International Journal of Engineering and Advanced Technology (IJEAT)
ISSN: 2249-8958 (Online), Volume-11 Issue-4, April 2022

Optimal Value for Number of Clusters in a Dataset for Clustering Algorithm

Jayashree, Shivaprakash T

Abstract: ***It is essential to know the parameters required to clustering the dataset. One of the parameters is the number of clusters k and it is very important to select the k value to get deficient results on clustering. There are few algorithms to find the k value for k-means algorithm and it requires specifying a maximum value for k or a range of values for k as an input. This paper proposes a novel method Optimal cluster number estimation algorithm (OCNE) to find the optimal number of clusters without specifying the maximum or range of k values or knee point detection in the graph. In the experiment, this method is compared with the different existing methods with deficient real-world as well as synthetic datasets and provides good performance.***

Keywords: Cluster, Convex Cluster, Data Mining, Feature Extraction, K-Means

I. INTRODUCTION

There is a revolutionary development in the technology of data storage and computing, to manage the large quantity of data from a variety of sources or fields such as medical, social, scientific data, etc. These data contain the most valuable information and it requires clustering algorithms to form groups or clusters. For conventional clustering algorithm needs prior knowledge [1] of the actual number of clusters. Since clustering is an unsupervised type of operation, it is not easy to detect the number of clusters from large complex high-dimensional datasets [2]. So it is very important and significant to find the number of clusters in the datasets to get an efficient and correct solution of clustering [3] algorithm. The combination of data mining with machine learning [4][5], artificial intel- ligence, deep learning [8][9], statistics, and the database system [10] provide the techniques for extracting useful information [6][7] which may be unknown from a large dataset. One of the tasks of data mining [11] is clustering or cluster anal- ysis to extract information from the dataset [12] that is for partition unlabelled data [13] into a number of clusters (groups or categories) [14]. From the aspect of optimization of clustering [15], the main characteristic of clustering is to magnify both the (internal) homogeneity [16] within a cluster and the (external) heterogeneity [17] with different clusters.

Manuscript received on February 15, 2022.
Revised Manuscript received on February 23, 2022.
Manuscript published on April 30, 2022.
*Correspondence Author
Jayashree*, Department of Computer Science and Engineering, Vijaya Vittala Institute of Technology, Bangalore (Karnataka), India. E-mail: jayashrees08@gmail.com
Dr. Shivaprakash T, Professor, Department of Computer Science and Engineering, Vijaya Vittala Institute of Technology, Bangalore (Karnataka), India. E-mail: shivprakasht@gmail.com

Specifically, an identical type of data point should belong to the same cluster, while the alternative type of data points should be belonging to different clusters. For measuring similar- ity between instances in the datasets for clustering tasks, there are a variety of dissimilarity measures [18] and it has got some advantages and disadvantages de- pending on the type of datasets. For example, some are advisable for text catego- rization [19] while other ones are suitable to image segmentation [20], biomedical imaging [21], document clustering [22], statistical analysis, and network separa- tion [23]. In unsupervised feature extraction with a k-means algorithm [24], fixing the parameter is an open problem to solve. Finding optimal value for cluster number is an important parameter concerning cluster quality and time efficiency [25] of the clustering technique [26][27].This paper presents one of the parameters, which is number of cluster value k for clustering algorithm [28] without specifying prior knowledge for k values.

II. BACKGROUND AND RELATED WORK

Hancer and Karaboga [3] introduced a comprehensive survey of different ap- proaches such as traditional, merge-split, and evolutionary methods proposed for the estimation of cluster numbers. The determination of cluster numbers involve three main approaches: traditional, merge-split, and evolutionary computation- based approach [29]. In traditional approaches, all parameters other than cluster number make it fixed and execute clustering algorithm, where there needs to provide a range of numbers for parameter cluster number k. While comparing different traditional approaches, Calinski-Harabasz index was achieved better performance with oth- ers. The split and merge approach is based on splitting large, clusters into smaller ones and merge smaller similar clusters considering some user-specific parame- ters. ISODATA (Iterative Self-Organizing Data Analysis Technique) [30], DYNOC (Dynamic Optimal Cluster Seeking), maximum variance cluster (MVC) method, Modified Linde-Buzo-Gray (MLBG) methods are few examples of split and merge approach. The Evolutionary Computation (EC) based approach provides the solution to find the cluster number starting with some possible number and encode to get some specific and nearly optimal cluster number. There are different encoding schemes such as centroid-based encoding, don't care-based encoding, partition-based encoding, etc.
Dinh et al. [31] proposed an algorithm for finding the optimal value for the number of clusters in categorical data clustering by silhouette coefficient. The first step of this algorithm is to find a cluster center followed by find the distance between the center and each element of clusters and then the silhouette method to find the best k.

