Nanotechnology for Environmental Engineering
https://doi.org/10.1007/s41204-022-00222-8

ORIGINAL PAPER

Detection of pre-cluster nano-tendency through multi-viewpoints cosine-based similarity approach

M. Suleman Basha[1] · S. K. Mouleeswaran[1] · K. Rajendra Prasad[2]

Received: 20 November 2021 / Accepted: 10 January 2022

Abstract

Pre-clusters assessment is a significant problem in data clustering. It found that visual cluster tendency assessment (VAT) is majorly focused on addressing the problem of pre-clusters assessment. This visual technique initially derives the similarity features of data objects using either cosine or Euclidean distance metrics. Cosine is considering both magnitudes and direction of the vectors; thus, it greatly succeeded in data clustering applications. Only a single viewpoint (i.e., origin) is used in the cosine metric. Finding the similarity features using multiple viewpoints is more accurate than a single viewpoint cosine metric. This paper presents the multi-viewpoints cosine-based similarity VAT (MVS-VAT) which considers the multi-viewpoints for an effective assessment of nano-pre-clusters (or nano-cluster tendency). Clustering accuracy (CA) and normalized mutual information (NMI) are taken for measuring the performance of the existing and proposed methods. It is proved that the efficiency of the proposed MVS-VAT is improved from 20 to 40% compared to VAT and cVAT concerning the parameters of CA and NMI. Therefore, the quality of data clusters is obtained through the proposed technique MVS-VAT. Experimental is conducted on several benchmarked datasets for illustration of an empirical study of the existing and proposed techniques.

Keywords Data clustering · Pre-clusters assessment · Viewpoints · Cluster tendency · VAT · MVS-VAT

Introduction

Data objects need to be clustered (or classified) using the similarity features for the data clustering problem. Data clustering [1] is widely used in applications and includes web clustering [2], text clustering [3], image and signal processing [4], video mining, spatial mining for weather forecasting [5], big data clustering [6, 7], etc. Top methods [8] of data clustering are k-means, hierarchical, etc., that efficiently generate the data clusters for any related applications. However, pre-clusters are intractable which affects on quality of data clusters. During the study of various pre-clusters assessment methods, it has been observed that visual-related data clustering techniques [9, 10] effectively solve the problem of cluster tendency. These techniques are visual assessment of cluster tendency (VAT) and cosine-based VAT (cVAT). With the information of magnitudes and direction of the vectors, cosine-based similarity features are derived in cVAT. Euclidean distance takes only the distance in the derivation of similarity features. It is the reason to assess the clustering tendency (or determining the clusters assessment) effectively in cVAT than VAT. In cVAT, all the similarity feature among the data objects is derived from a single viewpoint. More informative assessment is performed using a multi-viewpoint instead of a single viewpoint. The proposed work incorporates this technique, hence, called multi-viewpoints cosine-based similarity VAT (MVS-VAT).

The procedural steps of the proposed work are shown in Fig. 1. Illustrative steps of this architecture are shown mentioned as follows: initial step of the proposed work is to take the input data $D = \{o1, o2, \ldots..on\}$. Pre-clusters assessment through multi-viewpoints is the key objective of the

✉ K. Rajendra Prasad
krprgm@gmail.com

M. Suleman Basha
suleman.ndl@gmail.com

S. K. Mouleeswaran
meetmoulee@gmail.com

[1] Department of Computer Science and Engineering, Dayananda Sagar University, Bangalore, India

[2] Department of Computer Science and Engineering, Rajeev Gandhi Memorial College of Engineering and Technology, Nandyal, India

