Two Novel Techniques for Finding Optimal K-value in K-means Clustering

Dr. D. Mabuni
Assistant Professor, Dept. of Computer Science and Technology, Dravidian University, Kuppam, Andhra Pradesh, India

Abstract: **Two new techniques are proposed for determining optimal K-value in K-means clustering using decision tree classifier accuracy and its height. The first method is called Elbow Decision Tree Classifier (EDTC) created at elbow decision tree accuracy turning point and the second method is called decision tree classifier height (DTCH) determination at decision tree accuracy turning point. Standard UCI machine learning datasets are employed for experimentation purpose. Elbow turning point is a special K-value determined during decision tree accuracy starts to increase instead of usual accuracy decreasing. In EDTC, K-value at Elbow turning point is selected as the optimal K-value for K-means clustering. In the second proposed method (DTCH), decision tree height at the elbow turning point is taken as optimal K-value. The remarkable point is that Elbow K-value is approximately very close to the decision tree height. That is, approximately, equal optimal K-values in both the proposed methods is an indication that experiments are correct and consequently determined optimal K-values are also correct. Many standard UCI machine learning datasets are employed for experimentation purpose. Experiments results reveal that results are correct and optimal K-values determined in both the proposed methods are determined correctly.**

Index Terms: **EDTC, DTCH, elbow K-value, Optimal K-value, K-means clustering, Decision tree classifier, Machine learning, Data mining.**

I.INTRODUCTION

K-means clustering algorithm is very popular data clustering algorithm in machine learning as well as in data mining. but one must note that there are many unnecessary and redundant distance calculations in the traditional k-means clustering algorithm in the iterative process. In order to reduce redundant calculations and improve the efficiency of the k-means data clustering algorithm, this paper proposes to combine the triangle inequality principle in the distance calculation, so as to achieve the purpose of accelerating the clustering algorithm. This is particularly important in the case of data clustering of very large amounts of data. Many algorithms are available for numerical data clustering but limited number of algorithms are available for categorical data clustering. On must think that there is a need to cluster relational database management data.

The quality of the clustering solution is measured by the average quantization error (distortion and squared reconstruction error), q(C) and it is defined by using the mathematical equation,

$$q(C) = \frac{1}{n}\sum_{1}^{n} d(x_i, C_j) \ldots\ldots (1)$$

Note that lower values of q(C) are better.

There has been extensive database research applied on clustering very large datasets. In general, the K-means clustering algorithm requires two parameters the number of clusters and desired accuracy. K-means clustering algorithm is one of the most important data clustering algorithms. It is one of the best top-10 data mining algorithms. It is simple, fast, and scalable for many datasets. Number of clusters centres and initial cluster centres play an important role in K-means data clustering. Simplicity and efficiency are the two key features of K-means data clustering. Clustering performance is very sensitive about K-value and the selection of initial cluster Centres. Research community is trying for developing scalable clustering algorithms for the selected databases. K-Means is an iterative data clustering algorithm. Simplicity and fast convergence are the two very attractive ad powerful striking features of the K-means clustering algorithm. In the data clustering literature different variants of K-means clustering algorithms are continuously proposing by various research communities. Selection of K-value in K-means clustering algorithm is very difficult and the selected K-value directly influences the actual clustering performance.

A cluster is a collection of data objects such that objects within the cluster are very similar and objects between the clusters are dissimilar. Clustering is done using this basic principle. Many of the traditional clustering techniques are mainly based on scalability,

initial points selection, and the number of clusters. K-means clustering algorithm is frequently used to cluster data belonging to many real applications. Clustering algorithms are divided into partitions-based cluster types, hierarchical based clusters types, density-based clusters, and model-based cluster types. The working principle of K-means clustering is partitioned based technique. For data clustering selection of initial points and the number of clusters greatly influences clustering performance. One must be cautious in those selections. Minimum number of iterations and reduced time complexity are desired features for data clustering. For example, clustering analysis is one of the most important data analysis tasks in machine learning.

