Grenze International Journal of Engineering and Technology, Jan Issue

Analytical Review of K-Means based Algorithms and Evaluation Methods

Trushali Jambudi[1] and Savita Gandhi[2]
[1-2]Department of Computer Science, Gujarat University, Ahmedabad, India
Email: talk2trushali@gmail.com, drsavitagandhi@gmail.com

***Abstract*—The traditional k-Means clustering algorithm suffers from many limitations and various researchers have proposed many revisions to improve its performance. In this paper we describe the k-Means clustering algorithm and explain its major limitations. We further study the revisions introduced by the researchers to the k-Means algorithm and explore in detail the parameters that are used to evaluate the performance of k-Means based clustering algorithms through experiments on several data sets.**

***Index Terms*— k-Means, Data clustering, Data mining.**

I. INTRODUCTION

Clustering [1][2] which is the grouping of data into clusters, forms these clusters such that objects within a cluster are extremely similar and are dissimilar to the objects in other clusters. Clustering is often the first data mining task applied on a given collection of data and is used to explore if any underlying patterns exist in the data. Examples of clustering tasks include dividing the plants and animals into groups, symptoms into disease, student groups, crime pattern recognition, etc. Clustering algorithms should perform well according to the metrics of: Clustering time, Main memory usage, Scalability, Ability to handle noise/outliers, Data order independence, Ability to discover arbitrary shaped clusters, Minimal requirements of domain knowledge to determine input parameters, Ability to deal with different types of attributes, High dimensionality, Constraint based clustering.

A. The k-Means Algorithm

k-Means [3] [4] which is a popular and simple partition based clustering algorithm [5], attempts grouping data objects into *k* number of clusters based on their similarity.
The k-Means algorithm is as follows:

1. Input *k*, the number of clusters to be formed
2. *k* initial "centres" are initialized at random from the domain of the data.
3. *k* clusters are formed by placing every observation to its nearest centre.
4. The following steps are repeated until cluster stops changing:
    a. For each cluster, the mean of the observations within that cluster is taken as the centre.
    b. Each data point is reassigned to the cluster whose centre is nearest to that data point

The k-Means algorithm has a rich set of possible applications [4] [20]. In principle, k-Means can be applied where you have several objects and each object has several attributes and you want to classify the objects based on the attributes.

Grenze ID: 01.GIJET.7.1.8_1

B. Characteristics of k-Means Algorithm

Table I summarizes various characteristics of k-Means algorithm and gives a general idea regarding its various obvious limitations; from table I it becomes clear that the original k-Means algorithm can be applied to the data sets of only numeric data. Also, the algorithm requires the number of clusters to be formed and the initial cluster centres as input; the performance of the algorithm will depend on the correctness of these input parameters which again depend on the level of correctness of the users' domain knowledge. Another observation is that the algorithm is sensitive to the presence of outliers in the data being clustered and is incapable of detecting the outliers in the data and if an outlier gets selected as a cluster centre, it may affect the convergence of the algorithm. If the number of data points in the data set being clustered are meagre, the initial grouping will greatly influence the cluster formations [6] [25] [26] [20].

TABLE I. THE CHARACTERISTICS OF TRADITIONAL K-MEANS ALGORITHM

Data Type support	Initial k Value [User Input/ Automatic/ Methodical	Sensitive to the presence of Outlier [Yes/No]	Outlier Detection [Local/ Global/ Both/ None	Automatic Initialization Of values [Yes/No]	Inputs to the algorithm
Numeric	Input by user	Yes	None	No	No. of clusters K, Initial cluster centres

Various attempts for improvements to the k-Means algorithm have been proposed by researchers to remove/ decrease algorithm's dependency on input parameters and sensitivity towards outliers [31], and are the scope of study and analysis for this paper.

II. LITERATURE STUDIES

In this section we study the work done in the direction of improving the k-Means algorithm itself as well as the various extensions to the k-Means algorithm in order to improve the results of applying the algorithm on the data sets being studied.

