ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

LINEAR SUBSPAC31E LEARNING-BASED CLUSTER TENDENCY ASSESSMENT VISUAL MODELS FOR HIGH-DIMENSIONAL BIG DATASETS

ASWANI KUMAR UNNAM[1], DR BANDLA SRINIVASA RAO[2]

[1]Research Scholar, Department of CSE, Acharya Nagarjuna University, Guntur, India

[2]Professor, Department of CSE, Teegala Krishna Reddy Engineering College, Hyderabad, India

E-mail: [1]askphy@gmail.com, [2]sreenibandla@gmail.com

ABSTRACT

Many new applications, including traffic image trajectories and video surveillance, require big data clustering. These applications create enormous volumes of high-dimensional data by utilizing sensors or the Internet of Things (IoT). Traditional big-data clustering techniques, such as single-pass k-implies (spkm), scaled down clump k-implies (mbkm) are broadly used to make an information segment over the enormous information. To decide the nature of the bunches covered by the huge information, they should, notwithstanding, have advance information on the group assessment. The convenience of bunching inclination for gigantic information is made conceivable by the as of late evolved examining based multi-perspectives based cosine measure visual evaluation of (group) propensity (S-MVCM-Tank). For high-layered huge information, hybrid big data clustering visual models are proposed in this study. These models address the curse of dimensionality and determine the quality of data clusters by utilizing S-MVCM-VAT and linear subspace learning (LSL). The purpose of the experimental investigation is to demonstrate how well the suggested LSL-based S-MVCM-VAT approaches perform in comparison to alternative large data clustering strategies.

Keywords: *Big Data Clustering, Clustering Tendency, Linear Subspace Learning, Multiview Points, Dimensionality Problem*

1. INTRODUCTION

Numerous large data applications make for clustering large amounts of data, use of single-pass k-means (spkm) [1], mini-batch k-means [2], and other cutting-edge techniques [3] The evaluation of clusters presents a significant challenge when dealing with enormous amounts of data.. To evaluate the quality of the clusters over the vast volumes of data, they must first comprehend the cluster evaluation. Data partitioning (or groups) is an issue that may be effectively solved with the help of cluster analysis [4], which divides data items into groups according to shared criteria. The similarities between different data objects can be computed by utilizing a variety of distance metrics [5]. For spkm and mbkm, the user may be attempting an inaccessible 'k' worth (or group propensity), which can bring about unfortunate bunching results. It tends to be brought about by outer impedance. For instance, even when employing k-means, it is only occasionally possible to settle on a single 'k' number to use. It was discovered after a thorough analysis of the literature that visual models, or visual assessment of (cluster) tendency (VAT) [6], [7], effectively ascertain the clustering tendency for datasets without labels. As a result, this VAT model produces cluster tendency and good clustering results for unlabeled datasets. Recently, advanced visual models, such as ClusiVAT, and other models [8], have been implemented to assess cluster tendency and discover the quality of data clusters over big datasets. However, these techniques need help to handle the curse of dimensionality problems in high-dimensional big data. FensiVAT[9], [10] handles this issue with random projections; linear subspace learning is the best alternative to random projections

Journal of Theoretical and Applied Information Technology
31st May 2024. Vol.102. No. 10
© Little Lion Scientific
ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

in FensiVAT. Therefore, the hybrid large data clustering models proposed in this research (S-MVCM-VAT) combine sampling-based MVCM-VAT. Head part examination (PCA) [11], straight discriminant investigation (LDA) [12], and territory safeguarding projection (LPP) [13] are the four techniques utilized in LSL. Subsequently, proposed half breed huge information bunching visual models are determined with these three variations of LSL strategies: PCA-based-S-MVCM-Tank, LDA-based-S-MVCM-Tank, and LPP-based-S-MVCM-Tank. Fig. The steps of the proposed work are shown in Figure 1.

