International Journal of
INTELLIGENT SYSTEMS AND APPLICATIONS IN ENGINEERING
ISSN:2147-6799 www.ijisae.org Original Research Paper

An Extended Clusters Assessment Method with the Multi-Viewpoints for Effective Visualization of Data Partitions

Aswani Kumar Unnam[1], Bandla Srinivasa Rao[2]

Submitted: 21/10/2022 **Revised**: 16/12/2022 **Accepted**: 02/01/2023

Abstract: Cluster analysis is the most important for the data partitions of unlabelled data in various big data applications. It analyses the data based on similarity features of data objects. Two significant steps of the cluster analysis are as follows: assess the initial cluster tendency, and explore the data partitions. Top big data clustering techniques, such as k-means ++, single pass k-means (spkm), mini-batch-k-means (mbkm), and spherical k-means, effectively generate the big data clusters. However, they cannot get the initial knowledge about the clustering tendency. Estimation of the knowledge about the number of clusters is known as the clustering tendency. Various estimation methods of cluster tendency are surveyed and finally investigated that visual assessment of cluster tendency (VAT) accurately assesses the clustering tendency. Finding the accurate similarity features plays a vital role in accurately assessing clusters in the VAT algorithm. This paper proposes a novel computational similarity measure for the best assessment of big data clusters. The experiments are conducted on big synthetic and big real datasets to illustrate the proposed technique's efficiency.

Keywords: *Big Data, Cluster Analysis, Cluster Tendency, cVAT, Multi-Viewpoints, VAT*

1. Introduction

Big data clustering [1] is one of the prominent research areas that has progressed in many emerging applications, like social data computing [2], video analytics [3], fraud detection systems [4], and Artificial Intelligence (AI) -based data-driven systems [5], etc. Big data is the massive collection of data objects with higher dimensional data. Analysis of clusters for such big data needs to effective similarity computation technique. The k-means++ [6], spkm [7], and mbkm [8] are used the Euclidean distance for finding the similarity features among the data objects in the numerical big data clustering. The spherical k-means [9] uses the cosine measure in the object's similarity computation. The cosine metric records the two values, i.e., direction and magnitudes of the direction vectors, unlike only the distance in Euclidean. For this reason, the cosine largely succeeded, particularly in text clustering applications [10]. The cosine finds the similarity between two objects with a reference of a single viewpoint, i.e., origin. The similarity between two objects is measured from multiple viewpoints, i.e., other than the origin, and its proposed measure is called a multi-viewpoints-based cosine measure (MVCM). A novel measure is proposed with the multi-viewpoints for an accurate similarity computation than a traditional cosine measure. Big data clustering comprises two key sub-problems: assessing the clustering tendency (number of clusters) and data partitions. Many pre-clustering methods are studied and found that visual assessment of (cluster) tendency (VAT) [11] efficiently assesses the pre-cluster tendency knowledge. The VAT initially determines the dissimilarity or similarity features among the data objects and then reorders the data objects according to hierarchies of similarities of data objects. It shows the clusters in a visual form for any unlabelled data. Another implemented technique is cVAT [12], which uses the cosine measure to find the similarity values among the data objects. The cVAT is extended with our proposed similarity computation measure MVCM for performing the best-visualized clusters assessment. The proposed work is shown in Fig. 1, and its extended method is known as MVCM-VAT. In MVCM-VAT, cosine similarity computations are performed for the set of viewpoints, MV. For the n data objects, similarity features are computed between Xi and Xi concerning the remaining (n-2) data objects; that is, MV is the set of all data objects except for Xi and Xi. These objects in MV are considered as viewpoints. Similarities between Xi and Xj are performed concerning (n-2) viewpoints and the mean of (n-2) similarities in the similarity computation of Xi and Xj data objects. The significance of this methodology is to find the similarity between any two data objects with multiple viewpoints instead of a single origin and derive the similarities among the data objects is performed with a more informative assessment of multi-viewpoints. Therefore, MVCM is a more accurate similarity measure than others. Data partitions are also effectively assessed using the proposed MVCM-VAT than VAT and cVAT.