Aggregation of objects into similar groups of objects based on some similarity measure is called clustering. Similarity between two objects is measured based on distance. K-means data clustering algorithm uses a special measure called the sum of squared errors (SSE) measure as the objective function to measure the clustering quality.Data clustering is usually done in terms of iterations and by applying a suitable metric called sum squared errors as the clustering criterion function. Some clustering algorithms uses max-min principle for data clustering. The core part of the K-means clustering algorithm is finding the distance between the tuple and the cluster centre. Now-a-days MapReducing techniques are used for parallel data clustering. Sometimes the clustering algorithm must be able to handle data uncertainty. Also clustering techniques are required for both categorical data and numerical data. Clustering techniques are widely used in many domains such as medical, biology, zoology, physics, engineering, business, document clustering, object clustering and in many other related fields. K-means clustering, K-modes, and K-medoids are some of the variants of the existing traditional clustering algorithms. Rough set-based data clustering is also important in some applications.

Various clustering algorithms available are – K-means, K-medoids, hierarchical clustering, agglomerate clustering, Density based, grid-based algorithms, statistical based clustering algorithms. All these different categories of clustering algorithms follow known unsupervised technique for data clustering. In this paper, decision tree-based elbow turning point rule is proposed for optimal K-value determination and decision tree classifier height determination method is also proposed for optimal K-value determination. Two new proposed clustering methods are experimentally verified and proved by taking the standard UCI machine learning datasets.

II. LITERATURE SURVEY

K-means data clustering is a well-known, popular and frequently used standard clustering algorithm in machine learning. Distributed, scalable, fast, efficient and effective network management K-means clustering features are desired. In this paper, Datta et al. [1] have considered and thoroughly discussed about the distributed K-means clustering problem particularly in peer-to-peer network. Two variants of normal K-means clustering are proposed. The result of the first method is approximately equal to the standard centralized K-means clustering algorithm and the second proposed method is producing the more accurate clustering results. Measuring similarity is the fundamental criterion in any clustering algorithm including K-means clustering algorithm. Ting et al. [2] proposed a new similarity-based clustering algorithm called point-set algorithm that determines similarity between two tuples. The performance of the proposed algorithm is far better than the many existing state-of-the-art clustering algorithms.

Ordonetz and Omiecinski [3] proposed an efficient disk-based K-means clustering algorithm for relational databases. It is scalable across high dimensions. It is optimized to perform heavy I/O disk operations. Clustering results of the proposed algorithm are compared with the standard K-means clustering algorithm and scalable K-means clustering algorithm. Zhao et al. [4] Proposed fuzzy K-means clustering algorithm based on the shrunk patterns, which contains approximate to the original data and a penalty term. Nie et al. [5] proposed a modified version of K-means clustering algorithm. Objective function is replaced with a new formulation. The proposed method no need to calculate cluster centres in each iteration. Also, other re-weighted algorithm is proposed for faster convergence. Mahdi et al. [6] reviewed most recent developments in data cluster management. Many relevant clustering algorithms are systematically studied for handling large amounts of data. Important ideas are clearly explained for effective management of big data clustering.

One of the mostly used clustering algorithms now-a-days is K-means clustering algorithm because of its

many useful attractive features such as easy to understand, easy to implement, scalable, simple to interpret its experimental results, simple coding, and so on. Clustering is one way of data structuring. Direct K-means clustering problem [7] is NP-Hard as a result many variants of it are emerging continuously by many researchers. To date large number of improvements have been made to the K-means clustering algorithm. Modifications in the K-means clustering algorithm are done with respect to the features such as initialization, classification, centroid calculations, and convergence. The usage of machine learning techniques is increasing day by day in cloud computing environment. To solve many problems in cloud computing environment Wu et al. [8] proposed secure and efficient K-means clustering outsourced encryption algorithm. The proposed algorithm maintains privacy of the data. Machine learning is continuously producing scalable and cost-effective solution finding techniques for many real problems.

Li et al. [9] proposed a new fuzzy K-means clustering algorithm for clustering numeric data by introducing a penalty term in the objective function for decreasing the sensitivity of the initial cluster centres. The proposed algorithm is producing good clustering results. The proposed algorithm automatically determines number clusters at the beginning. Now-a-days, the demand for fast and scalable big data science and engineering-based knowledge discovery techniques are increasing rapidly. A new variant of the big data K-means clustering algorithm is proposed in this paper to overcome many of the limitations of the existing traditional K-means clustering algorithm [10]. Problems in the existing clustering techniques are – correct selection of initial cluster centres is difficult, determination of optimal K-value is difficult, and runtime time complexity is very high.