Published online: 02 February 2022

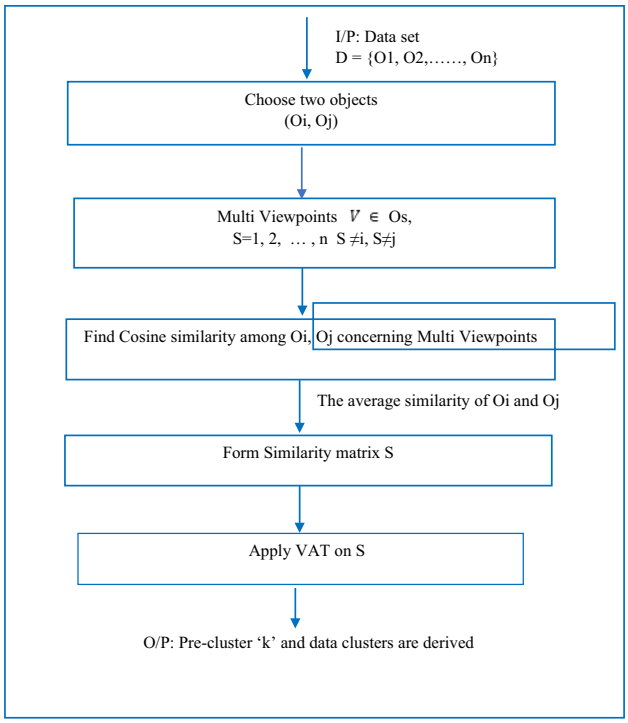


Fig. 1 Procedural steps of proposed work

work. Any two objects are selected, say, *oi* and *oj*, and other objects in *D* are considered as multi-viewpoints. The number of multi-viewpoints is restricted as $(n-2)$, i.e., taken the data objects in *D* except for *oi* and *oj*. Cosine similarity is applied between the data objects *oi* and *oj* for each multi-viewpoint $(n-2)$. Similarity values are derived between data objects *oi* and *oj*, taking an average of $(n-2)$ similarity values. The average value refers to the similarity feature between data objects *oi* and *oj*; for $i=1$ to n and $j=1$ to n nested iterations, compute the similarity feature between data objects *oi* and *oj*, for *i* is not equal to *j*. Finally, the similarity values among the data objects are placed in a single matrix, called a similarity matrix.

Nano-value of cluster tendency is required for the pre-clusters assessment problem. It is obtained using the approaches of VAT and cVAT. The cVAT finds the cosine-based similarity features among the data objects with reference to a single viewpoint. The aim of the proposed MVS-VAT is to find the cosine-based similarity features with reference to multiple viewpoints for the best assessment of pre-clusters in the data clustering.

The proposed work contributions are mentioned as follows:

1. Visualize the pre-clusters detections though lessen bright colored (i.e., dark-colored) blocks
2. Similarity features are extracted through multi-viewpoints
3. Develop the VAT procedure with a multi-viewpoints cosine-based similarity technique.
4. The quality of data clusters is assessed and generates the quality of data clusters.
5. Visualize the image clusters through MVS-VAT.

Sections 2, 3, 4, and 5 present the study of the work, MVS-VAT procedure, experimental study, and conclusion of the paper, respectively.

Study of the work

Post- and pre-clustering techniques are used for determining the number of partitions for the datasets priorly. Post-data clustering takes more computational time for finding the clustering tendency of unlabeled datasets [10]. Pre-clustering techniques [11] are the best for unlabeled data clustering when compared to post-data clustering techniques concerning computation time analysis. Bezdek et al. [9, 12, 13] proposed the visual assessment cluster tendency (VAT) and other models of VAT, improved VAT. These models use the Euclidean distance metric for finding either dissimilarity or similarity features for the pair of data objects over the dataset. Basic VAT steps are presented in Algorithm 1.

Algorithm 1: Visual Assessment of cluster tendency (VAT)

```
Step1:
        Obj_or = { };
        Obj_int_or ={0,1,.....n-1}
        Determine max of ObjdM [ ] [ ], and its index value are stored into (i,j)
        OrP(0) = i;
        Obj_or = {i};
        Obj_int_or=Obj_int_or - {Obj_or};

Step2:
        for (k=1;k<n; k++)
        {
         Find(i,j) from min {ObjdM[i][j]}
         where  i ϵ Obj_or, j ϵ Obj_int_or }
          Obj_or ={ Obj_or }U{j};
          OR ={Obj_int_or}-{Obj_or};
          Obj_int_or=OR;
          OrP(k) = j;
         }
Step3:
    /*Reordered Dissimilarity Matrix Computation*/
       for(i=0;i<n; i++)
            for (j=0;j<n; j++)
               RDM=ObjdM (OrderP[i],OrP[j]);
Step 4:
        Display Image (RDM)
```

For path-shaped data or any kind of complex dataset, iVAT is efficient for the generation of data clusters. It computes the path-based distances [14] for defining the dissimilarity matrix for the set of data objects over the datasets.