The contributions of the paper are summarized as follows:

1. Define the multi-viewpoints for a more informative assessment of the object's similarities
2. The clustering tendency problem is handled efficiently with the

1Research Scholar, Department of CSE, Acharya Nagarjuna University, Guntur, Andhra Pradesh, India, Email: askphy@gmail.com
2Professor in CSE, Department of CSE, Bhaskar Engineering College, Hyderabad, Telangana, India. E-mail: sreenibandla@gmail.com

proposed work

3. Develop the novel measure MVCM for computing the fundamental similarities between the data objects.

4. Design the pre-clusters assessment visual method, MVCM-VAT, to visually explore the number of partitions of big data.

5. Experiments are carried out to demonstrate the efficiency of MVCM-VAT using different performance parameters.

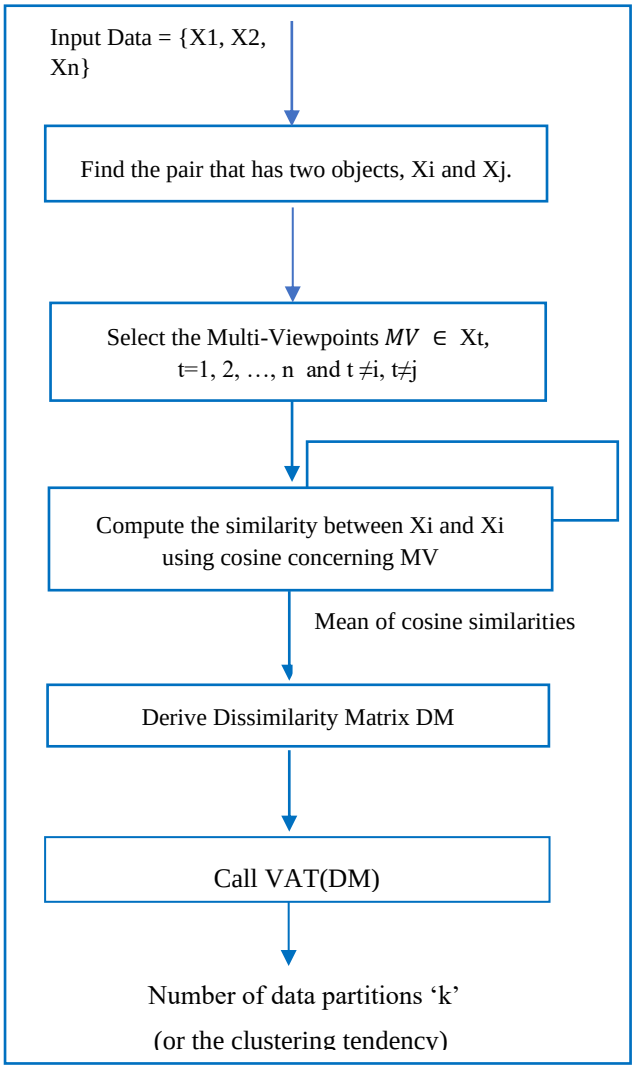


Fig. 1. Procedural steps for the Proposed Work

Other sections of the paper are presented as follows: literature study in Section 2, Proposed MVCM-VAT method in Section 3, Experimental Analysis and Discussion in Section 4, and Conclusions and Scope of Future Work in Section 5.

2. Literature Study

Popular data clustering techniques such as k-means [13], k-means++ [6], spkm [7], and mbkm [8] are effectively performed to determine the data partitions of the mid-range to big datasets. Its vast applications are social data clustering, video data clustering, medical image analysis, market analysis, fraud detection [14] etc. Any external interference should provide an initial k value for these algorithms. Sometimes, there is a chance to get the k value as intractable. Finding the number of clusters in the data clustering is a vital step in obtaining the quality of data clusters. Visual assessment of cluster tendency (VAT) is proposed by Bezdek et al. [11] for knowing the information about exploring the clusters of unlabelled data. The basic idea of VAT is formulated with three key steps. These steps are as follows: find the dissimilarity features of data objects in dissimilarity matrix (DM1) form; reorder the M using Prim's logic; and visualize the Image of reordered dissimilarity matrix (RDM1) for exploring the information of the number of clusters visually. The VAT procedural steps are shown in Algorithm 1.