K-means clustering algorithm is a very popular algorithm and there are many modified K-means clustering algorithms are available in the literature. Sinaga and Yang [11] proposed an unsupervised variant of K-means clustering algorithm that automatically finds optimal number of clusters without any initialization. Clustering is a method of finding greatest similarity within the cluster tuples and the greatest dissimilarity between the different clusters. The main requirements of the K-means clustering algorithm are initialization and K-value selection. In K-means clustering selection of initial clusters directly affects the clustering performance. Xu et al. [12] proposed a new variant of K-means clustering algorithm by using grid concept for removing the noise effects. Chi [13] used K-means data clustering algorithm for dividing marks data into clusters. The dataset consists of final grade details of software and information services subject. The clustering performance is useful for decreasing or increasing teachers or class hours.

Wilkin and Huang [14] proposed two different variants of K-means clustering algorithm and their experimental results are also noted down with parameters such as running times and distances. Qi et al. [15] proposed a new variant of K-means clustering algorithm called K*-clustering algorithm consists of three new features – hierarchical optimization principle, pruning strategy, optimized updating technique. Big data analysis methods are rapidly expanding in many applications such as medical, research, business, retail, and gene analysis etc. Wu et al. [16] proposed BigData processing model in combination of many other desired features. Yuan and Yang [17] have discussed four distinct methods-Elbow, Gap Statistic, Silhouette coefficient, and Canopy for selecting optimal K-value in data clustering applications. Authors [18] have tried to cluster the business decisions using K-means clustering algorithm by using various technologies that are used for this clustering purpose. That is, authors have tried to develop a business decision making tool using clustering technique.

Kanungo et al. [19] presented Lioyd data clustering algorithm using a special data structure called kd-tree. It is an efficient clustering algorithm. D. Xu and Y. Tian [20] have analysed data clustering algorithms with respect to two perspectives traditional clustering algorithms and modern clustering algorithms. Different types of data clustering algorithms are compared and the results are tabulated with selected important parameter values. Data clustering is considered to be the basic step for further learning. Almost all data clustering algorithms depends on two important measures called similarity and dissimilarity. In clustering, data is partitioned in such a way that intra-cluster distance is very small and inter-cluster distance is very large. Ghesmoune et al. [21] vigorously studied different stream data clustering algorithms and comprehensive details of those stream data clustering algorithms are presented in very simple

manner. Suarez et al. [22] have studied conceptual clustering techniques and their taxonomy is presented for better understanding of natural formation of clusters.

III. PROBLEM DEFINITION

Data clustering is a very useful and power full tool in machine learning. K-means clustering technique is one of the best top-10 data mining techniques. Many variants of K-means clustering are continuously being developed in the machine learning literature. K-means clustering technique has both advantages and disadvantages and it became one of the commonly and popularly used data clustering algorithm. Main problems of K-means clustering are – K value selection and initial cluster centres selection. Hence, finding a good optimal K-value, scalability, and cluster performance are main requirements of the K-means clustering algorithm.

IV. EXPERIMENTS

4.1 Determination of Optimal K-Value in K-Means Clustering using Distance Metric and Decision Tree Classifier Model

Experiments are conducted thoroughly by using many of the standard UCI machine learning datasets and then the experimental results are tabulated with many desired parameters and their values. At the beginning input data is taken without class labels and then by using suitable distance metric for randomly selected increasing K-values, K-clusters are created and then each cluster is classified with a distinct class label. If necessary, this step is repeated by selecting one or more iterations for a selected single and the same K-value. Experiments are repeated for increasing order of different selected K-values and then resulted data is classified and then classified data is submitted as input to the decision tree classifier. Now a decision tree is created along with classification accuracy value. Once the decision tree is created, its test accuracy is computed and then the height of the decision tree is taken as the true optimal K-value for K-means clustering. That is, in this paper, using proposed method highly efficient decision tree classifier model with very low time complexity, O(log n), is created and used for finding optimal K-value in K-means clustering.

In this paper, two methods are proposed for optimal K-value determination in K-means clustering. First method is called Elbow Decision Tree Classifier (EDTC) and the second method is called Decision Tree Classifier Height (DTCH). In the EDTC method for different increasing K-values decision tree classifier test data accuracies are computed and noted down in table. Decision tree test accuracy values decreases as the K-value increases up to a certain K-value, after that for a particular K-value, decision tree accuracy starts to increase by showing the elbow curve. This elbow curve special turning point is taken as the optimal K-value for K-means data clustering.

