An Extended Visual Technique for Exploratory Cluster Tendency Analysis

Veluru. Chinnaiah
Assistant Professor
Dept. of CSE, Vijaya Engineering College
Khammam, Telangana, India

B. V. RamNaresh Yadav
Associate Professor
Dept. of CSE, JNTUH
Hyderabad, Telangana, India

Abstract:

Cluster analysis is a popular technique for the classification of data objects based on similarity features. Top-clustering methods are majorly facing the problem of cluster tendency. Estimation of the number of clusters for unlabelled data is known as cluster tendency. Existing visual techniques, visual access tendency (VAT), and cosine-based VAT (cVAT) are used to enable the prior knowledge about the clustering tendency, in which similarity features are computed using distance metrics Euclidean and cosine. In a cosine metric, similarity features are computed using both the magnitude and direction of the data vectors for the pair of data objects. Thus, cluster tendency is also effectively determined using cVAT in some datasets. However, only a single viewpoint (i.e. origin) is taken as the reference in the computation of similarity features among data objects. For the more informative assessment, cosine-based spectral features are used in our proposed work to obtain a real cluster tendency visually to get an efficient clustering result. Performance analysis is conducted in the experimental study using various benchmarked datasets for demonstrating the efficiency of proposed cosine based spectral VAT (CS-VAT).

Abstract: Cluster Analysis, Cluster Tendency, Similarity Features, Visual Technique, Viewpoints

1. Introduction

Data clustering is an essential technique for making the data partitions of unlabelled data. Various applications of data clustering are text clustering [3], gene analysis [14], image clustering [15], video mining [13], e-commerce etc. It poses the key problem of cluster tendency. In top-clustering methods [17], the unknown value of cluster tendency 'k' is attempted (in k-means) with external interference i.e., the user must guess the correct 'k' value, some cases, it may be intractable. Bezdek et al. proposed the visual technique, VAT, for an assessment (or an estimation) cluster tendency, in which dissimilarity features are computed using distance measures [18], the matrix is reordered for finding the visual clusters in the form of black coloured square blocks in the diagonal of resulting visual image of VAT. Pre-cluster estimation is the significant idea of VAT, which useful for getting the quality of clusters. Euclidean or cosine metrics are used in VAT, cosine-based VAT (cVAT) [1], respectively, for the pre-estimation of clusters for the synthetic and real-life datasets. In some applications, like speech clustering [12], and text clustering [16], cVAT performs as the best compared to VAT due to the fact of both magnitude and direction of data vectors are considered while measuring either similarity or dissimilarity between any pair of data objects. Measuring the similarity between pair data objects are performed for origin in cosine distance measure. It means cosine distance measure uses only a single viewpoint (i.e., origin) and unable to use other reference points. Thus, for the significance of multi-viewpoints, visual technique, cVAT is extended with spectral features for accessing the more informative assessment in similarity (or dissimilarity) features computation, its proposed technique known as cosine based spectral VAT (CS-VAT). Its procedural steps are shown in Fig. 1. Initially, the data matrix is derived for the set of data objects and with their property values. Derive the spectral features for the set of data objects with the finding of normalized Laplacian matrix, which derived from two matrices, say, diagonal matrix 'M' of weighting matrix 'W' and it is shown in Eqn. (1)

$$L^1 = M^{-1/2} W M^{-1/2} \quad (1)$$

Find the cosine similarity for the spectral features of data objects with respect to a single viewpoint i.e., origin. The affinity or weighting matrix 'W' is computed as per the Eqn. (2), in which distance with nearest neighbors taken for smoothing the data clustering results.

$$W_{IJ} = \exp\left(-\frac{d_{ij}d_{ji}}{\sigma_i \sigma_j}\right) for\ i \neq j \quad (2)$$

Where d_{ij} refers to the dissimilarity between data objects di and dj, σ_i refers to the nearest k^{th} neighbour with data object di and σ_j refers to the nearest k^{th} neighbour with data object dj.

Reordered dissimilarity matrix 'RD' is derived for obtained Laplacian matrix using VAT, namely, Dissimilarity 'L^1'. With the RD, the indices of the objects are reordered according to multi-viewpoints similarity features.

