Soft Computing
https://doi.org/10.1007/s00500-022-07092-x

FOCUS

Hybrid visual computing models to discover the clusters assessment of high dimensional big data

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

Accepted: 25 March 2022

Abstract
Clusters assessment is a major identified problem in big data clustering. Top big data partitioning techniques, such as, spherical k-means, Mini-batch-k-means are widely used in many large data applications. However, they need prior information about the clusters assessment to discover the quality of clusters over the big data. Existing visual models, namely, clustering with improved visual assessment of tendency, and sample viewpoints cosine-based similarity VAT (SVPCS-VAT), efficiently perform the clusters assessment of big data. For the high-dimensional big data, the SVPCS-VAT is enhanced with the subspace learning techniques, principal component analysis (PCA), linear discriminant analysis (LDA), locality preserving projection (LPP), Neighborhood preserving embedding (NPE). These are used to develop hybrid visual computing models, including PCA-based SVPCS-VAT, LDA-based SVPCS-VAT, and LPP-based SVPCS-VAT, NPE-based SVPCS-VAT to overcome the curse of dimensionality problem. Experimental is conducted on benchmarked datasets to demonstrate and compare the efficiency with the state-of-the-art big data clustering methods.

Keywords Data clustering · Cluster tendency · Visual models · Big data · Subspace learning

1 Introduction

Cluster analysis (Jiang et al. 2004; Kumar et al. 2013) is the fabulous technique for the data partitioning problem in which a similar set of data substances are placed into the same cluster (or group). The similarity features of data objects are derived using the different distance metrics. The cluster analysis or data clustering is prominent, including two key steps, which are as follows: (1) clusters assessment for finding the prior knowledge about the total number of clusters (it known as cluster tendency problem) and (2) Generation of clusters for the set of data objects. Top-clustering methods (Wu et al. 2008), say, k-means and hierarchical clustering are an efficiently generates the data clusters in the applications, such as big data analysis (Deepak et al. 2021; Rajendra Prasad et al. 2019; Subba Reddy et al. 2022), social data clustering (Rui and Wunsch 2005), image clustering, speech and video clustering (Rajendra Prasad and Suleman Basha 2016), market research, pattern recognition (Bezdek 1981), web mining (Ramathilagam et al. 2013), biological data mining. Those were suffering from the issue of pre-clusters assessment or cluster tendency. Cluster tendency refers to the underlying practical assumption of several clusters. For example, in k-means also, it is not possible to assign the exact 'k' value in all cases. With an external interference, a user may attempt an intractable 'k' value (or cluster tendency), which attempts the poor clustering results. Pre-cluster assessment methods, i.e., visual computing models, say, visual assessment of (cluster) tendency (VAT) (Bezdek and Hathaway 2002), spectral-based VAT (SpecVAT) (Bezdek 2008), improved VAT (iVAT) (Havens and Bezdek May 2012), are widely used for determining the value of cluster tendency. Euclidean-based dissimilarity and re-ordered dissimilarity matrices are derived to assess cluster

Communicated by V Suma.

✉ M. Suleman Basha
suleman.ndl@gmail.com

S. K. Mouleeswaran
meetmoulee@gmail.com

K. Rajendra Prasad
krprgm@gmail.com

[1] Department of CSE, Dayananda Sagar University, Bangalore, India

[2] Department of CSE, RGM College of Engineering and Technology, Nandyal, India

Published online: 23 April 2022

tendency. Another distance metric, cosine, is more effective in finding the similarity features of data substances. It uses both the magnitude and direction of the data objects vectors, unlike Euclidean distance. Euclidean distance metric takes only the distance in finding either similarity or dissimilarity features of data substances.

For this reason, cosine-based VAT (cVAT) (Hu et al. 2012) and cosine-based spectral VAT (cSpecVAT) (Suleman Basha et al. 2019) are developed for the extraction of accurate clusters information. These methods determine the similarity features of data objects about a single viewpoint. Multi viewpoints-based closeness features using a cosine distance provide a more instructive assessment than a single viewpoint. Multi viewpoints-based cosine similarity VAT (MVS-VAT) is the technique in Suleman Basha et al. (2021) developed for the more appropriate clusters assessment. It may not be a cost-effective clusters assessment method for big data. SVPCS-VAT (Suleman Basha et al. 2021) is developed for big data, and it uses only sample viewpoints instead of (n-2) viewpoints. Here n refers to the number of data objects. For example, for any two data objects, t1, and t2, the SVPCS-VAT selects the sample viewpoints from remaining (n-2) data objects only. Thus, recent SVPCS-VAT is a more cost-efficient method concerning time and memory space values when compared to MVS-VAT. The basic idea of linear subspace learning (LSL) can be used for lowering the dimensionality of high-dimensional data. It maps the very higher number of dimensions into a very lower number of manifold dimensions. The LSL techniques (Jiang 2011), such as LDA, PCA, and LPP are taken to develop proposed hybrid visual computing models. Three variants of proposed models are implemented for addressing an emerging issue of dimensionality problem; which models are as follows: LDA-based-SVPCS-VAT, PCA-based-SVPCS-VAT, and LPP-based-SVPCS-VAT. The architecture of the proposed work is shown in Fig. 1.

The proposed framework is best suited for data partitions of high number dimensions of big data. It primarily uses the LSL and the min–max sampling strategy (Rajendra Prasad et al. 2021) for the best sample selection of viewpoints. The benefit of LSL is to obtain the low-dimensional manifolds of original big data. Finally, sampling strategy is applied for low-dimensional manifolds of big data, in which the inter clusters sample viewpoints are selected. This inter-cluster viewpoints-based cosine similarity is more accurate than others. The procedural steps are described in the relevant section of the proposed work. Finally, the dissimilarity matrix is obtained for the low-dimensional manifold big data, and it can be used as input of VAT. The VAT displays the output in a visual image

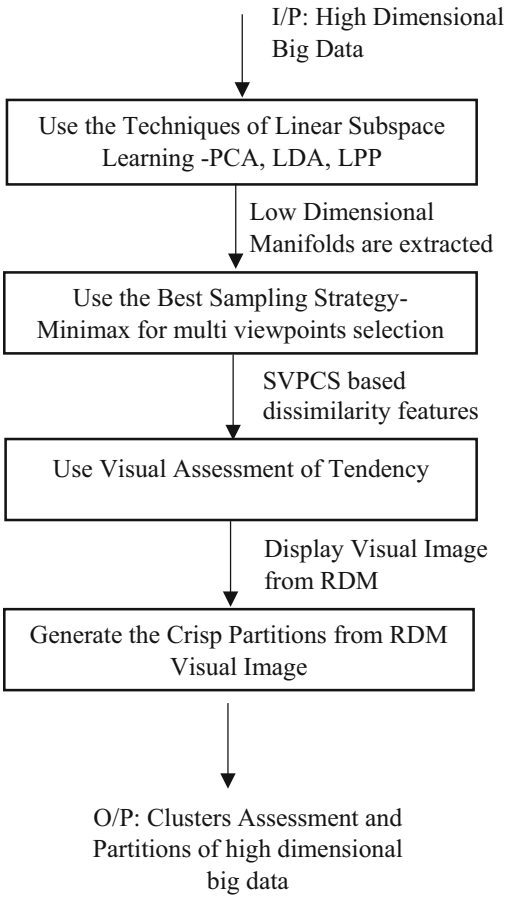


Fig. 1 Proposed work architecture

consisting of visual clusters in square-shaped dark-colored blocks. Important crisp partitions are derived with the information of diagonal and non-diagonal square blocks of visual images. These partitions are used for predicting the exact labels of data objects in big data partitioning problems. The classification of visual computing models is LDA-based-SVPCS-VAT, PCA-based-SVPCS-VAT, and LPP-based-SVPCS-VAT, which efficiency is greatly improved compared with the state-of-the-art of the other big data clustering methods to address the high-dimensional data partition problem as well as the size of the big datasets.

Summary contributions of the paper are described as follows:

1. Clusters assessment is effectively performed through the visual image information
2. Big dimensionality problem of big compact separated (CS) and non-compact separated (non-CS) datasets is handled in our proposed work.

3. The LSL-based cluster assessment models are developed for determining the clustering tendency of high-dimensional datasets.
4. The crisp partitions are derivative from the resulting images of visual computing models to discover complete clustering results over the high-dimensional big datasets.
5. The performance of the proposed models is demonstrated using CS and non-CS benchmarked datasets.

Other sections of the paper are mentioned as follows: Sect. 2 overviews the visual models for data clustering. Section 3 describes the linear subspace learning techniques for low-dimensional manifolds. Section 4 presents the proposed hybrid visual computing models. Section 5 shows the experimental results and discussion. Finally, the conclusion and scope of future work are described in Sect. 6.

