

Finding Prominent Clusters in VAT Family of Algorithms

Kartik Vishal Deshpande[1,2(✉)], Osmar R. Zaiane[1], and Dheeraj Kumar[2]

[1] University of Alberta, Edmonton T6G2R3, AB, Canada
{kvdeshpa,zaiane}@ualberta.ca
[2] Indian Institute of Technology, Roorkee 247667, Uttarakhand, India
{kvishaldeshpande,dheeraj.kumar}@ece.iitr.ac.in

Abstract. Identifying the prominent clusters is a vital factor that significantly impacts the accuracy and interpretability of the clustering results. Our proposed algorithm leverages MST-based VAT algorithms to detect and quantify distinct clusters within a given dataset, offering a fresh perspective in tackling this challenge. The visual analysis performed in this study on six varied real-world datasets reveals that the proposed algorithm consistently outperforms existing approaches, aVAT and ML-aVAT, making it a dependable tool for a wide range of data analysis tasks. We also compare the runtime efficiency and Dunn's Index values of all three algorithms to highlight the advantages of our proposed method. This novel approach has the potential to enhance clustering processes by minimizing the need for manual adjustments and improving the quality of clustering results across multiple domains.

Keywords: MST-iVAT · Clustering · VAT · Dark blocks · Prominent

1 Introduction

Clustering, an unsupervised data mining approach, is crucial for identifying patterns, enhancing decision-making, and delivering actionable insights. Clusters are groups of data points characterized by similarities in features. A significant challenge in clustering is determining the prominent clusters, as this choice has a profound influence on the quality and clarity of the findings. It is crucial to understand that the decision regarding the ideal number of clusters is not always easy, as it depends on specific objectives and limitations, and may necessitate a blend of quantitative analysis, expert knowledge, and subjective judgment.

Cluster analysis encompasses three fundamental aspects: 1) Clustering tendency—Are clusters present? If so, how many are there? 2) Partitioning—Which items belong to which cluster, and to what extent? 3) Cluster validity—Are the divisions effective? This paper addresses all three aspects. The VAT family of algorithms, as presented by [10], offers excellent tools for visually examining the clustering tendencies of different types of datasets. [1] made an effort to tackle the three aspects of cluster analysis by rearranging the dissimilarity

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2027
C. Ordonez et al. (Eds.): DaWaK 2026, LNCS 16861, pp. 3–17, 2027.
https://doi.org/10.1007/978-3-032-34896-8_1

matrix for simple datasets, so that the number of clusters is ideally indicated by the count of "dark blocks" along the diagonal, leading to the emergence of the VAT family of algorithms. VAT algorithms employ a minimum spanning tree (MST) strategy, yielding notable benefits for clustering. As highlighted in [5] [16], these advantages are: 1) finds the cheapest way to build a network that ensures connectivity, 2) MST is used for single-linkage clustering, and 3) spanning tree protocol prevents loops that can plan the most efficient backbone for a network. Real-world datasets tend to be more intricate in terms of finding prominent clusters. Thus, for a more generalized approach, we developed the **x** and **y** coordinates of the **iVAT** (**xyiVAT**) algorithm to find the prominent clusters for a given dataset. **The contributions of this research paper** include: (1) a swift and generalized algorithm to identify the prominent clusters and their coordinates by selecting the longest MST edges as done by [7], (2) visualization of all dark blocks while retaining gray regions in the xyiVAT image to depict clustering tendency, (3) the proposed algorithm highlights significant and minor dark blocks, which aid in identifying outliers, (4) capable to discover any number of clusters for big datasets, (5) can handle complicated iVAT/MST-iVAT images, and (6) experimental results indicate that xyiVAT performs excellently when bench-marked against key competitors regarding visual assessment, Dunn's index (DI), and runtime. We consider various types of real-world datasets for generalization, encompassing large-scale, high-dimensional, trajectory, spatio-temporal, and two challenging datasets, which are essential for comprehending diverse phenomena across multiple fields. The goal of xyiVAT is to determine the prominent clusters along with their coordinates and to identify their presence in the MST-iVAT image by using the longest MST edges.

2 Related Work

The VAT family of algorithms involves clustering using feature vectors, $\mathbf{z}_\mathrm{i} \in \mathbb{R}^\mathrm{p}$ (with $\mathbf{z}_\mathrm{i}$ being a p-dimensional feature vector, where $1 \leq i \leq N$) for the unlabeled dataset $\mathrm{X} = \{\mathrm{x}_1, \ldots, \mathrm{x}_\mathrm{N}\}$ into k clusters, where $1 < k < N$. The data is represented as a dissimilarity matrix, $D_{N\times N} = d_\mathrm{ij} = \|\mathbf{z}_\mathrm{i} - \mathbf{z}_\mathrm{j}\|$, which can utilize any norm on $\mathbb{R}^\mathrm{p}$. Subsequently, a variation of Prim's algorithm is applied to reorder the dissimilarity matrix, resulting in D^*. This D^* employs a graph-theoretic distance transformation to generate the iVAT matrix $D^{*'}$, ultimately producing the iVAT image. The sequence followed is $\mathrm{VAT}(D) = D^* \rightarrow \mathrm{iVAT}\,(D^*) = D^{*'}$. When working with large datasets, VAT/iVAT places limits on systems with restricted memory. The answer to this challenge is TkdT-VAT and MST-iVAT by [7], which employs MST edges to create an MST-iVAT image. **The objective of an iVAT/MST-iVAT image is clustering tendency, rather than to determine the prominent clusters**. Numerous studies focus on identifying the number of significant clusters, overlooking experiments on substantial real-world datasets with more complex iVAT/MST-iVAT images, algorithms with longer runtimes, and tend to disregard smaller clusters when using iVAT/MST-iVAT images, as discussed in this section.