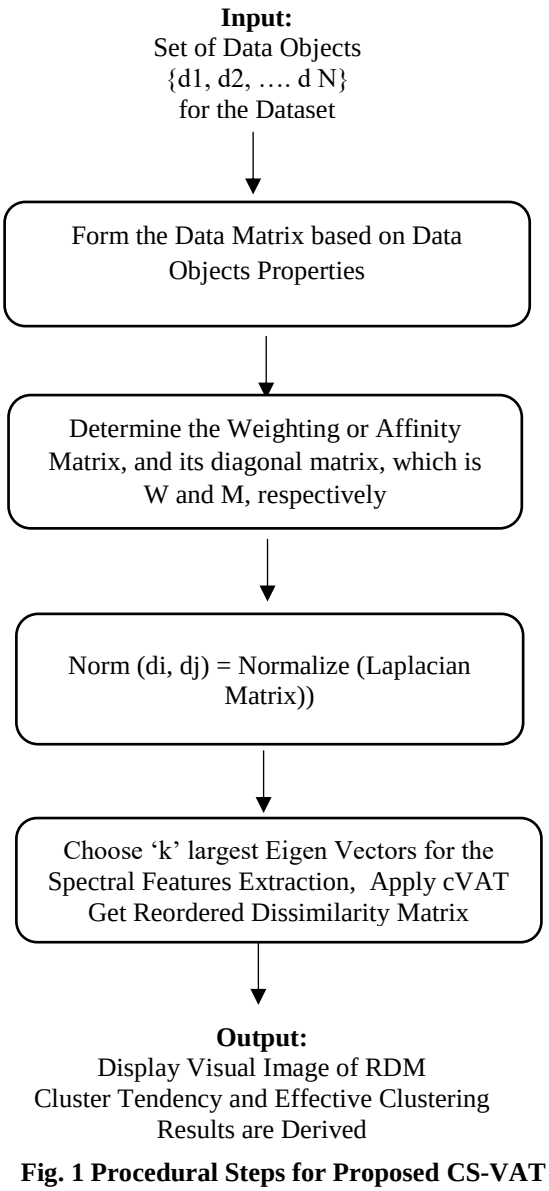


Fig. 1 Procedural Steps for Proposed CS-VAT

Finally, visual clusters are estimated with counting of appeared black coloured dark blocks found in the visual image's diagonal. Complete clustering results are also obtained with the crisp partitions of RD visual image.

Contributions of the proposed work are presented as follows:

1. Spectral features are derived for finding accurate similarity features among data objects
2. Visual image is displayed based on reordered spectral dissimilarity features of data objects
3. The best assessment of cluster tendency is performed with cosine based spectral VAT Image (CS-VAT Image)
4. Discover the robust clustering results by deriving the crisp partitions of CS-VAT Images
5. Demonstrate the efficiency of proposed work with several benchmarked synthetic and real-life datasets

The paper's remaining sections are described as follows: Section 2 presents the background study of the work; Section 3 discusses the proposed work; Experimental study is described in Section 4. Finally, the conclusion and scope of the work are presented in Section 5.

2. Background Study

Data clustering [8] is the technique of unsupervised classification. Popular techniques are k-means [9], minimum-spanning-tree (MST) based approaches [10], and other methods described in [1]. These techniques are recommended in real-time applications. These techniques need the prior value of 'k' for the quality of clustering results. The estimation of 'k' for the given dataset is known as cluster tendency. Pre-assessment of cluster tendency is the primary problem in such data clustering algorithms. Objects are classified into several clusters based on their similarity features. Pre-initialized clusters help for the quality of data clustering. Its issue is focused on this paper with the in-depth study of relevant work. Many algorithms are surveyed for determining the pre-cluster tendency. Bezdek et al. proposed a visual access tendency (VAT) [2] algorithm for the attracted visual based cluster tendency assessment results. In VAT, initially object's dissimilarity values are computed using an Euclidean distance and formed the dissimilarity matrix 'DM'; reordering the dissimilarity matrix is performed with basic Prim's [11] algorithm, in which objects indices are changed according to their similarity features. The image of RDM is displayed- the reordering of objects indices shown finally as a set of clusters, and it is reflected and shown with square black coloured blocks while displaying the RDM image. This image is known as VAT Image. Estimation of pre-clusters is determined with the count of the number of black or grey coloured blocks. Sample VAT outputs are shown in Fig. 2