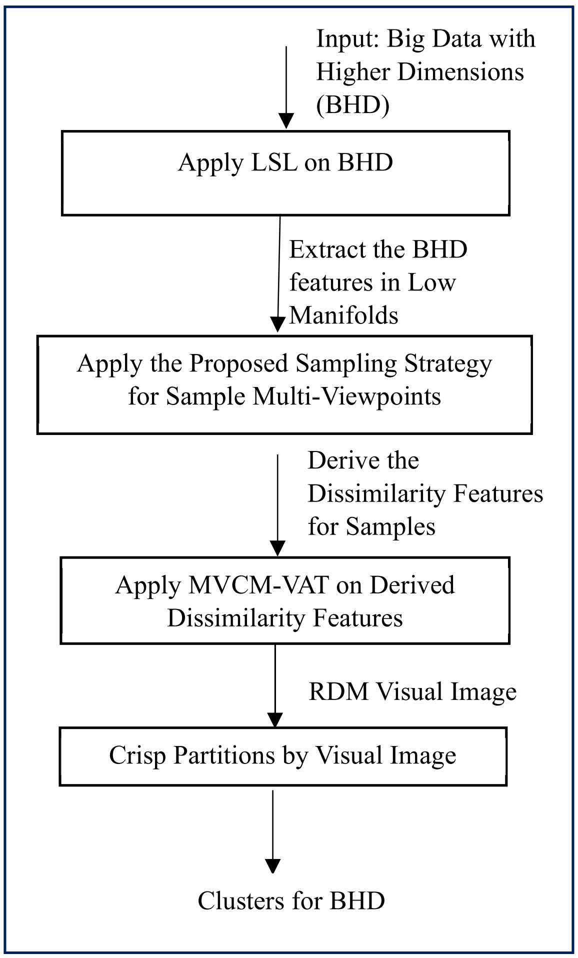

Fig. 1 Framework of the Proposed Technique

The proposed methodology really impresses when it comes to the bunching of high-layered information over a major dataset. The LSL and the min-max techniques are the principal ones utilized here to get the best example of perspectives. One benefit of utilizing LSL is that it makes it conceivable to get the low-layered manifolds of the gigantic information that was initially utilized. The inter-cluster sample points of view are the ones that are chosen for the sampling method that is later used for low-dimensional manifolds with a lot of data. The cosine similarity metric, which is based on how people in different clusters see things, is the most accurate way to measure how similar things are. The low-layered, enormous information is used to fabricate the divergence grid, which can then work as a contribution for the Tank. The Visual Examination Instrument (Tank) shows the outcomes as dull, square-formed blocks with visual bunches within them. Prior to building information groups, it requirements to track down clear divisions to get the projected bunch names for the information objects in enormous datasets with many aspects. The LDA-based MVCM-Tank, the PCA-based MVCM-Tank, and the LPP-based MVCM-Tank are the three proposed visual processing models examined in this work. The "scourge of dimensionality" and enormous dataset sizes are addressed by these methods. They are in this way definitely more powerful than other enormous information grouping strategies that are respected as state-of-the-art.

2. LITERATURE STUDY

In addition, it investigated the possibilities of using massive amounts of data. The "dimensionality curse" may be seen in many aspects of big data. This issue can be addressed by employing an additional cutting-edge visual model known as Fensi-VAT[13]. Both the cosine-based variance analysis technique (cVAT) [14] and the cosine-based spectral variance analysis technique (cSpecVAT) [15] are made to get exact information about clusters. These algorithms can determine what different data objects have in common when viewed from a single angle. Greater understanding may be gained from an analysis that employs cosine distance as a measure of similarity between many points of view than from an analysis that just utilizes one. It's possible that this method of analysing clusters in big data won't be cost-effective. Multi perspectives based cosine similitude Tank is the name of the methodology that was developed and established in [16], [17] for the more appropriate cluster assessment (MVCM-VAT) [17]. Using random projections reduces the high dimensions to a three-dimensional subspace that is easier to work.

When it comes to grouping big volumes of data, Festival works better than both the spkm and Mini-batch-k-means techniques. A well-established technique for clustering massive volumes of high-layered information is the FensiVAT. The procedure uses random projections to make the data less dimensional. LSL approaches [18] are utilized to pick the irregular projections and Eigenvectors that give the best outcomes while decreasing a high-layered space to subspaces. Direct subspace learning methods are utilized to track down the best subspace from high-layered information. As candidates, the largest or best Eigenvectors are chosen. (LSL). Instead of random projection, linear subspace learning techniques are used to construct the most effective low-dimensional manifolds.

One of the LSL approaches is principle component analysis (PCA) [19], whose output may be shown on a low-dimensional principal axis. Maintaining an adequate degree of class separation (or clusters) while converting high-layered information into a low-layered space is the objective of the regulated technique called linear discriminant analysis (LDA) [20]. All data that can be utilized to distinguish between classes is securely saved during the clustering process. An alternative method to principal component analysis (LPP) is to use locality-preserving projections [21]. High-dimensional object data maximizes the dimensionality space and yields the biggest variances in the projection vectors by utilizing the neighbourhood design of the information related to direct projections.

