International Journal on Recent and Innovation Trends in Computing and Communication
ISSN: 2321-8169 Volume: 11 Issue: 9
***Article Received:** 25 July 2023 **Revised:** 12 September 2023 **Accepted:** 30 September 2023*

An Enhanced Sampling-Based Viewpoints Cosine Visual Model for an Efficient Big Data Clustering

Aswani Kumar Unnam
Research Scholar, Department of CSE,
Acharya Nagarjuna University,
Guntur, Andhra Pradesh, India
askphy@gmail.com

Bandla Srinivasa Rao
Professor in CSE, Department of CSE,
Teegala Krishna Reddy Engineering College
Hyderabad, Telangana, India
sreenibandla@gmail.com

Abstract— Bunching is registering the item's similitude includes that can be utilized to segment the information. Object similarity (or dissimilarity) features are taken into account when locating relevant data object clusters. Removing the quantity of bunch data for any information is known as the grouping inclination. Top enormous information bunching calculations, similar to single pass k-implies (spkm), k-implies ++, smaller than usual group k-implies (mbkm), are created in the groups with k worth. By and by, the k worth is alloted by one or the other client or with any outside impedance. Along these lines, it is feasible to get this worth immovable once in a while. In the wake of concentrating on related work, it is researched that visual appraisal of (bunch) propensity (Tank) and its high level visual models extraordinarily decide the obscure group propensity esteem k. Multi-perspectives based cosine measure Tank (MVCM-Tank) utilized the multi-perspectives to evaluate grouping inclination better. Be that as it may, the MVCM-Tank experiences versatility issues in regards to computational time and memory designation. This paper improves the MVCM-Tank with the inspecting methodology to defeat the versatility issue for large information grouping. Trial investigation is performed utilizing the enormous gaussian engineered datasets and large constant datasets to show the effectiveness of the proposed work.

Keywords- Clustering Tendency, Similarity Features, Big Data Clustering, Sampling, MVCM-VAT.

I. INTRODUCTION

Big data [1] is becoming increasingly popular in various applications that analyse similarity characteristics [2] across multiple data items. Web and streaming data applications [3] exude massive data; analysing and grouping such dynamic vast data is a difficult topic in current big data clustering research. Unsupervised approaches such as mini-batch-k-means (mbkm) [4] and single pass k-means (spkm) [5] are ridiculous, k-means ++ [6] are the enhanced techniques of k-means. These are specifically developed for handling the issues of big data while performing the clustering of unlabelled data. These techniques are recommended in big data clustering applications, including social data clustering [7], video surveillance [8], Medical image processing [9], Internet of Things (IoT) based speech communications [10], etc. One of the essential steps in big data clustering approaches is computing the similarity (or dissimilarity) features.In their numerical significant data clustering, The Euclidean distance was employed by k-means++, spkm, and mbkm to determine the dissimilarity characteristics for the data items. The cosine measure is used by several clustering approaches to compute dissimilarity characteristics. for exploring the quality of data clusters.

The cosine similarity measure is appealing because it considers the magnitudes and directions of the data vectors when calculating how similar two objects are.Because of this, the cosine is mostly successful, especially in applications involving text (or social data)clustering. [11]. with a single reference point or origin, the cosine determines how similar two things are.Multiple perspectives (or viewpoints), i.e., perspectives other than the origin, are used to determine how similar two items are. A multi-viewpoints-based cosine measure (MVCM) has been developed for handling this issue. In contrast to the usual cosine measure, a novel measure with many views is suggested for an accurate similarity computation. The two basic sub-problems that comprise large data successfully evaluates information regarding the tendency prior to the cluster[12]. The Tank at first recognizes the distinctions or likenesses between the information objects before reordering them according on similarity hierarchies. For any unlabeled data, a visual representation of the clusters is displayed. Another method that has been used is cVAT [13], which finds the values of similarity between the data objects using the cosine measure. Our suggested similarity computation metric, MVCM, is added to the cVAT to accomplish the best-visualized clusters assessment. Although this MVCM-VAT precisely evaluates the grouping propensity, it requires more computational exertion and memory for monstrous information. Accordingly, the proposed study fosters the examining based MVCM-Tank to work on the versatility of enormous information bunching results (i.e., computational time and memory designation). Figure 1 portrays the proposed system's procedural stages for the proposed inspecting based MVCM-Tank. The spectral idea includes the Eigen concept, which first calculates the affinities between the data objects before building the Laplacian matrix to derive the spectral properties shown in the following flow diagram.

