Evaluation of Clustering Tendency by Enhanced Visual Access Tendency Method in K-Means Clustering Approach

P Kiran Kumar Reddy
Department of CSE
Narayana Engg.College, Gudur, A.P, INDIA

Abstract - **The data clustering is an emerging technique in various data mining applications and it plays vital role in classifying unlabeled data objects. K-means is one of the best and effective data partitioning method but it may suffer from unknown clustering tendency. The proposed method extends the k-means based clustering algorithm with Visual Access Tendency (VAT) procedure, called as VAT based k-means Clustering. This hybrid approach speeds up the clustering results. The k-means is unsupervised approach and it can solve the clustering problem of unlabeled data. VAT is a data visualization method and it determines the clustering tendency of unlabeled data. The existing system takes more run time when there are several iterations where as the proposed system takes single step with very less run time. Clustering validity is to be checked at every iterated k-means clusters by Dunn's Index. Higher Dunn's Index imposes the exact clustering. Key contribution of the paper is to find prior tendency in k-means Based Clustering by Visual Access Tendency and to find clustering results in a single step instead of several trails. Results are tested on synthetic data sets and real data sets to conclude the clustering results are improved by proposed method with respect to the runtime.**

Keywords - Clustering Analysis, Clustering Tendency, k-means, Dunn's Index, Visual Access Tendency (VAT).

I. INTRODUCTION

Clustering has been widely used in data analysis. The data clustering[26,27,28,29] is an emerging technique in various data mining applications[18,19], and it plays important role in classifying unlabeled data objects. Despite the exiting traditional clustering algorithms, k-means [21,23] is the best and effective data partitioning method. The k-means algorithm [1] is a classical approach and it is greatly succeed in many practical domains [10,11,14,15]. This clustering method may suffer from unknown clustering tendency [13]. Thus, several trails of execution (at k=2, 3, 4.., where k refers the number of clusters) is needed for k-means algorithm for obtaining of correct k value. The assessment of clustering tendency and finding the quality of clustering results is time-consuming in k-means. Several methods are investigated for automatic clusters detection (k), among these methods, the Visual Access Tendency (VAT) [20] is an optimal choice for detecting the clustering tendency or number of clusters. Therefore, the proposed work can assess the clustering tendency by VAT in k-means clustering algorithm and this proposed approach is known as VAT-based-k-means clustering algorithm. This paper carried out the experiments on several datasets for demonstrating the effectiveness of proposed hybrid approach. This paper contributes the proposed work on extensive ideas of k-means based clustering [22]aiming to extract the tight clusters in order to get two benefits; first is to reduce the time and second is to improve the time values since, we use the known tendency value in k-means clustering algorithms. The major objective of our proposed method is to make best usage of tendency value in k-means based clustering algorithms for improving performance values. Assessment of tendency is one of the important criteria during clustering analysis. Exact tendency values are inferred from VAT techniques.

Related work of k-means based clustering is presented in Section II. Concept of Enhanced Visual Access Tendency is discussed in Section III. Proposed method is described in Section IV. Section V describes Results & discussions and finally conclusion & future work is presented in the last Section.

II. K-MEANS BASED CLUSTERING

Among the clustering algorithms k-means is a simple and efficient clustering method. The aim of k-means is to generate the faster and efficient clustering results. However, it generates the data partition results without knowledge of prior clustering tendency i.e., the value of k is unknown and it is given by the user as approximately. This method may suffer from unknown clustering tendency. Therefore, it is required to run k-means clustering algorithms for several times as trails for finding the best data partitioning because the initial number of clusters (k) is unknown. Suppose the value of clustering tendency is known, and then it is enough to run the algorithm as a single time instead of several times.

The k-means is unsupervised approach and it can solve the clustering problem of unlabeled data. The procedure of k-means [2] follows a classical approach, where it classifies the dataset through an assumed number of clusters. The basic idea is to create the 'k' centers by selecting randomly 'k' initial objects, and one for each cluster. Assign each object at a time from the remaining objects to the closest cluster [3]. Update the centers or means after assigning each object into the cluster. It continued until assigning of all objects into their respective clusters. The k-means approach is minimizing the objective function, and it is given by the Eq.1

$$J = \sum_{j=1}^{k}\sum_{i=1}^{n}\left\|x_i^{(j)} - c_j\right\|^2 \qquad (1)$$

Where $||X_i^{(j)} - C_j||^2$ refers the distance between a data point $X_i^{(j)}$ and a cluster center C_j , it indicated the distance between data points, and a cluster center.