When using LPP, the grouping of the data points is preserved when creating the Laplacian matrix. This topic lends itself quite nicely to the linear subspace learning approach for tackling the "curse of dimensionality" problem. Regarding the production of ideal subspaces, the LSL approaches are preferable to random projection. In the field of research on dimensionality reduction, there has been much progress. These methods are purposely utilized to produce low-layered complex subspaces for high-layered information sets.

3. METHOD

The S-MVCM-VAT model is enhanced using LSL approaches to overcome the "dimensionality curse" problem in the proposed models. The three different LSL implementations are implemented in the three hybrid visual computing models (PCA, LDA, and LPP). Algorithm 1 presents models for hybrid big data clustering visual models.

It first invokes the LSL algorithm with value and BHD input parameters. First, read the BHD with a size of m x n; The number of data objects and dimensions, respectively, are denoted by m and n. Here, type demonstrates the sort of LSL technique utilized on the BHD to get the information's diminished dimensionality and store it in a LM, then involves LM as the contribution for the as of late evolved visual methodology S-MVCM-Tank.

The showed dim hued blocks are gotten from the S-MVCM-Tank picture and are viewed as along its inclining. These askew seen dull hued blocks give the premise of the fresh parcels. It is easy to peruse the bunch names of the data items after acquiring the crisp partitions. Finally, the proposed technique yields high-dimensional large data clustering results effectively.

Algorithm 5.1 : Proposed Hybrid Big Data Clustering Visual Model

Input : BHD-Big High-Dimensional Data

Output : C- Clusters Generation

Methodology :

// Identify the low-dimensional manifolds (LM) in the huge data with high dimensions.

1. *Read the "m" number of data objects and the "n" number of high dimensions in the "high-dimensional big data," or BHD.*
2. *Find LM by call the LSL function, LSL(BHD, type)*

// Create the precise partitions for the massive data clusters.

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

3. *Call S-MVCM-VAT to find the visual cluster images.*
4. *From visual representations or images, get the sharp partitions (known as the crisp partitions)*
5. *To find high-dimensional large data clusters, create cluster labels for data objects by derived crisp partitions in step 4. Generate the clusters 'C.'*

Function LSL(BHD, type)

{

If (type==1) // It is the procedure of PCA

{

i. *Ensure BHD uniformity*
ii. *Create the BHD covariance matrix.*
iii. *Finding the greatest k-eigenvectors requires using the Eigen decomposition method.*
iv. *Using the principal components derived from the greatest k-eigenvectors, determine the low-dimensional manifolds (LM) of BHD.*
v. *Return(LM)*

}

If (type==2) // LDA Procedure

{

i. *Creating n-dimensional mean vectors for data objects*
ii. *Find the Sw and Sb scatter matrices (called within-class scatter matrix and between the class scatter matrix, respectively)*
iii. *Find the $S_W^{-1}S_b$ for solving the Eigen decomposition problem*
iv. *The eigenvectors are sorted by decreasing eigenvalues*
v. *Select the largest k-eigenvectors when mapping n-dimensional data into a fold with a low dimension and save them in the 'LM' format.*
vi. *Return (LM)*

}

If (type==3) // LPP Procedure

{

i. *Create the neighborhood-based adjacency graph.*
ii. *Selecting the weights with the help of the heat kernel and figuring out the weighted matrix (W)*
iii. *Use a diagonal matrix to calculate the Laplacian matrix (L).*
iv. *Based on the L-ordered eigenvalues, determine the eigenvectors.*
v. *According to the chosen 'k' number of Eigenvectors and kn, choose the low-dimensional manifolds (LM)*
vi. *Return(LM)*

}

}

The three LSL models' phases and the decreased dimensions of the original high-dimensional large data are displayed in Algorithm 1. After the BHD has been standardised using the min-max normalisation approach, the covariance matrix is then built in PCA. The covariance matrix input is used as the input for the Laplacian matrix, where k is expected to be or to reflect the decreased number of dimensions "k," to find higher k- Eigenvectors. In LDA, the scattered structures Sw and Sb are developed including the n-layered mean vectors of the data for the article.

Then, at that point, as PCA in LDA, where k is the amount of reduced viewpoints, is the size of the BHD's diminished dimensionality. The weighted matrix W is used in LPP to calculate the Laplacian matrix "L". The adjacency network used to construct the W takes into account the affinities of surrounding data elements.