2 Overview of the visual models for data clustering

Data clustering algorithms require prior information about the clustering tendency or the clusters assessments in terms of several clusters. For the clustering tendency, many pre-clusters assessment methods are surveyed. Data visualization or visual models are the emerging techniques for solving the problem of cluster tendency. They aim to find the intractable value of cluster tendency (or k the value in k-means) in the topmost clustering methods. This cluster tendency problem is also referred to as the pre-clusters assessment problem. Visual model (visual assessment of tendency -VAT) is proposed by Bezdek et al. in Bezdek and Hathaway (2002), Kumar et al. (2016) for attempting the solution of a clusters assessment problem. It aims to assess the clusters' information through the visual image. It uses Prim's logic for re-ordering the dissimilarity matrix (RDM). Famous distance metrics are taken to find the dissimilarity features of data substances in the VAT, then re-orders the dissimilarity matrix based on the order of dissimilarities of the data substances. Objects are re-ordered according to the changed ordering of the dissimilarity matrix. The objects are re-ordered or moved into the respective clusters based on the distances or dissimilarities. All these steps of the VAT algorithm are illustrated in Algorithm 1.

Algorithm : VAT [7]
Input : Objdm[][]

Output : Total clusters or value of cluster tendency 'k'

Step (1)

```
L= { } ;
M={0,1,.....n-1}
Find maximum over the Objdm[ ] [ ], and store the corresponding row and column index values into (s,k)
P(0)=s;
L ={k};
M=M-L;
```

Step (2):

```
for (i=1;i<n; i++)
 {
for (j=1;j<n; j++)
{
  Take (i,j) by obtaining the minimum value form Objdm[i][j], here, i ϵ L, j ϵ M
   L = LU {j};
  M= M-L;
  P(i)=j;
 }
```

Step3:

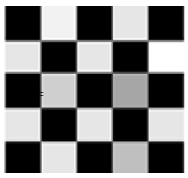
```
for(s=0;s<n;s++)
    for(k=0;k<n;k++)
      ObjRDM=Objdm(P[s],P[k]);
```

Step 4:

```
Display Visual Image(ObjRDM)
```

For example, Fig. 2 illustrates how the clusters information or cluster tendency is extracted for the set of data substances. The dissimilarity features of the data substances are mentioned in Fig. 2a, and using the visual model, VAT, the dissimilarity matrix is converted into a re-ordered dissimilarity matrix (RDM), and it is shown as another matrix form in Fig. 2b. Based on the observations, it is clear that clusters information is intractable in Fig. 2a Visual image; whereas in Fig. 2b. The clusters information is tractable through square-shaped dark colored blocks in the VAT Image.

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



(a) Dissimilarity Matrix of Dataset and intractable visual clusters in VAT Image

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

(b) After VAT, Re-ordered dissimilarity matrix (RDM) and tractable visual clusters in VAT Image.

Fig. 2 Tractable Clusters information from visual images

The VAT is a basic version of the pre-clusters assessment model, and it works for small-scale synthetic and other real-time datasets. For the complex datasets or any path-shaped datasets (Havens and Bezdek 2012), improved VAT (iVAT) is developed. It uses the path-based distance measures for the re-ordering of the data matrix. It is ideally suited for the data clusters assessment of path-shaped datasets. Other successive versions of visual models are Spectral VAT, cosine-based VAT (cVAT), cosine-based Spectral VAT(cSpecVAT) is developed for the better assessment of cluster tendency when compared to VAT and iVAT for the complex datasets. The spectral approach (Yang et al. 2015) finds the Eigenvectors by finding the affinity matrix and Laplacian matrix. The cSpecVAT discovers the dissimilarity features for the respective Eigenvectors of data objects using the cosine distance with a single viewpoint of origin. It finds either similarity (or dissimilarity) concerning a single origin (is also called a single viewpoint). Finding the closeness features concerning other multi viewpoints is more accurate than just using a single viewpoint. These enhancement steps are developed in another model MVS-VAT. The critical approach for finding the similarity features using the cosine, and multi viewpoints are illustrated with sample examples of five data objects (say, v1,v2, v3, v4, and v5) in Fig. 3. The similarity features between two data objects (i.e., v1 and v2) are derived using different viewpoints, shown in the same figure. Cosine-based similarity features are performed based on this example's three cases.

i. Similarity between (v1,v2) is derived based on viewpoint v3 and store the value in to a variable S1
ii. Similarity between (v1,v2) is derived based on viewpoint v4 and store the value in to a variable S2

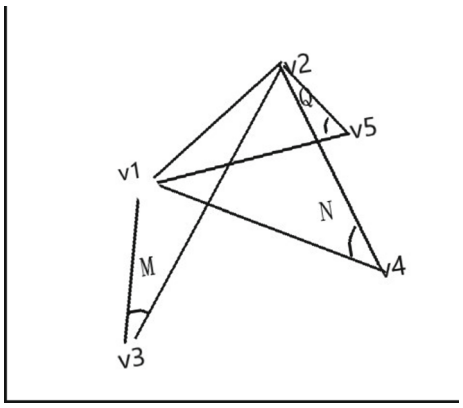


Fig. 3 Multi viewpoints cosine-based similarity computation

iii. Similarity between (v1,v2) is derived based on viewpoint v5 and store the value in to a variable S3

Finally, an average of three values, i.e., S1, S2, and S3, is taken as the similarity feature between the data objects v1 and v2. The same procedure is repeated for finding the similarity features for other pairs of data objects, say (v1, v3), (v1, v4).....(v2, v3), ...(v4, v5).

With the MVS-VAT (Suleman Basha et al. 2021), cluster tendency is automatically determined and generated the quality of data clusters. Big data demands a high amount of computation time and memory allocations. It is an expensive approach for big data clustering. Further enhancement of MVS-VAT is made with the sampling strategy to develop an efficient SVPCS-VAT (Suleman Basha et al. 2021) (Rajendra et al. 2021) visual model concerning the parameters of time and memory allocation. The big data is broadly classified into compact separated (CS) (Havens et al. 2009) and non-compact separated (non-CS) (Rathore et al. 2019) datasets. Traditional big data clustering methods, spkm (Hore et al. 2007; Bradley et al. 1998), Mini Batch-k-means (MB-K-means) (Sculley 2010; Subba Reddy et al. 2022), generate the quality of clusters for the CS type of big datasets. These methods cannot generate the quality of clusters for non-CS type of big datasets due to their irregular or arbitrary boundaries of the clusters. The SVPCS-VAT visual model effectively works for both CS and non-CS types of datasets. It investigated that big data may be available with high dimensions also. In these big data cases, the curse of dimensionality occurs. This challenge is solved with another state-of-the-art visual model, Fensi-VAT (Rathore et al. 2019). It uses random projections (Urruty et al. 2007; Achlioptas 2001) to transform the high-dimensions into a reduced subspace with fewer dimensions. It is applicable for both CS and non-CS type of datasets.

FensiVAT is the recent visual method for performing the biga clustering and its performance is impressive compared to other big data clustering methods. The

FensiVAT is a more effective visual big data extensive clustering method, and it has taken the random projections for dimensionality reduction.

Many advancements have been made in the research of dimensionality reduction. These techniques are specially used to create low-dimensional manifold subspaces for the high-dimensional data. These kinds of techniques are usually referred to as linear subspace learning techniques. These techniques are used to develop proposed hybrid visual computing models in this paper. Details of the linear subspace learning are described in the next following section.

3 Linear subspace learning techniques for low-dimensional manifolds

Nowadays, the technique of Eigen decomposition is used in most of the linear subspace learning techniques. In the Eigen decomposition, the Laplacian matrix is initially derived by finding the affinities matrix and diagonal matrix of the affinity matrix. Later, the best projections of high dimensional data are extracted by choosing the largest k-eigenvectors.

Various methods of LSL aim to produce the projections of lower dimensions which are alternate to the random projection method. The principal axis (Duda et al. 2001) is determined by finding the two largest Eigenvectors in principal component analysis (PCA) (Vidal et al. 2005). Maximizing the separability between the inter-cluster object's data showed better discriminations for high-dimensional data. The high-dimensional data is mapped with better discriminated low-dimensional data in the Linear discriminant analysis (LDA) (Blei et al. 2003) technique. Preserving the object's discriminant information is the ultimate objective of LSL methods, in which it is required to estimate the neighborhood affinities for the data objects. It is performed by locality preserving projections (LPP) (Belkin and Niyogi 2008) with the construction of neighborhood structure for the set of data objects.