Algorithm1: Visual Assessment of (cluster) Tendency (VAT) [11]

Input : The data objects – o1,o2, ….on

Methodology:

Step 1: Compute the dissimilarity matrix DM using the Euclidean measure

DM1=[d1ab] and 1 ≤ a,b ≤ n

Initially, fix L1 = { }, M1 = {1,2,…,n} and R1 = null row vector, i.e. (0, 0, …, 0);

Find (a,b) argmaxi,j {d1ij} ;

R(1) = b;

L1= L1 {b} ; M1 = M1 – {b}

Step 2: Reorder the DM using Prim's logic

while tr = 2,…, n:

{

Define (a,b) argmin(i L1,, j M1 ,{d1ij} };

R1(tr) = b;

L1 =L1 {b} and M1 = M1 – {a}

tr+=1

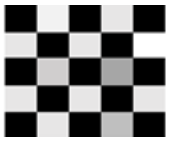
}

Step 3 : RDM1= [d1ab]= [d1R1(a)R1(b)] for 1 ≤ a,b ≤ n

The distances between the data objects are computed using a Euclidean distance measure which values are stored in the variable DM1. It is presented in step 1 of algorithm 1.

Reorder the dissimilarity matrix by finding the co-similar data objects which have maintained the minimum distance. It is explained in step2 of algorithm 1. In Prim's algorithm, adjacent nodes are connected by choosing the minimum distance. Thus, the VAT algorithm reorders the dissimilarity matrix using Prim's logic. Reflected reordered indices are stored in the variable R1. In step 3, R1 indices are used for finding the reordered dissimilarity matrix RDM1. Finally, display the Image of RDM1 that denotes the visual clusters in the form of square-shaped dark colored blocks.

$$DM1 = \begin{bmatrix} 0.00 & 0.73 & 0.19 & 0.71 & 0.16 \\ 0.73 & 0.00 & 0.59 & 0.12 & 0.78 \\ 0.19 & 0.59 & 0.00 & 0.55 & 0.19 \\ 0.71 & 0.12 & 0.55 & 0.00 & 0.74 \\ 0.16 & 0.78 & 0.19 & 0.74 & 0.00 \end{bmatrix} \quad I =$$



(a) Before VAT, the Image of DM1

$$RDM1 = \begin{bmatrix} 0.00 & 0.12 & 0.59 & 0.73 & 0.78 \\ 0.12 & 0.00 & 0.59 & 0.71 & 0.74 \\ 0.59 & 0.55 & 0.00 & 0.19 & 0.19 \\ 0.73 & 0.71 & 0.19 & 0.00 & 0.16 \\ 0.78 & 0.74 & 0.19 & 0.16 & 0.00 \end{bmatrix} \quad I =$$

(b) After VAT, the Image of RDM1

Fig. 2. VAT- Illustrative Example

Fig. 2 shows the sample illustration of the VAT algorithm. After applying the VAT algorithm, information about the number of clusters is displayed in dark-colored visual blocks [21] along the diagonal Image. In another existing algorithm cVAT [12], cosine-based distances are computed with a reference of a single viewpoint. These values are stored in the DM1. The remaining procedural steps of cVAT follow the VAT procedure. The cVAT is able to explore the visual clusters more effectively than VAT for the unlabelled datasets. Other developments in the visual methods are spectral VAT (SpecVAT) [16] and improved VAT (iVAT) [15]. Some datasets follow the path-shaped, these cases, iVAT works more efficiently since it computes the dissimilarity features of the data objects using a path-based distance measure. It is unable to detect the clusters for complex datasets. The SpecVAT determines the affinity matrix with the idea of Eigen decomposition. It also visualizes the clusters for complex datasets; however, its computational complexity is expensive. This paper introduces a novel pre-clusters assessment technique, MVCM-VAT. It initially determines the factual dissimilarity matrix with the finding of multi-viewpoints. The best assessment of pre-clusters is performed using this technique. The following section describes the procedural and algorithmic details of this proposed work.