Slobodan [32] provides a comparison between the Silhouette index and the Davies-Bouldin index in Labelling IDS Clusters. Here IDS stands for the intru- sion detection system. IDS uses clustering for determining the good and bad behaviors of the network/host. In that paper, the authors tested and compared the quality of the cluster with two indexes: the Silhouette index and the Davies- Bouldin index, and found that the silhouette index provides a slightly better result than Davies-Bouldin index [33][34].

2.1. Methods

There are few different methods to calculate index values and by using an index, k value can be estimated. Most of the methods required to point out knee point or elbow point in the graph of index values for optimal value k. Some time this elbow point finding is difficult in some methods, since there is no visible knee point in the graph. Few methods are as follows Elbow method Elbow method [35] computes total intra-cluster variation [or total within-cluster sum of square (WSS)] by using common clustering algo- rithm such as k-means. WSS should be minimum as possible, which indicates compactness of clustering. Silhouette Coefficient Silhouette Coefficient [36] [37] provides the idea about how the objects are classified in cluster output and it is computed as below formula 1.

$$SilhouetteScore = \frac{(b-a)}{max(a,b)} \tag{1}$$

where

- a is the average within-cluster distance i.e the mean of distance between each point within a cluster.
- b is the average between-cluster distance i.e the mean of distance between all clusters.

Silhouette scores range from -1 to 1, if the is -1 indicates wrong clusters and 1 indicates the number of clusters are well separated, and a higher the value for silhouette coefficient score indicates the model with better-defined clusters. Therefore the maximum value of the index indicates an optimal solution for k in clustering [38].

Davies-Bouldin Davies-Bouldin(DB) index [39] [40] is based on the finding of the value which represents the quality of partition and that can be computed by estimating the distances between clusters and their dispersions [41]. The DB index [42] can be calculated using formula 2,

$$DB = \frac{1}{n_c}\sum_{i=1}^{n_c} R_i \tag{2}$$

where

$$R_i = \max_{j=1...nc, i\neq j}(R_{ij}), i = 1...n_c$$

$$R_{ij} = \frac{s_i + s_j}{d_{ij}}$$

$$d_{ij} = d\left(v_i, v_j\right), s_i = \frac{1}{\|c_i\|}\sum_{x \epsilon c_i} d(x, v_i)$$

where

- d(x,y) is the euclidean distance between x and y
- c_i is cluster i.
- v_i is the centroid of cluster c_i
- $\|c_i\|$ refers to the norm of c_i

Lower the DB index[43] value, better is the clustering. It also has a drawback[44]. This is dependent both on the data as well as the algorithm.

Calinski-Harabaz Index(CH-Index) Calinski-Harabaz Index [45] is also called as the Variance Ratio Criterion and it is defined as below equation 3

$$CH = \frac{SS_B}{SS_W} * \frac{n-k}{k-1} \tag{3}$$

SS_W is the overall within-cluster variance, which is the same as total within sum of squares(Euclidean distance between the two vectors.) $\sum_i^k \sum_{x \epsilon C_i} \| x - m_i \|^2$ where x is the data point, C_i is the i^{th} cluster, m_i is the centroid of cluster i.

SS_B is the overall between-cluster variance, which can be calculated as the total sum of squares (tss) minus SS_W and(tss) is the squared distance of all the data points from the datasets centroid. $\sum_i^k n\|m_i - c\|^2$ n is the total number of observations (data points), k is the number of clusters,c is the global centroid.

Advantages include, the score can be computed quickly and the higher value indicates clusters are dense and well separated, which fulfill the standard concept of a cluster.

The limitations of this method are the CH index is generally higher for convex clusters[46] than other non-convex clusters, such as density-based clusters like those obtained through DBSCAN.

Krzanowski and Lai(KL) Krzanowski and Lai(KL)[47] [48] [49] index is defined by equation 4

$$KL = \frac{DIFF_q}{DIFF_{q+1}} \tag{4}$$

where $DIFF_q = (q-1)^{\frac{2}{p}} trace(W_{q-1}) - q^{\frac{2}{p}} trace(W_q)$

q is the number of clusters

p is the number of variables

W_q is within-group dispersion matrix for data clustered into q clusters

Gap static Gap Statistics [50] algorithm adopts the output of any clustering algorithm [51] and compares changes within-cluster variance and expected under some distribution [43]. Gap static can be defined by following equation 5

$$Gap_n(k) = E_n^* log(W_k) - log(W_k) \tag{5}$$

where k is the estimated optimal number of clusters,

n is the number of samples

W_k within-cluster sum of squares around the cluster means

E_n^* denotes the expectation under a sample of size n from the reference distribution

Depth difference (DeD) DeD stands for Depth difference and its unique way of finding the optimal k value [52]. In this the depth of each point is calculated by using Mahalanibois depth function [44]. DeD can be defined by using the formula 6.