A. Determining k, the number of clusters in k-Means

The traditional k-Means algorithm requires k, i.e., the number of cluster partitions as input from the user. This can be a challenging task if the user is not familiar with the data he is working on [5]. In the literature many researchers have proposed methods to overcome this limitation of the k-means algorithm wherein the value of k can be automatically determined rather than accepting it as input from the user. Das et al. [19], proposed algorithm ACDE which does not require any knowledge of the data being clustered and dynamically generates optimal number of clusters. However, ACDE uses the classical DE method which requires user inputs for determining the threshold values for predicting k. AM Mehar *et al.* [24] have used the attributes and variables from the data set being clustered to find the optimal value of k. The proposed method compares the cluster properties for a range of *K* values to determine optimal k value. Borah and Ghose [30], have proposed AIM-K-Means algorithm for automatic determination of the number of clusters to be formed and initialization of cluster centres. In this algorithm each data point is taken as a candidate for initial mean randomly from the data set and its average distance from the other initial means is computed and if this distance value satisfies the distance threshold then it is retained as a valid initial mean. The number of clusters is taken as the number of initial means that satisfy the distance threshold criterion. In reference [25] the user is not required to enter k, the number of clusters to be formed. Initial k is selected using the CascadeKM [35] [36] method and may change in the later stage to an optimal value as the algorithm proceeds. Global and local outliers are identified and removed from the data set unless they are of sufficient density to form clusters. At the end of each iteration the clusters that are close based on the distance threshold are merged. The algorithm converges when there is no change in the number of clusters. The algorithm partitions data into optimal clusters after outlier removal and without requiring k as input from the user. In [16] Syakur et.al., have applied the elbow method on data consisting of customer profiles from Indonesian SMEs to determine k, i.e., the number of clusters that can be formed. The k obtained from the elbow method is input in the k-Means algorithm to form that many clusters. In reference [15] four methods of determining k are analysed and evaluated which are Gap statistic, Silhouette Coefficient, Canopy and Elbow technique. These four methods are applied on the Iris data set to determine the k value and the resulting clustered data. The results of experiments reveal the characteristics of each of the four methods viz., the Elbow method will fail to determine the k value if the inflection point is not obvious, the Gap statistic and

Silhouette methods are not optimal in case of large data sets due to high computation complexity, whereas the Canopy method seems the best compared to the three aforementioned methods. Amelec Viloria *et al.* [12], have proposed an extension to the ACDE algorithm where they have replaced the DE method of finding the k activation threshold with the U Control Chart (UCC) method which is required for effectively determining the number of clusters to be formed.

Varying the k values results in changes in cluster formations (see Fig. 1) and also cluster quality. Too few clusters result in loosely coupled clusters where there are many points far away from the cluster centres. Too many clusters would create unnecessarily small partitions of data making it difficult for analysis. Hence, it is desirable to partition data into an optimal number of clusters.