The SVPCS-VAT is a hybrid big data clustering technique, which uses VAT and random projection for addressing the dimensionality problem. It is a recent technique, in which only the random projection is suggested for the dimensional problem. However, there are good alternatives are available for solving the problem of dimensionality reduction. The LSL techniques are used for optimal solutions of dimensionality reduction. With these techniques, proposed techniques are developed, which are the combinations of LSL and SVPCS-VAT. The following section presented algorithms of proposed hybrid visual computing models for performing of efficient high-dimensional big data clustering.

4 Proposed hybrid visual computing models

Proposed models extend the SVPCS-VAT model with LSL methods to overcome the problem of the curse of dimensionality. Three variants of LSL methods, i.e., PCA, LDA, and LPP, describe the three variants of hybrid visual computing models. Algorithm I illustrate the methodology of proposed hybrid visual computing models. Initial read the HBD data with the size of m * n, whereas m and n refer to the number of data objects and dimensions.

Initially, it calls the procedure of LSL with the input parameters of HBD and value. Here, value tells that type of LSL method is being applied on HBD to retrieve reduced dimensionality of the data and store it into a LM. Then calls the recent visual method SVPCS-VAT (Suleman Basha et al. 2021) with the input of reduced data LM. The SVPCS-VAT image is obtained and derives the visualized dark colored blocks along the diagonal of the SVPCS-VAT image. The crisp-partitions are derived from these visualized dark-colored blocks in the diagonal. After obtaining the crisp partitions, it is easy to read the data objects' cluster labels. Finally, the clustering results of high-dimensional big data are efficiently derived with the proposed methodology.

Algorithm I : Linear Subspace Learning - Based SVPCS-VAT

Input : High-dimensional big data (HBD) Int val;

Result : 'k' number of clustering results

Method :

1. Read HBD, which has to be 'm' number of data objects with the column dimensions of 'n'
2. LM = LSL(HBD, Val)

// Generate the crisp partitions for the big data clusters

3. Find the visual images using SVPCS-VAT [37]
4. Derive the crisp partitions from visual images
5. Predict the information of cluster labels towards data objects from the results of crisp partitions for discovering big data clusters.

Algorithm II : LSL-Technique (Val)

Input : HBD- High Dimensional Big Data

Output : k- clusters

Method:

1. Check the case for Val=1 or Not) // If Val==1 then applied LSL is PCA

{

a. Normalize the HBD
b. Determine the HBD covariance values
c. Derive the Eigen decomposition for finding the largest k-eigenvectors
d. Determine LDM (LDM refers to low-dimensional manifolds) of HBD by selecting the components, here the component vectors are referred to as the principal components.

Return LDM

}

2. Check Val=2 or Not // If the input of Val ==2 , then LSL is LDA

{

i. Computing the super mean vectors of n-dimensional big data
ii. Two LDA scaling matrices are derived which are S_w and S_b. These matrices are derived based on intra and inter distances of data objects.
iii. Eigen computation of LDA is performed with $S_W^{-1} S_b$
iv. Sorting the LDA component vectors with descending order of component vector values
v. Map the big dimensional data by selected k-component selected vectors and save the lo-dimensional features into LDM

Return LDM

}

3. Check Val=3 or Not

// If the input value of Val is 3, then LSL is LPP

{

i. Based on affinities of neighbors, the adjacency weighted matrix 'W' is constructed
ii. Each weight from W indicates the dissimilarity (or affinity) between the pair of objects.
iii. The diagonal matrix D is defined over the weighted matrix W; both W and D are used for the construction of Laplacian matrix 'L'
iv. Determine the Eigenvectors based on the ordering of Eigenvalues as per L
v. Finds the components based on the selection of k number of largest vectors whereas k should be less than or equal to a number of input dimensions 'n'. These components imposes the values of low-dimensional manifolds, LDM Return LDM

}

Table 1 Description of high-dimensional big datasets

S. Num	Name of the dataset	Size	Num. of dimensions
1	GD 1 (k = 2)	80,000	52
2	GD 2 (k = 3)	100,000	110
3	GD 3 (k = 6)	120,000	500
4	KC'99 (KDD CUP'99)	4,898,431	18
5	MiniBooNE (k = 2)	130,064	50
6	MNIST	70,000	784

**GD* Gaussian data

Algorithm II shows the procedures for the three models of LSL for obtaining the reduced dimensions of original high-dimensional big data. In PCA, the HBD is standardized using the min–max normalization technique, and then the covariance matrix is derived. The Laplacian matrix is derived using the covariance matrix input for finding the largest k- Eigenvectors, where k is assumed or represents the reduced number of dimensions 'k'. In LDA, the scattered matrices S_w and S_b are derived from the n-dimensional mean vectors of the object's data. After that, similar steps of PCA have applied in LDA also for determining the Laplacian and k-largest Eigenvectors. The low-dimensional manifolds 'LM' is determined with the most prominent 'k' Eigenvectors for n data objects. Thus, the size of the reduced dimensionality of HBD is n x k, where k refers to the reduced number of dimensions. In LPP, the Laplacian matrix 'L' is determined from the weighted matrix W. The W is constructed based on an adjacency graph considering neighborhood data objects' affinities.

5 Experimental results and discussion

The clusters assessment is performed by various visual methods for the experimental on various high-dimensional big datasets, and data sets are shown in Table 1. Synthetic big Gaussian data are generated with many dimensions in MATLAB. There different kinds of synthetic data are created for experimental demonstration purposes. Three big real-time datasets say, KC'99 (KDD CUP'99) (Tavallaee et al. 2009), MiniBooNE (Asuncion and Newman 2007), and MNIST (LeCun et al. 1998), are the high-dimensional big datasets, which are used for the experimental of cluster tendency and efficiency demonstration of proposed methods.

Initially, VAT images are generated using the existing SVPVS-VAT, Fensi-VAT, and proposed LSL-based SVPCS-VAT methods. The Fensi-VAT uses random projections for the reduction of dimensions. It enables the low-dimensional manifolds for handling the curse of

dimensionality with random projection mappings. It takes less time when compared to SVPCS-VAT; however, it may not be appropriate for the case of high dimensional datasets. Thus, FeniVAT gives less goodness (or clarity) of the visual images when compared to SVPCS-VAT for the high-dimensional datasets. Proposed LSL-based SVPCS-VAT i.e., PCA-based-SVPCS-VAT, LDA-based-SVPCS-VAT, and LPP-based-SVPCS-VAT achieved the best clarity of visual images shown in Figs. 4, 5, 6, 7, 8 and 9 for the datasets of GD 1, GD 2, GD3, KDD CUP'99, and MNIST. In all these cases, the proposed methods outperformed the others for determining the clustering tendency.

Figure 4 shows the experimental for obtaining the visual images using visual models, and it clears that all the LSL-based-SVPCS-VAT methods showed the best clarity of dark coloredcolored square blocks along the diagonal associated to other existing visual models. Figure 4b represents in the format of crisp partition.They showed that Gaussian data 1 is two clustered data accessing the two square-shaped dark colored blocks along with the diagonal. Goodness for the images of SVPCS-VAT and Fensi-VAT are retained as low. Thus, proposed visual models are effectively used for accessing the num. of clusters (or cluster tendency).

Three proposed models are shown the good clarity of visual images when compared to two existing methods. Figure 5 shows the assessment of cluster tendency for the Gaussian data2 using various models; here also, proposed models efficiently access the clustering tendency is 3 for the Gaussian data2 with their best clarity of the images. It also observed that LDA-based-SVPCS-VAT and LPP-based-SVPCS-VAT achieved the excellent clarity of visual images. Figure 6 also derives the same clarity of visual images for the Gaussian data 3 using the variants of visual models.