Another model of VAT, say cVAT [15, 16], is described in Algorithm 2, which uses the cosine metric for finding the similarity features among the objects concerning a single viewpoint. Justification of similarity features using a single viewpoint is not more appropriate. Calculating a similarity feature using multiple viewpoints is more accurate than a single viewpoint. Development of multi-viewpoints cosine-based similarity schema is presented in the next following section for the best data clustering results.

Algorithm 2: Cosine based Visual Assessment of cluster tendency (cVAT)

Step1:
Compute a local scale σi for each object oi using σi=d(oi, on) = din, where ok is the n-th nearest neighbour of oi.
Step2:
Construct the weighting matrix W € Rn×n
Defining wij= exp(−dijdij/(σiσj)) for i = j, and wii= 0.
Step3:
Let M to be a diagonal matrix with mii= ∑j=1wij
$L=M^{-1/2}W\,M^{-1/2}$*, is a normalized version of the Laplacian matrix.*
Step3a:
n=2;
repeat
Step4:
Choose v1, v2, · · · ,vk, the k largest eigenvectors of L to form the matrix V = [v1, · · · , vk] € R n×k by stacking the eigenvectors in columns.
Step5:
Construct new dissimilarity matrix Dnew between Pair of objects using cosine distance formula
Step6:
Apply VAT to Dnew to obtain I(Dnew),k=k+1
until k=kmax

MVS-VAT procedure

Multi-viewpoints are used as reference points in the cosine similarity metric. The similarity value is computed for every viewpoint.

Multi-viewpoints cosine-based similarity (MVS)

The MVS-VAT consists of two key procedural steps, which are as follows: compute the average similarity value over the $(n-2)$ multi-viewpoints, and VAT is applied on the resulting dissimilarity matrix of MVS for assessing the pre-clusters information of unlabeled data. The procedural idea of MVS is shown in Fig. 2.

The cosine-based similarity features for the pair of objects (e.g., ($v1$, $v2$)) are computed concerning $(n-2)$ viewpoints, here $n=5$, and $(n-2)=3$, treat all the data objects except the source pair values are considered as the multi-viewpoints. In this example, valid multi-viewpoints are v3, v4, and v5. The similarity features are computed in MVS for the given example as follows:

i. S1 = Cosine_similarity($v1$,$v2$) concerning the first viewpoint $v3$
ii. S2 = Cosine_similarity($v1$,$v2$) concerning the first viewpoint $v4$
iii. S3 = Cosine_similarity($v1$,$v2$) concerning the first viewpoint $v5$

Compute an average of S1, S2, and S3 for the final cosine base similarity for the pair of data objects ($v1$,$v2$). Similarly, cosine-based similarity is computed for other combinations of pair of data objects, i.e., ($v1$, $v3$), ($v1$, $v4$).....($v2$, $v3$), ...($v4$, $v5$).

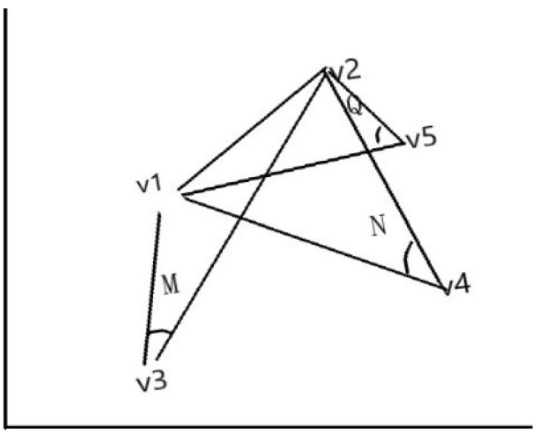


Fig. 2 Multi-viewpoints cosine-based similarity computation

MVS-VAT algorithm

Algorithm: MVS-VAT

Input : Dataset D = {o1,o2,.....on}

Output : Number of Clusters 'k' and their data objects

Methodology:

1. *Number of multi-viewpoints N1= (n-2);*
2. *For i= 1 to n*
   For j= 1 to n

   Choose the data objects pair (p,q)

   (p,q)=(o_i, o_j)

$$S = \frac{1}{N1} \sum_{m \in D \text{ and } m \neq p, m \neq q}^{N1} cosine(p,q) with\ the\ refernce\ of\ viewpoint\ 'm'$$

   MVS(i,j) = S;