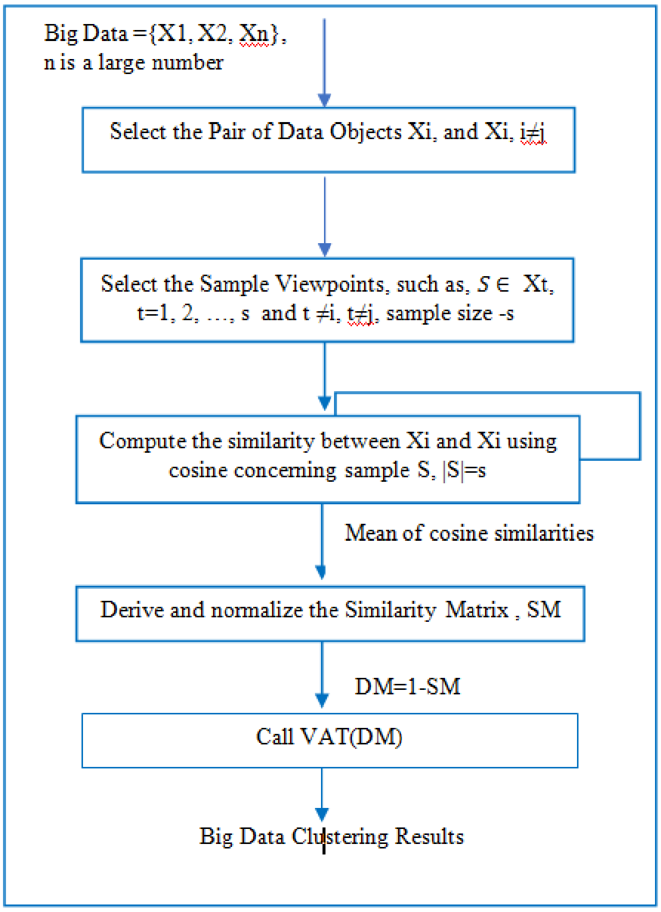

Big Data ={X1, X2, Xn},
n is a large number
Select the Pair of Data Objects Xi, and Xi, i≠j
Select the Sample Viewpoints, such as, $S \in$ Xt, t=1, 2, …, s and t ≠i, t≠j, sample size -s
Compute the similarity between Xi and Xi using cosine concerning sample S, |S|=s
Mean of cosine similarities
Derive and normalize the Similarity Matrix , SM
DM=1-SM
Call VAT(DM)
Big Data Clustering Results

Fig. 1 Proposed Framework of the Sampling-based MVCM-VAT

With the collection of sample viewpoints (or perspectives), S, in MVCM-VAT, cosine similarity calculations are made between two data objects. The remaining (n-2) data objects' Xi and Xi's similarities to the n data objects are computed; with the exception of Xi and Xi, S is the set of all data objects. In sample S, these data objects are referred to as perspectives or viewpoints. When comparing Xi and Xj Data items' similarities are determined using (n-2) perspectives. The goal similarity between the data items, Xi and Xj, is the mean of (n-2) similarities. This methodology's significance is in its ability to determine similarities between any two data objects using numerous viewpoints as opposed to a single origin and to determine these similarities using a more insightful examination of various viewpoints. Thus, compared to other similarity measures, MVCM is more accurate. The MVCM-VAT is more efficient at evaluating data partitions than VAT and cVAT. The sampling strategy is developed to select the best Over large data, sample representatives are selected, and this sample is then employed in MVCM. The suggested method solves the scalability problem by decreasing the computational time and memory allocations for big data clustering by performing similarity calculations utilizing sample perspectives rather than (n-2). Contribution highlights include summarized as follows:

1. Develop the sampling strategy for the selection of the best sample representative over the big data
2. Sampling-based MVCM measure is designed for an accurate.
3. Implement an enhanced sampling-based MVCM-VAT visual model for extraction of clustering tendency over the big synthetic and real-time datasets
4. Address the scalability issue of enormous information bunching utilizing the proposed framework
5. Demonstrate the proposed work's empirical analysis and illustrate the proposed work's efficiency using various performance measures.

Details of following sections are described: Section 2 describes the background study, Segment 3 portrays the proposed examining based MVCM-Tank visual model, Area 4 talks about the trial study and results, and Area 5 examines the ends and extent of future work.

II. BACKGROUND STUDY

Popular data clustering methods like k-means [14], k-means++ [15], spkm [16], and mbkm [17] help figure out how to divide up extensive data into different partitions based on similarity features. For these methods, any external disturbance should provide the first k value. The k value may be calculated accurately on occasion, or it may be intractable. The quantity of bunches in the information grouping is a significant stage in laying out the nature of information groups. Bezdek et al. [18] recommend using the VAT (visual evaluation of cluster tendency) to learn more about how to analyze unlabeled data clusters. Three basic advances help in the meaning of the fundamental thought of Tank. Coming up next are the activities: Find the difference highlights of the information components in the disparity grid (D.M.) structure; Improve the M utilizing Demure's thinking to get the reordered uniqueness lattice (RDM); Then, at that point, for visual examination, find the RDM picture [19], which shows the quantity of groups in dim/dark hued square blocks along the slanting. In another notable methodology known as cVAT, cosine-based distances are registered for a solitary viewpoint. These figures are saved in D.M. The succeeding cVAT stages are reliable with the Tank cycle. The cVAT is more fit than the Tank at examining visual bunches in unlabeled datasets. Extra advances in visual methods incorporate phantom Tank (SpecVAT) and improved Tank (iVAT) [20, 21]. iVAT performs better on path-shaped datasets because it uses a path-based distance metric to calculate the data items' dissimilarity features [22]. It is difficult to find the groups in confounded datasets. The fondness network is resolved utilizing the Eigen decay idea. It likewise shows groups for immense datasets, but at an extensive handling cost. The novel pre-bunches evaluation method MVCM-Tank is given in this work. It starts by perceiving various perspectives to develop the grid of authentic disparity. This system is utilized to research pre-groups as completely as plausible. Huge information makes a versatility trouble since developing the divergence lattice is time-consuming(D.M.) with different perspectives is very costly. All in all, immense information bunching requires handling time and costly memory. The sampling approach for the selected sample viewpoints that represent the original big data is developed in this research. The computation of D.M. is carried out using the sample perspectives using the best possible computation time and memory allocations. The specifics of this proposed work are described in the section that follows.