Figures 7, 8, and 9 show the experimental for the three real-time high dimensional big datasets, KDD CUP'99, MiniBooNE, and MNIST, respectively. Three proposed models showed the visual images of one big cluster, two moderate clusters, and six tiny clusters of the KDD CUP'99 datasets. Fensi-VAT shows the overlaps of visual clusters, and it is intractable to conclude the number of clusters. Another existing method, SVPCS-VAT derives the visual clusters with good visual images for KDD CUP'99. MiniBooNE is a boolean kind of high-dimensional big data. For this case, only two proposed variants, includes, LDA-based-SVPCS-VAT and LPP-based-SVPCS-VAT, achieved the best clarity of visual images. For the MNIST datasets, all the proposed visual variants achieved the best clarity of visual images. Overall experimental of visual images, proposed visual approaches achieved with high value of goodness. Both LDA-based-

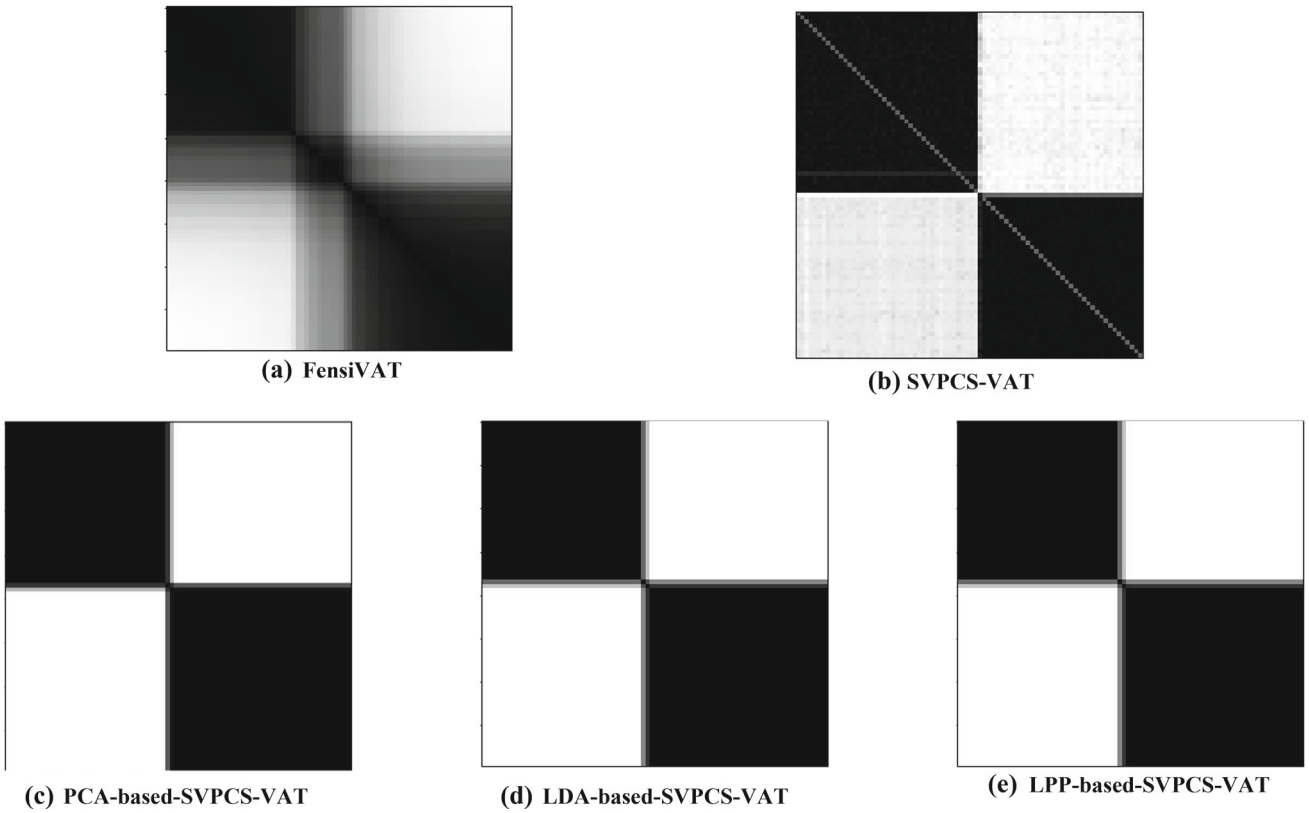


Fig. 4 Images of visual models for GD 1 ($k = 2$)

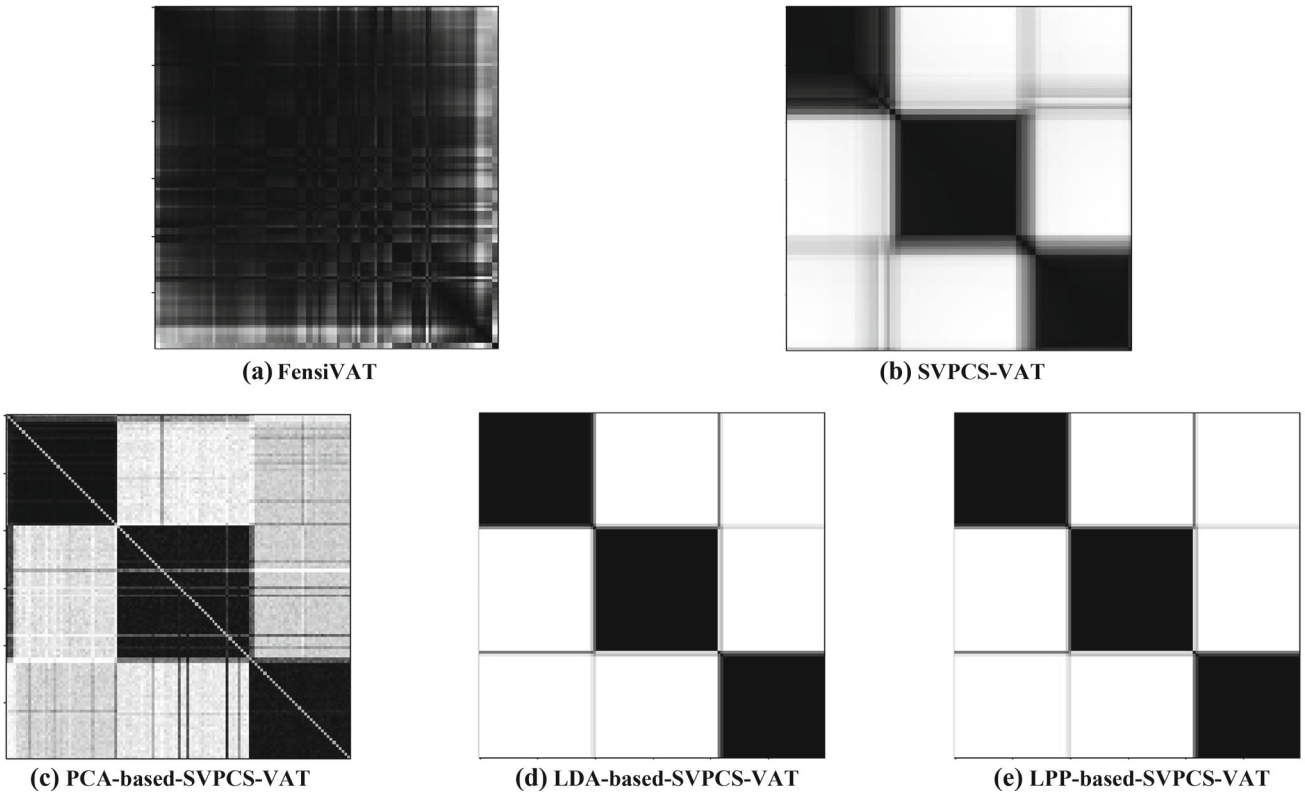


Fig. 5 Images of visual models for GD 2 ($k = 3$)