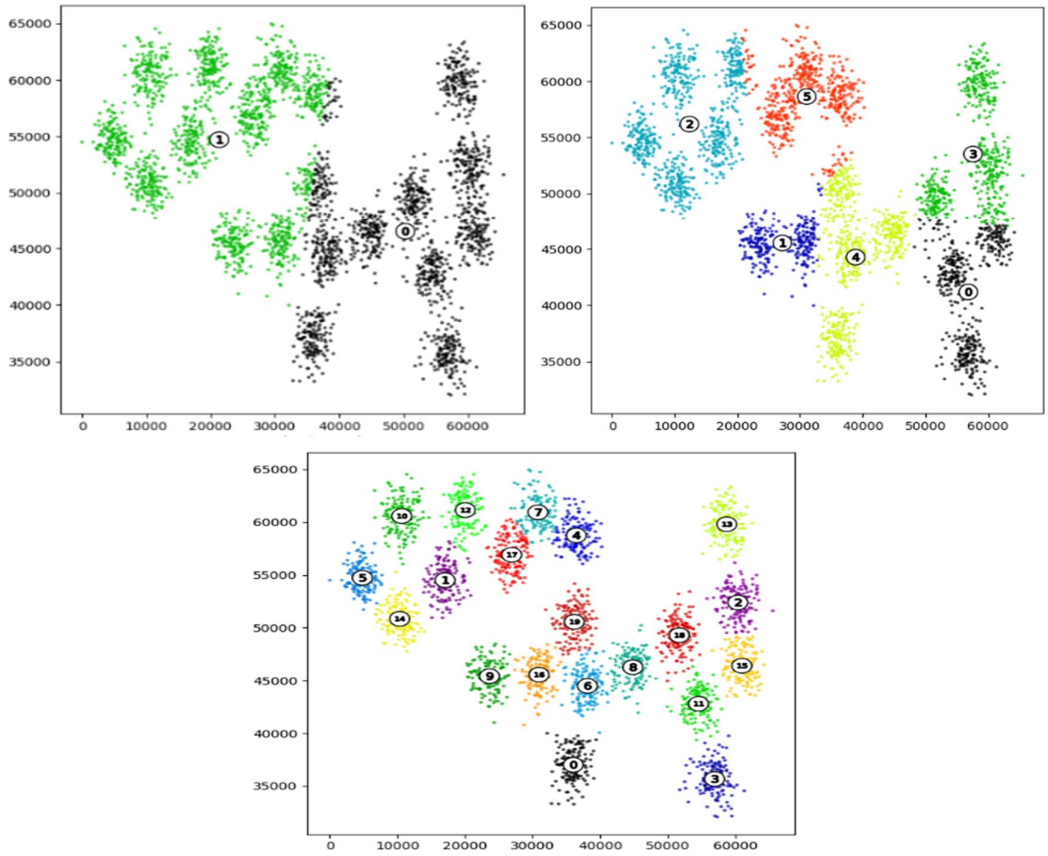

Figure 1. Cluster and cluster canter visualization on synthetic a1 data set [44] for k = 2, 6, 20 (from left to right and top to bottom)

B. Outlier detection in k-Means

Outliers are unusual and irregular patterns hidden within the data sets. These are data points which are extremely dissimilar from the majority of the data points in the data sets being clustered. Outliers can be broadly classified into a point outlier or a collective outlier depending on the number of data points identified as outliers [7]. Outliers are an important study matter so far as data clustering is concerned as their presence in the data sets can affect the clustering results in case the clustering algorithm cannot detect and separate them from non-outlying data points [3]. One of the shortcomings which impacts the performance of k-Means algorithm is its extreme sensitivity to the presence of outliers in the data [2] [13] as the presence of outliers can affect the determination of the number of clusters to be formed as well as the actual cluster formations. We can therefore form good quality clusters formed of only relevant data points by detecting and removing outliers from the data sets before the application of k-Means algorithm. In this section we review the various approaches for detecting and removing outliers from the data and thereby improving the performance of k-Means algorithm. In reference [8] the property of an outlier as being binary is extended by assigning to each object a degree of being an outlier. This degree is assigned to an object locally based on how isolated the object is from the other object in the