In [21], a cluster counting algorithm is introduced that applies frequency domain correlation and feature recognition, known as the cluster count extraction (CCE) algorithm. In the same year, [24] developed spectral VAT (specVAT) to create reordered dissimilarity matrix images that enhance the contrast between dark blocks, allowing for the visualization of complex cluster configurations using top eigenvalues. Subsequently, assessing specVAT (A-specVAT) selects the best images among specVAT images based on their grayscale histograms, taking into account "within-cluster distances" and "between-cluster distances". The method proceeds with partitioning specVAT (P-specVAT), which utilizes a genetic algorithm to divide the identified block structures while preserving the relationships and separations between clusters, thereby improving the clustering results. To accommodate large-scale datasets, extended specVAT (E-specVAT) efficiently scales while maintaining accuracy in cluster counting. Although this comprehensive study presents ample experiments on real-world datasets, the algorithm occasionally either overestimates or underestimates the number of clusters. [25] proposed the dark block extraction (DBE) algorithm, which employs various image processing and signal processing methods for cluster counting. This approach lacks sufficient experimental validation on large-scale real-world datasets to back its claims. [26] introduced the automatic VAT (aVAT) algorithm, which converts iVAT images into binary images by applying a 'square size' threshold to extract the coordinates of dark blocks. However, this algorithm is limited to datasets featuring clearly defined dark blocks within the iVAT images. [8] presented clustering in ordered dissimilarity data (CLODD), which formulates an objective function that recognizes the block-like characteristics of reordered data. However, this algorithm fails to yield results when VAT does not create distinct dark blocks along the diagonal. [15] proposed ML-aVAT, which features a two-stage classification process to identify cluster boundaries and hierarchies within clusters. This algorithm relies on a model trained with many iVAT/MST-iVAT images to deduce clustering hierarchies across various dataset types. Recent advances in deep clustering have shown strong results on image datasets by combining representation learning with cluster assignment. Methods like SCAN (nearest-neighbour learning followed by refined clustering) [23], Contrastive Clustering (CC) (joint instance- and cluster-level contrastive losses) [14], and DINO (self-supervised Vision Transformers) [3] have set new benchmarks and are widely used for downstream tasks. But these approaches come with heavy requirements: GPU-based training, large-scale pretraining on high-dimensional inputs, reliance on learned embeddings rather than raw distances. According to results related to CCE, DBE, specVAT, CLODD, HaVAT, CC, SCAN, and DINO, these algorithms are excluded as contenders due to their subpar performance or requirement for expensive resources. Therefore, aVAT and ML-aVAT are selected as viable contenders.

3 Proposed Approach

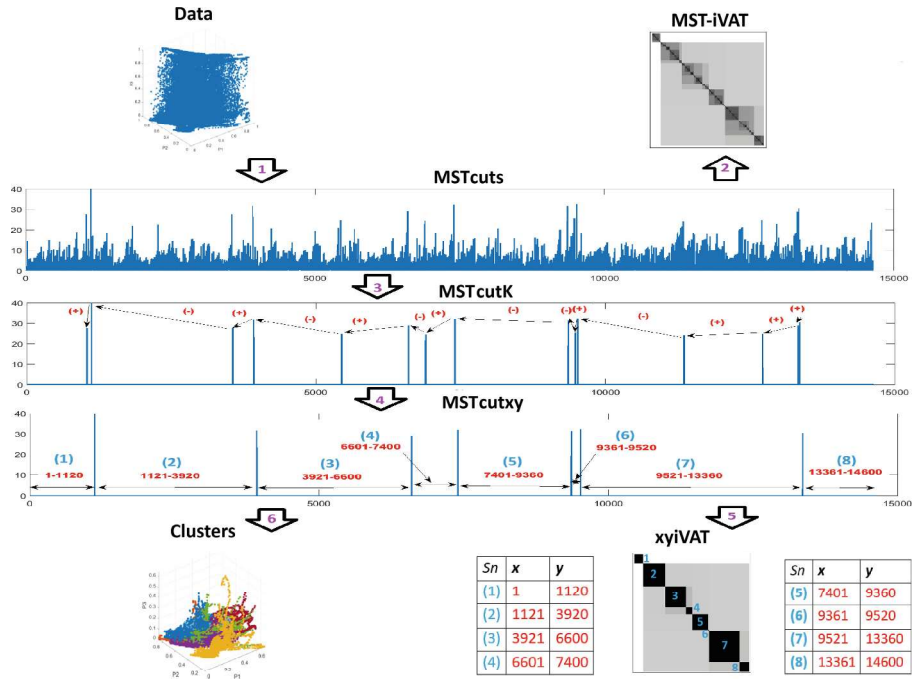


Fig. 1. Four new steps (3 to 6) for clustering in the VAT family of algorithms. 1) MST on the data provides $MSTcuts$. 2) MST-iVAT image formed using $MSTcuts$. 3) K longest $MSTcuts$ are chosen as $MSTcutK$. 4) $MSTcutK$ are processed to get $k-1$ $MSTcutxy$. 5) $MSTcutxy$ used in MST-iVAT to get the xyiVAT image. 6) $MSTcutxy$ is used to provide clusters of the data.