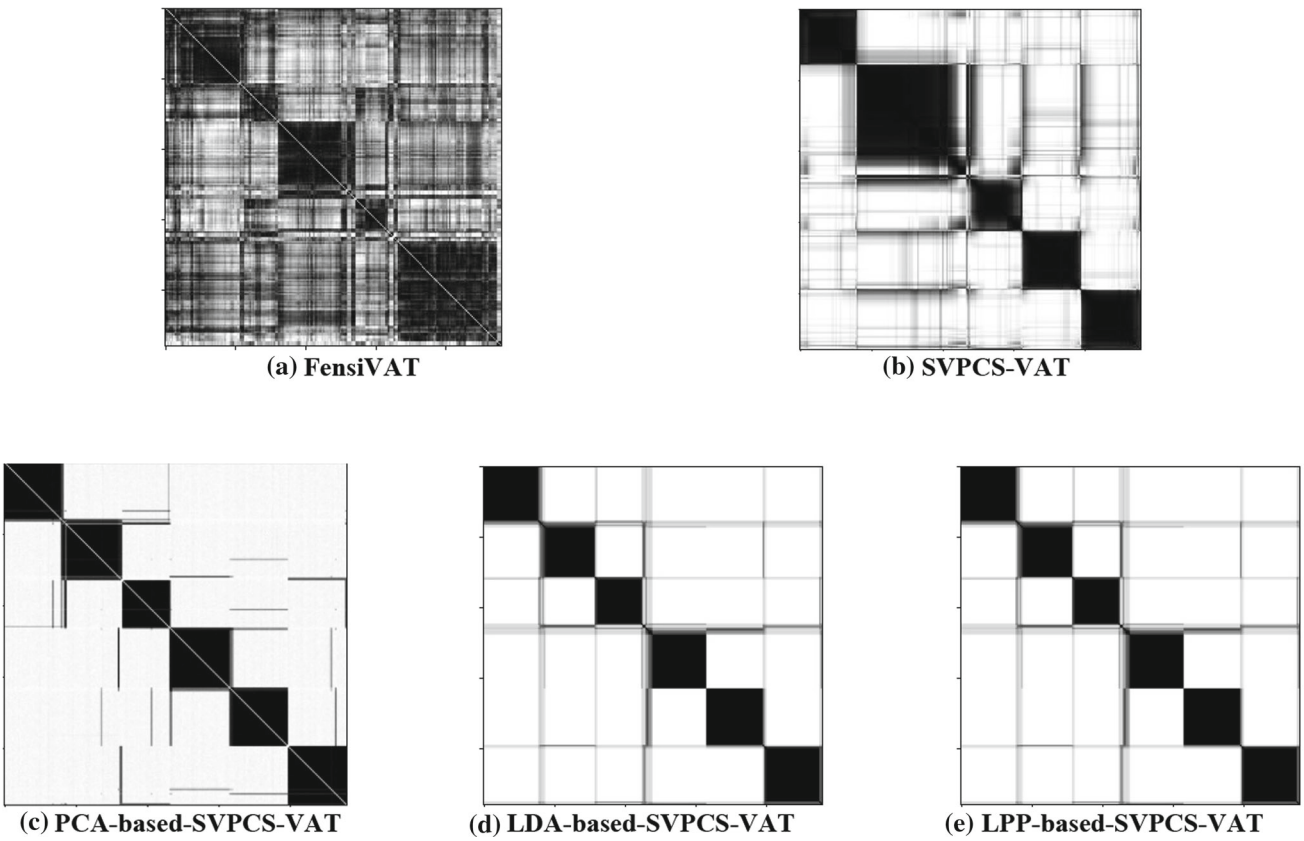


Fig. 6 Images of visual models for GD 3 ($k = 6$)

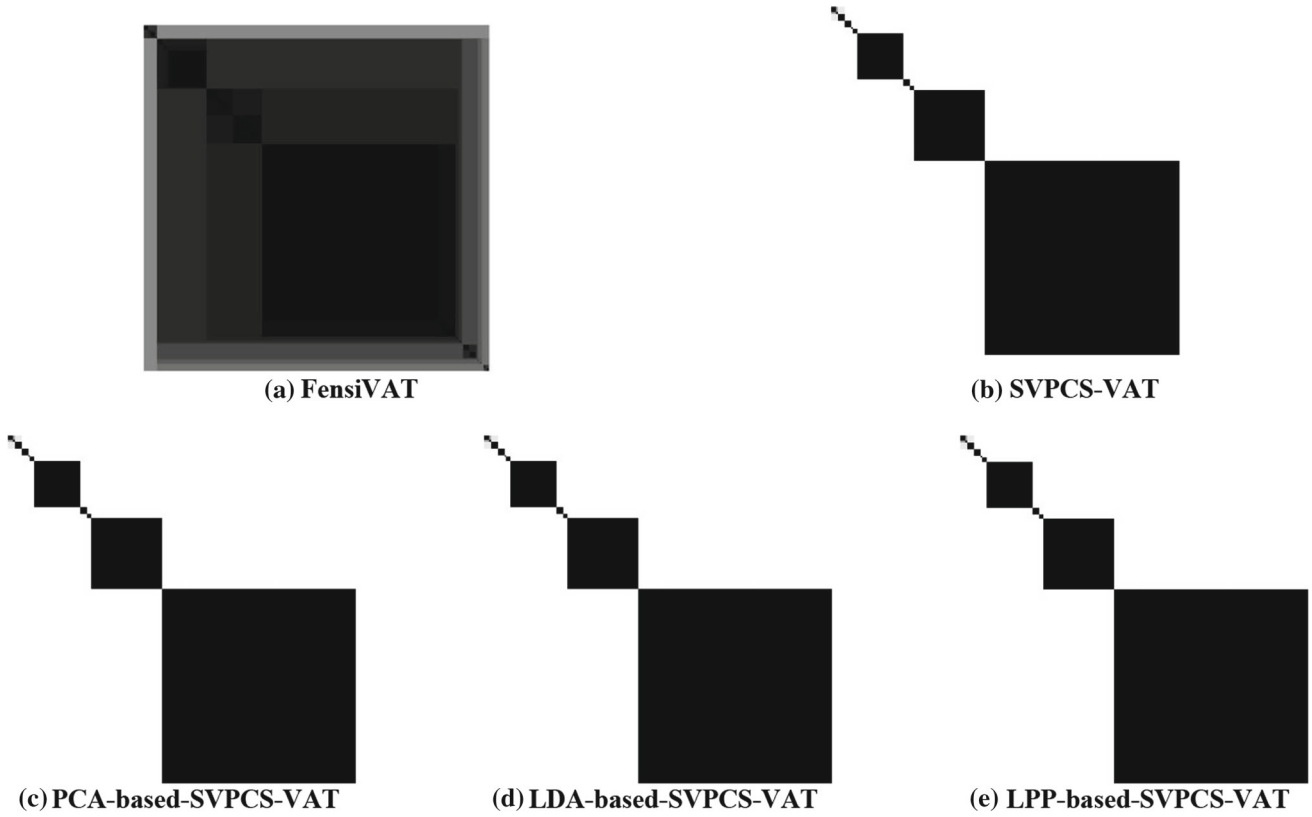


Fig. 7 Images of visual models for KC'99

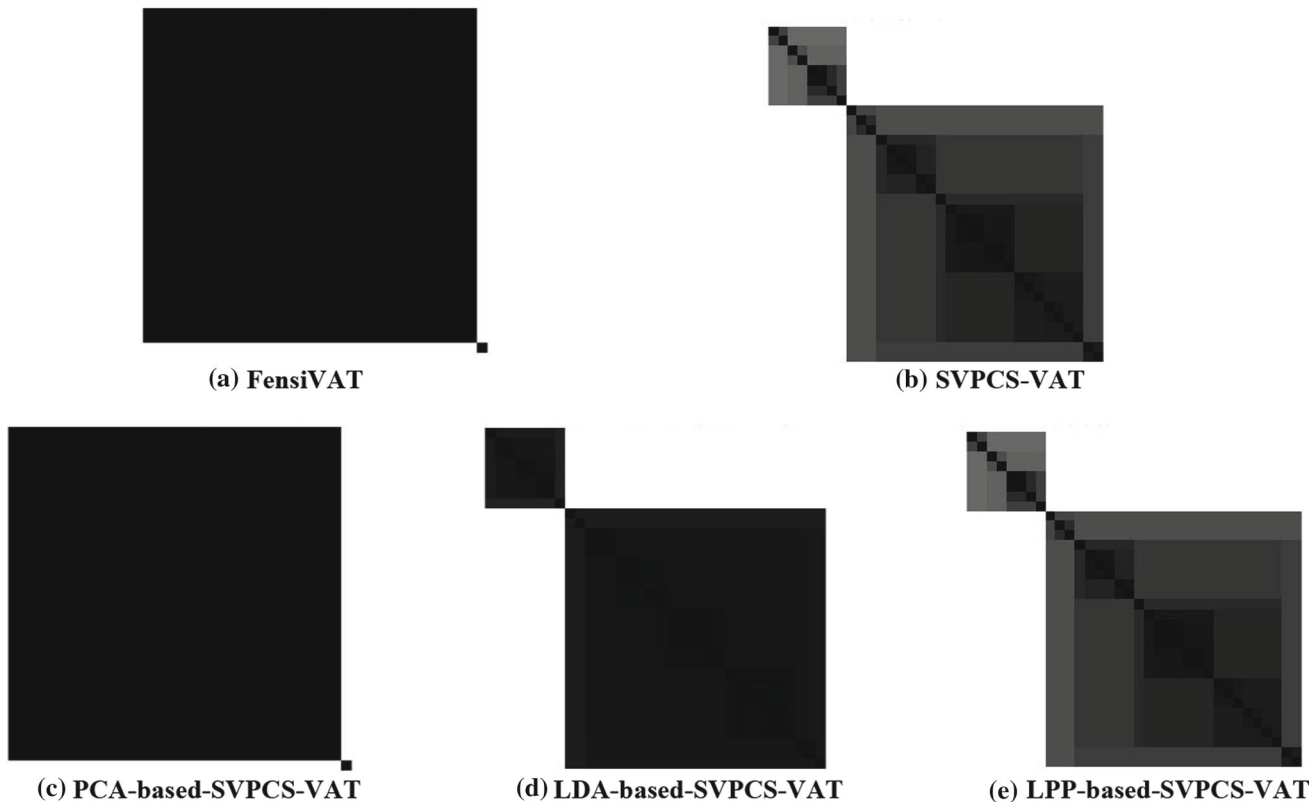


Fig. 8 Images of visual models for MiniBooNE ($k = 2$)