4. RESULTS AND DISCUSSION

Various For clustering, large, high-dimensional datasets are used. assessment to test existing and new methods. The large, high-

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

dimensional datasets are displayed in Table 1. Massive gaussian data in many dimensions are produced with MATLAB (details in Table 1). Different kinds of synthetic Gaussian data are made to illustrate empirical analysis. Three important real-time datasets—KDD CUP'99, MiniBooNE, and MNIST—are used to assess cluster tendency in order to demonstrate the efficacy of the suggested hybrid big data clustering visual model [22].

Table 1. Description of High-Dimensional Big Datasets

S.No	Gaussian / Real Data	No. of Data Objects	Total Dimensions
1	Gaussian Synthetic Data with Clusters=2	“80000”	“50”
2	Gaussian Synthetic Data with Clusters=3	“100000”	“100”
3	Gaussian Synthetic Data with Clusters=6	“120000”	“500”
4	Real time “KDD CUP’99”	“4898431”	“18”
5	Real time “MiniBooNE (k=2)”	“130064”	“50”
6	Real time “MNIST”	“70000”	“784”

The Fensi-VAT uses random projections in order to reduce dimensionality. Due to random projection mappings, the low-dimensional manifolds can deal with the dimensionality curse. Although it is faster than S-MVCM-VAT, high-dimensional datasets might not be a good fit. As a result, compared to S-MVCM-VAT, FeniVAT offers reduced.

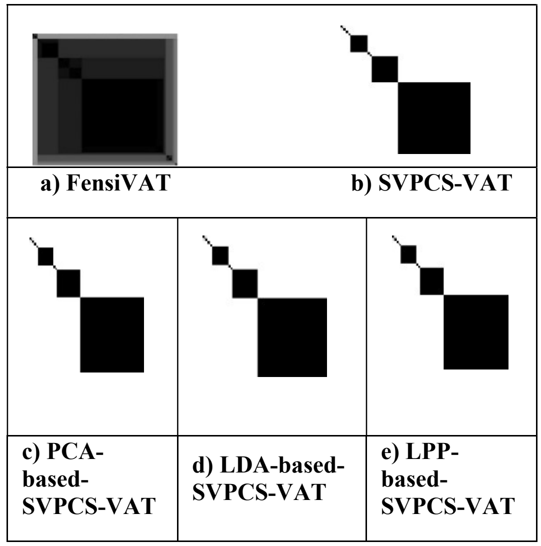

Fig. 2 Visual Clusters for KDDCUP’99

The experimental findings for the high-dimensional large data KDDCUP'99 are displayed in Fig. 2. The suggested LSL-based S-MVCM-VAT was found to perform better than the others. The KDD CUP'99 datasets, which included one major bunch, two middle groups, and six minuscule groups, were represented visually by three proposed models. It is difficult to count visual clusters as Fensi-VAT displays their overlaps. For the KDD CUP'99, a suggested method that is now in use offers visual clusters with excellent visual pictures.

Table 2: Partition Accuracy (PA) Analysis

Synthetic/ Real Data	Mini Batch k-means	spkm	Fensi VAT	MVCM-VAT	PCA-based-MVCM-VAT	LDA-based-MVCM-VAT	LPP-based-MVCM-VAT
	Partition Accuracy (PA)						
Gaussian Data with Clusters=2	0.23	0.26	0.26	0.32	1.00	1.00	1.00
Gaussian Data with Clusters=3	0.25	0.25	0.33	0.36	1.00	1.00	1.00
Gaussian Data with	0.21	0.22	0.29	0.34	1.00	1.00	1.00

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

Cluster s=6							
Real "KDD CUP'99"	0.33	0.15	0.50	0.52	0.59	0.83	0.83
Real "MNIST"	0.21	0.25	0.28	0.34	0.55	0.56	0.56
Real "MiniBooNE"	0.23	0.26	0.31	0.35	0.58	0.58	0.59

Table 3: Normalized Mutual Information (NMI) Analysis

Synthetic/ Real Data	Mini Batch k-means	spkm	Fensi VAT	MVCM-VAT	PCA-based-MVCM-VAT	LDA-based-MVCM-VAT	LPP-based-MVCM-VAT
Normalized Mutual Information (NMI)							
Gaussian Data with Clusters=2	0.22	0.24	0.26	0.32	1.00	1.00	1.00
Gaussian Data with Clusters=3	0.22	0.25	0.33	0.36	1.00	1.00	1.00
Gaussian Data with Clusters=6	0.22	0.25	0.28	0.35	1.00	1.00	1.00
Real "KDD CUP'99"	0.11	0.15	0.27	0.32	0.44	0.45	0.44
Real "MNIST"	0.23	0.23	0.25	0.35	0.44	0.46	0.46
Real "MiniBooNE"	0.22	0.17	0.13	0.27	0.43	0.44	0.45

The partition accuracy (PA) [23] and normalized mutual information [24] performance figures for the current and suggested approaches are displayed in Tables 2 and 3. According to this experimental assessment, when compared to other large data clustering methods already in use, the suggested LSL-based-MVCM-VAT performed the best.