respective cluster. The LOF is computed for each object in the cluster and the objects deep in the cluster the value of LOF is closer to 1, whereas for other objects upper and lower bound range is formed which helps determine whether the neighbours come from one or more clusters and this helps in identifying a point as belonging to another cluster or as an outlier. M. F. Jiang et al. [9], have proposed a k-Means based algorithm for outlier detection which comprises of two phases; In the first phase they check if the new input pattern is sufficient distance away from all the cluster centres then it is selected as a new cluster centre with the idea that almost all the data points in that cluster be either outliers or non-outliers. In the 2nd phase the small clusters are selected and considered outliers. In [10] a greedy algorithm is proposed which requires the desired number of outliers, k, as input and proceeds to greedily select outliers such that initially all points are marked as non-outliers and the outlier set is empty. Then k scans are performed on the data set to select k outliers. In each scan each point is identified as a non-outlier and removed from the outlier cluster and the entropy objective re-evaluated. The point that experiences maximum decrease in entropy due to the removal of a point is selected as outlier in that scan. In reference [11], an algorithm for outlier detection is proposed which forms clusters by selecting random nodes with proper procedure thereby reducing the number of iterations compared to CLARANS algorithm. G. Gan, M.K.-P. [14], propose KMOR algorithms which extend the k-Means algorithm to perform data segmentation and outlier detection. The algorithm forms an additional cluster which clasps all outliers which are the data points that are at least a threshold distance away from all the cluster centres. The KMOR algorithm requires k, the desired number of clusters as input and other two main parameters n0, which is the maximum number of outliers and γ which is used in the function to compute the distance threshold used in identifying outliers. In [17], an algorithm called k-Means-- is proposed which jointly performs clustering and outlier detection. The algorithm takes as input two values viz., k which is the desired number of clusters and l which is the total number of desired outliers. All the points in the input dataset are arranged in descending order of the distance from their nearest cluster centre and the top l points are considered outliers and removed from the data set. Each of the remaining points are assigned to their nearest centre and the algorithm proceeds as per the traditional k-Means algorithm till convergence is achieved. This algorithm has the limitations of the k-Means algorithm with the additional limitation that the user is required to know in advance the desired number of outliers present in the input data set. In [18], an algorithm for simultaneously clustering and outlier detection is proposed which filters data of outliers after the clustering process. In the first stage, clustering is performed using Genetic k-Means algorithm to form clusters and then identify the outliers from within each cluster based on the distance of each point from the cluster centroid; the farther the point from the centroid, the greater are its chances of being outlier. The algorithm identifies as outliers the farthest points from each cluster and removes them. In [23] Vaishali Patel and Rupa Mehta have proposed a modified k-Means algorithm in which they compute 5-95% from input data set and consider lower and upper 5% data as outliers and separate them from the original data set into a new cluster. The remaining data as well as the data in the outlier cluster are normalized and k-Means algorithm is applied on the outlier free data to form clusters. At the end of the first iteration of the k-Means algorithm, the cluster with the outlying points is introduced as an additional cluster and the k-Means algorithm is applied again on these clusters until convergence. In reference [25] global and local outliers are identified and removed from the data set unless they are of sufficient density to form clusters. The global outliers are identified using the Interquartile range(IQR) method [3] with 1.5 as outlier factor. The local outliers which are the outliers within the clusters are identified using LOF [22] method. The local outliers so identified are formed into a new cluster if they satisfy the density and nearness criteria else they are discarded as outliers.

C. Performance evaluation methods for k-Means algorithm

Clustering validation is one of the important issues that needs to be addressed for the success of clustering applications. In general, clustering validation measures can be categorized into two classes, external validation measures and internal validation measures. While the internal validation criteria measure compactness i.e., how closely the data points within a cluster are tied to each other and separation, i.e., how well-separated a cluster is from other clusters; The external validation criteria use external information which is not present in the given data, for example, entropy measures the cluster purity based on the given class labels [34]. k-Means algorithm is applied on the data sets for forming meaningful partitions of data such that the sum of the intra cluster variance is minimized while the sum of inter cluster variance is maximized, these are widely used to determine the quality of the clusters formed by the k-Means algorithm. The time taken by the algorithm to converge into a meaningful cluster along with the number of iterations required are other parameters by which we measure the algorithm's performance. Empirical methods show the plot of cluster centres on the graph and give an idea of the level of optimality of the solution centres [26]. In this section we discuss the relevant methods that can help measure the performance of k-Means algorithms and thereby the optimality of the clusters so formed. Celebi *et al.* [20], have