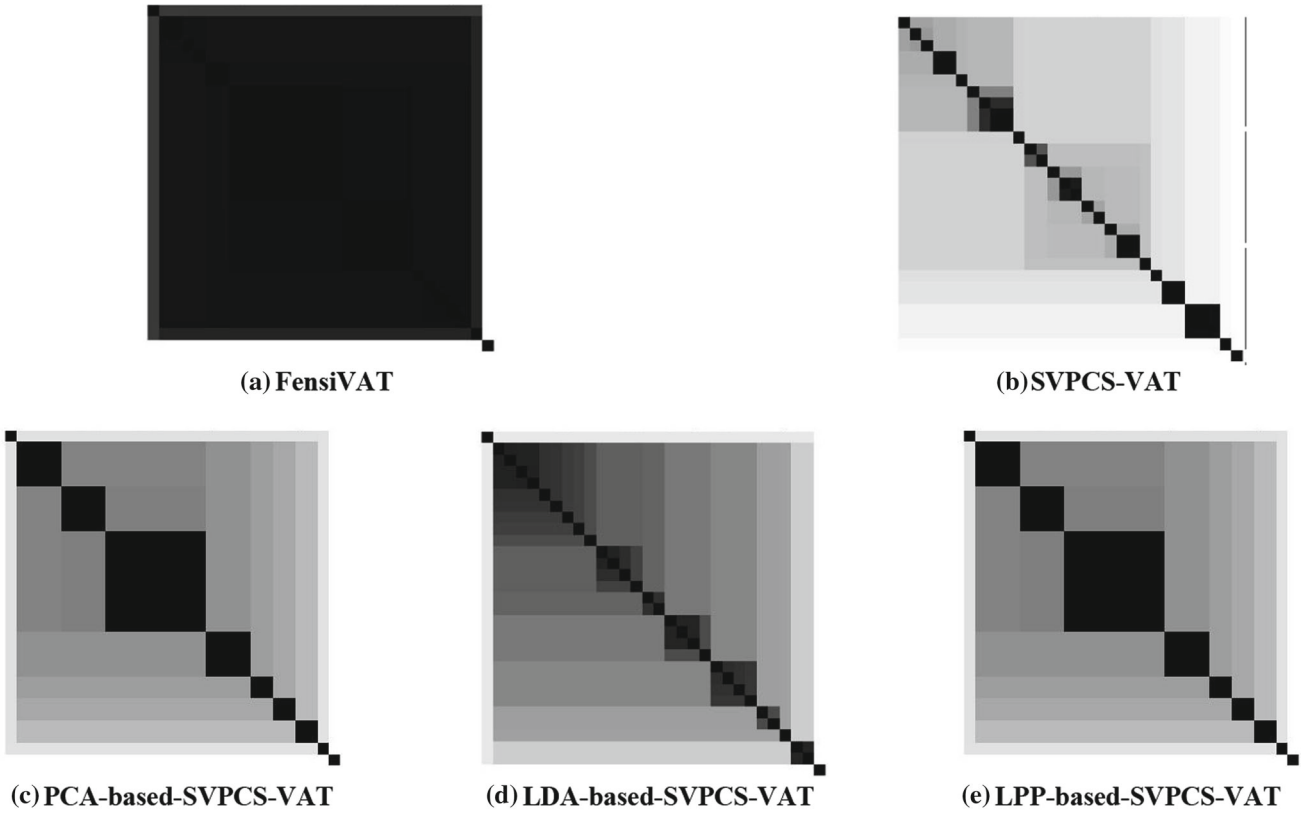


Fig. 9 Images of visual models for MNITS (k = 10)

SVPCS-VAT and LPP-based-SVPCS-VAT are most successful for all the experimental study datasets.

Table 2 presents the conduct of big data clustering methods. Performance of state-of-art and proposed methods are evaluated using the six parameters, includes, partition accuracy (PA) (Lastname et al. 2016), normalized mutual information (NMI) (Amelio and Pizzuti 2015), specificity, precision, recall, and sensitivity (Bhatnagar et al. 2018), in the experimental study. Based on observation of performance values, three proposed visual models, PCA-based-SVPCS-VAT, LDA-based-SVPCS-VAT, and LPP-based-SVPCS-VAT, achieved as excellent for the significant data clustering results. The proposed methods are the most accurate for generating big data clusters of synthetic Gaussian datasets. However, they maintain good accuracy for big data clusters associated to other big data clustering results. For other big real-time datasets, the performance values of proposed methods are significantly changed.

The speedup parameter (s) is evaluated with the running time of existing and proposed techniques. It is the quotient of these two running time values. The running time of visual models is presented in Table 3. The speedup parameter value is evaluated to compare the fastness of visual models relative to existing big data clustering methods, spkm, and Mini-Batch-k-means. Figures 10, 11, and 12 show the speedup value comparison for the proposed methods with existing spkm. In all these cases, proposed visual hybrid models proved faster techniques for the high dimensional big datasets compared with spkm. The same observation is also made from Figs. 13, 14, and 15, and proposed models are outperformed with other techniques of Min-Batch-k-means concerning the parameter of speedup. Figure 16 also observed that the proposed methods outperformed the others concerning memory allocation parameters (Figs. 17, 18).

6 Conclusion and scope of future work

Visual models play a vital role in high-dimensional big data clustering methods. According to the state-of-the-art visual models, it emerges to address the cluster tendency problem for high-dimensional datasets. The recent FensiVAT solves the problem of cluster tendency using random projections. This paper presented three hybrid visual computing models, which use LSL techniques to find the robust low-dimensional manifolds of high-dimensional big data. These techniques effectively explore the clustering tendency and discover the best significant data clustering results for high-dimensional datasets. The future work

Table 2 Performance assessment of proposed LSL-based SVPCS-vat approaches and other big data clustering approaches

Name of the datasets	MB-K-means	spkm	Fensi VAT	SVPCS-VAT	PCA-based-SVPCS-VAT	LDA-based-SVPCS-VAT	LPP-based-SVPCS-VAT
PA—partition accuracy							
GD1	0.231	0.255	0.258	0.321	1	1	1
GD2	0.245	0.249	0.327	0.359	1	1	1
GD3	0.211	0.219	0.285	0.341	1	1	1
KC'99	0.314	0.129	0.498	0.521	0.591	0.825	0.830
MNIST	0.210	0.252	0.277	0.342	0.550	0.559	0.562
MiniBooNE	0.225	0.268	0.318	0.357	0.587	0.589	0.591
NMI—normalized mutual information							
GD1	0.219	0.241	0.264	0.319	1	1	1
GD2	0.221	0.247	0.325	0.355	1	1	1
GD3	0.221	0.247	0.279	0.351	1	1	1
KC'99	0.110	0.152	0.265	0.315	0.441	0.452	0.441
MNIST	0.225	0.232	0.251	0.351	0.444	0.459	0.457
MiniBooNE	0.217	0.168	0.127	0.271	0.433	0.439	0.451
SP—specificity							
GD1	0.211	0.241	0.242	0.322	1	1	1
GD2	0.221	0.223	0.316	0.428	1	1	1
GD3	0.222	0.223	0.272	0.312	1	1	1
KC'99	0.227	0.246	0.343	0.315	0.515	0.518	0.531
MNIST	0.215	0.229	0.224	0.374	0.517	0.522	0.545
MiniBooNE	0.219	0.267	0.237	0.365	0.519	0.515	0.525
P—precision							
GD1	0.110	0.190	0.195	0.295	1	1	1
GD2	0.227	0.231	0.239	0.351	1	1	1
GD3	0.226	0.227	0.228	0.359	1	1	1
KC'99	0.222	0.237	0.254	0.315	0.428	0.434	0.438
MNIST	0.221	0.222	0.235	0.325	0.416	0.389	0.412
MiniBooNE	0.278	0.264	0.284	0.357	0.425	0.429	0.435
R—recall							
GD1	0.175	0.185	0.195	0.255	1	1	1
GD2	0.210	0.227	0.238	0.310	1	1	1
GD3	0.228	0.234	0.241	0.341	1	1	1
KC'99	0.259	0.268	0.298	0.358	0.487	0.491	0.495
MNIST	0.230	0.234	0.241	0.364	0.522	0.528	0.527
MiniBooNE	0.268	0.278	0.298	0.387	0.552	0.559	0.568
SN—sensitivity							
GD1	0.175	0.182	0.189	0.268	1	1	1
GD2	0.215	0.229	0.238	0.310	1	1	1
GD3	0.221	0.241	0.268	0.310	1	1	1
KC'99	0.251	0.258	0.298	0.358	0.498	0.521	0.524
MNIST	0.241	0.249	0.238	0.358	0.447	0.458	0.461
MiniBooNE	0.251	0.268	0.278	0.387	0.528	0.527	0.531