   Diss(i,j) = 1- Normalized (MVS(i,j));
3. *Apply VAT Procedure with the input of 'Diss' which displays the Image, known as MVS-VAT Image*
4. *Determine the pre-clusters information 'k' (also referred to as the cluster tendency 'k') with the counting of visible dark blocks that appeared along the diagonal of MVS-VAT Image*
5. *Generate the crisp-partitions to obtain the cluster labels of data objects*
6. *Derive the data clustering results based on the detected cluster labels of data objects.*

In MVS-VAT, the similarity between any two objects (*p*,*q*) is measured with cosine metric and multi-viewpoints. All the '*n*' number of data objects except p, and q are considered as the number of multi-viewpoints. Thus, the total number of viewpoints is taken ($n-2$) as N1, which is illustrated in Step 1. Cosine metric is used for finding the N1 similarity features for every pair of data objects. Here, average of N1 values is considered as the final multi-viewpoints-based cosine similarity and their procedure is mentioned in Step 2 of the algorithm. The dissimilarity is computed by subtracting the normalized similarity value from 1. Finally, all these dissimilarity values are stored in the form of a matrix (or two-dimensional array), called 'Diss.' The visual assessment of cluster tendency (VAT) is applied using the input of 'Diss' in Step 3. After applying the VAT, the 'Diss' matrix is reordered and displays the reordered Diss matrix in image form, which is known as MVS-VAT Image. This image shows each cluster as a dark-colored block in the diagonal. The count value of all these blocks is considered as pre-clusters '*k*' (or cluster tendency). The same is presented in Step 4 of the algorithm. Crisp partitions and obtaining the data clusters procedure are explained in Step 5 and Step 6. The MVS-VAT is an effective approach for data clustering and their experimental study is described in next following section.

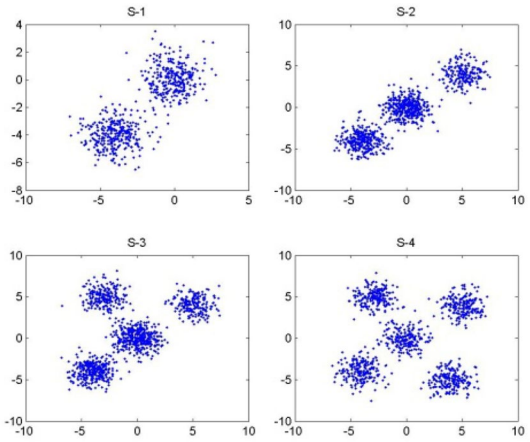


Fig. 3 Synthetic datasets (S-1 to S-2)

Experimental study

Pre-clusters assessment is performed on various types of clustered data. These clustered data are created manually, usually known as synthetic datasets, and they are shown in Fig. 3. The synthetic datasets are created with Gaussian parameters, mean, and variance. By fixing the different mean and variance values, different subsets of synthetic datasets are created in MATLAB 2020a. Other text datasets are retrieved through social media [17, 18]. Subsets of tweets datasets are generated concerning topics. Real-time datasets are also taken in the experimental study [19–21]. Details of these datasets are mentioned in Table 1.

Comparison of visual images

In the experimental, two existing methods, VAT and cVAT, are compared with MVS-VAT. Thus, their visual images are shown in Fig. 4.

With this observation, it noted that more clarity of dark blocks has appeared in MVS-VAT than other visual methods [18]. The quality of data clustering depends on the clarity of dark blocks.

Pre-clusters assessment is performed by counting dark blocks along the diagonal of visual images. If the clarity of the dark block is more, then it indicates the data objects are placed in the clusters without overlaps. That is, there are very few chances occur an irregular cluster boundary. Visual images through MVS-VAT are obtained with high quality compared with VAT and cVAT.

Table 1 Description of datasets

S. no.	Name of the datasets	No. of clusters
Synthetic data		
1	Synthetic data-1 (S1)	2
2	Synthetic data-2 (S2)	3
3	Synthetic data-3 (S3)	4
4	Synthetic data-4 (S4)	5
Real-time data		
5	Iris	3
6	Wine	3
7	Seeds	3
8	Voting	4
Twitter data		
9	Twitter data (2 topics)	2
10	Twitter data (5 topics)	5
11	Twitter data (10 topics)	10
12	Twitter data (15 topics)	15

Performance study

Majorly, two performance parameters are used for computing the quality of data clusters. These parameters are cluster accuracy (CA) [22] and normalized mutual information (NMI) [23]. Visual clustering results are presented in Tables 2 and 3 for the CA and NMI, respectively.