compared the performance of various k-Means initialization methods over a number of synthetic and real datasets using various performance evaluation criteria. Initial Sum of Squared Error (SSE) is the first quality measurement criteria where after initializing cluster centres and before clustering. Final SSE is another quality measurement criterion which is computed after data clustering to measure the competence of the initialization method. Number of Iterations required by the clustering algorithm till its convergence is used to measure algorithm's efficiency. CPU time is another efficiency measure; it is the total CPU time taken by both the initialization and clustering stages. In [21] SSE is used to determine the cluster quality and convergence speed is used to measure the efficiency of the algorithms. In [29] two cluster quality measurement methods are used namely: Silhouette and Sum of Squared Error(SSE). The results of experiments prove the reliability of outcomes of both Silhouette and SSE, i.e., and hence we can say that both Silhouette and SSE indicate a suitable number of clusters at the same k-value. In [27] a graphical presentation is proposed for partitioning methods where each segment is characterized by its silhouette which is based on that segment's tightness and separation. The silhouette helps identify which data points lie well within their cluster's centre and those which lie somewhere between the clusters. They have shown that average silhouette width gives assessment of clustering validity. In [28] A. Starczewski and A. Krzy˙zak have evaluated the performance of silhouette index as a measure of cluster validity assessment. The silhouette index can be computed in two ways where the first method uses the average of the average silhouette across all the clusters and in the second method silhouette index is computed by averaging the silhouette across the whole data set. The experiments show that both these methods of the index have substantial impact on determining the appropriate number of clusters in a data set. In [32], Two clustering algorithms are applied on data sets, namely k-Means and EM. The clustering results are validated for quality using logistic regression analysis in WEKA tool where the logistic regression gives the number of correctly and incorrectly classified instances for the clustered data. In [33] various Cluster validation indices are compared and analysed in detail through experiments on several synthetic as well as real data sets. In [37] two techniques are used for clustering validation, namely, SSE [3] and Silhouette [3]. From experiments it was found that the results of both SSE and Silhouette are consistent as both formed a consistent number of clusters at the same k value in case where there is no overlap in data. However, in case of overlapping data, the result of SSE is nearer to the truth value.

III. Experimental Setup

We have applied the k-Means algorithm on datasets R15, S3 and A1 available on [38] [39] and the code for the same was written in python. Performance of the k-Means algorithm on these data sets is measured using cluster validity indices namely, SSE and Silhouette and the results are summarized in Table II and the visualization of clustered data for R15, S3 and A1 datasets are shown in Fig. 2, Fig. 3 and Fig. 4 respectively.

Looking at the results of R15 data sets for k = 13, 14, 15 and 16, the Silhouette score is highest and SSE value is low at k = 15. For S3 datasets also, SSE and Silhouette score was computed for k = 13 to 16 and the results show that the Silhouette score is highest and SSE value is low at k = 15. For A1 data sets at k = 18, 19, 20 and 21, we can see that Silhouette score is highest and SSE value is low at k = 20. Also, the visualization of clustered data, for all the three datasets namely, R15, S3 and A1 show that cluster centres are appropriately placed.

The mean Silhouette coefficient *s(i)*, is a good measure of correctness of a data point's assignment to a cluster. The value of *s(i)* lies between the range -1 and 1. *s(i)* close to one means that the data is appropriately clustered. If *s(i)* is close to negative one, then by the same logic we see that *s(i)* would be more appropriate if it was clustered in its neighbouring cluster. An *s(i)* near zero means overlapping clusters and that the datum is on the border of two natural clusters. Varying the value of k will also affect the *s(i)* score and hence *s(i)* can also be used to verify the correctness of k [27] [28] [29]. SSE, which is the sum of the squared differences between each datum and its cluster's mean, is another cluster validation criteria of interest. SSE can be used as a measure of variation within a cluster. If all observations within a cluster are identical, then the SSE would be equal to 0. Hence a low SSE value is desirable as it would mean a low variation within a cluster and a high SSE value would mean a large variation within a cluster. However, as the value of k increases, SSE value decreases, so we cannot rely on SSE alone for forming optimal clusters [20] [26]. Our experiments also show consistent results for Silhouette score and SSE and the empirical plot of clustered data help verify the optimality of formed clusters. Hence, Silhouette score, SSE value and visualization of clustered data together are sufficient to ensure that the formed clusters are optimal in terms of data assignments to each formed cluster.

TABLE II. RESULTS OF SILHOUETTE SCORE AND SSE VALUE COMPUTED FOR A RANGE OF K VALUES ON THE DATA SETS

Data Set	*Dimensions*	*Tuples*	*K*	*Silhouette score*	*SSE value*
R-15	2	600	13	0.686	230.33
			14	0.7163	159.49
			15	***0.752***	***157.49***
			16	0.732	105.54
S3	2	5000	13	0.4695	22085667628581.26
			14	0.4799	19412942204118.65
			15	***0.492***	***18622166992730.344***
			16	0.482	16368882038508.115
A1	2	3000	18	0.575	18100547226.82
			19	0.585	14726993242.77
			20	***0.595***	***14487301345.25***
			21	0.583	14257623243.12