$$DM = \begin{bmatrix} 0.00 & 0.73 & 0.19 & 0.71 & 0.16 \\ 0.73 & 0.00 & 0.59 & 0.12 & 0.78 \\ 0.19 & 0.59 & 0.00 & 0.55 & 0.19 \\ 0.71 & 0.12 & 0.55 & 0.00 & 0.74 \\ 0.16 & 0.78 & 0.19 & 0.74 & 0.00 \end{bmatrix} \quad I =$$

(a) Dissimilarity matrix 'DM' and its dissimilarity image 'I' (before applying the VAT)

$$RDM = \begin{bmatrix} 0.00 & 0.12 & 0.59 & 0.73 & 0.78 \\ 0.12 & 0.00 & 0.59 & 0.71 & 0.74 \\ 0.59 & 0.55 & 0.00 & 0.19 & 0.19 \\ 0.73 & 0.71 & 0.19 & 0.00 & 0.16 \\ 0.78 & 0.74 & 0.19 & 0.16 & 0.00 \end{bmatrix} \quad I =$$

(b) After applying VAT, Re-ordered dissimilarity matrix (RDM), and its image 'I'

Fig. 2 Visual Method – VAT- Illustrative Example

Spectral VAT (SpecVAT) [4] uses data objects' spectral features for the better data clusters assessment and is expensive for large datasets. Another method, cVAT [19] uses the cosine measure for computing the similarity features among the data objects, which show impressive results in specific applications like text and speech clustering. State-of-the-art of VAT methods assess cluster tendency effectively with a Euclidean and cosine distance measures. It noted that cosine is preferable due to its stunning results in data clusters assessment—a single viewpoint used in cVAT for pre-clusters assessment for the synthetic datasets and real-life datasets.

Hybrid clustering methods are proposed in [2], which combines the visual techniques and traditional methods, k-means, and MST-based clustering approaches. These hybrid approaches are VAT-based-k-means, and VAT-based-MST-clustering algorithms [20] uses the cluster tendency value in k-means and MST-based clustering approaches. These techniques enable cluster tendency and clustering results. The problem of hybrid approaches is the processing complexity in terms of time and memory allocation. The problem is effectively handled in [1] with the visualized clustering approach (VCA), and it derives the crisp partitions instead of reusing the traditional algorithms, unlike k-means and MST-based clustering approaches. It is less expensive and produces the quality of data clustering results than hybrid clustering methods. Cosine based VCA is the best choice for data clustering. The extended idea of cosine based spectral VAT and the development of cosine based spectral- VCA (CS-VCA) are described in the following section.

3. Proposed Work

The proposed work uses spectral features while performing the similarity features computation. Here, the largest k-eigenvectors are computed from the Laplacian matrix, in which affinity values of data objects are computed to k-nearest neighbours; it has taken the more informative assessment with k-neighbours. Weighted matrix W and diagonal matrix are the primary computation sources for deriving the spectral features. The critical formulas involved in the work is shown in Eqn. (3)

$$CS\left(d_i, d_j\right) = \sum_{i,j=1 \text{ and } i \,!=j}^{N} \text{cosine}\left(d_i, d_j\right), \qquad (3)$$

Where d_i and d_j are collected from k-largest Eigenvectors of i^{th} and j^{th} rows of Eigen matrix of Laplacian matrix

Following algorithm 1 describes the procedural steps of cosine based spectral VAT (CS-VAT)

Algorithm 1: CS-VAT

Input :

Set of data objects with size of N, $\{d_1, d_2, \ldots\ldots d_N\}$

Output:

Cluster Tendency – k

Methodology:

1. Find the data matrix of DM with the properties (or values) for the set of documents $\{d_1, d_2, \ldots\ldots d_N\}$
2. For i = 1 to N
   Compute affinity scale σ_i for the i^{th} data object using nearest k^{th} nearest neighbour, $\sigma_i = d(d_i, d_k)$
   For j = 1 to N
   Compute affinity scale σ_j for the j^{th} data object using nearest k^{th} nearest neighbour, $\sigma_j = d(d_j, d_k)$
   Compute W_{ij} using Eqn. (3)
   End for
   End for
3. Determine the diagonal matrix 'M' of W and construct the Laplacian matrix L^1 using Eqn. (1)
4. Select the first k-largest Eigen vectors of L^1 and form the matrix spectral based data matrix, S-DM with size of N x k (for N data object with 'k' Eigen vectors)

5. Apply cVAT on S-DM in order to find the cosine based spectral re-ordered dissimilarity matrix (CS-RDM)
6. Display the image or heatmap of CS-RDM
7. Access the cluster tendency ‘k’ with the informative counting assessment of square-shaped black (or grey) coloured blocks appeared in the CS-RDM image
8. Return k

Based on data objects' properties, the data matrix is derived in step 1 of algorithm 1. Step 2 shows the computation of affinity values of data objects based on k-nearest neighbours. Further, weighted matrix and diagonal matrices are derived with reference to affinity values in order to determine the corresponding Laplacian matrix L^1 and it shown step 3. Spectral features are selected with the largest-k-Eigenvectors and respective matrix S-DM, which denotes the spectral based data matrix with size of N x k for the N documents and illustrated in Step 4. In step 5, cVAT is applied, in which cosine based spectral dissimilarity (or similarity matrix is derived and its heatmap image shown in Step 6 for the visual clusters information. Clusters estimation is accessed with black coloured blocks along the diagonal of CS-RDM image in step 7 and returns the value of numbering clusters 'k' in step 8.

Algorithm 2: CS-VCA

Input:
CS-RDM – cosine based spectral reordered dissimilarity matrix
Cluster tendency – k

Output: Data clusters

Methodology

1. Pick the cluster tendency ‘k’ and CS-RDM from CS-VAT for the set of data objects $\{d_1, d_2, \ldots\ldots d_N\}$
2. Display the heatmap or image of CS-RDM
3. Derive the crisp partitions of CS-RDM
4. Find the cluster labels of data objects with the information of crisp partitions
5. Discover the data clusters by placing the data objects into respective clusters according to cluster labels of data objects.

CS-VCA derives the complete clusters information with the accessed value of cluster tendency 'k' in CS-VAT algorithm. A visual heatmap image of step 2 is taken for the derivations of the crisp partition matrix and its explained in step 4 of algorithm 2. Associated cluster labels of data objects are derived from the crisp partition matrix information, then the complete clusters for the data objects are discovered, which is illustrated in step 5.

The proposed work develops the data clusters assessment methods based on spectral with two algorithms CS-VAT, and CS-VCA. Supporting experimental study is discussed in the following section.

4. Experimental Study

Performance of visual techniques, existing and proposed CS-VAT, are evaluated and compared in the demonstration of experimental conduced benchmarked synthetic and real-life datasets. Data sets, Goodness of visual images, and performance study are presented in the following sub-sections.

4.1 Datasets

Synthetic datasets shown in Fig. 3 are generated, and real-life datasets of [5] used (which are mentioned in Table 1) in our experiment. Synthetic datasets are generated by fixing the appropriate mean and covariance values between the data points in two-dimension space.

Table 1: Details of Real-life datasets

Real-life dataset	Name of the Dataset	Number of Clusters
R-1	Seeds	3
R-2	Iris	3
R-3	Voting	2
R-4	Wine	3