In the second proposed method of data clustering, decision tree height is noted down for each elbow turning point of the first proposed clustering method and after close inspection it is observed that elbow turning point is approximately very close to the decision tree height. After sufficient analysis of the experimental result proofs, it is clearly decided that optimal K-value in K-means clustering is exactly equal to the decision tree height. One must be cautious about the elbow turning point of the first proposed method. Elbow turning point is a selected special K-value for a particular decision tree test accuracy where decision tree test accuracy starts to increase instead of gradual decreasing up to that elbow point. This break up distinct point behaves distinctly in the elbow change.

Various UCI machine learning datasets used in experimentation are shown in Table-1. Both the proposed methods are showing the coincided output results.

Table-1 UCI machine learning datasets used for optimal K-value determination

S.No.	Dataset-Name	Number of training tuples	Test Accuracy	Optimal K-value through Proposed EDTC method-1	Optimal K-value by using proposed DTCH method-2
1	Iris	150	96.66	4	3
2	Glass	214	94.41	4	3
3	Breast Cancer 699	699	98.14	11	2
4	Wine Quality Red	1600	97.87	5	4
5	Haber man	306	96.09	5	4
6	Cloud	1023	94.33	7	6

7	Mamograph	961	91.98	4	5
8	Wine	178	96.4	4	2
9	Waveform	5000	79.88	7	8
10	Yeast	406	52.86	3	1

Details of the both the proposed clustering methods are shown in Table-1 with appropriate parameter values. Optimal K-values in both the proposed methods are almost equal. This equality in output results is an indication for the accurate optimal K-values of the K-means data clustering technique.

Table-2 Iris Dataset Results

S.No	Optimal K-value Through Proposed EDTC Method-1	Classification Accuracy	Leaf count	Optimal K-value Through Proposed DTCH method-2	Correctly classified tuples
1	2	96.66	3	2	145
2	3	91.33	5	3	137
3	**4**	**96.66**	**5**	**3**	**145**
4	5	95.33	5	3	143
5	6	84.00	4	2	126

Form the Table-2 second and third columns are separately selected and shown in Table 2.1 for graphically showing the relationship between the decision tree classifier accuracy and the optimal K-value selection.

Table2.1 Elbow curve data for Iris dataset

Optimal K-value Through Proposed EDTC Method-1	Classification Accuracy
2	96.66
3	91.33
4	**96.66**
5	95.33
6	84.00

Data in Table-2 (column-2 and column-3) is extracted separately and it is shown Table 2.1 just only for easy reference and understanding purpose only.Figure-1 shows data of proposed method-1, *Elbow Decision Tree Classifier (EDTC). Note that u*p to K = 3, decision tree classifier accuracy decreases and at K = 4, decision tree accuracy starts to increase. This point is called elbow point. Optimal K-value for Iris dataset is K = 3 or 4.

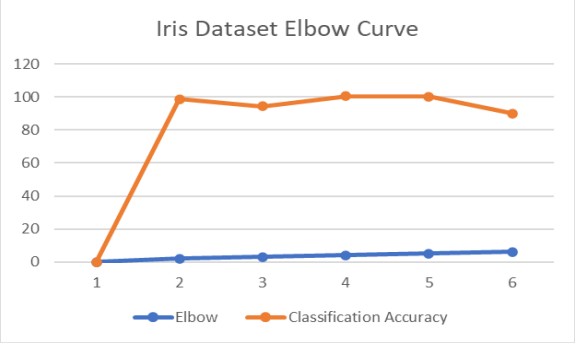

Figure-1 Iris Dataset Elbow curve

Table-3 Glass Dataset Results

S.No	Optimal K-value Through Proposed EDTC Method-1	Classification Accuracy	Leaf count	Optimal K-value by using proposed DTCH method-1	Correctly classified tuples
1	2	98.59	2	1	211
2	3	92.05	4	2	197
3	**4**	**94.41**	**5**	**3**	**203**

Table-3 shows optimal K-value for the dataset Glass is 4 using first proposed elbow method and K= 3 using decision tree classifier height method and these details are graphically shown in Figure-2. The main noticeable point is that the graph shown in Figure-2 is also following the same trend as in the graph shown in the Figure-1. This trending technique is same in almost all the graphs.