III. PROPOSED SAMPLING-BASED VISUAL MODEL

With a more extensive approach—a sampling-based views visual model—the proposed work seeks to obtain efficient big data clustering results. This proposed work tries to handle the MVCM-VAT problem by choosing representative perspectives from the regions between clusters, whose steps are outlined in Algorithm 1. This proposed visual model is known as S-MVCM-Tank. Fig. 2 portrays the plan representation of Testing Based-MVCM calculation. It portrays the closeness calculation with five example information objects, in particular, v1, v2, v3, v4, and v5. Assume the similitude is registered between two information objects, v1 and v2, concerning the three example perspectives v3, v4, and v5. The following three cases are taken during the MVCM similarity computation.

i. The cosine comparability of v1 and v2 is registered concerning the perspective v3 with a point of M.

ii. The cosine similitude of v1 and v2 is figured concerning the perspective v4 with a point of N.

iii. Cosine similarity of v1 and v2 is computed concerning the viewpoint v5 with an angle of Q

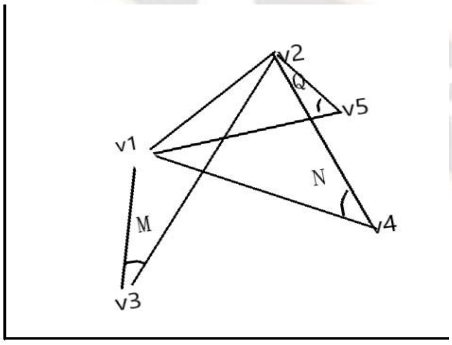


Fig. 2. Illustrative Example for Sampling-Based MVCM Computation

Finally, the Multiview points-based cosine similarity (MVCM) is computed by taking the average cosine similarity of (v1,v2) three viewpoints of v3,v4,v5 which we can observe in Eqn. (1)

$$MVCM(v1, v2) = 1/3 * (\cos(v1,v2) \text{ viewpoint } v3 + \cos(v1,v2) \text{ viewpoint } v4 + \cos(v1,v2) \text{ viewpoint } v5) \quad (1)$$

The similarity value, i.e., MVCM value, is normalized using Eqn. (2). Its value is normalized within the decimal scaling of (0,1)

$$Sim(v1,v2) = Norm(MVCM, (0,1)) \quad (2)$$

In the proposed sampling-based MVCM-VAT, the starting data object is chosen at random position r, and the index of the information object with the biggest separation from the underlying information object is chosen. The algorithm's first step explains it.The randomly selected data object is Xr, and the remaining data objects are XI, where I=1,2,…N, and I≠r. The argmax function is used to obtain information about a data object that retains the greatest separation from other data objects. Compared to other data items, the retained data objects become the most distinctive. After keeping this most notable data object, its status changes to the already visited data object. The distance value is changed by steps 2, 3, and 4 to allow the previously visited data objects to be ignored in subsequent iterative phases. Step 5 demonstrates how to include some of the most notable data objects. Step 6 describes how all the most notable data objects are chosen based on the threshold value of approximated data clusters. Other data objects are relocated to the closest centroids, as shown in Step 7.

```
Algorithm 1: Sampling-MVCM-VAT
    Input: Big data is defined with n number of objects
           {X1, X2,.....Xn}; where n is a large number
    Output: Extract the k value and clustering tendency

Methodology:
        // Find the initial distinguished data object over
        the big data

    1.  Pick r randomly from the range of 1, 2,..., N.
        Determine the distances from Xr to "X1, X2,...
        X.N." and select the index depending on the most
        significant distance. Maxindexshows the record of
        the information thing and has been set as the
        centroid utilizing the formulas Maxindex =
        argmax $_{I \,\epsilon\, 1,2,..N \text{ and } I \neq r}$ distance(Xr, XI), and
        Maxdist=distance(Fr, F.I.)

        // Determine the other distinguished data objects
        (called centroids) over the big data

    2.  Update the big data object's distances
    3.  i=1;
    4.  While (i< = N)
        {
            a.  Disti=min(max_dist,
                distance(Fmax_index, Fi))
            b.  i++
        }
    5.  From argmax function, the index of the centroid is
        determined,I ϵ {1,2,..N} {DistI},
6.  Steps 3 to 6 should be repeated until all centroids have been obtained to update maxindex and maxdist in light of some threshold value.
7.  Find the nearest centroids for the data objects.
8.  Select the equal ratio of representative samples from every approximated cluster with size s
9.  Let the number of sample multi-viewpoints SMVN= $n_s$-2; Let $n_s$ be the number of samples
10. i=1; j=1
11. while (i<=$n_s$)
        while(j<=$n_s$)
            if (i==j)
        SDM(i,j)=0
            else
                Find the pair of data objects, say, (a,b)
                (a,b)=(Xi, Xj)
```