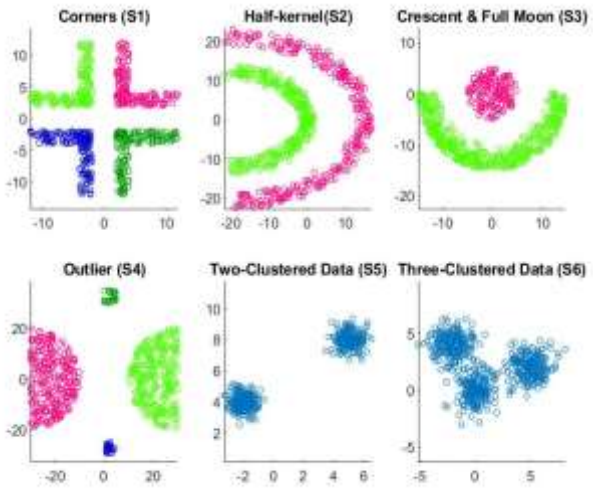


Fig. 3. Synthetic Datasets

4.2 The goodness of Visual Images

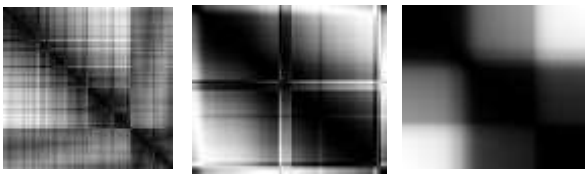
Visual images of VAT, cVAT (existing methods) and CS-VAT (proposed method) are displayed in Fig. 4 for comparative analysis. With the observation of visual images, it noted that CS-VAT had shown the excellent clarity of visual images compared to VAT, cVAT images.

VAT cVAT CS-VAT

(a) Synthetic Data – S1

VAT cVAT CS-VAT

(b) Synthetic Data – S2

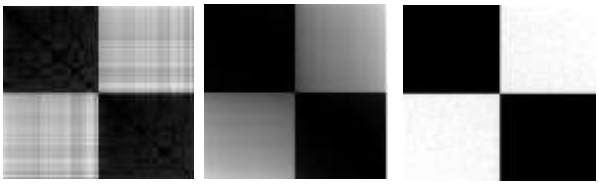


VAT cVAT CS-VAT

(c) Synthetic Data – S3

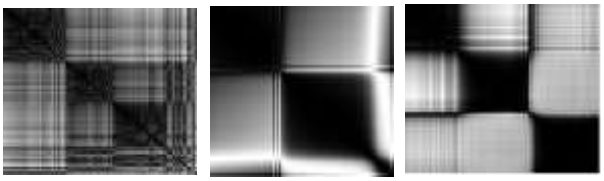
VAT cVAT CS-VAT

(d) Synthetic Data – S4



VAT cVAT CS-VAT

(e) Synthetic Data – S5



VAT cVAT CS-VAT

(f) Synthetic Data – S5

Fig. 4: Visual Image Clusters for the Visual Techniques (Top-Bottom for S1 to S6 Datasets, Left-to- Right for VAT, cVAT, CS-VAT Visualized Clustering Results)

Thus, there is possible for the good assessment of cluster tendency with CS-VAT images. Clarity or goodness of images are obtained through OTSU and presented in Table 2. The goodness of images is indicated with bold font. MV-VAT images score more goodness as per achieved results from this table.

Suppose the number of pixels are denoted with L grey levels i.e., {1,2,….L} and $N=n_1+ n_2 +…….+ n_L$, here n_i defines the number of pixels consists of 'i' intensity. Background and foreground (square-shaped clusters) objects information is accessed with two classes, namely, C_0 and C_1; whereas C_0 and C_1 represents the pixels with grey level sets {1,2,…k}, and {k+1,….L}respectively.

$$\mu_0 = \sum_{i=1}^{k} iPr(i|C_0) = \frac{\mu(k)}{w(k)} \quad (4)$$

$$\mu_1 = \sum_{i=1}^{k} iPr(i|C_0) = \frac{\mu_T - \mu(k)}{1 - w(k)} \quad (5)$$

$$w(k) = \sum_{i=1}^{k} p_i \; and \; p_i = {}^{n_i}/_{N}; \; \sum_{i=1}^{L} p_i = 1 \quad (6)$$

$$\mu(k) = \sum_{i=1}^{k} ip_i \quad (7)$$

$$\mu_T = \mu(L) = \sum_{i=1}^{k} ip_i \quad (8)$$

Variances are computed using Eqn. (9) and (10)

$$\sigma_0^2 = \sum_{i=1}^{k} (i - \mu_0)^2 ({}^{p_i}/_{w_0}) \quad (9)$$

$$\sigma_1^2 = \sum_{i=1}^{k} (i - \mu_1)^2 ({}^{p_i}/_{w_1}) \quad (10)$$

In the measure of cluster (or class) separability, Eqn. (11) to (13) are used and which indicate the within-cluster variance, between cluster-variance, and the total variance of levels.