3. Proposed MVCM-VAT Technique

The cosine-based VAT (cVAT) uses the cosine measure to find the data objects' dissimilarity (or similarity) features. Considering the directions and magnitudes of the data object vectors during the similarity computation impacts more on getting the best cluster tendency assessment.

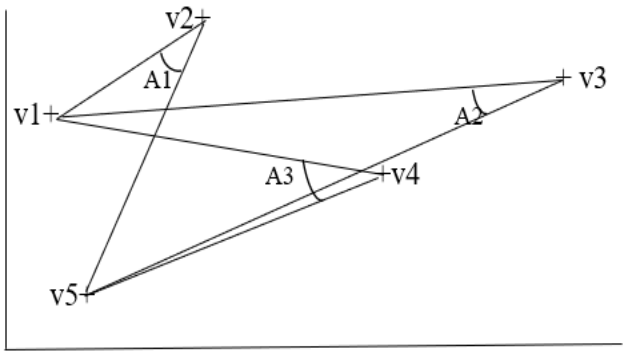


Fig. 3. Design of Multi-Viewpoints Illustration

Its clusters assessment results are achieved through the best clarity of visual image clusters than a VAT algorithm since VAT uses the Euclidean while finding the dissimilarity matrix. It computes the similarity between two data objects using a single reference point. It is required to justify the similarity between two data objects using more than one reference point. The proposed work is primarily focused on designing a novel similarity measure regarding multiple viewpoints. Any two data objects among the n data objects are taken to compute the similarity between these two objects. The remaining (n-2) data objects are taken as multiple viewpoints or reference points in the similarity computation between two data objects. This design idea of a novel multi-viewpoints-based cosine similarity (MVCM) is illustrated in Fig. 3, in which n=5 sample data objects are taken. Thus, several viewpoints would be (n-2), i.e., three multi-viewpoints are considered in the similarity computation of data objects.

Procedural steps of the proposed MVCM-VAT are presented in the following Algorithm 3.1, and is called the internal VAT procedure is given by Bezdek [20].

Algorithm 3.1: MVCM-VAT

Input: Data, D is defined with n number of objects, {X1, X2,…..Xn}

Output: Extracted k value and data partitions

Methodology:

1. *Let the number of multi-viewpoints MVN= n-2*
2. *i=1; j=1*
3. *while (i<=n)*
   while(j<=n)
   if (i==j)
   DM(i,j)=0
   else
   Find the pair of data objects, say, (a,b)
   (a,b)=(Xi, Xj)
   $MVS = \sum_{v \in D \text{ and } v \neq a \text{ and } v \neq b}^{MVN} cos(a,)$ *concerning viewpoint v*
   S(i,j) =Norm((MVS), (0,1));
   DM(i,j) = 1-S(i,j);
4. *Call VAT(DM), which displays the Visual Image that is the MVCM-VAT Image*
5. *Find the visible square-shaped dark blocks from MVCM-VAT Image and determine their count. Record and consider its count as the value of 'k.'*
6. *Explore the crisp-partitions of MVCM-VAT Image and determine the cluster labels of data objects*
7. *Save the clusters information k and its data partition results.*

In Step1, let assign the number of multiple viewpoints to the variable MVN as (n-2) since the all-n data objects are viewpoints except for the two targeted data objects. Given n number of data objects are, D= {X1, X2,….. \Xn}. This algorithm aims to find the similarity value for every pair of data objects so that two data objects are not identical. That is, it created the pair in which two data objects must be distinct. Step 2 and Step 3 illustrate the steps to initiate the process for taking such data objects; Xi and Xj and Xi should not be equal to Xj. Suppose the two data objects are the same in the pair, then the assigned dissimilarity between the same data objects is 0. If the two data objects, Xi and Xj, are different, then MVS computation is performed per the algorithm's indicated steps. The selection viewpoints followed the necessary condition,