5. CONCLUSION

There are several benefits to using the suggested visual models for high-layered enormous information bunching. The problem of cluster tendency in high-dimensional datasets is dealt with in a way that is in line with the most recent visual modeling methods. Utilizing LSL approaches, the three cross breed huge information bunching visual models that have been recommended recognize solid low-layered manifolds of high-layered expanded information. For high-dimensional datasets, these methods effectively analyze data clustering tendencies and identify the best clustering results. In contrast with other large information bunching procedures, proposed crossover visual figuring models obtain high accuracy rates for large gaussian synthetic datasets and an improvement in an accuracy rate of 10% to 30% for large real high-dimensional datasets

REFERENCES

[1] 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

[2] 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.

[3] M. A. Mahdi, K. M. Hosny and I. Elhenawy, "Scalable Clustering Algorithms for Big Data: A Review," in *IEEE Access*, vol. 9, pp. 80015-80027, 2021, doi: 10.1109/ACCESS.2021.3084057.

[4] A. Trisovic, "Cluster Analysis of Open Research Data and a Case for Replication Metadata," *2022 IEEE 18th International Conference on e-Science (e-Science)*, Salt Lake City, UT, USA, 2022, pp. 423-424, doi: 10.1109/eScience55777.2022.00069.

[5] M. Raeisi and A. B. Sesay, "A Distance Metric for Uneven Clusters of Unsupervised K-Means Clustering Algorithm," in *IEEE Access*, vol. 10, pp. 86286-86297, 2022, doi: 10.1109/ACCESS.2022.3198992.

[6] D. Kumar and J. C. Bezdek, "Visual Approaches for Exploratory Data Analysis: A Survey of the Visual Assessment of Clustering Tendency (VAT) Family of Algorithms," in *IEEE Systems, Man, and Cybernetics Magazine*, vol. 6, no. 2, pp. 10-48, April 2020, doi: 10.1109/MSMC.2019.2961163.

[7] Prasad, K.R., Reddy, B.E. & Mohammed, M. An effective assessment of cluster tendency through sampling based multi-viewpoints visual method. *J Ambient Intell Human Comput* (2021). https://doi.org/10.1007/s12652-020-02710-8

[8] D. Kumar, M. Palaniswami, S. Rajasegarar, C. Leckie, J. C. Bezdek and T. C. Havens, "clusiVAT: A mixed visual/numerical clustering algorithm for big data," 2013 IEEE International Conference on Big Data, Silicon Valley, CA, USA, 2013, pp. 112-117, doi: 10.1109/BigData.2013.6691561.

[9] P. Rathore, D. Kumar, S. Rajasegarar, M. Palaniswami and J. C. Bezdek, "A Scalable Framework for Trajectory Prediction," in *IEEE Transactions on Intelligent Transportation Systems*, vol. 20, no. 10, pp. 3860-3874, Oct. 2019, doi: 10.1109/TITS.2019.2899179.

[10] D. Kumar, J. C. Bezdek, M. Palaniswami, S. Rajasegarar, C. Leckie and T. C. Havens, "A Hybrid Approach to Clustering in Big Data," in *IEEE Transactions on Cybernetics*, vol. 46, no. 10, pp. 2372-2385, Oct. 2016, doi: 10.1109/TCYB.2015.2477416.

[11] T. -J. Chin and D. Suter, "Incremental Kernel Principal Component Analysis," in *IEEE Transactions on Image Processing*, vol. 16, no. 6, pp. 1662-1674, June 2007, doi: 10.1109/TIP.2007.896668.

[12] Tianwei Xu, Chong Lu and Wanquan Liu, "The matrix form for weighted linear discriminant analysis and fractional linear discriminant analysis," *2009 International Conference on Machine Learning and Cybernetics*, Hebei, 2009, pp. 1621-1627, doi: 10.1109/ICMLC.2009.5212309.