$$SMVS = \frac{1}{MVN} \sum_{v \in D \text{ and } v \neq a \text{ and } v \neq b}^{MVN} \cos(a,)\ concerning\ viewpoint\ v$$

```
                S(i,j) =Norm((SMVS), (0,1));
            SDM(i,j) = 1-S(i,j);
12. Call VAT(SDM), which displays the S-MVCM-Tank Visual Picture.
13. Locate and count the visible square-molded dim blocks from MVCM-Tank Picture. Make a note of it and analyse its worthof 'k.'
14. Investigate the fresh parcels of S-MVCM-Tank Picture and decide the group marks of information objects
15. Save the bunches data k and investigate the large information groups.
```

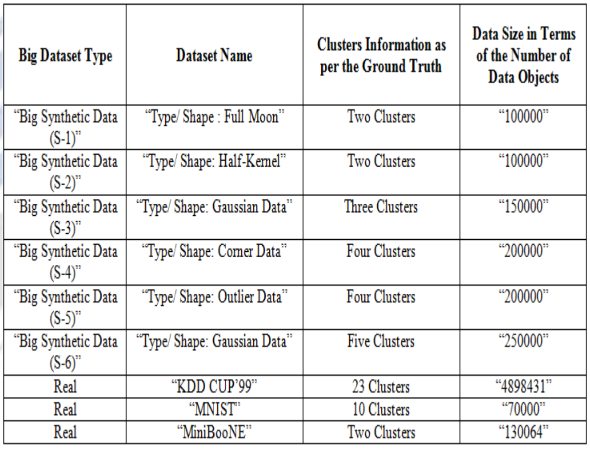
To choose the best sample with a size of s, a simple random sample without replacement is applied across the generated estimated clusters and is demonstrated in Step 8. Here the total number of sample data objects is ns; thus, the total number of viewpoints The variable SMVN=ns-2 defines the similarity computation between any two data items; the SMVN specifies the sample multi-viewpoints number. Steps 10 to 11 depict the similarity computing procedure for a pair of data items including sample multi-views rather than (n-2) viewpoints. The algorithm phases of normalising the similarity matrix values and then obtaining the dissimilarity matrix for the sample data items result in the sample-based dissimilarity matrix (SDM). The VAT method is invoked when SDM is entered, and SDM is reordered. The S-MVCM-VAT picture, which is a reordered SDM picture, is displayed. The clusters were shown as diagonal rows of square, dark-colored blocks in this picture. These blocks, which were described in stages 12 through 15, were responsible for the cluster labels of the data items by deriving the clean partitions.

IV. EXPERIMENTAL STUDY AND RESULTS DISCUSSION

In order to generate synthetic data for big data analysis, this research in the experimental study can adjust the values of gaussian parameters. It is first built using ground truth labels to play out the exploratory examination and correlation research between the traditional VAT, cVAT, and the suggested S-MVCM-VAT technique.

In addition, four benchmarked real-time datasets will be carried out as part of the experimental investigation. The information regarding massive synthetic data and real-time datasets [23] can be found in Table 1. Fig. 3 provides a graphical representation of the created amounts of synthetic data.

Table 1: Details of the Big Datasets Used for the Experimental Analysis



Big Dataset Type	Dataset Name	Clusters Information as per the Ground Truth	Data Size in Terms of the Number of Data Objects
"Big Synthetic Data (S-1)"	"Type/ Shape : Full Moon"	Two Clusters	"100000"
"Big Synthetic Data (S-2)"	"Type/ Shape: Half-Kernel"	Two Clusters	"100000"
"Big Synthetic Data (S-3)"	"Type/ Shape: Gaussian Data"	Three Clusters	"150000"
"Big Synthetic Data (S-4)"	"Type/ Shape: Comer Data"	Four Clusters	"200000"
"Big Synthetic Data (S-5)"	"Type/ Shape: Outlier Data"	Four Clusters	"200000"
"Big Synthetic Data (S-6)"	"Type/ Shape: Gaussian Data"	Five Clusters	"250000"
Real	"KDD CUP'99"	23 Clusters	"4898431"
Real	"MNIST"	10 Clusters	"70000"
Real	"MiniBooNE"	Two Clusters	"130064"