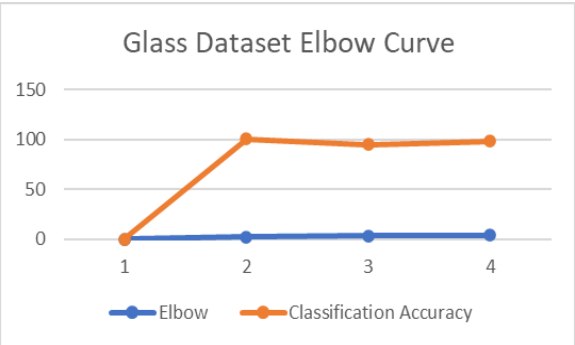

Figure-2 Glass Dataset Elbow curve

Table-4 Breast Cancer Dataset Results

S.No	Optimal K-value Through Proposed EDTC Method-1	Classification Accuracy	Leaf count	Tree height	Correctly classified tuples
1	2	100.0	2	1	699
2	3	100.0	3	2	699
3	4	99.85	3	2	698
4	5	99.57	3	2	696
5	6	99.28	3	2	694
6	7	99.14	3	2	693
7	8	98.99	3	2	692
8	9	98.71	3	2	690
9	10	93.13	6	4	651
10	**11**	**98.14**	**3**	**2**	**686**
11	15	97.13	3	2	679
12	20	96.28	3	2	673

Table-5 Wine Quality Red Dataset Results

S.No	Optimal K-value Through Proposed EDTC Method-1	Classification Accuracy	Leaf count	height	Correctly classified tuples
1	2	98.37	7	3	1574
2	3	98.12	8	4	1570
3	4	97.62	11	5	1562
4	**5**	**97.87**	**9**	**4**	**1565**

Figure-3 Wine quality red elbow curve with K = 5 or K = 4

Table-6 Haberman Dataset Results

S.No	Optimal K-value Through Proposed EDTC Method-1	Classification Accuracy	Leaf count	Optimal K-value by using proposed DTCH method	Correctly classified tuples
1	2	99.67	3	2	306
2	3	99.67	5	3	306
3	4	89.90	5	3	276
4	**5**	**96.09**	**7**	**4**	**295**

Figure-4 Haberman Elbow curve and decision tree height data. Elbow curve K-value = 5 and decision tree hight value = 4

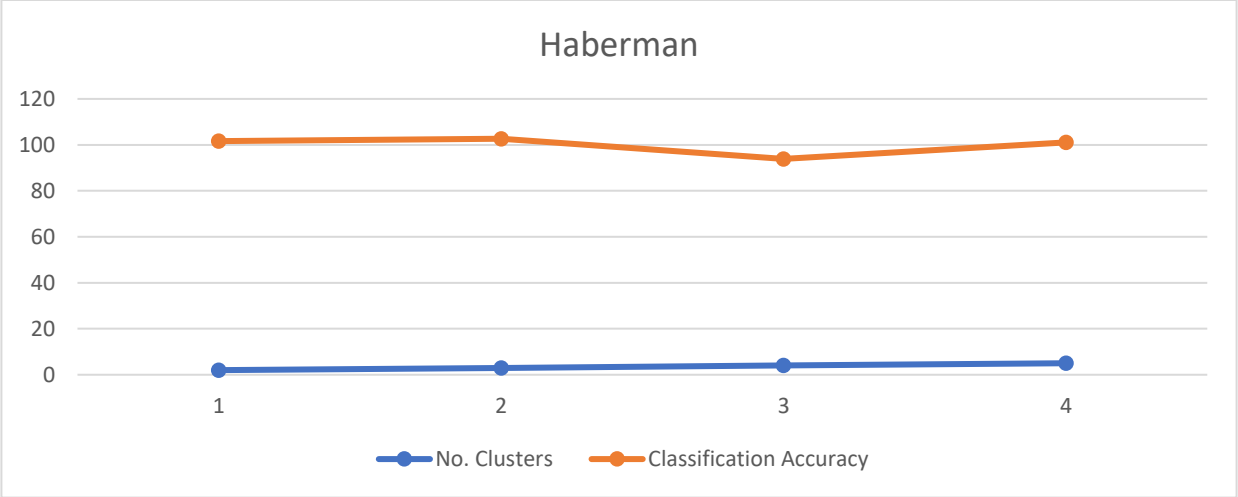

Table-7 Cloud Dataset Results

S.No	Optimal K-value Through Proposed EDTC Method-1	Classification Accuracy	Leaf count	Optimal K-value by using proposed DTCH method	Correctly 1026 classified tuples
1	2	99.70	4	2	1020
2	3	98.73	8	4	1011
3	4	97.36	9	5	997
4	**5**	94.62	**13**	**5**	**969**
5	**6**	93.91	**13**	**5**	**961**
6	**7**	**94.33**	**16**	**6**	**965**