[13] Yi Jin and Qiu-Qi Ruan, "An image matrix compression based supervised locality preserving projections for face recognition," *2007 International Symposium on Intelligent Signal Processing and Communication Systems*, Xiamen, China, 2007, pp. 738-741, doi: 10.1109/ISPACS.2007.4445993.

[14] P. Rathore, D. Kumar, J. C. Bezdek, S. Rajasegarar and M. Palaniswami, "A Rapid Hybrid Clustering Algorithm for Large Volumes of High Dimensional Data," in *IEEE Transactions on Knowledge and Data Engineering*, vol. 31, no. 4, pp. 641-654, 1 April 2019, doi: 10.1109/TKDE.2018.2842191.

[15] Rajendra Prasad, K., Mohammed, M., Narasimha Prasad, L.V. *et al.* An efficient sampling-based visualization technique for big data clustering with crisp partitions. *Distrib Parallel Databases* 39, 813–832 (2021). https://doi.org/10.1007/s10619-021-07324-3

[16] K. R. Prasad and B. E. Reddy, "Assessment of clustering tendency through progressive random sampling and graph-based clustering results," *2013 3rd IEEE International Advance Computing Conference (IACC)*, Ghaziabad, India, 2013, pp. 726-731, doi: 10.1109/IAdCC.2013.6514316.

[17] Subba Reddy, K., Rajendra Prasad, K., Kamatam, G.R. *et al.* An extended visual methods to perform data cluster assessment in distributed data systems. *J Supercomput* 78, 8810–8829 (2022). https://doi.org/10.1007/s11227-021-04243-z

[18] Yan, J., Cheng, Q., Yang, Q., Zhang, B. (2005). An Incremental Subspace Learning Algorithm to Categorize Large Scale Text Data. In: Zhang, Y., Tanaka, K., Yu, J.X., Wang, S., Li, M. (eds) Web Technologies Research and Development - APWeb 2005. APWeb 2005. Lecture Notes in Computer Science, vol 3399. Springer, Berlin, Heidelberg. https://doi.org/10.1007/978-3-540-31849-1_7

[19] R. He, B. -G. Hu, W. -S. Zheng and X. -W. Kong, "Robust Principal Component Analysis Based on Maximum Correntropy Criterion," in *IEEE Transactions on Image Processing*, vol. 20, no. 6, pp. 1485-1494, June 2011, doi: 10.1109/TIP.2010.2103949.

[20] Yurong Li, Guobo Xiang and Wei Xu, "A data analysis algorithm based on statistical filtration and linear discriminant analysis," *2006 6th*

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

World Congress on Intelligent Control and Automation, Dalian, 2006, pp. 4348-4351, doi: 10.1109/WCICA.2006.1713197.

[21] Lei Zhu and Lihua Li, "Gabor-based Discriminant Locality Preserving Projections for face recognition," *2008 IEEE Conference on Cybernetics and Intelligent Systems*, Chengdu, China, 2008, pp. 375-378, doi: 10.1109/ICCIS.2008.4670906.

[22] R. Yao, C. Liu, L. Zhang and P. Peng, "Unsupervised Anomaly Detection Using Variational Auto-Encoder based Feature Extraction," *2019 IEEE International Conference on Prognostics and Health Management (ICPHM)*, San Francisco, CA, USA, 2019, pp. 1-7, doi: 10.1109/ICPHM.2019.8819434.

[23] T. Limungkura and P. Vateekul, "Enhance accuracy of partition-based overlapping clustering by exploiting benefit of distances between clusters," *2016 Eighth International Conference on Knowledge and Systems Engineering (KSE)*, Hanoi, Vietnam, 2016, pp. 109-114, doi: 10.1109/KSE.2016.7758038.

[24] A K Unnam, Dr B.S Rao, “An Enhanced Sampling-Based Viewpoints Cosine Visual Model for an Efficient Big Data Clustering”, *International Journal on Recent and Innovation Trends in Computing and Communication,* 2023, pp. 3445-3452, https://doi.org/10.17762/ijritcc.v11i9.9553

[25] M. N. Nisha, S. Mohanavalli and R. Swathika, "Improving the quality of clustering using cluster ensembles," 2013 IEEE Conference on Information & Communication Technologies, Thuckalay, India, 2013, pp. 88-92, doi: 10.1109/CICT.2013.6558068.

[26] Aswani Kumar Unnam, Dr B. Srinivasa Rao, “An Extended Clusters Assessment Method with the Multi-Viewpoints for Effective Visualization of Data Partitions”, International Journal of Intelligent Systems And Applications in Engineering (IJISAE), 2023, 11(1s), pp. 51–56.