$$DeD = DW - DB \tag{6}$$

Curve method Basically the Curvature method is as shown in Algorithm 1, to determine the knee point in the evaluation graph and with an aspect of mathematics, curvature [53] is the amount by which a geometric object deviates from being flat or straight in the case of a line [45]. So the point with the maximum curvature in the graph represent the knee point [46]. The within-cluster variance evaluation graph can be calculated by using following formula 7 [47].

$$J(k) = \Sigma_{j=1}^{k} \Sigma_{x_i \epsilon C_j} ||x_i - \overline{x_j}||^2 \quad (7)$$

where C_j is the set of class j samples and $\overline{x_j}$ is the sample mean of class j. For a curve explicitly given as $y = f(x)$, the curvature is computed using the formula 8

$$K = \frac{|y^{"}|}{(1 + y^{'2})^{3/2}} \quad (8)$$

Algorithm 1 curvature method

Input: dataset X having objects as $x_1, x_2, x_3 ....x_n$ and k_max
Output: index d and k values
  for k=1 to $k_{max} + 1$ **do**
    for t=1 to 20 **do**
      Compute the within-cluster variance using k-means algorithm
    end for
    Get minimum of within-cluster variance
    if $k > 1$ **then**
      Compute the differences in within-cluster variance
    end if
  end for
  for k=2 to k_{max} **do**
    Compute the Curvature index
  end for
  The optimal number k is maximum value of Curvature index
  return k

III. OPTIMAL CLUSTER NUMBER ESTIMATION ALGORITHM

The proposed Optimal cluster number estimation algorithm(OCNE) is as shown in algorithm 2, the most common clustering algorithm k-means is used to find the within-cluster variance as the evaluation metric to form the graph for evaluation purposes. Let $X = x_1, x_2..., x_n$ each x_i, i=1,2, . . . , n is the data points of dataset X with n instances and within-cluster variance is defined by using Euclidean distance measure as shown in the formula

$$W = \Sigma_{j=1}^{k} \Sigma_{x_i \epsilon C_j} ||x_i - \overline{x_j}||^2 \quad (9)$$

where W is total with-in cluster variance, $C_j$is class j samples,k is the number of clusters that starts from value one,$\overline{x_j}$ is the mean of samples belong to j class.

The next main step in the algorithm is finding the validation index which can be calculated by finding the mean of difference in with-in cluster variance, which is defined as

$$VI(k) = \frac{1}{k} \Sigma_{i=1}^{k} delta_k \quad (10)$$

where $delta_k$ is the difference in with-in cluster variance. The algorithm stops execution when it reaches the condition $VI(k) < VI(k-1)$ and returns k-1 as an optimal k value.

3.1. Complexity

The time complexity of the proposed algorithm(Optimal cluster number estima-tion algorithm) is depend on two factors: the complexity of calculating with-in cluster variance which is nothing but k-means algorithm and complexity of com- puting validation index [48]. The time complexity [54] of k-means algorithm is $O(n^2)$.

Algorithm 2 Proposed method

Input: dataset X having objects as $x_1, x_2, x_3 ....x_n$
Output: k value
  k=1
  Initialize the validity index VI(k) to value zero
  Initialize scores and delta with value 0
  Do
    k=k+1
    Compute the within-cluster variance using k-means algorithm
    Get minimum of within-cluster variance
    Compute the differences in within-cluster variance and its $delta_k$
    $VI(k) = \frac{1}{k} \Sigma_{i=1}^{k} delta_k$
  Repeat the step from Do untill $VI(k) < VI(k-1)$
  return k-1

The time complexity of validation index[55] is the complexity of finding $delta_k$ which is $O(n*\log(n))$ where n is the size of dataset. Finally, the time complexity of the proposed algorithm is $O(n^2) + O(n*\log(n))$ where n is the size of dataset.

Similarly space complexity depends on k-means algorithm and validation index computation. The space complexity of validation index computation is $O(n)$ where n is the size of dataset. The space complexity of k-means algorithm is $O((m+k)*n)$ where m is the number of objects(samples) and n is the number of attributes considering n-dimensional objects. Therefore space complexity become $O((m+k)n) + O(n)$

IV. EXPERIMENT

Experiments consist of the dataset, different algorithms for comparison such as elbow method, Silhouette method, CH index, DB index method, and validation metrics such as relative error measure and k-value comparison table [49].

4.1. Experimental Settings

Datasets The datasets with the number of variables or attributes, number of samples or instances, and clusters are as shown in Table 1. The datasets include real-world datasets as well as synthetic datasets. The structure of few datasets are shown in Figure1 [55].