Algorithm 1 outlines the procedures involved in the xyiVAT. Let's denote the MST edges as $MSTcuts$ that result from any VAT family of algorithms. Initially, the K longest $MSTcuts$, along with their corresponding indices, labeled as $MSTcutK$ and $idxcutK$, are identified by locating the maximum values. Then, a zero is concatenated. Steps 2 to 10 serve the purpose of selecting a significant $MSTcutK$. The process begins by calculating the difference between consecutive $MSTcutK$, represented as $diff = MSTcutK_{i+1} - MSTcutK_i$. If the $diff$ is either negative or zero, the index of the current $MSTcutK$, $idxcutK_i$, is stored in $idxcutxy$; otherwise, $idxcutK_i$ is set to zero. This procedure is applied to every $MSTcutK$. Additionally, any non-K $MSTcuts$ are also assigned zeros. In step 12, non-zero and sorted indices of $MSTcutK$, referred to as $idxcutxy$, are identified as $idxy$. The subsequent step involves designating the starting coordinate, x, and the ending coordinate, y. Let's assume there are M entries in $idxy$, with the initial value of $x_1 = 1$. The for-loop in steps 14 to 17 assigns

coordinates such that $x_{i+1} = idxy_i$ and $y_i = idxy_i - 1$, where $1 \leq i \leq M$. In step 18, the ending coordinate is assigned as $y_{M+1} = N$. Furthermore, the indices in $idxy$ are utilized to highlight the notable $MSTcuts$ as $MSTcutxy$. The edges represented by these $MSTcutxy$ are employed in the MST-iVAT algorithm to generate the xyiVAT image. Consequently, $(M + 1)$ pairs of x and y coordinates are utilized to create $k = M + 1$ clusters and determine the coordinates of these clusters. The sizes of each cluster are calculated by taking the differences between the coordinates and adding one to each. Despite the simplicity, efficiency, and straightforward nature of this procedure, K is adjusted until all dark blocks produced by the xyiVAT image align with the darker regions of the MST-iVAT image, which is elaborated in Fig. 1.

Algorithm 1: xyiVAT

Input: $\{MSTcuts, K\}$ - N number of values

1 $[MSTcutK, idxcutK] \leftarrow maxK(MSTcuts_N)$, $MSTcutK = [MSTcutK; 0]_{K+1}$
2 **for** *i = 1:K* **do**
3 $\quad diff \leftarrow MSTcutK_{i+1} - MSTcutK_i$
4 $\quad$ **if** *diff <= 0* **then**
5 $\quad\quad idxcutxy \leftarrow idxcutK_i$
6 $\quad$ **end**
7 $\quad$ **if** *diff is positive* **then**
8 $\quad\quad$ continue
9 $\quad$ **end**
10 **end**
11 Replace other (non-K) $MSTcuts$ with zero.
12 Consider sorted $idxcutxy$ as $idxy$
13 Say M $(< K)$ number of $idxy$, $k = 1$, and $x_k = 1$
14 **for** *k = 1:M* **do**
15 $\quad x_{k+1} \leftarrow idxy_k$
16 $\quad y_k \leftarrow idxy_k - 1$
17 **end**
18 $y_{M+1} \leftarrow N$
19 $MSTcuts_{idxy} \rightarrow MSTcutxy$
20 $MSTcutxy \rightarrow MSTiVAT \rightarrow xyiVAT_{image}$
21 $(x, y)_{1:M+1} \rightarrow k_{clusters}, clusters_{coordinates}$
22 $(y - x)_{1:M+1} + 1 \rightarrow clusters_{size}$

Output: x, y and $k_{clusters}$

A visual step-by-step explanation of Algorithm 1 is depicted in Fig. 1. The xyiVAT begins by examining the $MSTcuts$, which are derived from any of the VAT clustering algorithms on a particular dataset. These $MSTcuts$ are typically employed to generate the MST-iVAT image, as shown in step 2. In step 3, we consider the K longest $MSTcuts$, designated as $MSTcutK$. If the difference between consecutive $MSTcutK$, specifically $MSTcutK_{i+1} - MSTcutK_i$ for $1 \leq i \leq K$, is greater than zero, then $idxcutK_i$ of that $MSTcutK_i$ is replaced with

zero; otherwise, $idxcutK_i$ of $MSTcutK_i$ is collected into $idxcutxy$. Additionally, all non-K $MSTcuts$ are replaced with zero. Thus, by utilizing $idxcutxy$, we can uncover the coordinates (x, y) and the prominent number of clusters (Sn) pertinent to a specific dataset, as seen in step 5. Clusters can also be identified in step 6. After conducting any VAT clustering algorithms, we also derive I, which represents a collection of data points' serial numbers that are connected in the order of MST formation. With these I values, clusters can be obtained as follows: cluster1=I(x_1:y_1), cluster2=I(x_2:y_2), and so forth. Although the choice of K is inherently subjective, prominent clusters are selected when all the dark blocks of the xyiVAT image align with the darker regions of the MST-iVAT image. This manual adjustment of K is preferred since the algorithm is simple and fast. For real-world data, using the K number of longest MST edges with xyiVAT will yield $k-1$ number of the most significant and longest MST edges, resulting in k prominent clusters. Therefore, the **definition** for determining the prominent clusters is: The minimum number of clusters 'k' to get by providing 'K' longest MST edges to xyiVAT, such that the dark blocks in the xyiVAT image coincide with the darker regions in the MST-iVAT image. This addresses clustering tendency and partitioning aspects of cluster analysis.

4 Numerical Experiments

We conduct a comprehensive evaluation of our algorithm's adaptability and generalizability by analyzing real-world datasets. The results from the experiments are systematically presented and analyzed, highlighting both the advantages and the potential limitations of our methods in relation to existing approaches. The experiments were conducted on a system equipped with an Intel Core i7 processor and 16GB of RAM, using MATLAB 2022b.