i.e., the viewpoint v is selected in such a way that it is neither Xi nor Xj. Thus, the remaining data objects are considered the viewpoints, and their total number is (n-2). The MVS of pair of data objects is derived by taking the average of cosine similarities concerning the (n-2) viewpoints. Its MVS value is normalized using the min-max normalization, in which minimum and maximum values are set with 0 and 1, respectively. It is calculated using the Norm( ) function (Norm(MVS, (0,1)). After normalizing the MVS-based similarity value for the pair of data objects, the dissimilarity value is derived from the subtraction of the normalized similarity value from 1. Finally, all these dissimilarity values for the data objects are stored in the dissimilarity matrix (DM). Next, in Step 4, the VAT is called using the input of DM for visualizing the Image, known as the MVCM-VAT Image. This Image presents the cluster's information in brilliant dark-colored blocks along the diagonal, and this value is recorded as the number of clusters k. It is explained in Step 5 of the algorithm. The crisp partitions of the Visual Image and the exploration of data partitions results are presented in Step 6 and Step 7 of the algorithm.

4. Experimental Analysis and Discussion

Big data synthetic data is generated by setting the different values of gaussian parameters. It is initially created with the ground truth labels for performing the experimental analysis and comparative study between existing VAT, cVAT, and proposed MVCM-VAT techniques. Benchmarked four real-time datasets are also carried out in the experimental study. Table 1 shows the details of big synthetic data and real-time datasets[19]. Fig. 4 shows the generated big synthetic data visually.

Table 1: Description of Datasets

S. No.	Name of the Synthetic / Real-time Dataset	Clusters as per the ground truth label
	Synthetic Data	
1	Two Clustered Gaussian Data-1 (S-1)	Two
2	Three Clustered Gaussian Data-2 (S-2)	Three
3	Four Clustered Gaussian Data-3 (S-3)	Four
4	Five Clustered Gaussian Data-4 (S-4)	Five
	Real-time Data	
5	Iris	Three
6	Wine	Three
7	Seeds	Three
8	Voting	Four

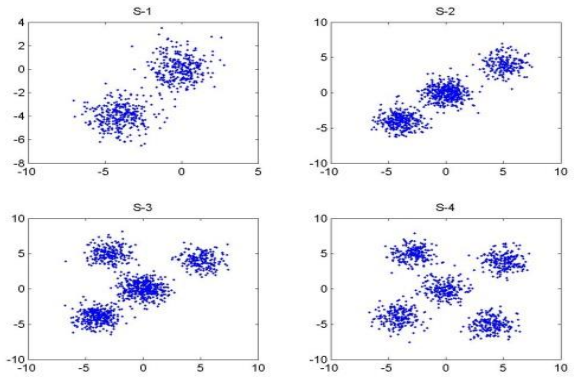


Fig. 4: Generated Gaussian Big Synthetic Data (Size(S-1) =100000; Size(S-2) =150000; Size(S-3) =200000; Size(S-4)=250000)

4.1. Comparative Analysis of Visual Images for the Data Partitions

This section experimentally demonstrates two existing visual methods, VAT and cVAT techniques, and proposed MVCM-VAT techniques. Derived visual images for these methods are presented in Fig. 5 for big synthetic data (S-2) and three real-time datasets (Iris, Seeds, and Voting) for comparative analysis Clustering tendency by MVCM-VAT for the big synthetic data is assessed through the obtained best clarity of MVCM-VAT images. It observed that the proposed MVCM-VAT produces the best visual image clarity compared to other visual techniques, VAT and cVAT. Therefore, clusters detection is performed more efficiently over the MVCM-VAT Image than others.

