12-14 Aug. 2009.
104) Y.K. Lam, P. W. M. Tsang, C.S. Leung, "PSO-based K-Means clustering with enhanced cluster matching for gene expression data", Neural Computing and Applications, June 2013, Volume 22, Issue 7-8, pp 1349-1355.
105) Y.M. Cheung, "k*-Means: A new generalized k-means clustering algorithm", Pattern Recognition Letters, Volume 24, Issue 15, November 2003, Pages 2883-2893.