$$\sigma_w^2 = w_0\sigma_0^2 + w_1\sigma_1^2 \quad (11)$$

$$\sigma_B^2 = w_0 w_1 (\mu_1 - \mu_0)^2 \quad (12)$$

$$\sigma_T^2 = \sum_{i=1}^{L} (i - \mu_T)^2 p_i \quad (13)$$

The optimal threshold is selected 'k' is selected in such a way that it maximizes the following OTSU Eqn. (14)

$$n(k) = {}^{\sigma_B^2(k)}\Big/_{\sigma_T^2} \quad (14)$$

Table 2: Goodness of Images of Visual Techniques

Dataset	VAT	cVAT	CS-VAT
R-1	0.465	0.481	**0.759**
R-2	0.396	0.512	**0.640**
R-3	0.291	0.443	**0.789**
R-4	0.458	0.532	**0.644**
S-1	0.465	0.509	**0.786**
S-2	0.489	0.612	**0.684**

S-3	0.479	0.521	**0.663**
S-4	0.496	0.332	**0.999**
S-5	0.917	0.651	**0.999**
S-6	0.496	0.765	**0.866**

Table 3: Clustering Accuracy (CA) of Images of Visual Techniques

Dataset	VAT-VCA	cVAT-VCA	CS-VAT-VCA
R-1	0.814	0.814	**0.885**
R-2	0.866	0.866	**0.921**
R-3	0.581	0.583	**0.791**
R-4	0.713	0.711	**0.730**
S-1	**0.375**	**0.375**	**0.375**
S-2	**1**	**1**	1
S-3	0.63	0.63	**0.72**
S-4	0.518	0.518	**0.52**
S-5	**1**	**1**	**1**
S-6	0.796	0.718	**0.961**

Table 4: Normalized Mutual Information of Images of Visual Techniques

Dataset	VAT-VCA	cVAT-VCA	CS-VAT-VCA
R-1	0.497	0.497	**0.647**
R-2	0.696	0.692	**0.712**
R-3	0.111	0.121	**0.221**
R-4	0.458	0.512	**0.644**
S-1	**0.251**	**0.251**	**0.251**
S-2	**1**	**1**	**1**
S-3	0.49	0.49	**0.59**
S-4	0.368	0.368	**0.465**
S-5	**1**	**1**	**1**
S-6	0.551	0.457	**0.856**

4.3 Performance Evaluation

Two measures are used, i.e. clustering accuracy (CA) [6], and normalized mutual information (NMI) [7] are taken for the performance evaluation of visual techniques. Formulas of CA and NMI are shown in Eqn. (15) and (16), respectively.

$$CA = \frac{\sum_{i=1}^{N} \delta(g_i, f_i)}{N} \quad (15)$$

where $\delta(g_i, f_i) = 1$ if and only if $g_i = f_i$

g_i, f_i is referred to the ground and obtained cluster labels for the i^{th} data object.

$$NMI = \frac{\sum_{i=1}^{c}\sum_{j=1}^{k} \frac{n_{ij}}{n} \log \frac{n_{ij} n}{n_{i.} n_{j}}}{\sqrt{\left(\sum_{i=1}^{c} \frac{-n_{i.}}{n} \log \frac{n_{i.}}{n}\right)\left(\sum_{j=1}^{k} \frac{-n_{j}}{n} \log \frac{n_{j.}}{n}\right)}} \quad (16)$$

Where, $n_{i.} = \sum_{j=1}^{k} n_{ij}$, $n_{.j} = \sum_{i=1}^{c} n_{ij}$, n, c, k are denoting the total number of data objects, clusters, and estimated classification of clusters respectively—performance of visual techniques described in Table 3 and Table 4 shown using CA and NMI respectively. Higher values of CA and NMI indicates the best clustering results. The values of both CA and NMI observed that CS-VCA scores the best values compared to VAT-VCA and cVAT-VCA. CS-VAT has a good visual image for showing the data clusters, and its crisp clustering partitions are also generated as the best. Data clustering of CS-VCA depends on CS-VAT visual crisp partitions; thus, CS-VCA is significantly improved compared to other visual techniques.