Table 3 Runtime of big data clustering approaches for large-dimensional data

Name of the Datasets	MB-K-means	spkm	Fensi VAT	SVPCS-VAT	PCA-based-SVPCS-VAT	LDA-based-SVPCS-VAT	LPP-based-SVPCS-VAT
GD1	18	21	15	14	11	11.1	11.25
GD2	20	22	16	15	13	12.2	11.9
GD3	22	22.4	17	15	12	12.9	13.2
KC'99	23	26.45	19	17	14	14.3	14.2
MNIST	25	27.4	21	19	15	14.8	15.3
MiniBooNE	27	29.5	23	20	16	15.8	15.9

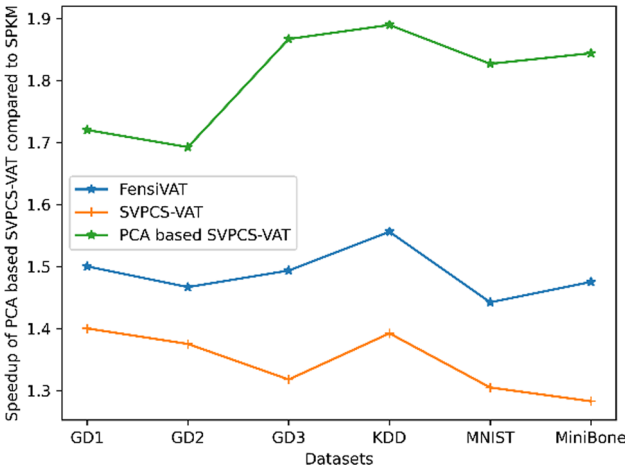


Fig. 10 Speedup relative to PCA based SVPCS-VAT compared to SPKM

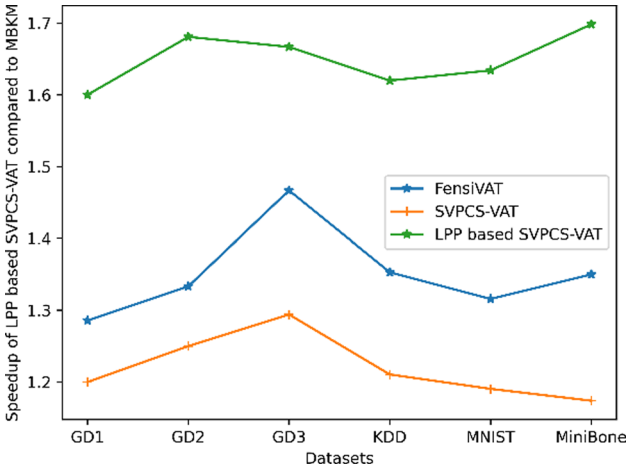


Fig. 12 Speedup relative to LPP based SVPCS-VAT compared to SPKM

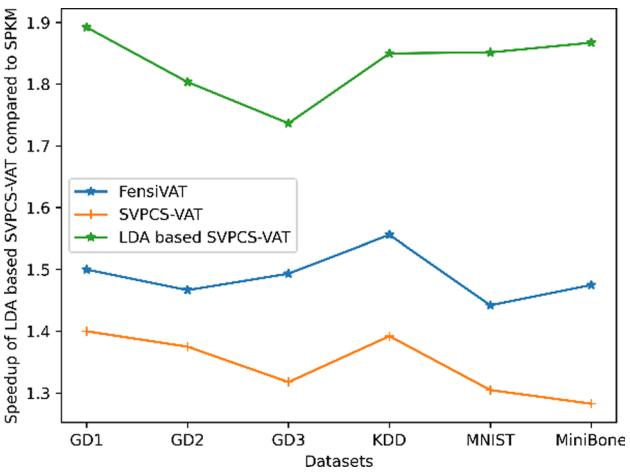


Fig. 11 Speedup relative to LDA based SVPCS-VAT compared to SPKM

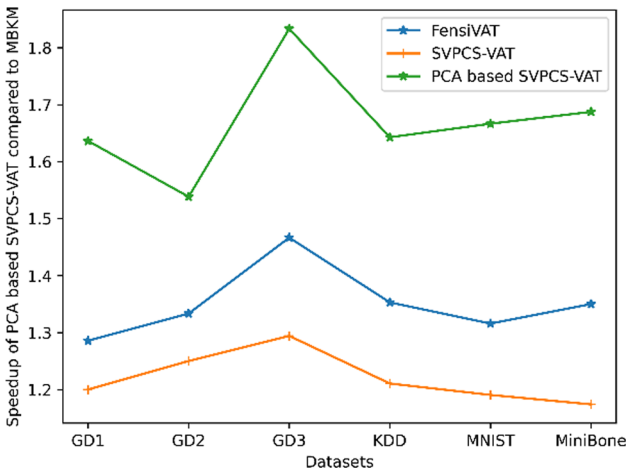


Fig. 13 Speedup relative to PCA based SVPCS-VAT compared to MBKM

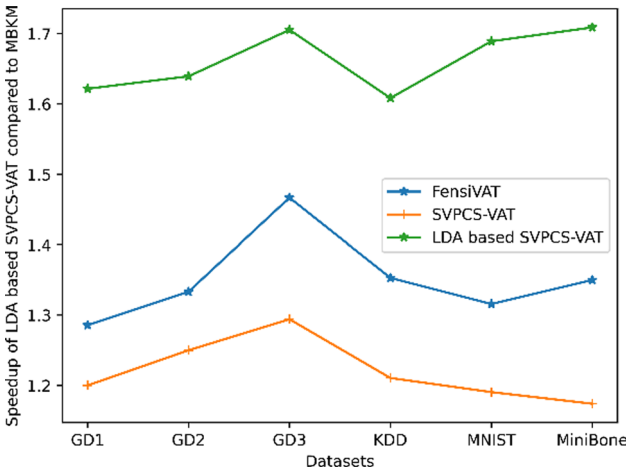


Fig. 14 Speedup relative to LDA based SVPCS-VAT compared to MBKM

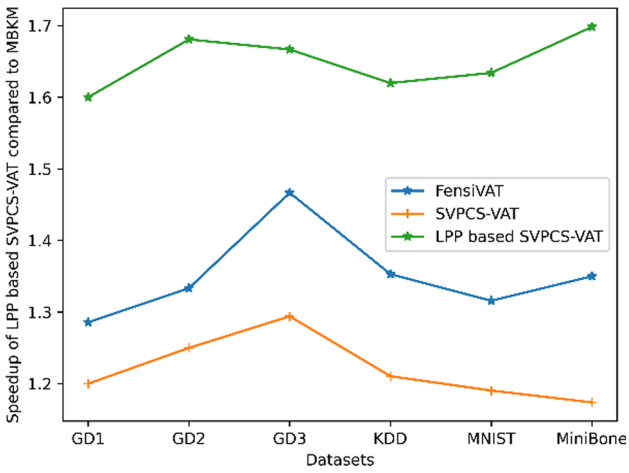


Fig. 15 Speedup relative to LPP based SVPCS-VAT compared to MBKM

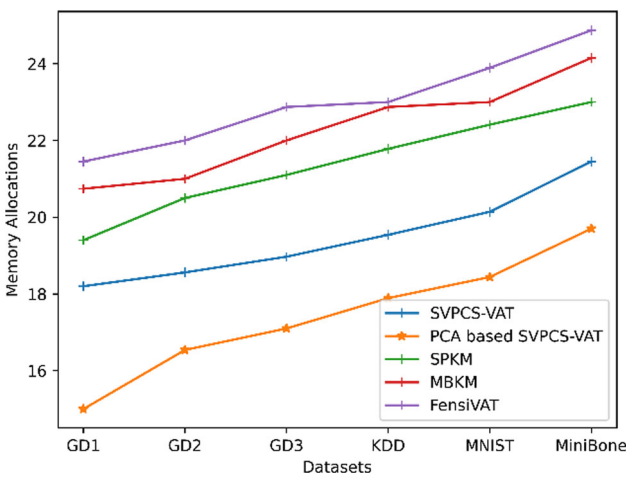


Fig. 16 Memory allocation—PCA based SVPCS-VAT to other methods

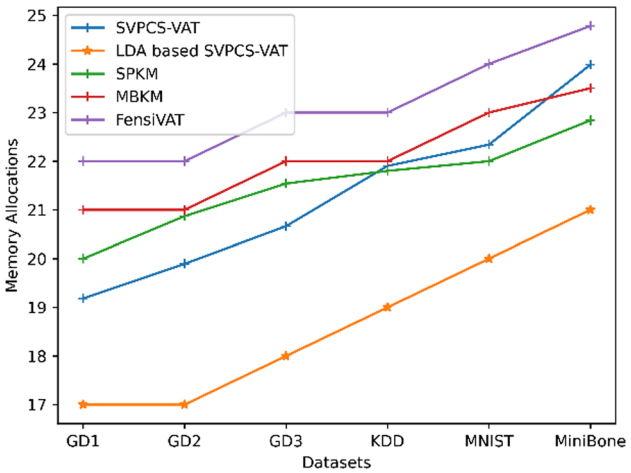


Fig. 17 Memory allocation—LDA based SVPCS-VAT to other methods

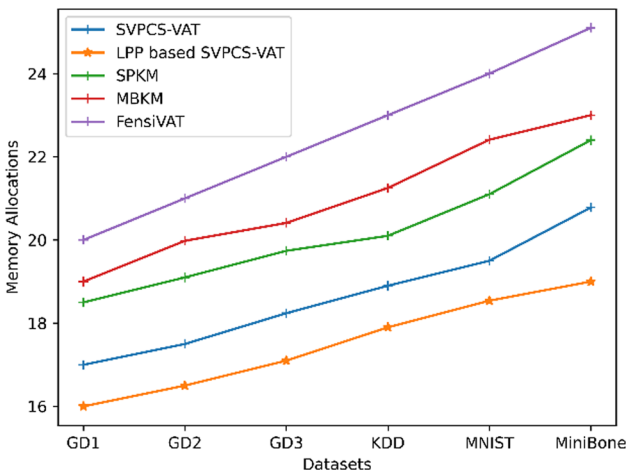


Fig. 18 Memory allocation—LPP based SVPCS-VAT to other methods

extends the proposed visual models for high-dimensional stream data clustering with scalable concepts.

Author Contributions MSB and SKM contributed toward designing hybrid visual computing models. MSB has collected the related study data of visual techniques for clusters assessment problems. KRP carried out data analysis and interpretation of clustering results analysis with indicate measures. He performed the critical investigations of the work in the experimental. MSB wrote the paper with the advice of other authors, and SM took the revision for the quality of the paper. MSB, SKM, KRP: Conceptualization; MSB, SKM: Data curation; MSB, SKM, KRP: Formal analysis; KRP: Funding acquisition, Funding—"Science and Engineering Research Board (SERB)" – Grant of DST (Department of Science and Technology), Government of India, Sanctioned File Number-ECR/2016/001556 MSB, SKM, KRP: Investigation; MSB, SKM, KRP: Three New Methods are developed they are, PCA-based SVPCS-VAT, LDA-based SVPCS-VAT, and LPP-based SVPCS-VAT; SKM: Project administration; MSB, RP: Resources; SK, KRP: Supervision; SB: Visualization; MSB, KRP: Writing—original draft; SKM: Writing—review and editing.

Funding There is No funding support for this work.

Data availability Enquiries about data availability should be directed to the authors.

Declarations

Conflict of interest There is No conflict of interest from any side.

References

Achlioptas D (2001) Database-friendly random projections. In: Proceedings of the twentieth ACM SIGMOD-SIGACT-SIGART symposium on principles of database systems, pp 274–281

Alessia Amelio, Clara Pizzuti (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

Asuncion A, Newman D (2007) Uci machine learning repository

Belkin M, Niyogi P (2008) Towards a theoretical foundation for Laplacian-based manifold methods. J Comput Syst Sci 74(8):1289–1308

Bezdek J (1981) Pattern recognition with objective function algorithms. Plenum, New York, NY, USA

Bezdek JL (2008) SpecVAT: enhanced visual cluster analysis. IEEE international conference on data mining, ICDM

Bezdek JC, Hathaway RJ (2002) VAT: a tool for visual assessment of (cluster) tendency. In Proceedings of. 2002 international joint conference on neural networks, Honolulu, HI, 2002, 2225–2230

Bhatnagar V, Majhi R, Jena PR (2018) Comparative performance evaluation of clustering algorithms for grouping manufacturing frms. Arab J Sci Eng 43:4071–4083

Blei DM, Ng AY, Jordan MI (2003) Latent Dirichlet allocation. J Mach Learn Res 3:993–1022