Three important limitations are identified in present system. These are clustering tendency is unknown, it doesn't generate the exact number of clusters without knowledge of tendency, and it requires external interference for specifying termination condition. Therefore, the purpose of assessing tendency, we propose the specific visualization method in k-means based clustering algorithms for detecting exact tendency value.

III. VISUAL ACCESS TENDENCY (VAT) AND ENHANCED VAT METHODS

Visual Access Tendency (VAT) [7, 8, 9] is a data visualization method and it determines the clustering tendency of unlabeled data. K-means algorithm suffers from unknown clustering tendency. Therefore, this method is merged with VAT for addressing the assessment of clustering tendency. The proposed hybrid approach is known as VAT based k-means clustering algorithm.

The clustering algorithm k-means use the VAT for determining the number of valid clusters. The VAT reorders the dissimilarity [25]matrix D to D^* (n × n , where n refers data objects) using Prim's logic, and it generates the image of D^*, $I(D^*)$ (it is known as VAT Image). The VAT is reveals the hidden clustering structures by squared-shaped dark blocks of I (D^*). In the VAT, the clustering tendency is accessed by assessing the information of a total number of square-shaped dark blocks. The VAT is an effective procedure for detecting the clustering tendency in a visual form by counting the number of square shaped dark blocks along the diagonal in a VAT image. The results of VAT depend on dissimilarity or similarity features of objects. Therefore, the way for finding the dissimilarities is most important in VAT algorithm. Generally, the dissimilarity matrix in VAT is computed in Euclidean space. From the statistical evidence of [4], it is noted that the cosine metric is more robust in similarity features or dissimilarity features computation. Therefore, the present paper proposed to use a cosine space instead of Euclidean in an Enhanced VAT (EVAT). The EVAT uses the New Dissimilarity (ND), in which the construction of ND is shown in Eq.2.

$$ND\,(x, y) = 1 - (xy \,/\, \|x\|\|y\|) \qquad (2)$$

Here clustering tendency results of VAT in k-means based clustering are used for achieving the effective clustering results. Fig 1 shows the steps for proposed approach. This approach addresses the clustering tendency by visual methods and produces the quality of clustering results at a known clustering tendency. The advantages of the proposed method are:

- This hybrid approach generates the quality of clustering results because the clustering tendency is known.
- The substantial amount of execution time is saved because this approach need not to check the clustering results at different k values, they generate explicitly at known k value.
- The clustering tendency problem is detected effectively by proposed EVAT for some typical datasets.

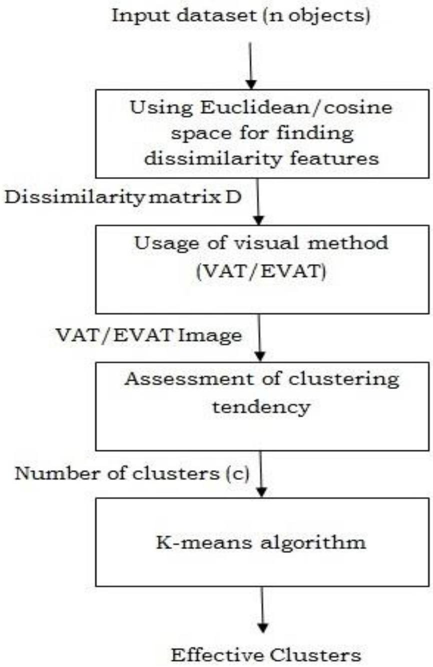

Input dataset (n objects)
Using Euclidean/cosine space for finding dissimilarity features
Dissimilarity matrix D
Usage of visual method (VAT/EVAT)
VAT/EVAT Image
Assessment of clustering tendency
Number of clusters (c)
K-means algorithm
Effective Clusters

Fig. 1 VAT-based-k-means-Clustering

IV. PROPOSED METHOD

The proposed hybrid approach consists of two major steps: The first step deals the method to find the value of cluster count by assessment of clustering tendency through data visualization method (VAT) for unlabeled data. The second step uses the standard clustering approach i.e. k-means based clustering algorithm, where it is applied on datasets for the known clustering tendency is obtained from the first step for discovering the faster and efficient clustering results.