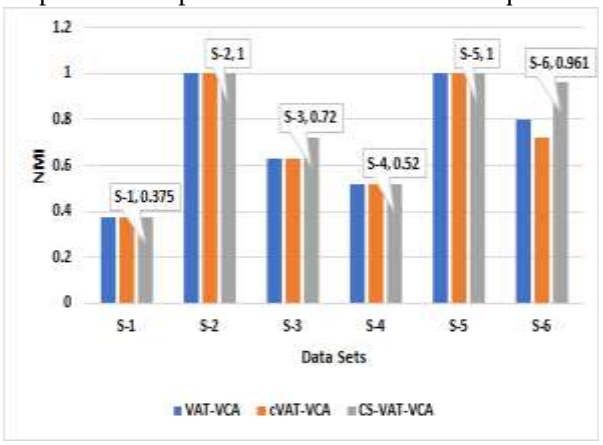


Fig. 5 Empirical Analysis of Visual Techniques (with CA) for Synthetic Data

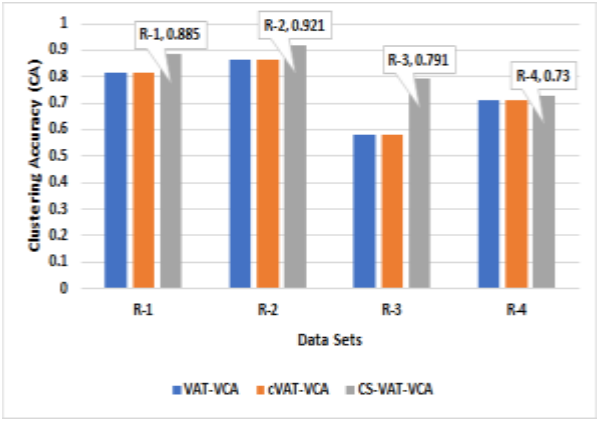


Fig. 6 Empirical Analysis of Visual Techniques (with CA) for Real-life Data

Fig. 7 Empirical Analysis of Visual Techniques (with NMI) for Synthetic Data

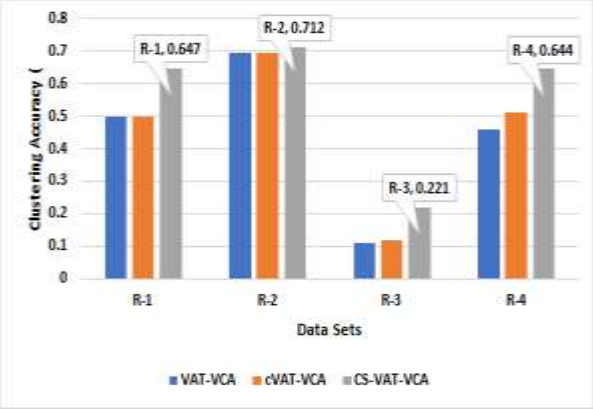


Fig. 8 Empirical Analysis of Visual Techniques (with NMI) for Real-life Data

A graph-based empirical analysis for the visual techniques shown in Fig. 5 and Fig. 6 for synthetic and real-life datasets, respectively. It stated that CS-VCA achieves good CA empirical values. The same observation is made in the empirical analysis of NMI values in Fig. 7 and Fig. 8, respectively.
In the experimental, cosine-based similarity features are computed with (N-2) viewpoints. For large datasets or big datasets, it demands more computational time and memory allocations. In future work, it recommends developing the scalable visual techniques for the large datasets.

5. Conclusion and Scope of the Work

The visual techniques are widely used for the estimation of cluster tendency in data clustering problems. Existing techniques use the Euclidean and cosine distance measures for finding the similarity features of data objects, in which cosine is recommended for the best assessment of cluster tendency in various applications. However, it uses a single viewpoint in the similarity features computation among the data objects. Our proposed work uses the spectral features; thus, it discovers the robust clustering results in the experimental study. Due to improving the scalability feature, it suggests that develop scalable visual techniques for large datasets.