Evaluation metrics The relative error (RE) for computed k number of clusters is by using the formula 11.

$$RE = \frac{|k_{act} - k_{est}|}{k_{act}} \quad (11)$$

where k_{est} is the number of clusters that is computed and k_{act} is the actual number of clusters in the known dataset. Table 3 shows the relative error of optimal k value calculated with different algorithms. There is the average highest relative error in the experiment is Gap static followed with DeD algorithm. The lowest relative error is DB algorithm followed by CH algorithm. The proposed algorithm(Optimal cluster number estimation algorithm) provides a moderate level of relative error. It is observed that the algorithm does not depend on the number of instances or a size of dataset, since it is providing approximately correct k values to the dataset which is having a large

Table 1: List of datasets and it's characteristics

Dataset	No of instances	No of attributes	No of clusters
Wine(WN)	178	13	3
Breast-cancer	286	10	2
Breast-tissue	106	10	4
Ionosphere	351	34	2
Parkinsons	195	23	2
Seeds	210	8	3
Pima	768	9	2
Hill	1212	101	2
Haberman	306	4	2
Heart-va	200	13	2
Musk	476	167	2
Cardiotocography	476	167	10
Wine-quality-red	1599	12	6
Transfusion	748	5	2
Soybean	683	36	19
Gas-drift	13910	130	6
Iris	150	5	3
A1	1500	2	3
A2	1500	4	3
A3	1500	8	4
A4	1500	8	2
A5	1500	8	3
A6	1200	3	3

Published By:
Blue Eyes Intelligence Engineering and Sciences Publication (BEIESP)

number of instances such as gas-drift or wine-quality-red dataset.

4.2. Experimental Results

The results are shown in Table 2. It consists of different datasets such as wine, breast-cancer, ionosphere, etc., and *k_known* in the Table is known k values.

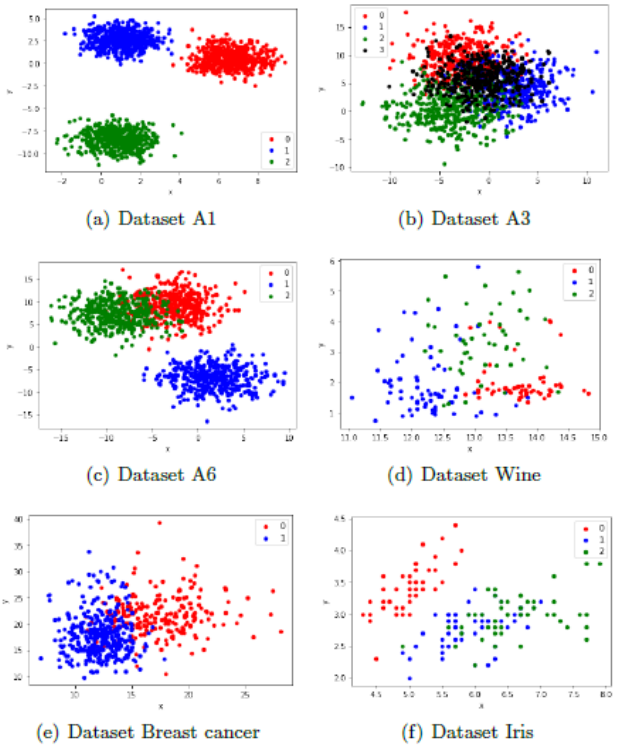


Fig. 1: Structure of the Different Datasets.

for each dataset. Also, it contains k values of the proposed algorithm for each dataset along with different algorithm k values for comparison purposes. The proposed method(Optimal cluster number estimation algorithm) achieved to provide better k values for cluster number than the other existing algorithms. In the experiment with different datasets, good results are getting followed with proposed algorithm are curve method, DB method and Gap static methods. The experiment with real-world datasets such as wine, hill, seeds, transfusion and iris provides the same k values as known k values. It is observed in the experiment that the k values getting as result are nearly equal to the known k value or the k values are near to the known k value,in the case of datasets where k values are not the same as known k value.

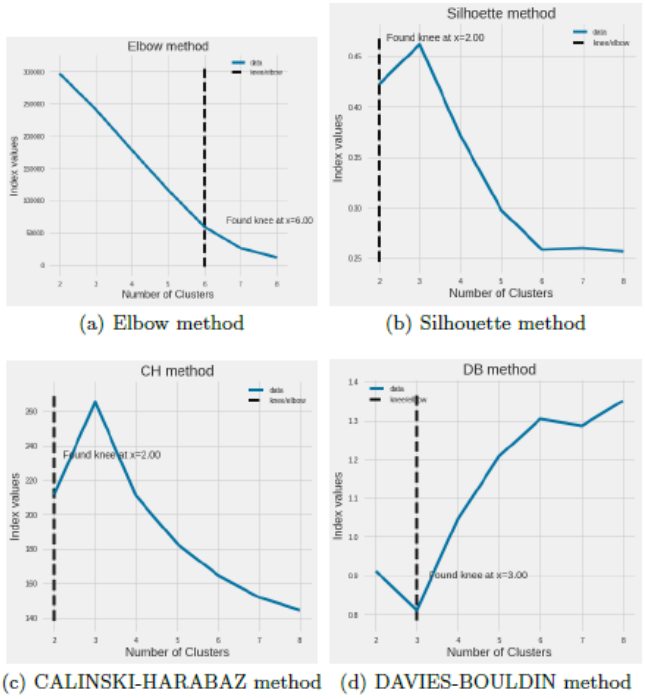
In the experiment, proposed algorithm on cardiotocography dataset provides k value as 7 which is near to the known k value 10. Similarly proposed algorithm with breast-tissue and haberman datasets provide k values as 3 which is near to the value 4 and 2 respectively. In the case of the synthetic dataset also similar kinds of results were observed, such as three datasets out of six provide the same k values as known k values and two datasets provides k values near to known k values.