Algorithm presents proposed hybrid approach in which the clustering tendency is detected from Step 1 and to discover the clustering results from Step 2. The Step 2 illustrates the k-means procedure. Hence, this hybrid approach is known as a VAT-based-k-means clustering algorithm.

```
Algorithm: VAT-Based-k-means-Clustering (For unlabeled
           dataset)
Step 1:
a)  Find Re-ordered dissimilarity image (I) using
    VAT/EVAT.
b)  Apply Image threshold on I.
```

c) Find histograms by applying consecutive operations of 2D FFT, Inverse of FFT and Correlation.

d) Extract the cluster count ‘k’ either from the number of histograms or square-shaped dark blocks of VAT/ EVAT Image.

Step 2:

a) Place ‘k’ (k is known from step1) number of initial points into the space represented by the objects that are being clustered i.e. initial group centroids.

b) Assign each data object into the group that has the closest centroid.

c) If the objects assigned into closest clusters, then recalculate the positions of the k- centroids.

Repeat above b and c Steps until the centroids are no longer move. This produces a separation of the objects into groups

V. RESULT AND DISCUSSIONS

The proposed study has presents the experimental results based on various synthetic data sets from S_1 to S_4 as shown in Fig. 3 collected from the UC Irvine Machine Learning Repository [5] for evaluating the performance of the method with respect to quality and run time.

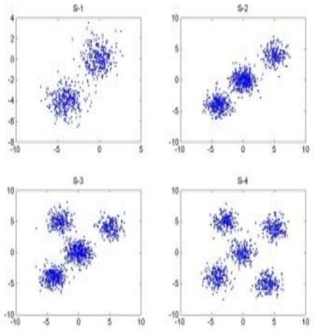


Fig. 2 Synthetic Datasets (S_1 to S_4)

In evaluation of result analysis, the existing system doesn’t have the prior value of clustering tendency. External interference is required for clustering tendency, so obtained results of existing system may or may not have good Dunn’s Index[12]. The higher value of Dunn’s Index indicates the good number of valid clusters for given data. Dunn’s Index is a metric for evaluating of correct partitioning [6]. K-means clustering algorithm is experimented several times until getting the good Dunn’s Index. So, the problem of existing system is runtime. Therefore, the proposed work first solves the problem of tendency by extracting of obtained square shaped dark blocks, secondly it retrieves k-means based clustering results based on tendency. This procedure output the VAT image for input dataset. After that we apply 2D FFT, inverse FFT, and correlation[16,17] on getting VAT image. Finally, the clustering number is extracted that which is referred as clustering tendency.

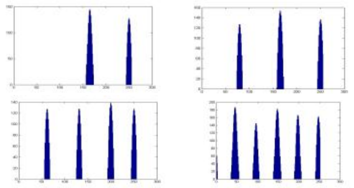
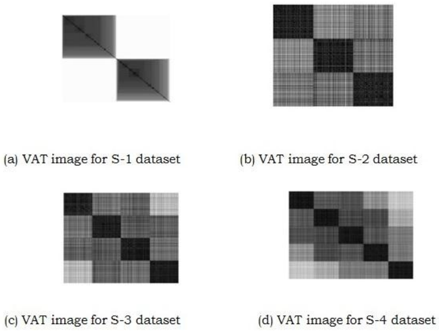


Fig. 3: Histograms for Synthetic data for clusters count (k)

The clustering algorithm cannot detect the clustering tendency. External user interference is required for finding the clustering tendency. However, the user is intractable to detect the suitable number of clusters. Therefore, the proposed hybrid systems use the VAT for detecting the number of clusters. Figure 4 and 5 shows the outputs of VAT for synthetic and real datasets respectively. The VAT Image gives the more informative assessment of clustering tendency by square-shaped dark blocks. VAT can access the number of clusters by square-shaped dark blocks. Each square-shaped dark blocks represent as a single cluster in VAT Image.