Empirical analysis of the proposed and existing data clustering techniques are presented with bar graphs, which are shown in Figs. 5 and 6. From this empirical analysis, it observed that MVS-VAT is outperformed with others considering the parameters of CA and NMI. For Twitter (social data) clustering also, MVS-VAT produces excellent data clustering when compared with another visual method, VAT and cVAT.

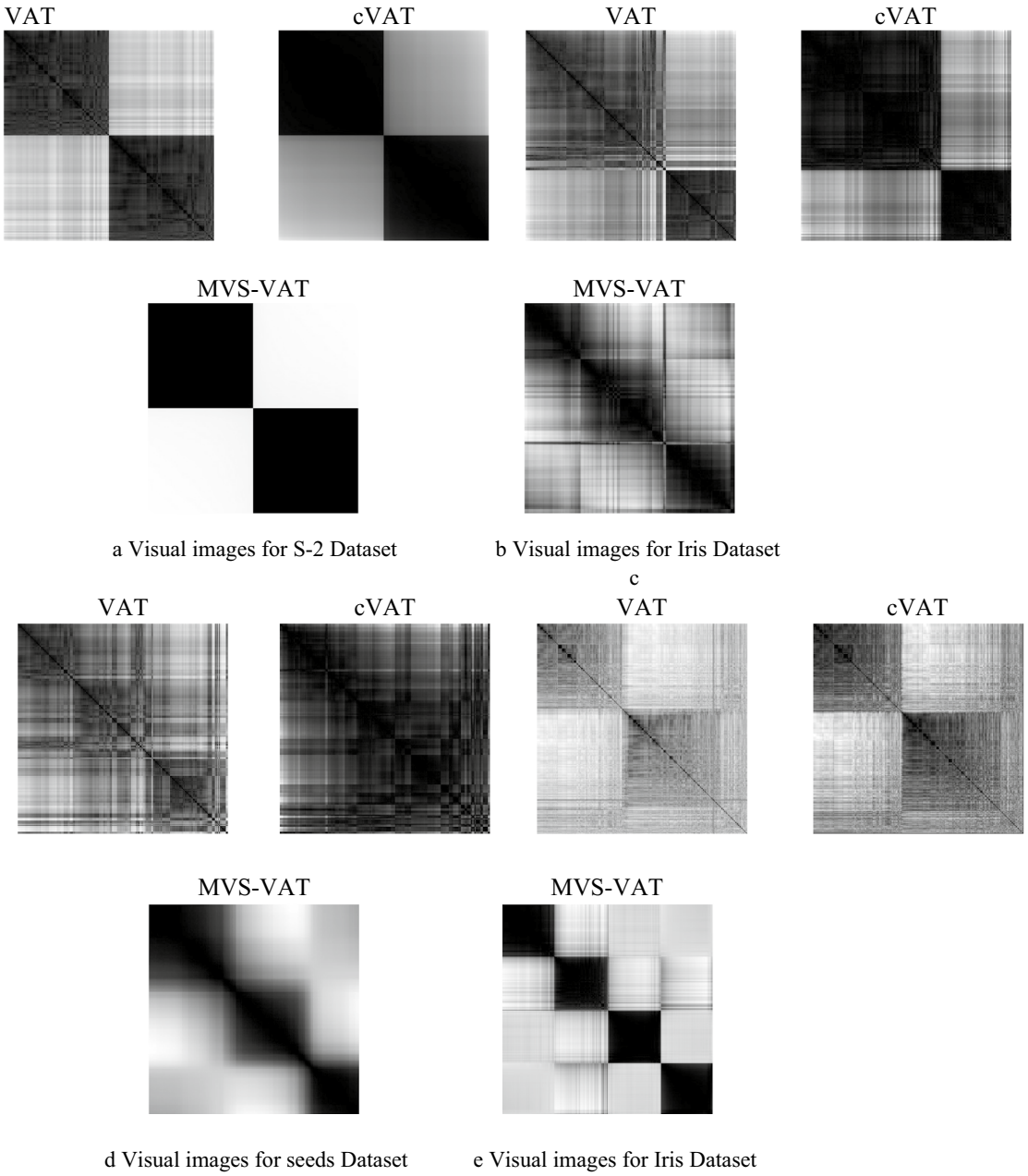
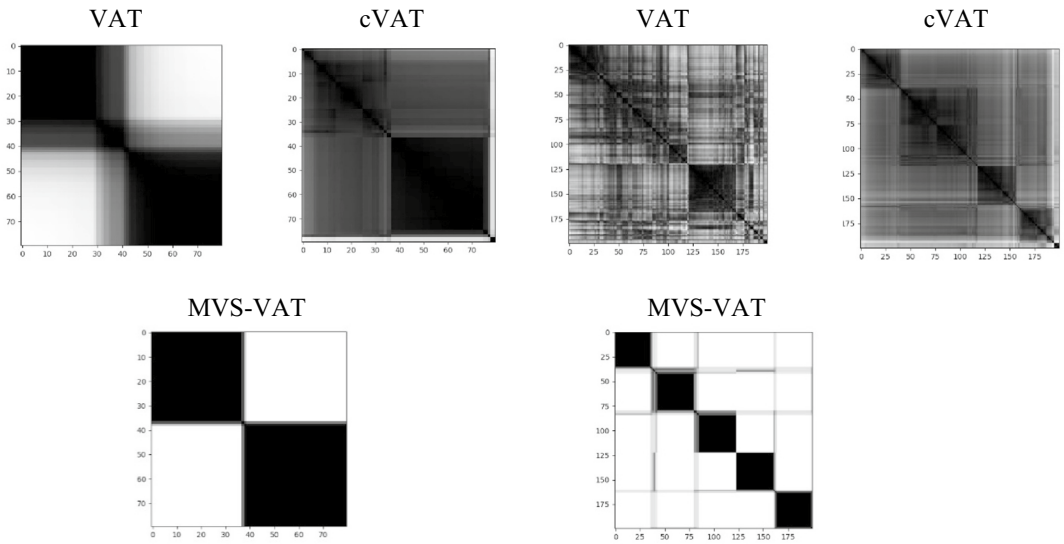
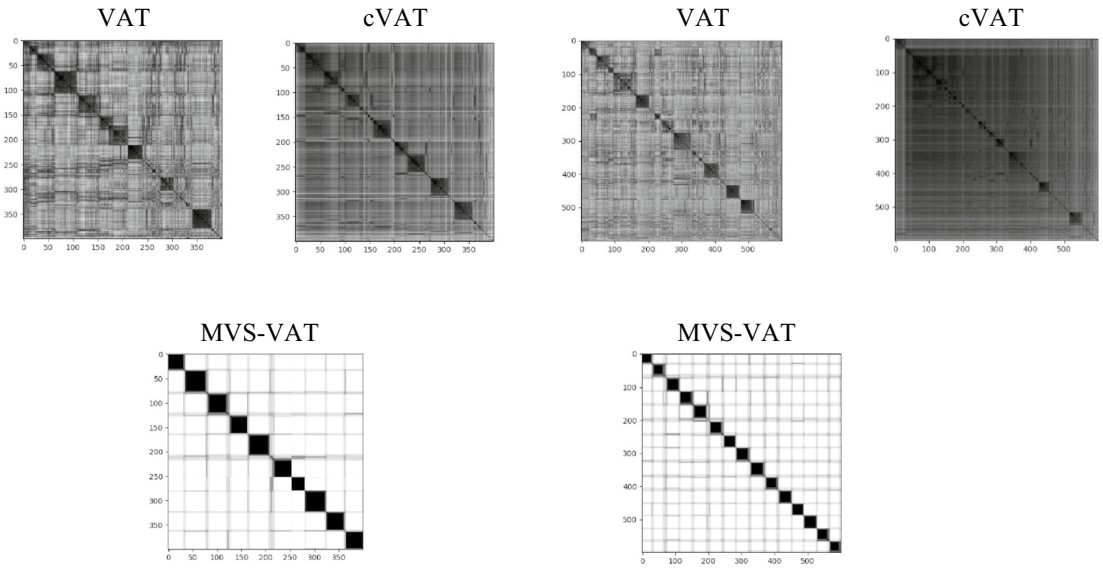