Relation Between K and k:

It is important to recognize how MST edges influence the clustering tendency in Fig. 1. The relationship between K and k is visually explored in Fig. 2. Smaller values of K focus only on the longer MST edges, which segregate the data into major clusters, while higher values of K navigate through hierarchies of clusters. Either labeled or unlabeled datasets, the first condition to achieve is to vary K, which is a quick and simple process, so that the dark blocks of the xyiVAT image align with the darker regions of the MST-iVAT image. Alignment is considered even if k exceeds the number of labels in labeled datasets. In this case, K should be adjusted until the number of data points of some clusters closely matches the number of data points associated with certain labels. For unlabeled datasets, the minimum K value is chosen until the dark blocks of the xyiVAT image align with the darker regions of the MST-iVAT image. Although Fig. 2 indicates that the number of clusters may reach thousands, this encompasses both significant and minor clusters (outliers), which cannot be extracted by any other algorithm to date.

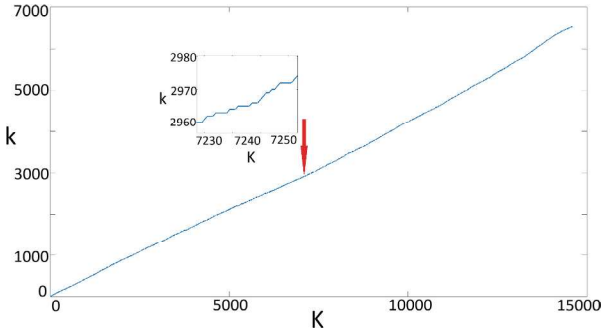


Fig. 2. Relation between K and k with a zoom in the middle (not a straight line).

Computational Complexity:

The aVAT algorithm is based on image processing derived from iVAT images, resulting in longer runtimes. ML-aVAT also has a longer runtime due to the need for training and testing various iVAT images to identify significant clusters within a dataset. The computational complexity of both algorithms is influenced by the iVAT/MST-iVAT images, being particularly useful when there are well-defined dark blocks along the diagonal. Algorithm 1 outlines the xyiVAT method, which operates with only K iterations over K ($<< N$) of the longest MST edges, selecting only $k-1$ ($< K$) of them to create the xyiVAT image. This approach enables relatively quick experimentation with different K values manually, specifically $\mathcal{O}(K)$, until the desired outcomes are achieved, and can process data of any size within a second.

Datasets:

A labeled real-world large-scale dataset, the KDDCup-99 computer network intrusion detection dataset by [22], contains 4,292,637 instances across 41 dimensions. Each instance categorizes the attack type as either normal or one of 22 simulated attack types, resulting in a total of 23 categories. **A labeled real-world high-dimensional dataset**, the MNIST handwritten digit dataset by [13], includes data points represented by numerous features. The MNIST dataset consists of images of handwritten digits divided into 60,000 training images and 10,000 testing images, with each image measuring 28 $\times$ 28 pixels and covering 10 categories representing the numerical digits from 0 to 9. **An unlabeled real-world spatio-temporal dataset**, the STData dataset by [6], features 350,400 hourly temperature recordings spanning a year from 1 April 2002 to 31 March 2003 at two locations with a total of 40 weather stations (Location 1 with 15 weather stations and Location 2 with 25 weather stations). **An unlabeled real-world MIT trajectory dataset** by [27] is employed for analyzing trajectories from a single camera viewpoint. Over a period of five days, 40,453 trajectories

were gathered from a car park scenario. **Two unlabeled real-world challenging datasets**, the MIT-IBRL dataset by [2] and the IIITH dataset by [19], are analyzed due to the resultant MST-iVAT images being complicated. The MIT-IBRL dataset comprises readings from 51 operational sensors that measure temperature, humidity, light intensity, spatial coordinates (x and y), and temporal coordinates (epoch time), with values averaged every 10 min, resulting in 76,500 readings. The IIIT-H dataset comprises readings from 9 active sensors that track temperature, relative humidity, and $PM2.5$ concentration parameters over five months, resulting in 84,800 readings.

4.1 Evaluation

The paper [17] addresses the computational challenge of Dunn's Index (DI), which has a quadratic time complexity, making it impractical for large datasets, and proposes six approximation methods to make it feasible for big data partitions. Among these, four methods are based on maximin sampling, which works by identifying only the boundary points of each cluster to form a skeleton of the full partition, significantly reducing the number of distance computations required without compromising the quality of the index. We choose the αMMSR DI version for evaluation. This makes it compatible with MST-based VAT clustering, which also operates purely on distances without any centroid assumption. Runtime is a critical metric because MST-based VAT clustering is specifically designed for large and complex datasets, so demonstrating that the algorithm scales efficiently proves its practical superiority over centroid-based methods. Visual evaluation through iVAT/MST-iVAT images is uniquely suited for MST-based methods since the MST itself is a visual structure, allowing direct inspection of where edge cuts occurred and whether clusters formed at natural boundaries. Together, these three metrics cover mathematical validity, computational efficiency, and structural correctness. Hence, we assess the effectiveness of aVAT, ML-aVAT, and xyiVAT by examining DI, runtime, and the visual evaluation.

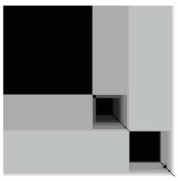
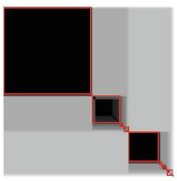
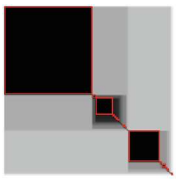
Silhouette Coefficient [4] needs to compute distances between every pair of points in the dataset, which means storing an N×N matrix in memory—for large datasets that MST algorithms are designed to handle, this simply runs out of RAM before even producing a result. Davies-Bouldin Index [4] tries to measure how tight and well-separated clusters are by using their centroids, but MST clusters can be crescent-shaped, chain-like, or ring-shaped, meaning the centroid often lands in empty space and no longer represents the cluster, so the index is essentially measuring distances to a ghost point. Calinski-Harabasz Index [4] was originally built on the assumption that clusters look like round, Gaussian blobs, and it rewards clustering that keeps points tightly packed around a center—but MST is specifically good at finding elongated and irregular clusters, which will always look "spread out" to this index even when the clustering is actually correct. Xie-Beni Index [20] comes from the fuzzy clustering world where every cluster has a well-defined prototype or center point, and it measures quality based on how close points are to that center—but in MST-based clustering,

there is no such prototype since clusters are formed by cutting weak edges in a graph, not by grouping points around a focal point.