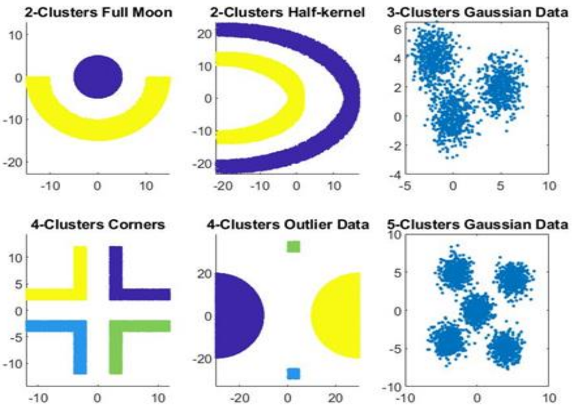


Fig. 3: Big Synthetic Data for S-1: Two-Clusters Full Moon, S-2: Two Clusters Half-Kernel, S-3: Three-Clusters Gaussian Data, S-4: Four Clusters Corners Data, S-5: Four Clusters Outlier Data, S-6: Five Clusters Gaussian Data

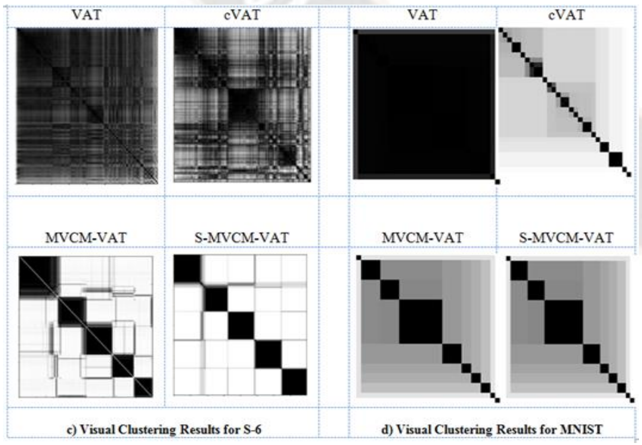
***Article Received:** 25 July 2023 **Revised:** 12 September 2023 **Accepted:** 30 September 2023*

A. Comparative and Evaluation Analysis

This section shows two current visualisation approaches, VAT and cVAT, as well as suggested Sampling-based-MVCM-VAT (S-MVCM-VAT) techniques. The following criteria are used to assess the performance of various data clustering algorithms for enormous volumes of data.

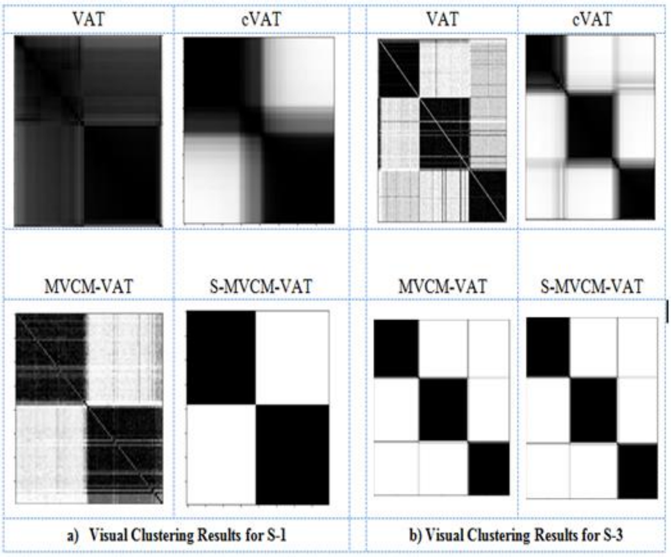
i) Evaluation of Visual Images by Goodness Measure

The quality or clarity of the visual image is always going to be the best indicator of whether or not there is a tendency to cluster. OTSU [24] is successfully used for calculating the clarity or goodness parameter in the evaluation of visual images of existing and proposed techniques. Figure 4 depicts the visual images of the three big synthetic datasets (S-1, S-3, and S-6) and one big real-time dataset (MNIST). According to these statistics, it was found that distinct clarity of blocks colored dark or grey appeared in S-MVCM-VAT. Therefore, the S-MVCM-VAT cluster analysis method provides a more accurate result than VAT, cVAT, and MVCM-VAT. Table 2 contains the goodness value evaluations that were performed on the visual images (shown the visual image results for S-1, S-3, S-6, and MNIST real-time dataset).A high goodness value shows that the visual images have a sufficient level of clarity, which is helpful for the practical computation of cluster tendency for the datasets. Assessment and examination of the nature of visual pictures are done for a sum of six a lot of manufactured information and three a lot of genuine information. Every aspect of the proposed method has a high value of goodness scenario; hence, spectral features have a positive impact on the computation of the clustering tendency that is best.



c) Visual Clustering Results for S-6

d) Visual Clustering Results for MNIST



a) Visual Clustering Results for S-1

b) Visual Clustering Results for S-3

Fig. 4 Visual Cluster Image Results for Three Big Synthetic Datasets (S-1, S-3, and S-6) and One Big Real-Time MNIST Dataset

It would appear that the proposed S-MVCM-VAT successfully produced a clear visual representation to assess clusters utilizing the sample perspectives. The more recent MVCM-VAT approach also gave the best visual clustering images for the larger datasets.On the other hand, S-MVCM-VAT is superior in terms of goodness, normalized mutual information (NMI), partition accuracy (P.A.) [25], speed, and memory efficiency of visual images.