Fig. 4 Images of visual methods



f Visual images for Twitter Data (2-Topics) g Visual images for Twitter Data (5-Topics)



h Visual images for Twitter Data (10-Topics) i Visual images for Twitter Data (15-Topics)

Fig. 4 (continued)

Table 2 CA for visual methods

Name of the datasets	VAT	cVAT	MVS-VAT
S-2	1	1	1
S-3	0.36333	0.585	0.97667
S-4	0.38125	0.66125	0.98875
S-5	0.384	0.664	1
Iris	0.40667	0.81667	0.86333
Wine	0.31348	0.71348	0.73034
Seeds	0.31429	0.81429	0.88571
Voting	0.15161	0.55862	0.58161
Twitter data (2 topics)	0.225	0.575	0.8
Twitter data (5 topics)	0.165	0.765	0.8
Twitter data (10 topics)	0.1525	0.4425	0.715
Twitter data (15 topics)	0.25	0.276	0.513333

Conclusion

Visual methods effectively perform the data clustering results by detecting the pre-cluster tendency. The existing methods say VAT and cVAT use Euclidean and cosine distance metrics, respectively. The basic model of cosine metric is used for finding similarity features with a single reference (or viewpoint) in cVAT. In the proposed work, similarity features are measured using a cosine distance metric with the reference of multi-viewpoints for achieving more accurate nano-cluster tendency and its data clustering results. Experimental results of the existing methods VAT and cVAT are compared with the proposed MVS-VAT for demonstrating the efficiency of the proposed method. The efficiency is computed using the CA and NMI parameters and concluded that the proposed MVS-VAT is improved with an increased rate of 20–40% accuracy compared to the existing methods.

Table 3 NMI for visual methods

Name of the datasets	VAT	cVAT	MVS-VAT
S-2	1	1	1
S-3	0.30151	0.87076	0.90794
S-4	0.50471	0.92465	0.95786
S-5	0.55857	0.95269	1
Iris	0.3969	0.59461	0.6405
Wine	0.31961	0.32152	0.41971
Seeds	0.46555	0.49705	0.64744
Voting	0.3102	0.33256	0.45672
Twitter data (2 topics)	0.4578	0.4785	0.5589
Twitter data (5 topics)	0.673193	0.671266	0.7125
Twitter data (10 topics)	0.636414	0.636414	0.674946
Twitter data (15 topics)	0.45	0.486	0.498

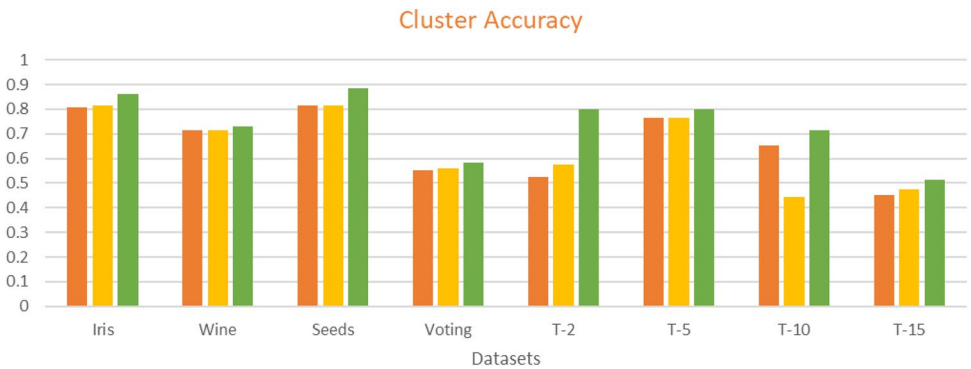


Fig. 5 Empirical analysis of cluster accuracy (CA) with different datasets

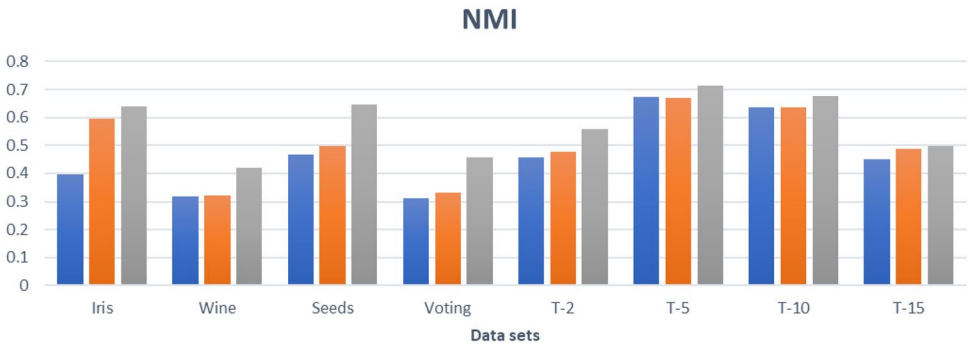


Fig. 6 Empirical analysis of normalized mutual information (NMI) with different datasets