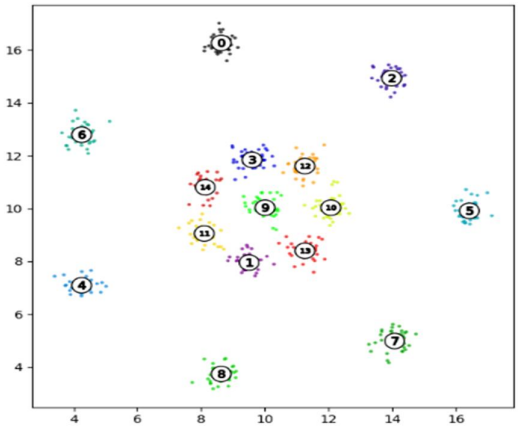

Figure 2. Clustered data for R15 data set

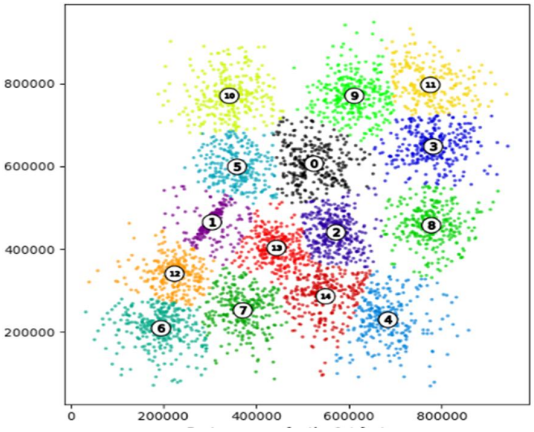

Figure 3. Clustered data for S3 data set

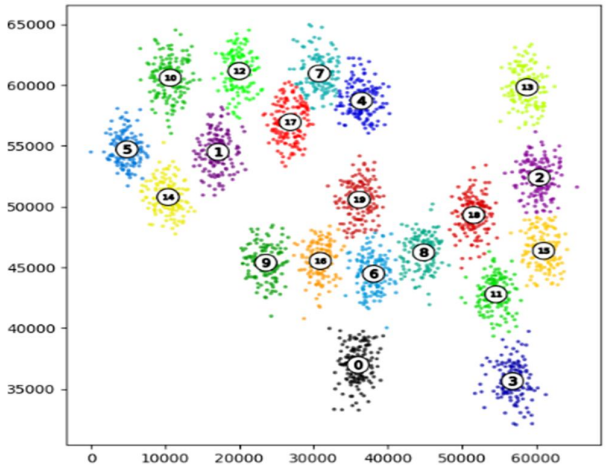

Figure 4. Clustered data for A1 data set

IV. CONCLUSION AND FUTURE WORK

We have analysed the work done in literature towards improving the performance of k-Means algorithm in areas of automatically determining k - the number of clusters to be formed and removing sensitivity of k-Means

algorithm towards the presence of outliers in the data set being clustered. We have also reviewed various cluster quality validation criteria. In this research, we applied the k-Means algorithm on the data sets for a range of k values and computed SSE and Silhouette scores along with plots of clustered data. Through our experiments we are able to form conclusions with respect to clustering quality as well as identify the clustering solution with most optimal clusters. From research we have proved that SSE, Silhouette and visualization of clustered data together can be used for verifying the optimality of the assignment of each point to its closest centre. Cluster visualization plots of the clustered data points and cluster centre visually help in the verification of cluster quality.

From the above given study of related work, it is clear that the limitations of k-Means algorithm have been addressed but with limited results and it becomes necessary to address the scope for improvements in order to achieve good quality clusters.

REFERENCES

[1] S. Revathi and Dr.T.Nalini, “Performance Comparison of Various Clustering Algorithm”, International Journal of Advanced Research in Computer Science and Software Engineering, Volume 3, Issue 2, February 2013 ISSN: 2277 128X.