(a) VAT image for S-1 dataset

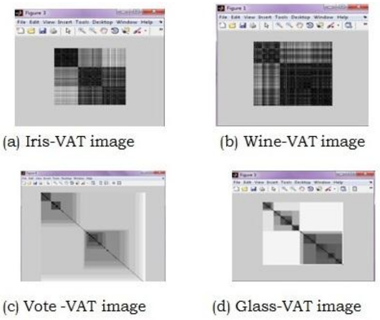
(b) VAT image for S-2 dataset

(c) VAT image for S-3 dataset

(d) VAT image for S-4 dataset

Fig. 4 VAT Images for Synthetic Datasets

The VAT can also detect number of clusters using VAT histograms. The VAT histograms are constructed by applying a series of three steps on VAT Image, which are 2D FFT, Inverse FFT, and correlation.



(a) Iris-VAT image

(b) Wine-VAT image

(c) Vote -VAT image

(d) Glass-VAT image

Fig.5 VAT Images for Real datasets

The clustering tendency is unknown in k-means clustering algorithm, hence, this section conduct the experiments for this algorithm at different k-values and perform the post-validation of these clustering results by Dunn’s Index (DI) for determining the best clustering. Table 1 and 2 illustrates k-means clustering results for synthetic and real datasets respectively, here the clustering tendency i.e. k value is unknown. However, the best k values in existing methods are found using the Dunn's Index value. The maximum value of DI indicates the best clustering.

The Dunn's Index computes the distance for the pairwise data objects, and it calculates the distance between two inter clusters. We use the Eq.3.

$$DI(c) = \min_{i \in c}\left\{ \min_{j \in c, j \neq i}\left\{ \frac{\delta(A_i, A_j)}{\max_{k \in c}\{\Delta(A_k)\}} \right\}\right\} \quad (3)$$

The problem of this approach is that it take more runtime for validating the clustering tendency and clustering results, because it is required to run the present clustering algorithms at multiple times for different k values. Since, 'k' value is unknown.

The proposed hybrid method obtains the value of clustering tendency by VAT. Thus, this hybrid method, namely, VAT-based-k-means clustering are faster. Table 3 & 4 shows the Dunn's Index and runtime results for VAT-based-k-means algorithm for synthetic and real datasets where Clustering Tendency is known.

Table 1: Dunn's Index and Runtime Results for Synthetic datasets (k-means Based Clustering-Clustering Tendency is Unknown)

Datasets	Dunn's Index for Number of clusters (C)				Run time (Sec)
	C=2	C=3	C=4	C=5	
S_1	0.44	0.03	0.02	0.02	0.19
S_2	0.41	0.79	0.04	0.03	0.25
S_3	0.19	0.31	0.37	0.01	0.26
S_4	0.02	0.02	0.01	0.22	0.30

Table 2: Dunn's Index and Runtime Results for Real-time Datasets (k-means Clustering-Clustering Tendency is Unknown)

Datasets	Dunn's Index for Number of clusters (C)				Run time (Sec)
	C=2	C=3	C=4	C=5	
Iris	0.06	0.05	0.01	0.08	0.59
Wine	0.02	0.01	0.02	0.02	0.25
Vote	0.19	0.20	0.14	0.14	0.35
Glass	0.03	0.04	0.01	0.06	0.27

Table 3: Dunn's Index and Runtime Results for Synthetic Datasets (Proposed Approach- Clustering Tendency is Known)

Synthetic Datasets	Dunn's Index (Clustering Tendency 'C' is extracted from proposed approach)	Runtime (Sec)
S_1	0.44(C=2)	0.09
S_2	0.79 (C=3)	0.14
S_3	0.37 (C=4)	0.21
S_4	0.22 (C=5)	0.24

Table 4: Dunn's Index and Runtime Results for Real Time Datasets (Proposed Approach- Clustering Tendency is Known)

Synthetic Datasets	Dunn's Index (Clustering Tendency 'C' is extracted from proposed approach)	Runtime (Sec)
Iris	0.06(C=2)	0.19
Wine	0.02 (C=2)	0.14
Vote	0.20 (C=3)	0.18
Glass	0.04(C=3)	0.10

Fig. 6 shows the runtime comparison between the proposed VAT-based-k-means and k-means clustering. This comparison analysis shows that hybrid methods are faster.