References

1. P. Rathore, D. Kumar, J. C. Bezdek, S. Rajasegarar and M. Palaniswami, "A Rapid Hybrid Clustering Algorithm for Large Volumes of High Dimensional Data," in *IEEE Transactions on Knowledge and Data Engineering*, vol. 31, no. 4, pp. 641-654, 1 April 2019, doi: 10.1109/TKDE.2018.2842191.
2. Kumar D, Bezdek JC, Palaniswami M, Rajasegarar S, Leckie C, Havens TC, A hybrid approach to clustering in big data. IEEE Trans Cybern 46(10):2372–2385, (2016)
3. C. C. Aggarwal and C. K. Reddy, Data clustering: algorithms and applications. CRC Press, 2013.
4. Bezdek, James Leckie, SpecVAT: Enhanced visual cluster analysis, IEEE International Conference on Data Mining, ICDM, 2008.
5. A. Asuncion and D. Newman, "Uci machine learning repository," 2007.
6. Pattanodom, M., I am-On, N., and Boongoen, T. " Clustering data with the presence of missing values by ensemble approach," 2016 Second Asian Conference on Defense Technology (ACDT). doi:10.1109/acdt.2016.7437660
7. Alessia Amelio and Clara Pizzuti, " Is Normalized Mutual Information a Fair Measure for Comparing Community Detection Methods? , " IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining, 2015.
8. M Suleman Basha, S K Mouleeswaran, K Rajendra Prasad, Cluster Tendency Methods for Visualizing the Data Partitions, International Journal of Innovative Technology & Exploring Engineering, 2019
9. JS Low, Z Ghafoori, JC Bezdek, C Leckie, Seeding on samples for accelerating k-means clustering, Proceedings of the 3rd International Conference on Big Data and Internet of Things, 2019.
10. S. Singh, R. Srivastava, V. Kumar and S. Agarwal, "An approximate algorithm for degree constraint minimum spanning tree," 2010 International Conference on Computer and Communication Technology (ICCCT), Allahabad, Uttar Pradesh, 2010, pp. 687-692, doi: 10.1109/ICCCT.2010.5640455.
11. S. Chopade and P. More, "Effective bug triage with Prim's algorithm for feature selection," *2017 International Conference on Signal Processing and Communication (ICSPC)*, Coimbatore, 2017, pp. 217-220, doi: 10.1109/CSPC.2017.8305842.
12. K Rajendra Prasad, M Suleman Basha, Improving the performance of speech clustering method, IEEE- 10th International Conference on Intelligent Systems and Control (ISCO), 2016.
13. D. Saravanan and S. Srinivasan, "Data mining framework for video data," *Recent Advances in Space Technology Services and Climate Change 2010 (RSTS & CC-2010)*, Chennai, 2010, pp. 167-170, doi: 10.1109/RSTSCC.2010.5712827.
14. J. Xie et al., "Prediction of Essential Genes in Comparison States Using Machine Learning," in IEEE/ACM Transactions on Computational Biology and Bioinformatics, doi: 10.1109/TCBB.2020.3027392.
15. A. Tariq and H. Foroosh, "T-clustering: Image clustering by tensor decomposition," *2015 IEEE International Conference on Image*

Processing (ICIP), Quebec City, QC, 2015, pp. 4803-4807, doi: 10.1109/ICIP.2015.7351719.

16. Lin, Yung ShenJiang, Jung Yi,Lee, Shie Jue, A similarity measure for text classification and clustering, IEEE Transactions on Knowledge and Data Engineering (2014)
17. Rajendra Prasad, K., Mohammed, M. ,Noorullah R.M. "Visual topic models for healthcare data clustering", *Evolutionary Intelligence, 2019*
18. S. Mahallati, J. C. Bezdek, D. Kumar, M. R. Popovic, and T. A. Valiante, "Interpreting cluster structure in waveform data with visual assessment and dunn's index," in Frontiers in Computational Intelligence. Springer, 2018, pp. 73–101
19. Hu Y, John A, Wang F, Kambhampati S (2012) Et-LDA: joint topic modeling for aligning events and their twitter feedback. In: AAAI conference on artifcial intelligence (AAAI 2012), vol 12, Toronto, ON, Canada, pp 59–65
20. E. J. Otoo, A. Shoshani, and S.-w. Hwang, "Clustering high dimensional massive scientific datasets," Journal of Intelligent Information Systems, vol. 17, no. 2-3, pp. 147–168, 2001