[2] Rui Xu, Student Member, IEEE and Donald Wunsch II, Fellow, IEEE, “Survey of Clustering Algorithms”, IEEE TRANSACTIONS ON NEURAL NETWORKS, VOL. 16, NO. 3, MAY 2005.

[3] Jiawei Han and Micheline Kamber. “Data Mining Concepts and Techniques", Reference Book, Elsevier, Morgan Kaufmann Series, Second Edition, 2011, pg. no. 383-386.

[4] MacQueen, J. 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. https://projecteuclid.org/euclid.bsmsp/1200512992.

[5] Wu X, Kumar V, Quinlan JR, Ghosh J, Yang Q, Motoda H, et al. Top 10 algorithms in data mining. Knowledge and information systems. 2008;14(1):1–37.

[6] Shraddha k. Popat and Emmanuel M., “Review and Comparative study of Clustering Techniques”, (IJCSIT) International Journal of Computer Science and Information Technologies, Vol. 5 (1) , 2014, 805-812.

[7] Ji Zang, “Advancements of Outlier Detection: A Survey”, 2nd ed., ICST Transactions on Scalable Systems, January-March 2013, Volume 13, Issue 01-03.

[8] Breunig, M.M., Kriegel, H.P., and Ng, R.T., “LOF: Identifying Density-Based Local Outliers”, ACM conference Proceedings, 2000, pp. 93-104.

[9] M. F. Jiang, S.S. Tseng, C.M.Su., “Two-phase Clustering Process for Outlier Detection”, Pattern Recognition Letters, 200122(6-7):691-700.

[10] Zengyou He, Xiaofei Xu, “A Fast Greedy Algorithm for Outlier Mining”, ACM Digital Library, April 2006.

[11] S. Vijayarani and S. Nithya, “An Efficient Clustering Algorithm for Outlier Detection”, International Journal of Computer Applications (0975-8887), October 2011, Volume 32-No. 7.

[12] Viloria, Amelec, and Omar Bonerge Pineda Lezama. "Improvements for determining the number of clusters in k-means for innovation databases in SMEs." Procedia Computer Science 151 (2019): 1201-1206.

[13] A. K. Jain, “Data Clustering: 50 years beyond k-means”, Pattern Recognition Letters 31 (2010) 651–666, Elsevier, 2009. [Online]. Available: http://www.journals.elsevier.com/pattern-recognition-letters.

[14] Guojun Gan and Michael kwok-Po Ng, “K-means clustering with outlier removal”, Pattern Recognition Letters 90 (2017) 8–14, Elsevier, 2017.

[15] Yuan, Chunhui, and Haitao Yang. "Research on K-value selection method of K-means clustering algorithm." J—Multidisciplinary Scientific Journal 2, no. 2 (2019): 226-235.

[16] Syakur, M. A., B. K. Khotimah, E. M. S. Rochman, and B. D. Satoto. "Integration k-means clustering method and elbow method for identification of the best customer profile cluster." In IOP Conference Series: Materials Science and Engineering, vol. 336, no. 1, p. 012017. IOP Publishing, 2018.

[17] Chawla, Sanjay, and Aristides Gionis. "k-means–: A unified approach to clustering and outlier detection." In Proceedings of the 2013 SIAM International Conference on Data Mining, pp. 189-197. Society for Industrial and Applied Mathematics, 2013.

[18] Marghny, M. H., and Ahmed I. Taloba. "Outlier detection using improved genetic k-means." arXiv preprint arXiv:1402.6859(2014).

[19] Das, Swagatam, Ajith Abraham, and Amit Konar. "Automatic clustering using an improved differential evolution algorithm." IEEE Transactions on systems, man, and Cybernetics-Part A: Systems and Humans 38, no. 1 (2007): 218-237.

[20] Celebi, M. Emre, Hassan A. Kingravi, and Patricio A. Vela. "A comparative study of efficient initialization methods for the k-means clustering algorithm." Expert systems with applications40, no. 1 (2013): 200-210.

[21] J.M. Peña, J.A. Lozano, P. Larrañaga, An empirical comparison of four initialization methods for the k-means algorithm, Pattern Recognit. Lett. 20 (10) (1999). 1027–1040.