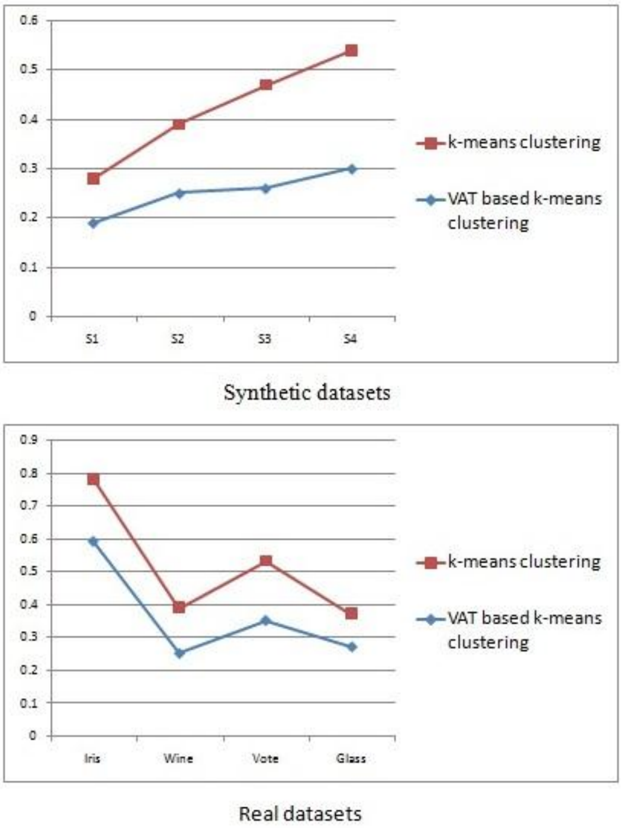


Fig.6. Runtime Comparison: between k-means and VAT-based-k-means clustering

VI. CONCLUSIONS

This proposed hybrid approach perform well toward to both visual assessment of clustering tendency and clustering results. This work evaluates the performance of proposed method by two measures i.e., clustering accuracy and normalized mutual information. The experimental results demonstrate that VAT-based-k-means is efficient and faster. However, the VAT requires more runtime for assessment of clustering tendency. Experimental results are tested on various synthetic datasets. Runtime and Dunn's Index values are evaluated and compared in both existing and proposed systems. According to the results analysis, the proposed system requires less time than existing system and also produces high quality of clustering results after observing of Dunn's Index value. Higher value of Dunn's

Index concludes the good clustering results. The future scope of the work is to obtain best indexed clustering results by techniques of sampling method and spectral approach in our proposed method.

REFERENCES

[1] Jain A.K.;Murty, M.N.; Flynn, P.J.; Data Clustering: A Review, Journal Acm Computing Surveys,31(3), 1999, 264-323.

[2] Jain, A.K.; Data clustering: 50 years beyond k-means, Pattern Recognition Letters 31(8), 2010, 651-666.

[3] Qinpei, Z.; Pasi, F.; Centroid Ratio for a Pairwise Random Swap Clustering Algorithm, IEEE Transactions on Knowledge and Data Engineering, 26(5) ,2014.

[4] Senoussaoui, M.; Kenny, P.; Stafylakis, T.; A study of the cosine distance-based mean shift for telephone speech diarization, IEEE Transactions on Audio, Speech, and Language Processing,22(1),2014,217-227.

[5] http.// archive.ics.uci.edu/ml/datasets.html.

[6] Cai, D.; He X.; Han, J.; Document clustering using locality preserving indexing, IEEE Transactions on Knowledge and Data Engineering, 17(2),2005, 1624-1637.

[7] L. Wang, T.Nguyen, J.Bezdek, C. Leckie , and K.Rammohanarao, "iVAT and aVAT: Enhanced visual analysis for clustering tendency assessment" in Proc PAKDD,India, Jun 2010.