It would appear that the proposed S-MVCM-VAT successfully produced a clear visual representation to assess clusters utilizing the sample perspectives. The more recent MVCM-VAT approach also gave the best visual clustering images for the larger datasets. On the other hand, S-MVCM-VAT is superior in terms of goodness, normalized mutual information (NMI), partition accuracy (P.A.) [25], speed, and memory efficiency of visual images.

Table 2: Goodness Evaluation by OTSU [24]

Big Dataset Type	Dataset Name	VAT	cVAT	MVCM-VAT	S-MVCM-VAT
"Big Synthetic Data (S-1)"	"Type/ Shape : Full Moon"	0.673	0.720	0.809	0.812
"Big Synthetic Data (S-2)"	"Type/ Shape: Half-Kernel"	0.625	0.712	0.754	0.822
"Big Synthetic Data (S-3)"	"Type/ Shape: Gaussian Data"	0.643	0.682	0.688	0.767
"Big Synthetic Data (S-4)"	"Type/ Shape: Corner Data"	0.630	0.675	0.697	0.798
"Big Synthetic Data (S-5)"	"Type/ Shape: Outlier Data"	0.742	0.812	0.822	0.856
"Big Synthetic Data (S-6)"	"Type/ Shape: Gaussian Data"	0.543	0.622	0.709	0.876
Real	"KDD CUP'99"	0.673	0.688	0.689	0.887
Real	"MNIST"	0.475	0.545	0.556	0.615
Real	"MiniBooNE"	0.782	0.842	0.876	0.912

Article Received: *25 July 2023* ***Revised:*** *12 September 2023* ***Accepted:*** *30 September 2023*

B. Evaluation of Visual Models by Performance Parameters

In this experimental investigation, we analyze the existing and proposed S-MVCM-VAT approaches using two different data clustering performance indicators. Both partition accuracy (P.A.) This performance evaluation of visual models employs and normalised mutual information (NMI). Tables 2 and 3 show how the P.A. and NMI are used to assess the quality of massive data divisions. These tables are presented below.

Table 2. P.A. for the Visual Models for Big Data Clustering

Big Dataset Type	Dataset Name	VAT	cVAT	MVCM-VAT	S-MVCM-VAT
"Big Synthetic Data (S-1)"	"Type/ Shape : Full Moon"	0.643	0.655	0.715	0.779
"Big Synthetic Data (S-2)"	"Type/ Shape: Half-Kemel"	0.552	0.577	0.612	0.623
"Big Synthetic Data (S-3)"	"Type/ Shape: Gaussian Data"	0.534	0.552	0.622	0.691
"Big Synthetic Data (S-4)"	"Type/ Shape: Comer Data"	0.562	0.611	0.712	0.782
"Big Synthetic Data (S-5)"	"Type/ Shape: Outlier Data"	0.611	0.709	0.756	0.789
"Big Synthetic Data (S-6)"	"Type/ Shape: Gaussian Data"	0.432	0.511	0.699	0.767
Real	"KDD CUP'99"	0.312	0.422	0.453	0.563
Real	"MNIST"	0.452	0.511	0.522	0.623
Real	"MiniBooNE"	0.712	0.767	0.789	0.823

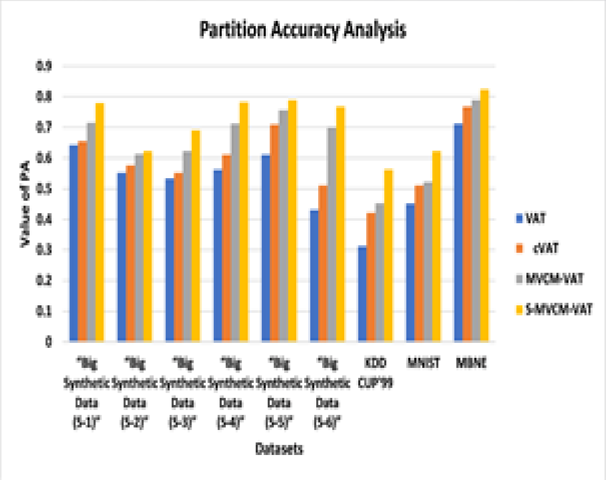


Fig. 5 Empirical Analysis of Visual Techniques Using the P.A.

The PA and NMI values obtained experimentally are shown in Fig. 5 and Fig. 6, respectively, and are empirically evaluated. After comparing our proposed S-MVCM-VAT to the more common VAT, cVAT, and MVCM-VAT, we found that our method had the highest values for P.A. and NMI. Higher values of both P.A. and NMI imply better data partitions or achieving an exceptional quality. Higher PA and NMI values also indicate that more data clusters have been successfully created.The empirical examination of P.A. in Figure 5 and NMI in Figure 6 reveals that the S-MVCM-VAT model scored the highest values compared to the other models. It was discovered through this investigation that the S-MVCM-VAT performed better than the VAT, the cVAT, and the S-MVCM-VAT.