The iVAT/MST-iVAT images display all potential hierarchical clusters within a given dataset. In the aVAT image, the diagonal of the iVAT/MST-iVAT image is highlighted with red boxes to indicate the selected clusters. ML-aVAT images utilize red, green, and blue boxes on the diagonal of the iVAT/MST-iVAT image for cluster selection. The xyiVAT algorithm identifies k prominent clusters by selecting the K longest MST edges, and these prominent clusters are represented by dark blocks along the diagonal of the xyiVAT image. For all three algorithms, k denotes a set of prominent clusters. As xyiVAT uses the longest MST edges as inter-cluster edges to obtain the xyiVAT image, indicating that clusters are well-separated, and the best metric to evaluate this is DI. Depending on the separation of the dark blocks along the diagonal and the complexity of the iVAT/MST-iVAT images are the main factors contributing to runtime. All of these factors are tested using six real-world datasets.

Visual Evaluation:

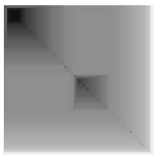
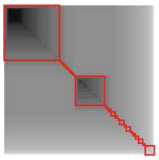
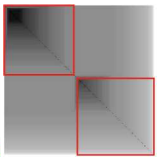
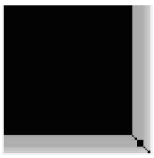
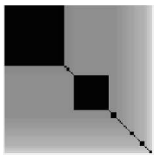
Table 1. (b) aVAT provides 10 red boxes as clusters, (c) ML-aVAT provides 15 red boxes as clusters, and (d) for $K = 65$, xyiVAT provides 23 dark blocks as clusters, same as the number of labels, for KDDCup-99 large-scale dataset.

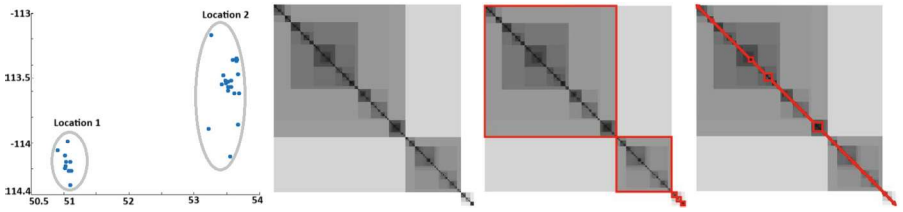
Dataset	MST-iVAT	aVAT	ML-aVAT	xyiVAT
KDDCup-99 4,292,637				
	(a)	(b) k=10	(c) k=15	(d) K=65,k=23

The clusiVAT algorithm by [12] effectively samples large-scale data to yield an MST-iVAT image from which clusters can be extracted and subsequently applied to the remaining unsampled data using the nearest-prototype rule. The KDDCup-99 dataset has 23 labels. As displayed in Table 1, aVAT and ML-aVAT identify fewer dark blocks marked with red boxes ($k < 23$) as clusters. To determine the suitable number of clusters, xyiVAT uses $K = 65$ to get $k = 23$ dark blocks. [18] proposed FensiVAT, which employs quick memory reduction and an effective sampling method to cluster extensive high-dimensional data. Darker regions in the MST-iVAT image shown in Table 2 for the MNIST dataset indicate the presence of clusters, and these regions are not effectively captured in aVAT or ML-aVAT. The selection of the dark blocks highlighted by red boxes in aVAT fails to maximize the inclusion of all points associated with the relevant clusters, and ML-aVAT's red boxes extend too far by including points from other clusters. In xyiVAT, to ensure the number of prominent clusters equals the number

of labels ($k = 10$), we set $K = 80$. But dark blocks in the xyiVAT image do not align with darker regions of the MST-iVAT image. Instead, we opt for $K = 250$ to align with the anticipated darker regions in the MST-iVAT image, resulting in $k = 46$ clusters (2 big, 8 small, and other tiny clusters). A possible reason for $k > 10$ is that the images in the dataset do not resemble digits between 0 and 9. Here, a relatively lighter gray shade in the bottom right of the MST-iVAT image suggests the presence of longer MST edges. For alignment of darker regions in the MST-iVAT image, we choose a higher K value.

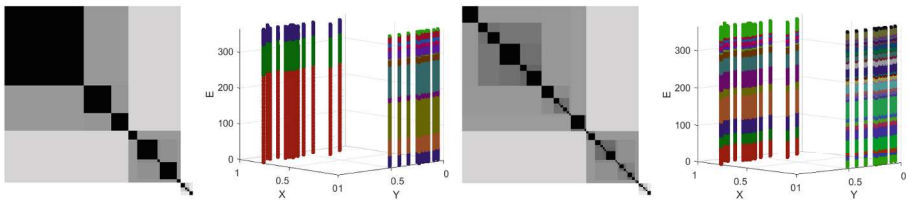
Table 2. (b) aVAT provides 35 red boxes as clusters, (c) ML-aVAT provides 2 red boxes as clusters, and (d) xyiVAT provides 10 dark blocks as clusters for K=80, which is misleading. Based on darker regions in MST-iVAT, (e) for K=250, xyiVAT provides 46 dark blocks as clusters. Two of these dark blocks align with darker regions in MST-iVAT for the MNIST dataset.