[8] Timothy C. Havens, James C. Bezdek, " An efficient formulation of the improved visual assessment of cluster tendency" IEEE Trans on Knowledge and Data Engineering,Nov,2011.

[9] J.Bezdek and R.Hathaway, "VAT: A tool for visual assessment (cluster) tendency", in Proc. IJCNN, Honolulu, Hi,2002, pp.2225-30.

[10] M.Ester; P. Kriegel; J. Sander; X.xu," A density based algorithm for discovering clusters in large databases with noise" ,Int Conference on knowledge discovery and data mining 1996 ,pp 226-231.

[11] W. Wang; J. Yang; R. Muntz," STING: A statistical information grid approach to spatial data mining", Int Conf on very large data bases, pp 186-195.

[12] T.C. Havens, J.C.Bezdek, J.M.Keller,M. Popescu, " Dunn's Cluster Validity Index as Contrast Measure of VAT Images" Int Conf IEEE 2008.

[13] J.C. Bezdek, R.J. Hathaway, and J. Huband, "Visual Assessment of Clustering Tendency for Rectangular Dissimilarity Matrices". IEEE Trans. vol. 15, no. 5, pp. 890-903, 2007.

[14] I Dhillon, D. Modha, and W. Spangler, Proc. 30th Symp. Interface: Computing Science and Statistics, 1998, "Visualizing Class Structure of Multidimensional Data".

[15] T. Tran-Luu, PhD dissertation, Univ. of Maryland, College Park, "Mathematical Concepts and Novel Heuristic Methods for Data Clustering and Visualization, 1996.

[16] M. Breitenbach and G. Grudic, "Clustering through Ranking on Manifolds," Proc. 22nd Int'l Conf. Machine Learning (ICML), 2005.

[17] R.B. Cattell, "A Note on Correlation Clusters and Cluster Search Methods," Psychometrika, vol. 9, no. 3, pp. 169-184, 1944.

[18] P. Sneath, "A Computer Approach to Numerical Taxonomy," J. General Microbiology, vol. 17, pp. 201-226, 1957.

[19] Rousseeuw, P. J.: A Graphical Aid to the Interpretations and Validation of Cluster Analysis. J. Computational and Applied Math., Vol. 20, 1987, pp. 53–65.

[20] T.C. Havens, J.C. Bezdek, J.M. Keller, M. Popescu, and J.M. Huband, "Is VAT Really Single Linkage in Disguise?" Pattern Recognition Letters, 2008.

[21] Pena J. M., Lozano J. A. and Larranaga P., 1999. An empirical comparison of four initialization methods for the k-meansalgorithm, Pattern Recognition Letters, Vol. 20, No. 10, pp. 1027-1040.

[22] Xu R. and Wunsch D., 2005. Survey of clustering algorithms, IEEE Trans. Neural Networks, Vol. 16, No. 3, pp. 645-678.

[23] Xu Junling, Xu Baowen, Zhang Weifeng, and Hou Jun, Stable initialization scheme for K-means clustering, IEEE Trans.,Vol. 6, No.1. 2009.

[24] Savitzky, A.—Golay, M. J. E.: Smoothing and Differentiation of Data by Simplified Least Squares Procedures. Analytical Chemistry, Vol. 36, 1964, No. 8, pp. 1627–1639.

[25] Havens, T.—Bezdek, J.—Keller, J.—Popescu, M.: Clustering in Ordered Dissimilarity Data. Technical report, Univ. of Missouri 2007.

[26] Maimon, O.—Rokach, L.: Decomposition Methodology for Knowledge Discovery and Data Mining. World Scientific 2005, pp. 90–94.

[27] Bezdek, J. C.—Pal, N. R.: Some New Indices of Cluster Validity. IEEE Trans. System, Man and Cybernetics, Vol. 28, 1998, No. 3, pp. 301–315.

[28] Tibshirani, R.—Walther, G.—Hastie, T.: Estimating the Number of Clusters in a Data Set via the Gap Statistics. J. Royal Statistical Soc. B, Vol. 63, 2001, pp. 411–423.

[29] Calinski, R. B.—Harabasz, J.: A Dendrite Method for Cluster Analysis. Comm. In Statistics, Vol. 3, 1974, pp. 1–27.