Table 3. NMI for the Visual Models for Big Data Clustering

Big Dataset Type	Dataset Name	VAT	cVAT	MVCM-VAT	S-MVCM-VAT
"Big Synthetic Data (S-1)"	"Type/ Shape : Full Moon"	0.345	0.532	0.544	0.599
"Big Synthetic Data (S-2)"	"Type/ Shape: Half-Kemel"	0.441	0.449	0.492	0.509
"Big Synthetic Data (S-3)"	"Type/ Shape: Gaussian Data"	0.422	0.453	0.459	0.478
"Big Synthetic Data (S-4)"	"Type/ Shape: Comer Data"	0.309	0.412	0.422	0.475
"Big Synthetic Data (S-5)"	"Type/ Shape: Outlier Data"	0.322	0.423	0.436	0.456
"Big Synthetic Data (S-6)"	"Type/ Shape: Gaussian Data"	0.223	0.322	0.432	0.467
Real	"KDD CUP'99"	0.208	0.213	0.224	0.267
Real	"MNIST"	0.256	0.267	0.278	0.311
Real	"MiniBooNE"	0.454	0.522	0.578	0.623

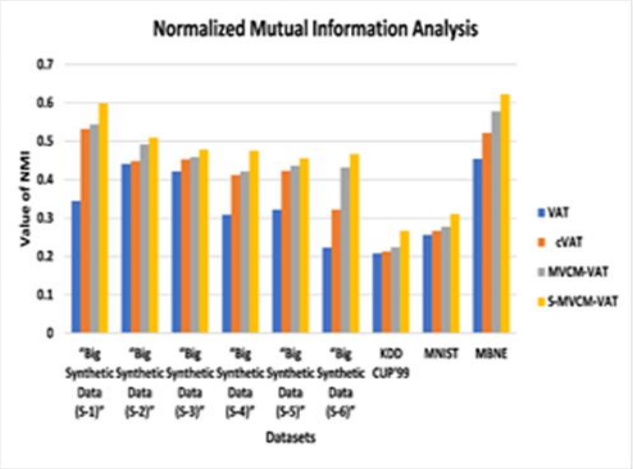


Fig. 6 Empirical Analysis of Visual Techniques Using the NMI

Six big (thousands of data items) datasets comprising of numerous data objects on a two-dimensional plane are produced. The experimental inquiry is carried out with the help of the three large real-time datasets and the six unique gaussian-generated datasets. According to the findings that were obtained, an innovative concept of sampling-based-multi-viewpoints cosine measure contributes to an improvement in the significant growth rate in the process of establishing the quality of data clusters. It was found that the overall accuracy of the proposed S-MVCM-VAT improved by an average rate of 5% to 10% compared to previous strategies for analyzing clustering tendencies and examining the quality of data clusters. This finding was based on P.A. and NMI experimental values.

CONCLUSION AND SCOPE OF THE FUTURE WORK

Discovering the pre-clustering tendency is essential in developing high-quality clusters from big or regular data. The required knowledge regarding the tendency to cluster can be determined using visual techniques, which are well-suited for the task. Measures of Euclidean and cosine distance are utilized in the currently available methods, which include VAT and cVAT. These methods considerably evaluate the usefulness of clustering tendency. The MVCM-VAT is the most recent visual technique that effectively assesses the initial knowledge about the clustering tendency concerning multi-viewpoints. Due to its expensiveness, it is further enhanced with the sampling strategy, which proposed technique is S-MVCM-VAT. It uses the smaller size of sample viewpoints instead of taking the original big data for determining the results over the big data.

REFERENCES

[1]. Shirkhorshidi, 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. Oyelade et al., "Data Clustering: Algorithms and Its Applications," 2019 19th International Conference on Computational Science and Its Applications (ICCSA), 2019, pp. 71-81, doi: 10.1109/ICCSA.2019.000-1.

[3]. V. Gulisano, R. Jiménez-Peris, M. Patiño-Martínez, C. Soriente and P. Valduriez, "StreamCloud: An Elastic and Scalable Data Streaming System," in IEEE Transactions on Parallel and Distributed Systems, vol. 23, no. 12, pp. 2351-2365, Dec. 2012, doi: 10.1109/TPDS.2012.24.

[4]. 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.

[5]. 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

[6]. Kapoor and A. Singhal, "A comparative study of K-Means, K-Means++ and Fuzzy C-Means clustering algorithms," 2017 3rd International Conference on Computational Intelligence & Communication Technology (CICT), 2017, pp. 1-6, doi: 10.1109/CIACT.2017.7977272.