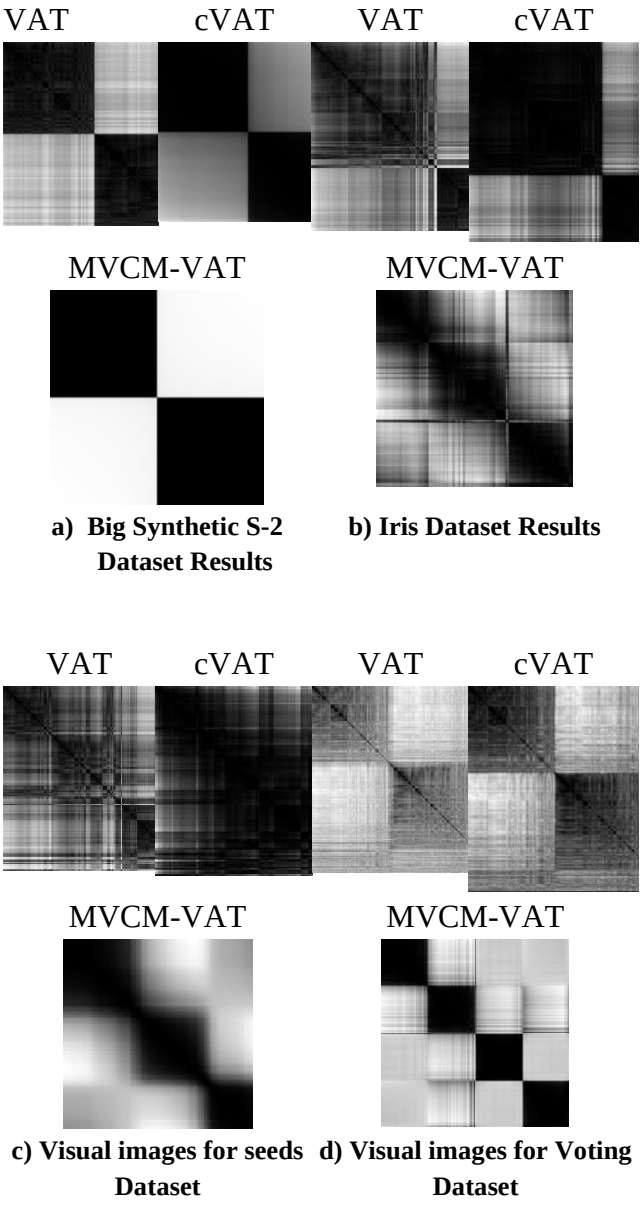


Fig. 5 Comparative Analysis of the Resulting Images between the Proposed and Existing Visual Techniques

The proposed MVCM-VAT smoothly performs clustering tendency assessment for the unlabelled datasets. It aims to explore the best-visualized images in terms of brilliant dark blocks and other non-diagonal images. The MVCM-VAT generates more clarity of visual images than VAT and cVAT; thus, it can better assess the crisps or clusters from the resulting images than others.

It also shows that MVCM-VAT classifies the best crisp partitions for both the real-time and synthetic datasets. From each crisp partition, cluster labels of the corresponding data objects are derived accurately.

4.2 Performance Discussion

This experimental study uses two data clustering performance measures to evaluate the existing and proposed MVCM-VAT techniques. These measures are partition accuracy (PA) [17] and normalized mutual information (NMI) [18]. Evaluation of the quality of big data partitions is presented in Table 2 and Table 3 using the PA and NMI, respectively.

Table 2. PA for the Visual Techniques

Synthetic / Real-time datasets	VAT	cVAT	MVCM-VAT
Two Clustered Gaussian Data-1 (S-1)	1	1	1
Three Clustered Gaussian Data-2 (S-2)	0.465	0.577	0.875
Four Clustered Gaussian Data-3 (S-3)	0.421	0.672	0.812
Five Clustered Gaussian Data-4 (S-4)	0.401	0.524	0.783
Iris	0.316	0.527	0.763
Wine	0.432	0.511	0.699
Seeds	0.422	0.632	0.785
Voting	0.456	0.721	0.822