Dataset	MST-iVAT	aVAT	ML-aVAT	xyiVAT_1	xyiVAT_2
MNIST 70,000x784					
	(a)	(b) k=35	(c) k=2	(d) K=80, k=10	(e) K=250, k=46



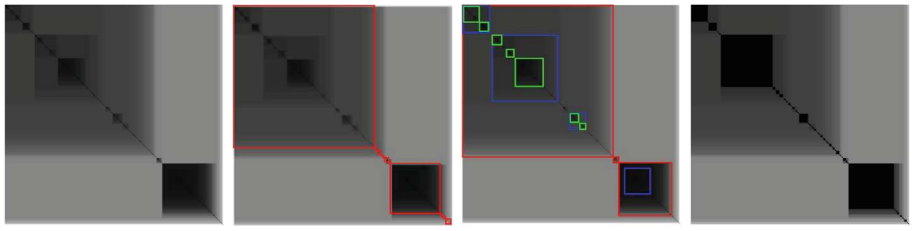
(a) Two locations (b) MST-iVAT (c) aVAT (d) ML-aVAT

Fig. 3. (a) Scatter plot of sensors' locations. (c) provides aVAT with k=5 and (d) provides ML-aVAT with k=93 on the MST-iVAT image for the STData dataset.



(a) K=30, k=14 (b) ST plot (c) K=100, k=49 (d) ST plot

Fig. 4. For the STData dataset, (a, b) and (c, d) pairs provide the xyiVAT image and spatio-temporal data clusters for two values of K.



(a) MST-iVAT (b) aVAT (c) ML-aVAT (d) xyiVAT

Fig. 5. (b) aVAT provides k=15 clusters. (c) ML-aVAT provides k=3 red, 7 green, or 4 blue clusters. (d) xyiVAT with $K = 500$ yields k=98 clusters corresponding to the MST-iVAT image.

Table 3. (a) is the MST-iVAT image and (b, c, d) are the xyiVAT images with k=516, 1746, 4739 number of clusters respectively for the MIT-IBRL dataset. (e) is the MST-iVAT image, and (f, g, h) are the xyiVAT images with k=44, 268, 1045 number of clusters respectively for the IIIT-H dataset.

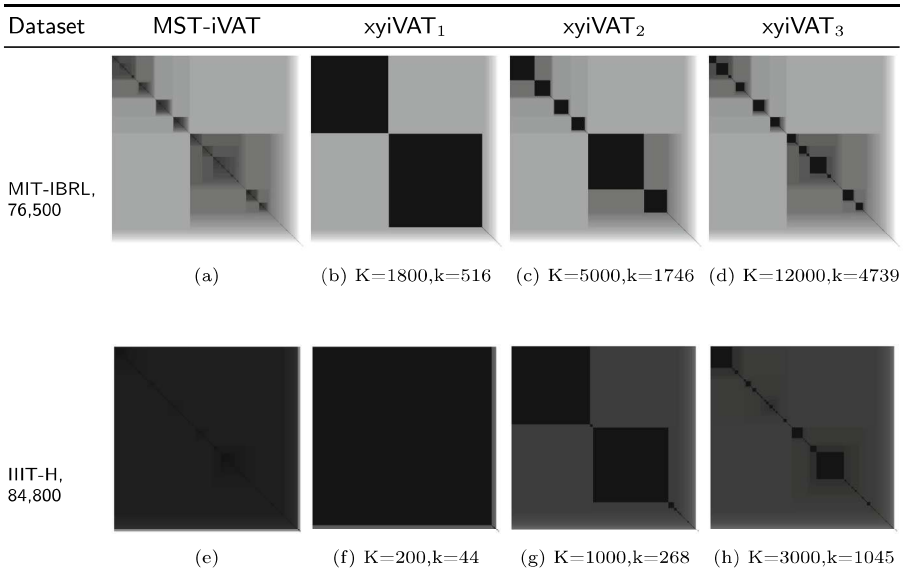


Table 4. Dunn's Indices of six datasets for aVAT, ML-aVAT, and xyiVAT.

Algorithm	KDDCup-99	MNIST	STData	MIT-Trajectory	MIT-IBRL	IIITH
aVAT	3.62e-04	0.95e-07	1.29e-03	0.69e-03	3.36e-06	0.23e-04
MLaVAT	1.26e-03	0.67e-07	0.54e-03	0.46e-04	9.42e-06	6.21e-05
xyiVAT	**0.38**	**6.8e-05**	**0.422**	**5.9e-02**	**6.3e-04**	**0.47e-03**

Figure 3 displays spatial data from weather stations in two locations by [6], as well as the MST-iVAT, aVAT, and ML-aVAT images for the STData dataset. Failure to select proper dark blocks in aVAT and ML-aVAT, as indicated by the red boxes, shows suboptimal performance. Figure 4 shows xyiVAT images for two different K values. Given that the dataset lacks labels to determine the number of clusters to identify, this highlights the adaptability of the xyiVAT algorithm. Figures 4c and 4d are considered the most favorable visualizations of clusters for the dataset, as the darker regions in the MST-iVAT image align with the dark blocks in the xyiVAT image, which influences the preference in this case. Therefore, for different K values, we delve into the cluster hierarchies to obtain more detailed information about the data, which is not achievable with aVAT and ML-aVAT.

[11] introduced a dynamic time warping (DTW) based clusiVAT clustering algorithm for trajectory datasets. A visual assessment of the MST-iVAT image for the MIT-Trajectory dataset, illustrated in Table 5, clearly indicates that aVAT identifies only $k = 15$ red boxes as notable clusters. In contrast, ML-aVAT shows three red boxes, seven green boxes, and four blue boxes in the MST-iVAT image. In these three hierarchies of clusters, we can see that not all data points are selected for clustering. Figure 5d shows the xyiVAT image for $K = 500$, showcasing $k = 98$ dark blocks (all data points considered) categorized as clusters, which correspond to the darker regions in the MST-iVAT image. Therefore, xyiVAT offers improved trajectory data clustering compared to aVAT and ML-aVAT.