[7]. Rajendra Prasad, K., Surya Prabha, I., Rajasekhar, N., Rajasekhar Reddy, M. (2018). Social Data Analytics by Visualized Clustering Approach for Health Care. In: Saeed, K., Chaki, N., Pati, B., Bakshi, S., Mohapatra, D. (eds) Progress in Advanced Computing and Intelligent Engineering. Advances in Intelligent Systems and Computing, vol 564. Springer, Singapore. https://doi.org/10.1007/978-981-10-6875-1_15

[8]. Sreenu, G., Saleem Durai, M.A. Intelligent video surveillance: a review through deep learning techniques for crowd analysis. J Big Data 6, 48 (2019). https://doi.org/10.1186/s40537-019-0212-5

[9]. Rajendra Prasad, K., Praveen Kumar, C., Multi-ROI segmentation for effective texture features of mammogram images, Journal of Discrete Mathematical Sciences and Cryptography,Vol. 24, Issue No. 8, pp: 2461-2469 https://doi.org/10.1080/09720529.2021.2016192

[10]. M. Mehrabani, S. Bangalore and B. Stern, "Personalized speech recognition for Internet of Things," 2015 IEEE 2nd World Forum on Internet of Things (WF-IoT), 2015, pp. 369-374, doi: 10.1109/WF-IoT.2015.7389082.

[11]. P. Sachar and V. Khullar, "Social media generated big data clustering using genetic algorithm," 2017 International Conference on Computer Communication and Informatics (ICCCI), 2017, pp. 1-6, doi: 10.1109/ICCCI.2017.8117716.

[12]. Wang, L., Nguyen, U.T.V., Bezdek, J.C., Leckie, C.A., Ramamohanarao, K. (2010). iVAT and aVAT: Enhanced Visual Analysis for Cluster Tendency Assessment. In: Zaki, M.J., Yu, J.X., Ravindran, B., Pudi, V. (eds) Advances in Knowledge Discovery and Data Mining. PAKDD 2010. Lecture Notes in Computer Science(), vol 6118. Springer, Berlin, Heidelberg. https://doi.org/10.1007/978-3-642-13657-3_5

[13]. K. R. Prasad and B. E. Reddy, "An efficient visualized clustering approach (VCA) for various datasets," 2015 IEEE International Conference on Signal Processing, Informatics, Communication and Energy Systems (SPICES), 2015, pp. 1-5, doi: 10.1109/SPICES.2015.7091373.

[14]. K. P. Sinaga and M. -S. Yang, "Unsupervised K-Means Clustering Algorithm," in IEEE Access, vol. 8, pp. 80716-80727, 2020, doi: 10.1109/ACCESS.2020.2988796.

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

[16]. A. Elsayed, O. Ismael and H. M. O. Mokhtar, "Distributed single pass clustering algorithm based on MapReduce," 2017 Eighth International Conference on Intelligent Computing and Information Systems (ICICIS), 2017, pp. 160-165, doi: 10.1109/INTELCIS.2017.8260047.

[17]. A. Feizollah, N. B. Anuar, R. Salleh and F. Amalina, "Comparative study of k-means and mini batch k-means clustering algorithms in android malware detection using network traffic analysis," 2014 International Symposium on Biometrics and Security Technologies (ISBAST), 2014, pp. 193-197, doi: 10.1109/ISBAST.2014.7013120.

[18]. J. C. Bezdek, R. J. Hathaway and J. M. Huband, "Visual Assessment of Clustering Tendency for Rectangular Dissimilarity Matrices," in IEEE Transactions on Fuzzy Systems, vol. 15, no. 5, pp. 890-903, Oct. 2007, doi: 10.1109/TFUZZ.2006.889956.

[19]. T. B. Iredale, S. M. Erfani and C. Leckie, "An efficient visual assessment of cluster tendency tool for large-scale time series data sets," 2017 IEEE International Conference on Fuzzy Systems (FUZZ-IEEE), 2017, pp. 1-8, doi: 10.1109/FUZZ-IEEE.2017.8015587.

[20]. T. C. Havens and J. C. Bezdek, "An Efficient Formulation of the Improved Visual Assessment of Cluster Tendency (iVAT) Algorithm," in IEEE Transactions on Knowledge and Data Engineering, vol. 24, no. 5, pp. 813-822, May 2012, doi: 10.1109/TKDE.2011.33.

[21]. 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.

[22]. M. Palaniswami, A. S. Rao, D. Kumar, P. Rathore and S. Rajasegarar, "The Role of Visual Assessment of Clusters for Big Data Analysis: From Real-World Internet of Things," in IEEE Systems, Man, and Cybernetics Magazine, vol. 6, no. 4, pp. 45-53, Oct. 2020, doi: 10.1109/MSMC.2019.2961160.

[23]. https://archive.ics.uci.edu/ml/datasets/

[24]. Q. Wang, H. Zhang, Qi Dong, Q. Niu, G. Xu and Y. Xue, "Otsu thresholding segmentation algorithm based on Markov Random Field," 2011 Seventh International Conference on Natural Computation, 2011, pp. 969-972, doi: 10.1109/ICNC.2011.6022194.

[25]. Z. Chen, D. Chang and Y. Zhao, "An Automatic Clustering Algorithm Based on Region Segmentation," in IEEE Access, vol. 6, pp. 74247-74259, 2018, doi: 10.1109/ACCESS.2018.2881230.