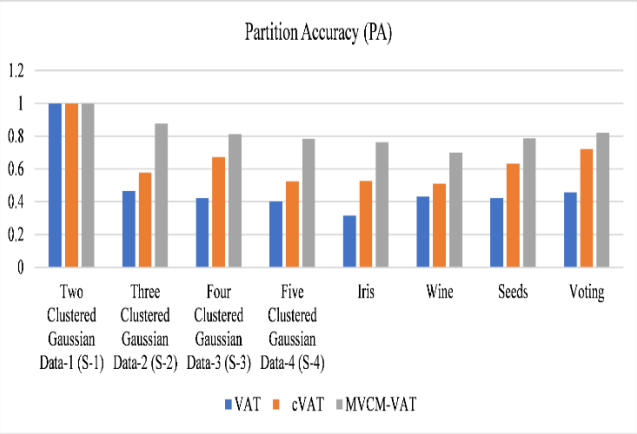


Fig. 6. Empirical Analysis of Visual Techniques Using the PA

Table 3. NMI for the Visual Techniques

Synthetic / Real-time datasets	VAT	cVAT	MVCM-VAT
Two Clustered Gaussian Data-1 (S-1)	0.92	0.934	0.965
Three Clustered Gaussian Data-2 (S-2)	0.354	0.466	0.864
Four Clustered Gaussian Data-3 (S-3)	0.391	0.561	0.709
Five Clustered Gaussian Data-4 (S-4)	0.377	0.413	0.672
Iris	0.222	0.234	0.345
Wine	0.332	0.409	0.519
Seeds	0.202	0.567	0.678
Voting	0.345	0.612	0.722

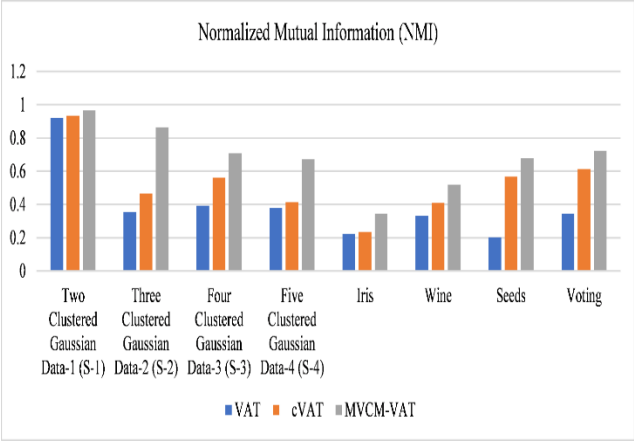


Fig. 7. Empirical Analysis of Visual Techniques Using the NMI

Four big synthetic datasets are created with many data objects in the two-dimensional plane (i.e., lakhs of data objects). The experimental analysis is conducted using the four different gaussian-generated datasets and four real-time datasets. Data clustering tendency and its data partitions results are analyzed over the four big datasets and four normal ranges of real-time datasets. The obtained results stated that a novel idea of multi-viewpoints cosine measure helps improve the significate growth rate in achieving the quality of data clusters. The PA and NMI are improved with an average rate of 12.22% and 10.5% in the proposed MVCM-VAT compared to other techniques for assessing clustering tendency and exploring the quality of data clusters.

5. Conclusion and Scope of the Work

Finding the pre-clustering tendency is vital step for producing the quality of clusters over the big data or regular data. Visual techniques are well-suited for determining the required knowledge about the clustering tendency. Existing methods, VAT, and cVAT, significantly assess the value of clustering tendency using Euclidean and cosine measures. The cosine measure is ideally suited for data clustering applications. However, it computes the similar features of the data objects with the justification of a single reference viewpoint. The proposed algorithm designed the multi-viewpoints-based cosine measure for the accurate similarity computations of data objects. The proposed MVCM-VAT uses this measure to assess the clustering tendency as the best and explore the data partitions outperformed the others.

Author contributions

Aswani Kumar Unnam : Conceptualization, Literature Study, Problem Definition, Design of Multi-viewpoints based cosine

similarity measure, Experimental Work

Bandla Srinivasa Rao: Synthetic Data Analysis, Data Study, Writing-Original draft preparation, Algorithm Design of MVCM-VAT, Visualization, Investigation, Writing-Reviewing and Editing.

Conflicts of interest

The authors declare no conflicts of interest.