Bradley PS, Fayyad UM, Reina C et al (1998) Scaling clustering algorithms to large databases. KDD, pp 9–15

Deepak V, Khanna MR, Dhanasekaran K, Prakash PGO, Babu DV (2021) An efficient performance analysis using collaborative recommendation system on big data. In: 2021 5th international conference on trends in electronics and informatics (ICOEI), pp 1386–1392. https://doi.org/10.1109/ICOEI51242.2021.9452737

Duda RO, Hart PE, Stork DG (2001) Pattern classification. Wiley, New York

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

Havens TC, Bezdek JC, Keller JM, Popescu M, Huband JM (2009) Is VAT really single linkage in disguise? Ann Math Artif Intell 55(3–4):237–251

Hore P, Hall L, Goldgof D (2007) Single pass fuzzy C means. In: Proceedings of IEEE international Fuzzy system conference, London, UK, pp 1–7

Hu Y, John A, Wang F, Kambhampati S (2012) Et-LDA: joint topic modelling for aligning events and their twitter feedback. In: AAAI conference on artificial intelligence (AAAI 2012), Vol 12, Toronto, Ontario, Canada, pp 59–65

Jiang D, Tang C, Zhang A (2004) Cluster analysis for gene expression data: a survey. IEEE Trans Knowl Data Eng 16(11):1370–1386

Xudong Jiang, Linear Subspace learning based dimensionality reduction, IEEE Signal Processing Magazine, 2011

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. https://doi.org/10.1109/TCYB.2015.2477416

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, CA, pp 112–117. https://doi.org/10.1109/BigData.2013.6691561

Pattanodom et al. (2016) Clustering data with the presence of missing values by ensemble approach. In: Second Asian conference on defense technology

LeCun Y, Cortes C, Burges CJ (1998) The mnist dataset of handwritten digits. http://yann.lecun.com/exdb/mnist

Rajendra Prasad K, Reddy BE, Mohammed M (2021) An effective assessment of cluster tendency through sampling based multi-viewpoints visual method. J Amb Intell Human Comput. https://doi.org/10.1007/s12652-020-02710-8

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)

Rajendra Prasad K, Mohammed M, Noorullah RM (2019) Visual topic models for healthcare data clustering. Evol Intell

Ramathilagam S, Devi R, Kannan SR (2013) Extended fuzzy c-means: an analyzing data clustering problems. Cluster Comput

Rathore P, Kumar D, Bezdek JC, Rajasegarar S, Palaniswami M (2019) A rapid hybrid clustering algorithm for large volumes of high dimensional data. In: IEEE transactions on knowledge and data engineering 31(4): 641–654. https://doi.org/10.1109/TKDE.2018.2842191

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

Sculley D (2010) Web-scale k-means clustering. In: Proceedings of the 19th international conference on world wide web. ACM, pp 1177–1178

Subba Reddy K, Rajendra Prasad K, Kamatam GR et al (2022) An extended visual methods to perform data cluster assessment in distributed data systems. J Supercomput. https://doi.org/10.1007/s11227-021-04243-z

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. https://doi.org/10.1007/s11227-021-03618-6

Suleman Basha M, Mouleeswaran SK, Rajendra Prasad K (2019) Cluster tendency methods for visualizing the data partitions. Int J Innov Technol Explor Eng

Tavallaee M, Bagheri E, Lu W, Ghorbani A (2009) A detailed analysis of the KDD'99 CUP data set. In: Proceedings of 2nd IEEE symposium on computer intelligence conference on security defense applications (CISDA), Vol 40, Ottawa, ON, Canada, pp 44–47

Urruty T, Djeraba C, Simovici DA (2007) Clustering by random projections. In: Industrial conference on data mining. Springer, pp 107–119

Vidal R, Ma Y, Sastry S (2005) Generalized principal component analysis (GPCA). IEEE Trans Pattern Anal Machine Intell 27(12):1945–1959

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

Yang Y, Ma Z, Yang Y, Nie F, Shen HT (2015) Multitask spectral clustering by exploring intertask correlation. IEEE Trans Cybern 45(5):1069–1080