Figure-5 Cloud data Elbow curve

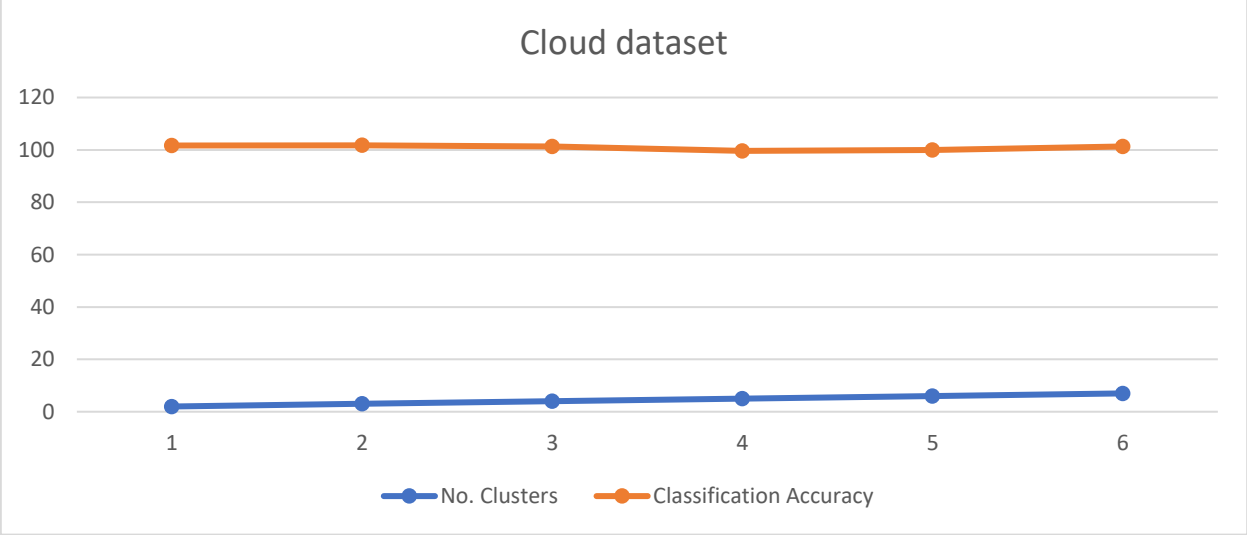

Table- 8 Mammograph Dataset Results

S.No	Optimal K-value Through Proposed EDTC Method-1	Classification Accuracy	Leaf count	Optimal K-value by using proposed DTCH method-2	Correctly classified tuples
1	2	93.34	5	3	897
2	3	68.36	3	2	657
3	4	91.98	8	5	884

Table-9 Wine Dataset Results

S.No	Optimal K-value Through Proposed EDTC Method-1	Classification Accuracy	Leaf count	Optimal K-value by using proposed DTCH method-2	Correctly classified tuples
1	2	100.0	2	1	179
2	3	96.08	2	1	172
3	4	96.04	3	2	170
4	**5**	96.04	**3**	**2**	**170**
5	**6**	91.52	**3**	**2**	**162**

Every table contains data values as expected. Experimentally determined values are correct and in every table, trend is continuing in the same manner as expected without any deviations. So, proposed methods are working correctly in a systematic way. Same clarity is appearing in graphically displayed data also. Except breast cancer dataset all results of the selected datasets are following the smooth and the same pattern results. Only breast cancer data is somw what special data clustering.