Optimal cluster numbers of clusters with the graph using few different ex- isting algorithms with seeds dataset is as shown in Figure2. Here the x-axis indicates k values and the y-axis indicates indexes of respective algorithms on seeds dataset. In the proposed OCNE algorithm, the estimated k value is 3 which is shown in Figure2(e). As per the algorithm k=3 is an optimal valve and after that algorithm does not iterate with the next k values and returns the k value. This signifies validation index decides the execution of iterations in the algo- rithm. Here elbow point or knee points are not considered such as other existing algorithms.

There are other 4 different algorithms shown Figure 2(a) 2(b) 2(c) 2(d) which uses knee point or elbow point to find optimal k value. These algorithms find indexes for each k value and then the user has to find optimal value by finding elbow point. In some datasets, it is a challenge to find knee points in the graph.

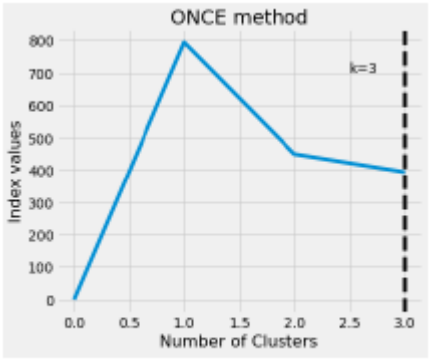
Table 2: Experimental comparison for k values of proposed method OCNE with other different methods on different datasets.

Dataset	k_known	CH	DB	SIL	Gap	DeD	Curve	OCNE
Wine(WN)	3	2	2	3	14	2	2	3
breast-cancer	2	5	3	3	14	3	3	6
breast-tissue	4	3	4	3	13	2	2	3
ionosphere	2	3	2	2	14	6	12	7
parkinsons	2	2	2	3	14	11	3	4
Seeds	3	2	3	2	13	3	3	3
Pima	2	2	4	2	13	19	3	5
Hill	2	2	2	3	14	2	2	2
Haberman	2	2	4	2	7	2	5	3
Heart-va	2	5	2	3	13	4	2	6
Musk	2	5	2	2	14	2	3	7
Cardiotocography	10	5	2	2	14	2	3	7
wine-quality-red	6	3	4	3	14	2	7	5
transfusion	2	2	3	4	14	13	4	2
Soybean	19	3	2	3	14	3	2	5
Gas-drift	6	4	2	4	14	2	2	3
Iris	3	4	2	5	14	2	2	3
A1	3	2	2	6	3	5	3	2
A2	3	2	2	2	3	2	3	3
A3	4	2	2	2	4	2	4	3
A4	2	4	2	4	2	2	2	8
A5	3	2	2	2	3	2	3	3
A6	3	4	2	5	3	7	2	3



(a) Elbow method (b) Silhouette method

(c) CALINSKI-HARABAZ method (d) DAVIES-BOULDIN method



(e) OCNE method

Fig. 2: Different algorithms to find Optimal value for number of clusters.

V. CONCLUSION

The paper presents a novel approach, the optimal cluster number estimation al- gorithm, to find the optimal value for k number of clusters. This method selects

Table 3: List of relative errors for k value with respect to a real number of clusters

Dataset	CH	DB	SIL	Gap	DeD	Curve	OCNE
Wine(WN)	0.33	0.33	0	3.66	0.33	0.33	0
breast-cancer	1.5	0.5	0.5	6	0.5	0.5	2
breast-tissue	0.5	1	0.5	6	0	0	0.5
ionosphere	0.5	0	0	6	2	5	2.5
parkinsons	0	0	0.5	6	4.5	0.5	1
Seeds	0.33	0	0.33	3.33	0	0	0
Pima	0	1	0	5.5	8.5	0.5	1.5
Hill	0	0	0.5	6	0	0	0
Haberman	0	1	0	2.5	0	1.5	0.5
Heart-va	1.5	0	1.5	6	1	0	2
Musk	1.5	0	0	6	0	0.5	2.5
Cardiotocography	0.5	0.8	0.8	0.4	0.8	0.7	0.3
wine-quality-red	0.5	0.33	0.5	1.33	0.66	0.16	0.16
transfusion	0	0.5	1	6	5.5	1	0
Soybean	0.84	0.89	0.84	0.26	0.84	0.89	0.73
Gas-drift	0.33	0.66	0.33	1.33	0.66	0.66	0.5
Iris	0.33	0.33	0.66	3.66	0.33	0.33	0
A1	0.33	0.33	1.0	0.0	0.66	0.0	0.33
A2	0.33	0.33	0.33	0.0	0.33	0.0	0.0
A3	0.33	0.33	0.33	0.0	0.33	0.33	0.0
A4	1.0	0.0	1.0	0.0	0.0	0.0	3.0
A5	0.33	0.33	0.33	0.0	0.33	0.0	0.0
A6	0.33	0.33	0.66	0.0	1.33	0.33	0.0