References

[1] Hirkhorshidi, A.S., Aghabozorgi, S., Wah, T.Y., Herawan, T. (2014). Big Data Clustering: A Review. In: , *et al*. Computational Science and Its Applications – ICCSA 2014. ICCSA 2014. Lecture Notes in Computer Science, vol 8583. Springer, Cham. https://doi.org/10.1007/978-3-319-09156-3_49

[2] J. Tong, L. Shi, L. Liu, J. Panneerselvam and Z. Han, "A novel influence maximization algorithm for a competitive environment based on social media data analytics," in *Big Data Mining and Analytics*, vol. 5, no. 2, pp. 130-139, June 2022, doi: 10.26599/BDMA.2021.9020024.

[3] Balasundaram, A., Chellappan, C. An intelligent video analytics model for abnormal event detection in online surveillance video. *J Real-Time Image Proc* **17**, 915–930 (2020). https://doi.org/10.1007/s11554-018-0840-6

[4] G. J. Priya and S. Saradha, "Fraud Detection and Prevention Using Machine Learning Algorithms: A Review," *2021 7th International Conference on Electrical Energy Systems (ICEES)*, 2021, pp. 564-568, doi: 10.1109/ICEES51510.2021.9383631.

[5] Buxmann, P., Hess, T. & Thatcher, J.B. AI-Based Information Systems. *Bus Inf Syst Eng* **63**, 1–4 (2021). https://doi.org/10.1007/s12599-020-00675-8

[6] Kłopotek, M.A. An Aposteriorical Clusterability Criterion for *k*-Means++ and Simplicity of Clustering. *SN COMPUT. SCI.* **1**, 80 (2020). https://doi.org/10.1007/s42979-020-0079-8

[7] SARMA, T.H., VISWANATH, P. & REDDY, B.E. Single pass kernel ***k***-means clustering method. *Sadhana* **38**, 407–419 (2013). https://doi.org/10.1007/s12046-013-0143-3

[8] K. Peng, V. C. M. Leung and Q. Huang, "Clustering Approach Based on Mini Batch Kmeans for Intrusion Detection System Over Big Data," in *IEEE Access*, vol. 6, pp. 11897-11906, 2018, doi: 10.1109/ACCESS.2018.2810267.

I. Sharma and H. Sharma, "Recognizing Patterns in Text Data through Effective Initialization of Spherical K-means," *2018 Second International Conference on Electronics, Communication and Aerospace Technology (ICECA)*, 2018, pp. 327-331, doi: 10.1109/ICECA.2018.8474766.

[9] Fidan, H., Yuksel, M.E. A Novel Short Text Clustering Model Based on Grey System Theory. *Arab J Sci Eng* **45**, 2865–2882 (2020). https://doi.org/10.1007/s13369-019-04191-0

[10] J. C. Bezdek and R.J. Hathaway (2002). VAT: A tool for visual assessment of (cluster) tendency. In Proc. 2002 International Joint Conference on Neural Networks, Honolulu, HI, 2002, 2225-2230.

[11] B Eswara Reddy, K Rajendra Prasad: Improving the performance of visualized clustering method. International Journal of System Assurance Engineering and Management (Springer), Volume 7(1), pp 102–111 (2016)

[12] Rui X, Wunsch D (2005) Survey of clustering algorithms. IEEE Trans Neural Netw 16(3):645–678

[13] Rudolf Scitovski et al. cluster analysis and applications, springer, 2021

[14] T. C. Havens and J. C. Bezdek, "An efficient formulation of the improved visual assessment of cluster tendency (iVAT) algorithm," IEEE Trans. Knowl. Data Eng., vol. 24, no. 5, pp. 813–822, May 2012

[15] L. Wang, X. Geng, J. Bezdek, C. Leckie and R. Kotagiri, "SpecVAT: Enhanced Visual Cluster Analysis," *2008 Eighth IEEE International Conference on Data Mining*, 2008, pp. 638-647, doi: 10.1109/ICDM.2008.18.

[16] Pattanodom, et al. "Clustering data with the presence of missing values by ensemble approach," 2016 Second Asian Conference on Defence Technology.

[17] 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.

A. Asuncion and D. Newman, "UCI machine learning repository," 2007.

[18] 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.

[19] Rathore P., Bezdek J.C., 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