Table-10 Waveform Dataset Results

S.No	Optimal K-value Through Proposed EDTC Method-1	Classification Accuracy	Leaf count	Optimal K-value by using proposed DTCH method-2	Correctly classified tuples
1	2	95.86	61	9	4794
2	3	87.32	165	13	4366
3	4	86.94	147	12	4348
4	**5**	84.54	**161**	**12**	**4228**
5	**6**	79.88	**182**	**12**	**3995**

Table-11 Yeast Dataset Results

S.No	Optimal K-value Through Proposed EDTC Method-1	Classification Accuracy	Leaf count	Optimal K-value by using proposed DTCH method-2	Correctly classified tuples
1	2	52.86	2	1	785
2	3	52.86	2	1	785
3	4	52.86	2	1	785
4	**5**	52.86	2	1	785

Second proposed method for optimal K-value selection can also be presented in the manner shown in the respective tables - Table-12, Table-13, Table-14, and Table-15.

Table-12

Dataset Iris correctly classified tuples	Accuracy	Tree height	Leaf nodes count	Pruning threshold
147	94.23	4	8	2
147	94.23	4	8	5
151	96.79	3	7	10
149	95.51	3	7	20
149	95.51	3	5	30
149	95.51	3	5	40
143	91.66	2	4	50
128	82.05	2	3	60
128	82.05	2	3	75

Table-13

Dataset Glass correctly classified tuples	Accuracy	Tree height	Leaf nodes count	Pruning threshold
183	83.18	4	11	2
183	83.18	4	11	5
172	78.18	4	9	10
154	70.0	4	8	20

157	71.36	3	6	30
154	70	3	5	40
154	70	3	5	50

Table-14

Dataset Breast Cancer correctly classified tuples	Accuracy	Tree height	Leaf nodes count	Pruning threshold
705	100	4	5	2
700	99.29	2	3	5
700	99.29	2	3	10
700	99.29	2	3	20
700	99.29	2	3	30
700	99.29	2	3	40
700	99.29	2	3	50
700	99.29	2	3	75
700	99.29	2	3	100
700	99.29	2	3	150

Table-15

S.No	Mamographic correctly classified tuples	Accuracy	Tree height	Leaf nodes count	Pruning threshold
1	874	90.38	7	12	2
2	874	90.38	7	12	5
3	874	90.38	7	12	10
4	874	90.38	5	10	20
5	873	90.27	5	9	30
6	873	90.27	5	9	40
7	873	90.27	5	9	50
8	873	90.27	5	9	60
9	873	90.27	5	9	70
10	873	90.27	5	9	80
11	873	90.27	5	9	90
12	873	90.27	5	9	100
13	873	90.27	5	9	125
14	873	90.27	5	8	150
15	873	90.27	5	8	175
16	873	90.27	3	6	200
17	873	90.27	3	6	250
18	769	79.52	3	5	300
19	670	69.28	2	3	400
20	670	69.28	2	3	500

CONCLUSIONS

Two new variants of K-means data clustering techniques are proposed for determining optimal K-value in K-means clustering using a decision tree classifier accuracy and its height. The first method is called Elbow Decision Tree Classifier (EDTC) created at elbow decision tree accuracy turning point and the second method is called decision tree classifier height (DTCH) determination. Both the methods are hybrid methods in which two internal techniques are used in each of the proposed methods. First given data is pre-processed by assigning distinct class labels for each distinct cluster but same class label to each tuple

present in the same cluster and these labelled data is then used to generate decision tree classifier, which is used for optimal K-valuedetermination in K-means data clustering. In the future, these proposed methods will be enriched with the state-of-the-art hybrid clustering techniques.

REFERENCE

[1] S. Datta, C. Gainnella and H. Kargupta, "Approximate Distributed K-Means Clustering over a Peer-to-Peer Network", Published in: IEEE Transactions on Knowledge and Data Engineering (Volume: 21, Issue: 10, October 2009)

[2] Kai Ming Ting; Jonathan R Wells; Ye Zhu, "Point-Set Kernel Clustering", Published in: IEEE Transactions on Knowledge and Data Engineering (Early Access), Date of Publication: 25 January 2022

[3] C. Ordonetz and E. Omiecinski, "Efficient disk-based K-means clustering for relational databases", September 2004 IEEE Transactions on Knowledge and Data Engineering 16(8):909 - 921

[4] X. Zhao, F. Nie, R. Wang, X. and Li, "Robust Fuzzy K-Means Clustering With Shrunk Patterns Learning", IEEE Transactions on Knowledge and Data Engineering2023.03

[5] F. Nie, Z. Li, R. Wang, and X. Li, "An Effective and Efficient Algorithm for K-means Clustering with New Formulation", Published in: IEEE Transactions on Knowledge and Data Engineering (Early Access), Date of Publication: 01-March-2022

[6] M.A. Mahdi, K.M. Hosny, and Elnawy, "Scalable clustering algorithms for big data: A review", Received May 11, 2021, accepted May 22, 2021, date of publication May 26, 2021, date of current version June 8, 2021., IEEE Access.