the k value for a number of clusters of a dataset by executing the k-means cluster- ing algorithm over a dataset to compute within-cluster variance and returns the k value on satisfying validation index. This method does not require an extra in- put parameter, unlike other existing algorithms. The proposed method provides efficient results on finding k value on different real and synthetic datasets. On a comparative study on different datasets on finding k value in the experiment, the proposed algorithm outperforms the other existing algorithms. From the point of view of relative error on k value computation, the proposed algorithm provides moderately good results. In the experiment, it is found that the algorithm does not depend on the number of cluster in dataset or the data size, on computing k value. In the case of a complex dataset, it is difficult to select the correct or nearly correct value of k for clustering data. The different algorithms provide different values for the same data set on each execution. It is a hard problem to the calculate number of clusters k, because of unknown labels or features in a dataset. Although this method approximately provides good results on different known dataset on executions. The limitation with the algorithm is time efficiency and the relative error is quite high for some datasets. Future work includes finding efficiency with respect to time and experimenting on different types of datasets. This method uses Euclidean distance to measure the distance between data points and k-means algorithm for clustering, future work can also include other distance measures and other clustering algorithms to get efficient results.

ACKNOWLEDGMENT

This material is based upon work supported by the Google Cloud Research Credits program with the award GCP19980904

REFERENCES

1. E. P. Simoncelli, “Optimal estimation in sensory systems,” The Cognitive sswNeu- rosciences, IV, pp. 525–535, 2009.
2. H. Wang and M. Song, “Ckmeans. 1d. dp: optimal k-means clustering in one di- mension by dynamic programming,” The R journal, vol. 3, no. 2, p. 29, 2011.
3. E. Hancer and D. Karaboga, “A comprehensive survey of traditional, merge- split and evolutionary approaches proposed for determination of cluster number,” Swarm and Evolutionary Computation, vol. 32, pp. 49–67, 2017.
4. S. Angra and S. Ahuja, “Machine learning and its applications: A review,” in Big Data Analytics and Computational Intelligence (ICBDAC), 2017 International Conference on. IEEE, 2017, pp. 57–60.
5. P. Louridas and C. Ebert, “Embedded analytics and statistics for big data,” IEEE software, vol. 30, no. 6, pp. 33–39, 2013.
6. “Artificial intelligence what it is and why it matters,” https://www.sas.com/enus/insights/analytics/what-is-artificial-intelligence.html.
7. T. A. C, “Artificial intelligence hacker noon,” https://hackernoon.com/understanding-understanding-an-intro-to-artificial-intelligence-be76c5ec4d2e, Jun 2017.
8. F. Shaikh, P. Dar, and P. Srivastava, “Comparison between deep learning & machine learning,” https://www.analyticsvidhya.com/blog/2017/04/comparison-between-deep-learning-machine-learning/, Apr 2017.
9. S. Devarajan, “A quick introduction to ai,” http://www.iamwire.com/2017/07/quick-introduction-ai/ 155885, Jul 2017.
10. “What is deep learning? how it works, techniques and applications,” https://in.mathworks.com/discovery/deep-learning.html.
11. S. Gheware, A. Kejkar, and S. Tondare, “Data mining: Task, tools, techniques and applications,” International Journal of Advanced Research in Computer and Communication Engineering, vol. 3, no. 10, 2014.
12. K. Kameshwaran and K. Malarvizhi, “Survey on clustering techniques in data mining,” International Journal of Computer Science and Information Technologies, vol. 5, no. 2, pp. 2272–2276, 2014.
13. N. Matatov, L. Rokach, and O. Maimon, “Privacy-preserving data mining: A fea- ture set partitioning approach,” Information Sciences, vol. 180, no. 14, pp. 2696– 2720, 2010.
14. J. Shen, S. I. Chang, E. S. Lee, Y. Deng, and S. J. Brown, “Determination of cluster number in clustering microarray data,” Applied Mathematics and Computation, vol. 169, no. 2, pp. 1172–1185, 2005.
15. M. Teboulle, “A unified continuous optimization framework for center-based clus- tering methods.” Journal of Machine Learning Research, vol. 8, no. 1, 2007.
16. Q. Liu, W. Liu, J. Tang, M. Deng, and Y. Liu, “Two-stage permutation tests for determining homogeneity within a spatial cluster,” International Journal of Geographical Information Science, vol. 33, no. 9, pp. 1718–1738, 2019.
17. S. Zheng, K. A. Hume, H. Able, S. L. Bishop, and B. A. Boyd, “Exploring de- velopmental and behavioral heterogeneity among preschoolers with asd: A cluster analysis on principal components,” Autism Research, vol. 13, no. 5, pp. 796–809, 2020.
18. A. S. Shirkhorshidi, S. Aghabozorgi, and T. Y. Wah, “A comparison study on similarity and dissimilarity measures in clustering continuous data,” PloS one, vol. 10, no. 12, p. e0144059, 2015.
19. R. Bekkerman, R. El-Yaniv, N. Tishby, and Y. Winter, “On feature distributional clustering for text categorization,” in Proceedings of the 24th annual international ACM SIGIR conference on Research