The MST-iVAT image of the MIT-IBRL dataset in Fig. 3a shows a considerable number of long gray lines (i.e., long inter-cluster edges) in the bottom right corner, indicating that the entire dataset can be considered as a single cluster. To select prominent clusters, the K value needs to be in thousands to exceed the number of long gray lines, as illustrated in Figs. 3b, 3c, and 3d of Table 3. The dark MST-iVAT image of the IIIT-H dataset in Fig. 3e indicates that data points are dense, suggesting that clustering may or may not be required. When xyiVAT is applied to the IIIT-H dataset, the images shown in Figs. 3f, 3g, and 3h of Table 3 show the presence of darker blocks by varying K values. In some instances, the absence of darker blocks may occur when K is varied. This situation demonstrates that clustering is unnecessary. Among all the clustering algorithms published to date, this paper uniquely establishes the condition under which clustering is not required.

Evaluation by DI and Runtime:
In [15], the evaluation was carried out using partition accuracy as the main metric. As xyiVAT can produce hierarchies of clusters for various K values, employing partition accuracy as an evaluation criterion becomes unfeasible. An alternative is DI, which offers insights into the compactness and separation between clusters. The value ranges between 0 and 1. The comprehensive breakdown in Table 4 displays the DI values for all three algorithms across the six datasets. The corresponding K values for each dataset used to compute DI are: KDDCup-99 is 65, MNIST is 250, STData is 100, MIT-Trajectory is 500, MIT-IBRL is 5000,

and IIITH is 3000. It is noteworthy that the DI values for xyiVAT surpass those of aVAT and ML-aVAT, indicating that the clusters are effectively partitioned in xyiVAT, thus addressing the issue of cluster validity in cluster analysis. The elevated DI values in xyiVAT can be attributed to its methodology of selecting the longest MST edges as inter-cluster edges. In contrast, aVAT and ML-aVAT aim to approximate the selection of prominent clusters in iVAT/MST-iVAT images, which may result in less accurate coordinate selections, leading to lower DI values. xyiVAT shows a notably higher DI for the KDDCup-99 and STData datasets, as the data points are forming proper clusters. On the contrary, the MNIST, MIT-trajectory, MIT-IBRL, and IIIT-H datasets yield lower DI values, as data points are densely packed, indicating that inter-cluster edges are significantly smaller than the possible maximum cluster width. Examining the workflow, ML-aVAT requires both training and testing of iVAT images, resulting in a longer runtime (up to an hour) compared to aVAT, which focuses on processing iVAT images (taking 10 to 15 min). xyiVAT solely depends on the K longest MST edges with K iterations, thus providing rapid results (less than a second). The combination of visual and numerical evaluations of the algorithms offers strong evidence supporting the superior performance of xyiVAT. Such thorough evaluations are crucial for advancing clustering algorithms, as they yield valuable insights into the strengths and weaknesses of various approaches.

5 Limitations and Future Works

Although contending algorithms automatically select clusters (not all clusters), xyiVAT involves manually selecting the longest MST edges to identify prominent clusters. This creates an opportunity for developing an automatic cluster counting algorithm using a method similar to xyiVAT. DI is considered for clustering validation because it measures both the compactness and separation of clusters without relying on average calculations or centroid-based assumptions, which can obscure poor intra-cluster structure—a significant weakness of the Davies-Bouldin Index and the Calinski-Harabasz index as provided by [9]. Unlike the Silhouette coefficient, which can favor convex clusters, DI focuses on the worst-case inter-cluster separation, making it more robust to irregular cluster shapes. Additionally, while the Adjusted Rand Index and Normalized Mutual Information require ground truth labels and are thus limited for unsupervised evaluations, DI operates solely on the inherent structure of the dataset, making it an effective internal validation measure. Although DI measures compactness and separation, we can develop a new evaluation metric to assess the effectiveness of xyiVAT, where some datasets may require more than one set of clusters.

6 Conclusion

We conducted experiments on xyiVAT using six publicly available real-world datasets to capture a wide array of applications and scenarios. xyiVAT involves selecting the K longest MST edges to align the dark blocks of the xyiVAT image

with the darker regions of the MST-iVAT image. Our method reduces dependency on user expertise and mitigates potential biases arising from the manual selection of coordinates in MST-iVAT images. In contrast to aVAT, ML-aVAT, and other algorithms, xyiVAT operates independently of image processing, signal processing, and machine learning methods. The experiments illustrate the versatility of xyiVAT, which is unattainable with aVAT and ML-aVAT. Furthermore, the paper addresses and resolves three aspects of cluster analysis along with specific challenges in MST-iVAT images for real-world datasets. This contribution highlights the potential to enhance analytical capabilities and improve decision-making processes across various scientific and industrial domains.