[22] Breunig, M.M., Kriegel, H.P., and Ng, R.T., “LOF: Identifying Density-Based Local Outliers”, ACM conference Proceedings, 2000, pp. 93-104.
[23] Patel, Vaishali R., and Rupa G. Mehta. "Impact of outlier removal and normalization approach in modified k-means clustering algorithm." International Journal of Computer Science Issues (IJCSI) 8, no. 5 (2011): 331.
[24] Mehar, Arshad Muhammad, Kenan Matawie, and Anthony Maeder. "Determining an optimal value of K in K-means clustering." In 2013 IEEE International Conference on Bioinformatics and Biomedicine, pp. 51-55. IEEE, 2013.
[25] Jambudi, Trushali, and Savita Gandhi. "A New K-Means-Based Algorithm for Automatic Clustering and Outlier Discovery." In Information and Communication Technology for Intelligent Systems, pp. 457-467. Springer, Singapore, 2019.
[26] Blömer, Johannes, Christiane Lammersen, Melanie Schmidt, and Christian Sohler. "Theoretical analysis of the k-means algorithm–a survey." In Algorithm Engineering, pp. 81-116. Springer, Cham, 2016.
[27] Rousseeuw, Peter J. "Silhouettes: a graphical aid to the interpretation and validation of cluster analysis." Journal of computational and applied mathematics 20 (1987): 53-65.
[28] Starczewski A., Krzyżak A. (2015) Performance Evaluation of the Silhouette Index. In: Rutkowski L., Korytkowski M., Scherer R., Tadeusiewicz R., Zadeh L., Zurada J. (eds) Artificial Intelligence and Soft Computing. ICAISC 2015. Lecture Notes in Computer Science, vol 9120. Springer, Cham. https://doi.org/10.1007/978-3-319-19369-4_5.
[29] Thinsungnoena, Tippaya, Nuntawut Kaoungkub, Pongsakorn Durongdumronchaib, Kittisak Kerdprasopb, and Nittaya Kerdprasopb. "The clustering validity with silhouette and sum of squared errors." learning 3, no. 7 (2015).
[30] Borah, Samarjeet, and Mrinal Kanti Ghose. "Performance analysis of AIM-K-means & K-means in quality cluster generation." arXiv preprint arXiv:0912.3983 (2009).
[31] Rodriguez, Mayra Z., Cesar H. Comin, Dalcimar Casanova, Odemir M. Bruno, Diego R. Amancio, Luciano da F. Costa, and Francisco A. Rodrigues. "Clustering algorithms: A comparative approach." PloS one 14, no. 1 (2019): e0210236.
[32] Jung YG, Kang MS, Heo J. Clustering performance comparison using K-means and expectation maximization algorithms. Biotechnology & Biotechnological Equipment. 2014;28(sup1):S44–S48.
[33] Arbelaitz O, Gurrutxaga I, Muguerza J, Pérez JM, Perona I. An extensive comparative study of cluster validity indices. Pattern Recognition. 2013;46(1):243–256.
[34] Liu Y, Li Z, Xiong H, Gao X, Wu J. Understanding of internal clustering validation measures. In: Data Mining (ICDM), 2010 IEEE 10th International Conference on. IEEE; 2010. p. 911–916.
[35] Calinski, T. and J. Harabasz. 1974. A dendrite method for cluster analysis. Commun. Stat. 3: 1--27.
[36] rdocumentation.org/packages/vegan/versions/2.4- 2/topics/cascadeKM.
[37] IThinsungnoena, Tippaya, Nuntawut Kaoungkub, Pongsakorn Durongdumronchaib, Kittisak Kerdprasopb, and Nittaya Kerdprasopb. "The clustering validity with silhouette and sum of squared errors." learning 3, no. 7 (2015).
[38] Dua, D. and Graff, C. (2019). UCI Machine Learning Repository [http://archive.ics.uci.edu/ml]. Irvine, CA: University of California, School of Information and Computer Science.
[39] P. Fränti and S. Sieranoja, K-means properties on six clustering benchmark datasets Applied Intelligence, 48 (12), 4743-4759, December 2018. https://doi.org/10.1007/s10489-018-1238-7 BibTex.