and development in information retrieval, 2001, pp. 146–153.
20. S. Naz, H. Majeed, and H. Irshad, “Image segmentation using fuzzy clustering: A survey,” in 2010 6th international conference on emerging technologies (ICET). IEEE, 2010, pp. 181–186.
21. R. Xu and D. C. Wunsch, “Clustering algorithms in biomedical research: a review,” IEEE reviews in biomedical engineering, vol. 3, pp. 120–154, 2010.
22. O. Zamir and O. Etzioni, “Web document clustering: A feasibility demonstration,” in Proceedings of the 21st annual international ACM SIGIR conference on Research and development in information retrieval, 1998, pp. 46–54.
23. S. Bandyopadhyay and U. Maulik, “An evolutionary technique based on k-means algorithm for optimal clustering in rn,” Information Sciences, vol. 146, no. 1-4, pp. 221–237, 2002.
24. W. Suhui, C. Ying, Z. Yanning, and P. Yuntao, “Survey on k-means algorithm,” Data Analysis and Knowledge Discovery, vol. 27, no. 5, pp. 28–35, 2011.
25. E. R. Hruschka, R. J. Campello, and L. N. de Castro, “Improving the efficiency of a clustering genetic algorithm,” in Ibero-American Conference on Artificial Intel- ligence. Springer, 2004, pp. 861–870.
26. A. Shafeeq and K. Hareesha, “Dynamic clustering of data with modified k-means algorithm,” in Proceedings of the 2012 conference on information and computer networks, 2012, pp. 221–225.
27. C. Rosenberger and K. Chehdi, “Unsupervised clustering method with optimal estimation of the number of clusters: Application to image segmentation,” in Pro- ceedings 15th International Conference on Pattern Recognition. ICPR-2000, vol. 1. IEEE, 2000, pp. 656–659.
28. N. M. Patil and D. V. Patil, “A survey on k-means based consensus clustering,” International Journal of Engineering Trends and Technology, vol. 1, no. 3, 2016.
29. Z. Aliniya and S. A. Mirroshandel, “A novel combinatorial merge-split approach for automatic clustering using imperialist competitive algorithm,” Expert Systems with Applications, vol. 117, pp. 243–266, 2019.
30. A. W. Abbas, N. Minallh, N. Ahmad, S. A. R. Abid, and M. A. A. Khan, “K- means and isodata clustering algorithms for landcover classification using remote sensing,” Sindh University Research Journal-SURJ (Science Series), vol. 48, no. 2, 2016.
31. D.-T. Dinh, T. Fujinami, and V.-N. Huynh, “Estimating the optimal number of clusters in categorical data clustering by silhouette coefficient,” in International Symposium on Knowledge and Systems Sciences. Springer, 2019, pp. 1–17.
32. S. Petrovic, “A comparison between the silhouette index and the davies-bouldin index in labelling ids clusters,” in Proceedings of the 11th Nordic Workshop of Secure IT Systems. Citeseer, 2006, pp. 53–64.
33. A. Hamdi, N. Monmarch´e, M. Slimane, and A. M. Alimi, “Intelligent hybrid algo- rithm for unsupervised data clustering problem,” in International Conference on Hybrid Intelligent Systems. Springer, 2016, pp. 442–455.
34. A. Kassambara, Practical guide to cluster analysis in R: Unsupervised machine learning. Sthda, 2017, vol. 1.
35. M. Syakur, B. Khotimah, E. Rochman, and B. Satoto, “Integration k-means clus- tering method and elbow method for identification of the best customer profile cluster,” in IOP Conference Series: Materials Science and Engineering, vol. 336, no. 1. IOP Publishing, 2018, p. 012017.
36. H. B. Zhou and J. T. Gao, “Automatic method for determining cluster num-ber based on silhouette coefficient,” in Advanced Research on Intelligent System, Mechanical Design Engineering and Information Engineering III, ser. Advanced Materials Research, vol. 951. Trans Tech Publications Ltd, 7 2014, pp. 227–230.
37. S.-X. Xia, W.-C. Li, Y. Zhou, L. Zhang, and Q. Niu, “Improved k-means clustering algorithm,” Journal of Southeast University (English Edition), vol. 23, no. 3, pp. 435–438, 2007.
38. P. J. Rousseeuw, “Silhouettes: a graphical aid to the interpretation and validation of cluster analysis,” Journal of computational and applied mathematics, vol. 20, pp. 53–65, 1987.