[7] J. Ortega, N. Almanza, Vega, Pazos, Rangale, Zavala, "The K-means Algorithm Evaluation", Submitted: December 18th, 2018 Reviewed: February 25th, 2019 Published: April 3rd, 2019, DOI:10.5772/intechopen.85447

[8] Wei Wu; Jian Liu; Huimei Wang; Jialu Hao; Ming Xian, "Secure and Efficient Outsourced k-Means Clustering using Fully Homomorphic Encryption With Ciphertext Packing Technique", Published in: IEEE Transactions on Knowledge and Data Engineering (Volume: 33, Issue: 10, 01 October 2021)

[9] J. Li, K. Ng, Y. Cheung, and J. Z. Huang, "Agglomerative Fuzzy K-Means Clustering Algorithm with Selection of Number of Clusters", Published in: IEEE Transactions on Knowledge and Data Engineering (Volume: 20, Issue: 11, November 2008)

[10] K.E. Dierckens, A. B. Harrison, C. K. Leung, and A. V. Pind, "A Data Science and Engineering Solution for Fast K-Means Clustering of Big Data", Published in: 2017 IEEE Trustcom/ BigDataSE/ICESS

[11] K.P. Sinaga, S. Yang, "Unsupervised K-Means Clustering Algorithm", Published in: IEEE Access (Volume: 8), Date of Publication: 20 April 2020

[12] Hui Xu, Shunyu Yao, Qianyun Li, Zhiwei Ye, "An Improved K-means Clustering Algorithm", Date of Conference: 17-18 September 2020, Date Added to IEEE Xplore: 22 December 2020

[13]D. Chi, "Research on the Application of K-Means Clustering Algorithm in Student Achievement", Published in: 2021 IEEE International Conference on Consumer Electronics and Computer Engineering (ICCECE)

[14] A. Wilkin and X. Huang, "K-Means Clustering Algorithms: Implementation and Comparison", Published in:Second International Multi-Symposiums on Computer and Computational Sciences (IMSCCS 2007)

[15]J. Qi, Y. Yu, L. Wang, and J. Liu, "K*-Means: An Effective and Efficient K-Means Clustering Algorithm", Published in: 2016 IEEE International Conferences on Big Data and Cloud Computing (BDCloud),

[16] X. Wu, X. Zhu, G.-Q. Wu, and W. Ding, ''Data mining with big data,'' IEEE Trans. Knowl. Data Eng., vol. 26, no. 1, pp. 97–107, Jan. 2014.

[17] C. Yuan and H. Yang, "Research on K-Value Selection Method of K-Means Clustering Algorithm", Multidisciplinary scientific journal, received: 21 May 2019; Accepted: 15 June 2019; Published: 18 June 2019

[18] M.A. Hamada and Lyazat, "Decision Support System with K-Means Clustering Algorithm for Detecting the Optimal Store Location Based on Social Network Events", Published in: 2020 IEEE European Technology and Engineering Management Summit (E-TEMS)

[19] Kanungo, T.; Mount, D.M.; Netanyahu, N.S.; Piatko, C.D.; Silverman, R.; Wu, A.Y. "An efficient k-means clustering algorithm: analysis and

implementation", IEEE Trans. Pattern Anal. Mach. Intell. 2002, 24, 0–892.
[20] D. Xu and Y. Tian, "A comprehensive survey of clustering algorithms,'' Ann. Data Sci., vol. 2, no. 2, pp. 165–193, Jun. 2015.
[21] M. Ghesmoune, M. Lebbah, and H. Azzag, ''State-of-the-art on clustering data streams,'' Big Data Anal., vol. 1, no. 1, p. 13, Dec. 2016.
[22] A. Pérez-Suárez, J. F. Martínez-Trinidad, and J. A. Carrasco-Ochoa, ''A review of conceptual clustering algorithms,'' Artif. Intell. Rev., vol. 52, no. 2, pp. 1267–1296, Aug. 2019.