References

1. Wu X, Kumar V, Quinlan JR et al (2008) Top 10 algorithms in data mining, knowledge information system, vol 14. Springer, Heidelberg, pp 1–37
2. Rui X, Wunsch D (2005) Survey of clustering algorithms. IEEE Trans Neural Netw 16(3):645–678
3. Bhatnagar V, Majhi R, Jena PR (2018) Comparative performance evaluation of clustering algorithms for grouping manufacturing frms. Arab J Sci Eng 43:4071–4083
4. Rajendra Prasad K, Suleman Basha M (2016) Improving the performance of speech clustering method. In: IEEE-10th international conference on intelligent systems and control (ISCO)
5. Bezdek J, Pattern recognition with objective function algorithms. Plenum, New York
6. Kumar D, Palaniswami M, Rajasegarar S, Leckie C, Bezdek JC, Havens TC (2013) clusiVAT: a mixed visual/numerical clustering algorithm for big data. In: 2013 IEEE international conference on big data, Silicon Valley, pp 112–117
7. Rathore P, Bezdek JC, Palaniswami M (2021) Fast cluster tendency assessment for big, high-dimensional data. In: Lesot MJ, Marsala C (eds) Fuzzy approaches for soft computing and approximate reasoning: theories and applications. Studies in fuzziness and soft computing, vol 394. Springer, Cham. https://doi.org/10.1007/978-3-030-54341-9_12
8. Rathore P, Kumar D, Bezdek JC, Rajasegarar S, Palaniswami M (2019) A Rapid hybrid clustering algorithm for large volumes of high dimensional data. IEEE Trans Knowl Data Eng 31(4):641–654
9. Bezdek JC, Hathaway RJ (2002) VAT: a tool for visual assessment of (cluster) tendency. In: Proceedings of the 2002 international joint conference on neural networks, Honolulu, pp 2225–2230
10. Suleman Basha M, Mouleeswaran SK, Rajendra Prasad K (2019) Cluster tendency methods for visualizing the data partitions. Int J Innov Technol Explor Eng 8(11):2978–2982
11. Suleman Basha M, Rajendra Prasad K (2018) Efficient cluster tendency methods for discovering the number of clusters. ARPN J Eng Appl Sci 13(4):1327–1334
12. Havens TC, Bezdek JC (2012) An efficient formulation of the improved visual assessment of cluster tendency (iVAT) algorithm. IEEE Trans Knowl Data Eng 24(5):813–822
13. Kumar D, Bezdek JC, Palaniswami M, Rajasegarar S, Leckie C, Havens TC (2016) A hybrid approach to clustering in big data. IEEE Trans Cybern 46(10):2372–2385
14. Rahamat Basha S et al (2020) A comparative approach of text mining: classification, clustering and extraction techniques. J Mech Continua Math Sci. https://doi.org/10.26782/jmcms.spl.5/2020.01.00010
15. Narasimhulu K et al (2021) An enhanced cosine-based visual technique for the robust tweets data clustering. Int J Intell Comput Cybern 14(2):170–184
16. Rajendra Prasad K, Mohammed M, Noorullah RM (2019) Visual topic models for healthcare data clustering. Evolut Intell. https://doi.org/10.1007/s12065-019-00300-y
17. Suleman Basha M, Mouleeswaran SK, Rajendra Prasad K (2021) Sampling-based visual assessment computing techniques for an efficient social data clustering. J Supercomput 77:8013–8037
18. Prasad K, Mohammed M, Prasad L, Anguraj DK (2021) An efficient sampling-based visualization technique for big data clustering with crisp partitions. Distrib Parallel Databases. https://doi.org/10.1007/s10619-021-07324-3
19. https://www.webmd.com/
20. Eswara Reddy B, Rajendra Prasad K (2016) Improving the performance of visualized clustering method. Int J Syst Assur Eng Manag 7(1):102–111
21. Asuncion A, Newman D (2007) UCI machine learning repository
22. Pattanodom et al (2016) Clustering data with the presence of missing values by ensemble approach. In: 2016 Second Asian conference on defence technology
23. Amelio A, Pizzuti C (2015) Is normalized mutual information a fair measure for comparing community detection methods? In: IEEE/ACM international conference on advances in social networks analysis and mining