References

1. Bezdek, J., Hathaway, R.: Vat: a tool for visual assessment of (cluster) tendency. In: International Joint Conference on Neural Networks. IJCNN'02 (2002)
2. Buonadonna, P., Gay, D., Hellerstein, J., Hong, W., Madden, S.: Task: sensor network in a box. In: European Workshop on Wireless Sensor Networks (2005)
3. Caron, M., Touvron, H., Misra, I., J'egou, H., Mairal, J., Bojanowski, P., Joulin, A.: Emerging properties in self-supervised vision transformers. In: IEEE/CVF International Conference on Computer Vision (ICCV), (2021)
4. Chicco, D., Campagner, A., Spagnolo, A., Ciucci, D., Jurman, G.: The silhouette coefficient and the Davies-Bouldin index are more informative than dunn index, Calinski-Harabasz index, shannon entropy, and gap statistic for unsupervised clustering internal evaluation of two convex clusters. PeerJ Comput. Sci. (2025)
5. Cormen, T.H., Leiserson, C.E., Rivest, R.L., Stein, C.: Introduction to Algorithms, Third Edition, The MIT Press (2009)
6. Corporation, V.C.: Weather Data Services — Visual Crossing — visualcrossing.com (2002-2003). https://www.visualcrossing.com/weather/weather-data-services
7. Deshpande, K.V., Kumar, D.: Time and memory scalable algorithms for clustering tendency assessment of big data. Inf. Sci. (2024)
8. Havens, T.C., Bezdek, J.C., Keller, J.M., Popescu, M.: Clustering in ordered dissimilarity data. Int. J. Intell. Syst. **24**(5), 504–528 (2009)
9. Ikotun, A., Habyarimana, F., Ezugwu, A.: Cluster validity indices for automatic clustering: A comprehensive review. Heliyon (2025). https://doi.org/10.1016/j.heliyon.2025.e41953
10. Kumar, D., Bezdek, J.C.: Visual approaches for exploratory data analysis: A survey of the visual assessment of clustering tendency (vat) family of algorithms. IEEE Syst. Man Cybern. Mag. **6**(2), 10–48 (2020)
11. Kumar, D., Bezdek, J.C., Rajasegarar, S., Leckie, C., Palaniswami, M.: A visual-numeric approach to clustering and anomaly detection for trajectory data. Vis. Comput. (2017)
12. Kumar, D., Palaniswami, M., Rajasegarar, S., Leckie, C., Bezdek, J.C., Havens, T.C.: clusivat: A mixed visual/numerical clustering algorithm for big data. In: 2013 IEEE International Conference on Big Data, pp. 112–117 (2013)
13. LECUN, Y.: The mnist database of handwritten digits (1998). http://yann.lecun.com/exdb/mnist/
14. Li, Y., Hu, P., Liu, Z., Peng, D., Zhou, J.T., Peng, X.: Contrastive clustering. ArXiv **abs/2009.09687** (2021). https://api.semanticscholar.org/CorpusID:221819451

15. Mittal, H., Laxman, J.S., Kumar, D.: Ml-avat: A novel 2-stage machine-learning approach for automatic clustering tendency assessment. Big Data Res. (2023)
16. Perlman, R.: An algorithm for distributed computation of a spanningtree in an extended LAN. Association for Computing Machinery, New York, NY, USA (1985). https://doi.org/10.1145/319056.319004
17. Rathore, P., Ghafoori, Z., Bezdek, J., Palaniswami, M., Leckie, C.: Approximating dunn's cluster validity indices for partitions of big data. IEEE Trans. Cybern. (2018). https://doi.org/10.1109/TCYB.2018.2806886
18. Rathore, P., Kumar, D., Bezdek, J.C., Rajasegarar, S., Palaniswami, M.: A rapid hybrid clustering algorithm for large volumes of high dimensional data. IEEE Trans. Knowl. Data Eng. **31**(4), 641–654 (2019)
19. Reddy, C.R., et al.: Improving spatio-temporal understanding of particulate matter using low-cost iot sensors. In: IEEE Annual International Symposium on Personal, Indoor and Mobile Radio Communications (2020)
20. Singh, M., Bhattacharjee, R., Sharma, N., Verma, A.: An improved xie-beni index for cluster validity measure. In: 2017 Fourth International Conference on Image Information Processing (ICIIP) (2017). https://doi.org/10.1109/ICIIP.2017.8313691
21. Sledge, I.J., Huband, J.M., Bezdek, J.C.: (automatic) cluster count extraction from unlabeled data sets. In: Fifth International Conference on Fuzzy Systems and Knowledge Discovery (2008)
22. Tavallaee, M., Bagheri, E., Lu, W., Ghorbani, A.A.: A detailed analysis of the kdd cup 99 data set. In: IEEE Symposium on Computational Intelligence for Security and Defense Applications (2009)
23. Van Gansbeke, W., Vandenhende, S., Georgoulis, S., Proesmans, M., Van Gool, L.: SCAN: Learning to Classify Images Without Labels. In: Vedaldi, A., Bischof, H., Brox, T., Frahm, J.-M. (eds.) ECCV 2020. Van Gansbeke, W., Vandenhende, S., Georgoulis, S., Proesmans, M., Van Gool, L.: Scan: Learning to classify images without labels. In: Vedaldi, A., Bischof, H., Brox, T., Frahm, J.M. (eds.) Computer Vision - ECCV 2020, pp. 268–285. Springer, Cham (2020), vol. 12355, pp. 268–285. Springer, Cham (2020). https://doi.org/10.1007/978-3-030-58607-2_16
24. Wang, L., Geng, X., Bezdek, J., Leckie, C., Kotagiri, R.: Enhanced visual analysis for cluster tendency assessment and data partitioning. IEEE Trans. Knowl. Data Eng. (2010)
25. Wang, L., Leckie, C., Ramamohanarao, K., Bezdek, J.: Automatically determining the number of clusters in unlabeled data sets. IEEE Trans. Knowl. Data Eng. (2009)
26. Wang, L., Nguyen, U.T.V., Bezdek, J.C., Leckie, C.A., Ramamohanarao, K.: ivat and avat: Enhanced visual analysis for cluster tendency assessment (2010)
27. Wang, X., Ma, K.T., Ng, G.W., Grimson, W.E.L.: Trajectory analysis and semantic region modeling using nonparametric hierarchical bayesian models. Int. J. Comput. Vision (2011)