39. J. C. R. Thomas, M. S. Pen˜as, and M. Mora, “New version of davies-bouldin index for clustering validation based on cylindrical distance,” in 2013 32nd International Conference of the Chilean Computer Science Society (SCCC). IEEE, 2013, pp. 49–53.
40. G. P. Coelho, C. C. Barbante, L. Boccato, R. R. Attux, J. R. Oliveira, and F. J. Von Zuben, “Automatic feature selection for bci: an analysis using the davies- bouldin index and extreme learning machines,” in The 2012 international joint conference on neural networks (IJCNN). IEEE, 2012, pp. 1–8.
41. J. Leskovec and A. Rajaraman, “Clustering algorithms,” CS345a: Data Mining, Standford University, United State, 2010.
42. J. Xiao, J. Lu, and X. Li, “Davies bouldin index based hierarchical initialization k-means,” Intelligent Data Analysis, vol. 21, no. 6, pp. 1327–1338, 2017.
43. D. L. Davies and D. W. Bouldin, “A cluster separation measure,” IEEE transac-tions on pattern analysis and machine intelligence, no. 2, pp. 224–227, 1979.
44. K. Gomonov, S. Ratner, I. Lazanyuk, and S. Revinova, “Clustering of eu countries by the level of circular economy: An object-oriented approach,” Sustainability, vol. 13, no. 13, p. 7158, 2021.
45. Y. Wang, Y. Liu, L. Li, D. Infield, and S. Han, “Short-term wind power forecasting based on clustering pre-calculated cfd method,” Energies, vol. 11, no. 4, p. 854, 2018.
46. G. K. Chen, E. C. Chi, J. M. O. Ranola, and K. Lange, “Convex clustering: An at-tractive alternative to hierarchical clustering,” PLoS computational biology, vol. 11, no. 5, p. e1004228, 2015.
47. P. Novianti, D. Setyorini, and U. Rafflesia, “K-means cluster analysis in earthquake epicenter clustering,” International Journal of Advances in Intelligent Informatics, vol. 3, no. 2, pp. 81–89, 2017.
48. M. S. Pramono, S. Linuwih, and W. Lestari, “Hartigan, krzanowski-lai, silhouette, gap statistic’s estimate cluster number (k) methods and the implementation on health’s sector,” IEESE International Journal of Science and Technology, vol. 3, no. 4, p. 1, 2014.
49. M. Charrad, N. Ghazzali, V. Boiteau, A. Niknafs, and M. M. Charrad, “Package nbclust,” Journal of statistical software, vol. 61, pp. 1–36, 2014.
50. M. Yan and K. Ye, “Determining the number of clusters using the weighted gap statistic,” Biometrics, vol. 63, no. 4, pp. 1031–1037, 2007.
51. R. Tibshirani, G. Walther, and T. Hastie, “Estimating the number of clusters in a data set via the gap statistic,” Journal of the Royal Statistical Society: Series B (Statistical Methodology), vol. 63, no. 2, pp. 411–423, 2001.
52. C. Patil and I. Baidari, “Estimating the optimal number of clusters k in a dataset using data depth,” Data Science and Engineering, vol. 4, no. 2, pp. 132–140, 2019.
53. Y. Zhang, J. Man´dziuk, C. H. Quek, and B. W. Goh, “Curvature-based method for determining the number of clusters,” Information Sciences, vol. 415, pp. 414–428, 2017.
54. M. K. Pakhira, “A linear time-complexity k-means algorithm using cluster shift- ing,” in 2014 International Conference on Computational Intelligence and Com- munication Networks, 2014, pp. 1047–1051.
55. J. Arvidsson, “Finding delta difference in large data sets,” 2019.

AUTHOR PROFILE

Jayashree received her BE degree in 2007 and M.Tech in 2010 from Visvesvaraya Technological University, Belagavi, India. She is working towards PhD registered with Vijaya Vittala Institute of Technology, Research center of VTU, Belagavi. Her research interest include machine learning, data science and algorithms.

Shiva Prakash T is a Professor in the Department of Computer Science and Engineering at Vijaya Vittala Institute of Technology, Bangalore, India. He obtained his B.E, M.S and PhD degrees in Computer Science and Engineering from Bangalore University, Bangalore. His research interest include Wireless Sensor Network, Embedded Systems and Digital Multimedia.